CALIFORNIA
TECHNICAL FORUM

Cal TF Technical Position Paper No. 5: Reducing Measure Complexity

I. Background

This document provides guidelines for Cal TF assessment of ex ante measure development. The guidelines include considerations for the appropriate level of measure complexity that should be employed in ex ante measure development, as well as considerations for determining when the "best available information" is sufficient to support measure development.

"Measure complexity" in this context generally refers to a) how many different "measure combinations" should be developed for a measure to account for differences in how a measure is deployed, where it will be installed, and how it will be used, and b) the engineering approach used to generate savings estimates, either through building energy computer simulations or through simpler engineering calculations that require fewer inputs and assumptions, and c) the application of additional factors such as HVAC "interactive effects" to more accurately estimate energy savings.

Regulatory Context

The California Public Utilities Commission issued the following guidance for ex ante measure development:

1. Use the "best available information" for measure development[1]
2. DOE-2 modeling is appropriate for weather-sensitive measures[2]
3. Balance accuracy with the need to enable promising new technology to enter IOU portfolios[3]
4. Balance accuracy and precision, cost, and certainty[4]

Measure development, including the number of combinations, must be supported by data and information. The directive to use "best available information" in developing ex

[1] D. 12-05-015 at 320; Ordering Paragraph 26 of D. 09-09-047 at 356, D. 11-07-030
[2] D.12-05-015, May 18, 2012 footnote 64
[3] D.12-05-015 at 297
[4] D.09-09-047 at 299

ante savings estimates is a recurring mandate in CPUC decisions. Commission Staff are tasked with using “best available Information” for DEER ex ante updates, and similarly the IOUs are expected to use “best available information” when developing non-DEER ex ante estimates.[1] However, CPUC decision language does not specifically define “best available information.” Thus it is not always clear what “best available information” means, and how to consider the accessibility, applicability, credibility, and cost of various potential data/information sources in determining whether the “best available data” standard has been met. This ambiguity may lead to suboptimal use of ratepayer dollars to fund superfluous research when existing data may provide a reasonable level of ex ante accuracy.

In seeking to populate DEER with the best available information, the CPUC has recognized that DOE-2 modeling is an appropriate source for weather-sensitive estimates. Yet, this should not be interpreted as limiting data inputs to those used by the DEER team. In fact, Commission Decision 12-05-015 clarifies that “Staff should continue to seek input from parties to determine where and when to use a particular analytical approach.”[2] For example, while staff has mandated that measures implemented in conditioned spaces that are not directly weather-dependent include interactive effects derived from building simulations for lighting measures, this requirement has been questioned by several workpaper developers for certain measures such as notebook computers. Thus, this approach should be reconsidered by the Cal TF for appropriateness and applicability moving forward.

Since 2008, the Commission has called for the use of the best available data and the most recent published studies, a tighter loop between evaluation and ex ante value development, and for staff to balance accuracy with the IOUs’ need to “continuously augment the portfolios with new technologies that offer promise.”[3] D.12-05-015 encourages staff to not let “the perfect be the enemy of the good” and to consider risk-sharing approaches when assigning ex ante values to promising new technologies.[5]

The Commission has also repeatedly reminded staff, the IOUs, and interested parties that “savings measurement and verifications should reflect a reasonable balance of accuracy and precision, cost, and certainty, and be designed for incorporation into procurement process.”[4] While this directive was targeted at EM&V activities, ex-post analysis should arguably be more accurate and certain than forecasted ex ante

[5] *Ibid.*

analysis. Therefore, it is reasonable to assume that the directive to balance accuracy, cost, and certainty in ex post analysis also applies to ex ante savings development.

II. Current Process

Measure Complexity

The development of multiple measure combinations through energy modeling and application of multipliers (such as interactive effects), favored by the current DEER, is intended to provide accurate savings estimates. However, in many cases there is little evidence to show whether the distribution of many measure combinations among program participants, climate zones, and building types supported by energy modeling, contributes to greater accuracy over simpler approaches. While the greater number of measure combinations may not lead to greater accuracy, it does lead to increased costs, and furthermore may not be easy to implement. Another issue that should be addressed is proper ways to scale, extrapolate or interpolate data.

Finally, consideration of how program (and measure) complexity may affect customer understanding of program and savings and ultimately their willingness to participate should be considered.

Best Available Information

The level of accuracy and complexity of measure ex ante estimates is determined by the information used to support measure development. In lieu of an official definition of "best available information," both Commission Staff and the IOUs are left to interpret the meaning of "best available" at their own discretion, leading to differing opinions and inconsistencies in many cases. The Cal TF needs a consistent definition of "best available information" to properly determine if proposed energy efficiency savings estimates are developed in accordance with CPUC directives. [6]

Challenges

[6] To maximize the effectiveness of these guidelines, the Cal TF will need to work with key stakeholders in the ex-ante measure development process to ensure stakeholders understand the recommendations put forth in these guidelines.

The California IOUs face many challenges resulting from the level of complexity expected of ex ante measure savings development and reporting, as well as the ambiguity surrounding the “best available information” standard. High levels of measure complexity due to multiple measure combinations and complex building models make it time-consuming and costly for IOUs to plan a measure’s portfolio impact, perform QA/QC on measure values, update measures, track/report, and reconcile program delivery strategy with information collection needs during program implementation.

An Example, as a result of a 2013 Title 24 update to building standards, and 2014 updates to supporting weather data, all IOU direct and indirect climate sensitive measures required revisions. While not all measures were affected by updates, the vast majority of IOU measures are considered weather-sensitive due to the application of HVAC interactive effects factors for products operating in conditioned building spaces such as lighting. The cost and time to update these workpapers by all IOUs was extensive, while most measure values changed by less than 5%. This is an example of situation where new, more recent data may come available (such as new weather data), but the resulting difference in measure estimates may not be statistically valid due to measure estimate uncertainty of greater than 5%, in which case it may not be worth the time or cost to incorporate newly available data until the next update.

The lack of clarity surrounding what constitutes “best available information” makes it difficult for the IOUs to determine when existing information is sufficient, what level of precision is necessary for measure development, and when it is necessary to pay for additional data collection when some information is already available. Data collection efforts can cost tens of thousands of dollars to support a subset of the parameters required to be reported for ex ante measures. For example, a study may be done to determine the operating hours of a product; however, a separate study may need to be done to determine the incremental cost of the product.

III. Purpose of this Memo

In the absence of existing guidance for some specific components of ex ante measure development, this document recommends some *guidelines* and some admittedly-arbitrary criteria values to focus the discussion and to simplify decisions that need to be made for future measure development. This memo outlines a process for consideration of the following criteria in ex ante development:

- Determine what end goal is (kWh/kW/therm)

- What calculation approach to use: engineering equations, statistics, building simulations, ex post evaluation results, etc.
- Use of measure approval expiration dates/interim approval
- Level of statistical rigor
- Determining appropriate number of measure combinations
- Evaluate relative impact of various inputs on the desired goal to determine which are critical to determine at high level of accuracy
- The impact of the level of complexity for various levels of measure savings on the portfolio goals
- Cost of obtaining information, updating workpapers, etc.
- When and how additional data needs to be collected

IV. Recommended Process

a) Calculation Approach

Measure Definitions

- Normal impact measure: A measure predicted to be normal impact, or that has demonstrated normal portfolio impact through the course of implementation.
- Low impact measure: A measure that is predicted to have a lower impact on the portfolio than average.
- Demonstrated high impact measure: A measure predicted to immediately be high impact, or has demonstrated high portfolio impact through the course of implementation.
- Interim measure: A measure for which sufficient information is anticipated but not yet available that would satisfy the level of rigor for a measure predicted to be normal or high impact. Interim measures must be re-examined after 1 year or another duration determined by the Cal TF.

Calculation Approaches

- Engineering equation with documented inputs: a standard engineering equation is employed with inputs that can be traced back to sources, or justified using professional judgment. No additional empirical or statistical justification of inputs is required.

- Engineering equation or curve fit with calibrated results, with statistically-justified inputs: a standard engineering equation is employed with inputs that are supported by statistical analysis, and outputs that are calibrated using existing data if appropriate and possible.
- Direct measured data: Empirical approaches that utilize on site measurement, including AMI, ET studies or other approaches that are not explicitly part of an ex post study or engineering equation with a curve fit
- Calibrated building model: Use of an accepted building energy simulation to estimate savings. All model inputs must be documented and traceable, and statistically justified to the extent that is practical and possible. Outputs must be calibrated as appropriate and possible.
    - DOE-2.2 (used by CPUC for DEER measures)
    - Energy Plus (used by CEC for Title 24, non-residential)
    - *List of building models, and when they are appropriate to use*
    - *Perform appropriate parametric analyses to assess the key input variable to focus on*
- Evaluation results: Statistically valid ex-post evaluation results that may be directly applied or modified for measure application.

b) Use of Parametric Analysis to Identify Key Inputs, Number of Measure Combinations Needed

Measure complexity and costs associated with developing and updating measures can and should be addressed through parametric analysis – understanding what parameters most impact key outputs (savings and cost-effectiveness analysis) and what measure combinations are significantly different. Open Studio, one of the EnergyPlus interfaces, allows for high-speed, high volume parametric analysis. The parametric analysis function can be used to identify which input parameters are drivers of key outputs, so that the greatest scrutiny and follow-up data collection and analysis can be focused in impactful input parameters. Furthermore, parametric analysis can be used to identify which measure combinations are truly distinct – if two different measure combinations do not yield truly distinct energy savings values when applied to a building prototype in a parametric analysis, then there the measure combinations should be combined into one.

Table 1. Recommended Calculation Approaches

Approval Type	Low Impact Measure	Normal Impact Measure	Demonstrated High Impact Measure	Interim Measure
Short Term (Measure expires after 1 year, then "Long Term" review required)				**Weather-sensitive measures:** 1. TF judgment for: a. Building model, or **b.** Engineering equation with calibrated results, statistically valid inputs **Non-weather sensitive:** 1. Engineering equation with documented inputs
Long Term (Measure expires at discretion of Cal TF)	**All measures:** 1. Engineering equation with documented inputs	**Weather-sensitive measures:** 2. TF judgment for: c. Building model, or **d.** Engineering equation with calibrated results, statistically valid inputs **Non-weather sensitive:**	**Weather-sensitive measures:** 1) Evaluation results 2) TF judgment for: a. Calibrated building model, or **b.** Engineering equation with calibrated results, statistically valid inputs **Non-weather sensitive:** 1. Evaluation results 2. Engineering equation with calibrated results, valid inputs	

		1. Engineering equation with calibrated results, statistically valid inputs		

c) Criteria to Evaluate Applicability and Validity of Available Information

- It is up to the Cal TF to determine if the information available for a given measure meets "best available information" criteria.
- Criteria for the TF to consider include:
  - Applicability of sample data to the anticipated program participation population
  - Data collection and analysis methods
  - Measurement error
  - Confidence intervals (uncertainty and precision)
  - Sample size relative to statistical results
  - Study age
- If the TF determines that additional data must be collected, the study results should follow the conditions outlined in Table 2 for recommended statistical rigor. A Research Plan Template is attached in Appendix A.

d) Level of Statistical Rigor

- The TF will need to use professional judgment based on the data collection method, study results, and measure characteristics as to whether the statistical analysis and results are adequate. The TF should consider the following criteria, as appropriate. These terms are discussed in more detail in the Appendix.
  - Confidence interval bounds for mean estimates
  - Standard error of the mean
  - Coefficient of variation
  - Sample size with respect to confidence level and results
- In considering statistical results to support input parameters for engineering calculations, the TF should consider the cumulative uncertainty of the inputs
- The TF will use professional judgment to determine if data that is sampled is applicable to the measure-targeted population and is sufficient to support measure calculations.
  - Data from non-California jurisdictions will require higher level of scrutiny.
- In the event that the Cal TF feels that the statistical rigor of existing data is inadequate, additional data collection efforts should consider the recommendations for statistical rigor outlined in Table 2.

Table 2. Recommended Considerations for Statistical Rigor of Studies done Pursuant to Cal TF Recommendation

Approval Type	Low Impact Measure	Normal Impact Measure	Demonstrated High Impact Measure	Interim Measure
Short Term (Measure expires after 1 year, then "Long Term" review required)				• Statistical analysis for highest impact parameters in engineering calculation • 80% confidence level • Sampling and measurement error within suitable range
Long Term (Measure expires at discretion of Cal TF)	TF judgment	• Statistical analysis for highest impact parameters in engineering calculation • 80% confidence level • Sampling and measurement error within suitable range	• Statistical analysis for highest impact parameters in engineering calculation • 90% confidence level • Sampling and measurement error within 10%	

e) Determining the number of measure combinations

Some unnecessary measure complexity can be controlled with the high-speed, high-volume parametric analysis capabilities offered by the National Renewable Energy Laboratory's OpenStudio platform. OpenStudio enables measure modelers to quickly run high volumes of energy simulations to identify how much small variations in single inputs affect final energy savings estimates. This allows the modeler to quantify the added value and/or uncertainty of creating additional measure combinations and only create those that are truly significantly different.

Low Impact, Normal Impact, and Interim Measures: Measure combination values should differ by more than 10%. It is assumed that, for differences in measure combinations of less than 10%, eliminating multiple measure combinations will result in a negligible portfolio impact for low and normal impact measures. This will reduce the time and cost necessary for measure development, implementation, and maintenance. Additionally, measures should be updated only if the new value is more than 10% different from the old value to minimize time and cost associated with negligible changes to the portfolio savings. Additional criteria should also be developed for how to address if some fraction of the overall dataset exceeds the 10% threshold.

High Impact Measures: If the differentiation in measure combinations changes portfolio savings by more than 1%, then measure combination options should be more carefully examined, regardless of the % difference between measure combination values. Measure combination differences should be statistically justified and supported by the certainty of both the data and calculation methods used to determine values.

- If values are based on building simulation results, then requiring measure combinations to differ by more than 10% is justifiable. A literature review of DOE-2 building simulation uncertainty was conducted by Texas A&M. Of about 24 building simulations compared with utility bills, there was on average a 10% difference in modeled vs. actual energy usage. [7]

[7] The Texas A&M literature review indicates an average difference of 10% between building simulations and billing data, with a range of 1% – 26% amongst 31 building simulation/billing data comparisons. Because building

- Measure combinations should differentiate by only the most impactful parameters used to determine energy savings.[8]
- The TF should consider the uncertainty of statistical analyses for assumptions that are used for engineering calculations that do not employ building modeling. The accuracy of an engineering calculation output is only as strong as the inputs, so the 10% threshold may be reconsidered if statistical analyses regularly show uncertainty of less than 10%.
- If data supporting an input parameter has a confidence interval with upper and lower bounds differing by more than 10%, multiple measure combinations for that input should be considered.[9]
- A statistically significant difference of greater than 10% between sample means for data sets supporting a given parameter (such as weather data for different climate zones) should warrant different measure combinations.

modeling of deemed ex ante measures seeks to estimate average building usage across sectors, climate zones and time with input assumptions that are based on professional judgment or measured parameters with embedded uncertainty, we feel that changes to building model outputs for deemed measures of 15% or more is within the uncertainty range of the building model outputs. However, 10% is used as a conservative estimate.

[8] Tornado diagrams are useful for deterministic sensitivity analysis - comparing the relative importance of variables

[9] Example: If the operating hours amongst all building types sampled for a measure results in a mean of 1200 hours with a confidence interval of (1000, 1800), then multiple building types should be considered due to the 57% difference in the values. If the confidence interval is (1200, 1300) then the sample mean should be used for all building types due to a percentage difference of 8% between the upper and lower confidence bounds.

Approval Type	Low Impact Measure	Normal Impact Measure	Demonstrated High Impact Measure	Interim Measure
Number of Measure Combinations	Different measure combination values should differ by more than 10%.	Different measure combination values should differ by more than 10%.	If the differentiation in measure combinations changes portfolio savings by more than 1%, then measure combination options should be more carefully examined, regardless of the % difference between measure combination values. Measure combination differences should be statistically justified and supported by the certainty of both the data and calculation methods used to determine values.	Different measure combination values should differ by more than 10%.
Measure Updating (Aligning with rolling cycles, triggers, etc.)	Measure values should only be updated if new values differ by 10% or more from old values.	Measure values should only be updated if new values differ by 10% or more from old values.	If the change in measure value results in a change to portfolio savings by more than 1%, then the supporting evidence for a new measure value should be carefully examined. The new value should be statistically justified and supported by the certainty of both the data and calculation methods used to determine values.	Measure values should only be updated if new values differ by 10% or more from old values.

f) Administrative considerations

- Cost of workpaper development, both initial development and maintenance
- Cost of processing measure data internally for reporting purposes
- Risk of human error due to number of measure combinations, frequency of updates, etc.
- Infrastructure required for developing, tracking, and reporting
- Customer impacts due to requirements for many measure types (where the values are not dealt with behind the scenes)

g) Other issues to address

- Guideline for running parametric analyses
- Calibration guidelines
- Appropriate application of interactive effects[10]
- Definition/consideration of bias
- Best practices for measure documentation
- Early retirement
- Add on measures
- Baseline determination and justification

[10] The Northwest Regional Technical Forum guidelines consider "interaction" effects to be significant if the interaction changes a measure's savings estimates by more than +/- 10%. Roadmap for the Assessment of Energy Efficiency Measures, page 8. Accessed at http://rtf.nwcouncil.org/subcommittees/Guidelines/RTF%20Guidelines%20%28revised%206-17-2014%29.pdf

CALIFORNIA
TECHNICAL FORUM

APPENDIX A

Cal TF Research
Plan Template_ver 1.

APPENDIX B

Statistics Refresher

Statistical Term	Formula	Definition	Use
Standard Deviation (s)	$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ where x = [illegible] $\bar{x}$ = [illegible]	The standard deviation measures the dispersion of the data set, or how much individual data points vary from the sample mean. (Related: variance, which equals s^2).	Indicates the distribution of data about the mean.
Standard Error (SE)	$SE = \frac{s}{\sqrt{n}}$	The standard error (SE) is the standard deviation of the sampling distribution of the mean. As sample size increases, standard error decreases.	Important for indicating the uncertainty of a mean estimate

CALIFORNIA
TECHNICAL FORUM

Confidence Interval (CI)	$Lower\ Bound = \bar{x} - (SE \times t_{\alpha/2})$ $Upper\ Bound = \bar{x} + (SE \times t_{\alpha/2})$ Where t is a function of the confidence level (1-α). A significance level α=0.1 implies a confidence level of 90%	The confidence interval is the upper and lower bounds of a statistic (such as the mean), calculated based on an assumed significance level or level of confidence (such as 90%). A 90% confidence interval indicates that there is a 90% probability that the true mean of the population falls within the upper and lower bounds of the interval. The more narrow the confidence interval, the more precise the estimate.	Indicates both the uncertainty and precision of a sample mean as representative of population mean. A narrower confidence interval indicates less certainty and more precision.
Coefficient of Variability (CV)	$CV = \frac{s}{\bar{x}} \times 100$	The coefficient of variability measures the relative variance of sample data. The CV is equal to the sample standard deviation divided by the sample mean, multiplied by 100. The smaller the CV, the closer sample data points are to the sample mean. The standard deviation (SD) is a measure of the dispersion of data around the mean. The smaller the SD, the narrower the range of values and the closer most values fall to the mean. The CV is most useful for samples with positive values (negative values may result in a mean close to zero, generating a misleading CV).the TF may determine if a threshold for the coefficient of variability should be established.	Indicative of how much data points vary from the mean. Because VC is expressed as a percent, it is unitless and can be used to compare variation of two similar data sets with different units (e.g., one data set stating hours per year vs. another dataset stating days per year).

Confidence Level	Confidence Level = (1 – α) A confidence level of 90% has a significance level of α = 0.1.	For a 90% confidence interval, if infinite samples of the same size were drawn from the population and the sample mean calculated, 90% of the time the confidence interval would contain the true population mean. Alternatively, there is a 90% chance that the confidence interval calculated from the sample contains the true population mean. Generally speaking, for a given sample size, the higher the confidence level, the wider the confidence interval. To increase confidence that the interval contains the true mean, the bounds must be larger (increase confidence, decrease precision). Alternatively, for the same confidence level, the bounds may get smaller as sample size is increased (increase sample size, increase precision).	The TF must determine if a minimum confidence level is required (e.g., 90% for high impact measures and 80% for normal impact measures)