Energy Efficiency Evaluation Recommendations

Primary Causes of Reduced Reliability of All the Studies:

1. Late start due to contracting problems and the further delay created by the wholesale shift from program evaluation to "high-impact measure evaluation" resulted in:
   a. inability to collect accurate baseline and net-to-gross data (because too much time had elapsed),
   b. too little time to collect complex data with adequate sample sizes, supplemental data collection as needed, and good quality control, and
   c. too little time to complete the complex analyses with through assessment and testing of alternative specifications and good quality control.

2. The largest energy efficiency program evaluation effort in the US made a wholesale shift halfway through the process away from the standard, well-understood program evaluation approach to a completely untested approach called the "High-Impact Measure" approach.

Parameter Name	ED/Consultant Result	Alternate Result	Rationale for the alternate result, including why alternate result is more reliable than study result
1. Use of Net-to-Gross	Ex post NTG	Two possibilities: Use *ex ante* NTG. Eliminate the use of NTG altogether.	The rules governing Net-to-Gross (NTG) estimates need to be revisited. As energy-efficient technologies become more accepted in the marketplace, NTG values decline. Consequently, declines in NTG values indicate program success: there are higher levels of free ridership that result in lower NTG values. Although lower NTG values indicate increased acceptance of energy-efficient technologies, the IOUs are penalized since NTG is a key factor in calculating program attribution. Specifically, several issues lead to questioning the validity of NTG calculations and their use in the 2006-2008 program cycle. · The validity of results from the self-report NTG survey used for most of the mass market (residential and small nonresidential) programs suffered from several issues. These problems made the method unreliable. · Improper NTG ratio construction: A percentage probability of being a free rider was created from respondents' 1-10 scores on multiple questions that aren't about whether they would have purchased the product without the program. · It was often administered years after a customer purchased a product. · In multiple-decision maker (nonresidential) cases a single

			respondent does not have sufficient perspective to understand organizational decision-making that occurs over time and involves multiple people and/or departments. These types of problems make self-report an unreliable method to determine NTG. · NTG attribution is limited to program cycle efforts only, resulting in pre-cycle program efforts being attributed to free-ridership and the current program cycles efforts that will bear fruit in later cycles never being credited in this or the later cycle. For example, a community college's EE efforts that were partly attributable to conversations with a PG&E program manager led to changes in internal policies to foster move to more EE buildings. When the college finally built a project, the evaluator notices "green policies" but ignores role IOU programs played before current program cycle by classifying entire project as free-ridership. · For the large nonresidential programs, it is difficult if not impossible for respondents to tease apart the energy efficiency aspect of a larger project when responding to a long battery of questions posed by an interviewer. Responses concerning timing and what "would have happened absent the program" may lead responses concerning an entire project, not just the EE portion. · Selective use of collected data suggests negative bias in the calculation of NTG. One clear-cut example is the Fabrication evaluation. In this evaluation, in 20 of top 60 sites, the evaluator dropped highest score (usually the program influence score). Never were any of the lower scores dropped. In some cases, the average score was further reduced by ½. These reductions were applied to largest site evaluated. The reason given was that it was the only project considered. This practice runs contrary to the evolutionary nature of these large projects: although the end result is that only one project is completed and evaluated, the reality is that many variants were considered. · The authors of the studies themselves are also, at times, very concerned about the reliability of the NTG estimates. On page 82 of the Upstream Lighting impact evaluation for instance, the authors state that "Given the timing of this evaluation we are concerned that none of the NTGR results derived from the various methods can be considered representative of the 2006-2008 program…" and, "In the end, the final recommended NTGR estimates represent our best judgment based on a preponderance of evidence". Obviously,

			“judgment” is very difficult to vet and verify. These concerns cast doubt not only on this study but all studies facing either of the same issues: being conducted far too late in time to capture the conditions prevailing throughout the program period or finding vastly different answers from different approaches.
2. Adjustments Made without Final Studies	Changed savings estimates for programs or measures without conducting studies	No changes should be made to program savings estimates unless there are updated studies to justify them. Ex ante savings estimates should be used for programs and measures for which studies were not done.	Examples: 1) Residential Interactive Effects. There is no study available for the utilities (or anyone else) to review related to the calculation of residential interactive effects. But many of the measures now found in DEER include such “simulated” effects, with no study to support that. No study result using this unstudied DEER data should be accepted, and no Evaluation Report Tool should use it until a study is made available and fully vetted. 2) In large part because of shifting from evaluating programs to high-impact measures, many small measures and small programs were not included in the studies. In these cases, the ex ante estimates should be used, as was understood at the beginning of the program cycle. Instead, there are now plans to subjectively determine what are “similar” programs and measures, and apply new DEER or study results to them, with minimal, poorly-informed analysis to determine whether this can be justified. That is not ex post evaluation! . Example of problem: No updated studies were conducted on the Local Business EE and Energy Savings BID programs yet DEER updates were arbitrarily applied to these programs based upon SPC updates. Energy Division is expected to make adjustments to measures that have not been studied in their modeling process.
3. Confidence Intervals and Sample Sizes	Many programs with small sample sizes with very large confidence intervals	Ex-ante estimates are based on either engineering estimates or previous reliable studies. These should be used in cases where it is determined that the confidence intervals are very large (for instance encompassing the	Confidence Intervals and Sample Sizes: Many of the studies have extremely wide confidence intervals. Many, in fact, are so large that they include the ex-ante value or a greater value as well as zero. This is generally the result of small sample sizes while estimating large populations, often due to limited funding or limited time to gather data. Whatever the cause, the consequence is unreliable data. There is a strong argument for retaining the ex ante estimates in all such cases. EXAMPLES: A. Program SCG3513 In the “Major Commercial Contract Group Final Impact Evaluation Report” in table 24 the program SCG3513 has a gross savings realization rate of .72. However, the 90% confidence interval for the

		ex-ante estimate and zero or 1.0) so the results are unreliable.	program is .53! This means that the results of the study indicate that the true realization rate for this program falls somewhere between .19 (.72-.53) and 1.25 (.72+.53), a huge range which could make the program either *extremely cost effective or not cost effective at all*! When questioned about the size of the confidence intervals, the evaluator response stated that "An analysis of the confidence intervals around the UES estimates shows that, over all 11 combinations of program fuel estimates reported, in 7 of them the confidence interval included the IOU clam; in 9 of them the confidence interval included zero"! This result seriously questions the reliability of the study. Ex-ante values are for the most part based upon sound engineering estimates or previously vetted studies. It stands to reason that these values, if they fall within the 90% confidence ranges of the study and the impact studies lack reasonable reliability, should be used instead of the mean estimates. In these cases, the study has not refuted their value, and in fact provides support for the ex-ante claims. B: Program SCE2517 (Standard Performance Contract) A sample size of 18 for about 1,400 participants in a major savings program. Only 9 of 13 cases in the certainty stratum (the set of largest savings cases that should be sampled at 100%) were completed. For the remaining four other strata, only 2 or 3 were sampled from the remaining 1,384 measures. The fact is that there are no credible results for the 4 lower strata, leaving the ex ante estimates as the only alternative credible data source. Because of the small numbers, ex ante results remains the more reliable data source even for the 5 major participants not reviewed. And for all the sampled cases, these cases should be handled in line with the recommendations for the Baseline Issues problem. C.. SCE Industrial and Agricultural Programs, SCE2509 and 2510 For SCE2509 (Industrial) and SCE 2510(Agricultural), Itron confesses: "As a result of re-directing resources to the analysis of Steam Traps and Tank Insulation HIMs, the M&V scope for programs SCE2509 and SCE2510 was limited to the samples drawn in March 2008," a heroic 30 out of a program population of 264 for Industrial and 10 for Agricultural. There are no signs that the largest sites were sampled at 100%, which could have given the results more reliability. They had plans to continue sampling, so it appears they would have reached a more defensible size

			except for the forced shift from program evaluation to "high-impact measure evaluation." This calls into question the impact results of SCE2509. To their credit, the evaluators refused to provide SCE2510 impact estimates due to the sample size of 10. D. Appliance Recycling Program This study collected no usage data for refrigerators recycled during the program cycle. Instead, it collected data from a too-small sample of 137 refrigerators recycled in 2009. Instead of building on the data collected over several past program cycles and the strong multi-faceted approach developed in the 2004-5 study, this project failed to collect any controlled, lab-metered data at all. It metered two weeks of energy usage of 137 refrigerators whose participant owners agreed to delay the pickup of their refrigerator. This small amount of data from homes recycling their refrigerators was used to project the full-year usage of all 2006-8 program refrigerators in the different locations and different uses they would have gone to if not recycled.
4. CFL Study Errors			In multiple areas, the CFL HIM study selected analysis approaches that would yield lower savings estimates than alternative approaches that have equal or stronger justifications for use. In some cases, the selected method is a very indirect and inexact way to produce an estimate. In others, the details of the particular analyses done contain significant flaws.
A. Net-to-Gross Ratio	PG&E: 0.48 SCE: 0.64 SDG&E: 0.48	PG&E: 0.71 SCE: 0.80 SDG&E: 0.71	The recommended approach is to use one of the 5 alternative methods explored by the HIM study, namely the one that was also used in the 2004-5 CFL study, in place of the judgmental combination of two other methods recommended by the evaluator. That is the supplier self-report approach Rationale: 1) That's the NTG approach that the program used for planning, 2) Using it creates a consistent approach over time, which is important for monitoring program performance over time. 3) The other four methods used by the HIM evaluation were good experimental approaches to explore, but they were not well executed. The preferred, self-report based methods do not capture what the suppliers know about how the program changed what was available for the customers to select.
B. Installation Rate	Residential: PG&E: 0.67 SCE: 0.77	Residential: PG&E: 0.80 SCE: 0.89	The HIM study completely ignores any installations during the program cycle beyond those estimated to happen during the first year. In other words, zero savings are counted for the CFLs purchased and held in

	SDG&E: 0.67 Nonresidential: All: 0.81	SDG&E: 0.80. Nonresidential: All: 0.92	reserve for more than a year. The alternative is calculated using the study's rate of deferred first year installations and applying it also as a second-year rate for bulbs purchased in 2006 and 2007 (but not for 2008 to limit savings to within the cycle time-frame). The alternate result does not include any CFLs installed post -2008 which is a policy issue still not addressed by the CPUC.
C. Residential/ Nonresidential Split	PG&E: 0.94/0.06 SCE: 0.94/0.06 SDG&E:0.95/0.05	All: 0.80/0.20 Alternative: PG&E: 0.92/0.08 SCE: 0.81/0.19 SDG&E: 0.87/0.13	The HIM study's estimates are based on the numbers of incandescents and CFLs found in its residential and commercial on-site surveys. It was done so late that about a quarter of all the program-rebated CFLs installed in commercial facilities would have burned out, having reached their lifetime hours. This strange method of estimating the split could only be justified if no other data was available. But other, better data sources are available. Using data from interviews of retailers, with sales-weighted proportions, indicate that about 20% of the sales are for business. When a different study surveyed residential customers, customers reported that 13% of their bulb purchases were going into their businesses. Even this is a lower-bound estimate, because it ignores the business customers who bought their CFLs through their businesses. Despite it being a lower-bound estimate, we recommend using the results of the HIM study, which actually asked CFL users whether they purchased CFLs for the home or a business.
D. Wattage Reduction per CFL Installed	PG&E: 44.2 W SCE: 44.8 W SDG&E: 44.4 W	Retain Energy Star Guidelines and apply them to the actual mix of CFLs rebated through the program	Instead of the obvious and only widely accepted approach (lumen-to-lumen equivalency), the HIM study used a strange method of establishing this parameter also: take the difference between the average incandescent wattage and the average CFL wattage, comparing "similar" sockets in the home. Then report out a single average wattage difference for all incandescent vs. CFL bulbs in place in 2008. Don't even attempt to compare the wattage differences by lighting level provided or recognize that these numbers ignore customer choices on which fixtures are most valuable for installing CFLs The obvious method is to assume that people install CFLs that match the incandescent wattage rating that they want to replace. This will tend to follow the information on the CFL packaging, which relies on the Energy Star Guidelines. The 2004-05 evaluation results, which estimated true delta watts, not just the difference between the averages, agreed with the implied wattage reductions from the Energy Star Guidelines. This is far better justified result.
E. Hours of Use	PG&E 1.9	2.34	Because of the combined problems of the data and the analysis in the

	SCE: 1.9 SDG&E: 1.5		HIM study, we should go back to the most recent study, as DEER did: the KEMA 2005 CFL metering study. This yields 2.18 hours for interior CFLs and relies on a 1999 HMG study for exterior lighting, yielding a total average of 2.34 hours per day. Note, DEER only includes an estimate for interior CFLs, while the program includes exterior-installed CFLs, so the overall KEMA study average should be used. The HIM study made a major effort to gather new data and do new, complex analysis with it. Unfortunately, the metering data had problems and the complex statistical analysis is unreliable, being very unstable and misspecified. For example, the metering data under-represents the highest-use lighting and there were problems with the metering equipment. The regression analysis produces bizarre results: if a customer moves from San Diego to Los Angeles, their hours of use change dramatically. And it excludes obvious determinants of usage, such as such as dwelling type, fixture type and lamp type.
5. Baseline Issues	Ex post baseline calculations	Use *ex ante* as the baseline for more savings estimates.	The 2006-8 nonresidential evaluations went too far in developing “more accurate” baselines, with the result that free ridership was measured (often incorrectly) and subtracted twice in creating the savings estimates for the projects and ultimately the programs Several evaluations used “What would have happened in the absence of the program?” as the baseline question. But that question simultaneously addresses both the starting situation for the EE retrofit and the free ridership issue. Unfortunately, these evaluations didn’t recognize that and also produced Net-to-Gross ratios to apply to the savings estimates based on answers to this question. The long-tested, traditional method is to keep these two questions separated. For the baseline, it uses a few simple cases to identify how much of the total change occurring is within the program’s scope to influence: Replace on Burnout; Early Replacement, Discretionary Replacement, or New Construction, with the applicability of codes and standards considered for each project. Then the NTG analysis takes care of the question of whether the customer would have made that amount of change without the program (which is what the” industry

			standard" concept addresses). The double-counting problem was compounded by wrong identification of the base case, not using the "as-is" condition (Discretionary Replacement) when it was the appropriate one. The study classifies far too many cases, involving industrialized processes and highly costly equipment, as "Replace on Burnout" cases, rather than "Discretionary Replacement" cases. In practice, older equipment is used long after typical estimated life, especially in challenging economic conditions. Common example: Since a customer was contemplating extending the life of equipment through maintenance and repairs (e.g., re-winding motors) rather than replacement, the "as-is" situation should be considered a valid baseline. In certain evaluations, data was not collected early enough in the program cycle to provide a realistic baseline. As noted in the *Compact Fluorescent Lamps Market Effects Draft Final Report*, baseline estimate studies were not conducted sufficiently early in the program cycle to identify quantifiable market effects that occurred early in a program's life. The lack of such baseline data, coupled with the rapid increase in CFL sales throughout the U.S. during the first part of the 2006-2008 program cycle and the more recent national downturn in sales, makes it extremely difficult for any program to claim or quantify savings from cumulative market effects induced by these programs alone.