Ew, Gross! Cleaning Up Gross Baselines

Michael W. Rufo, Itron Inc.

Abstract

This paper addresses key issues associated with recent discussions and debates over how to set gross baselines for estimation of energy efficiency program impacts and cost effectiveness. For much of the history of efficiency programs, baselines used to estimate gross impacts for non-new construction programs were typically set as either replace-on-burnout (ROB) or retrofit (RET). Net impacts were then often estimated as a fraction, the net-to-gross ratio (NTGR), of impacts attributable to the program and multiplicatively applied to the gross impact estimate. In addition, gross impacts were usually calculated only for an initial, first year. Recently, increasing attention has been paid by evaluation analysts, regulators, and program administrators to refining and expanding the range of gross baseline methods to: a) include the use of a dual baseline approach alongside the ROB and RET designations; and b) to move from a code, or minimum efficiency, baseline to a "standard" or "common" practice definition for setting the baseline efficiency level for estimating gross impacts. In this paper, we provide: 1) a brief framing of the different approaches to setting gross baselines; 2) an explanation of the dual baseline approach; 3) analysis of some of the issues associated with using standard or common practice baselines; 4) results from a simulation model developed by the author to inform considerations for setting common practice baselines to be internally consistent with net impact estimation; and 5) conclusions and recommendations for selecting and applying gross baselines in combination with a NTGR.

Introduction

This paper addresses key issues associated with recent discussions and debates over how to set gross baselines for estimation of energy efficiency program impacts and cost effectiveness. For much of the history of efficiency programs, baselines used to estimate gross impacts for non-new construction programs were typically set as either replace-on-burnout (ROB) or retrofit (RET). For replace-on-burnout measures, baselines were often set at code, where code, such as appliance standards, was applicable, or at the minimum efficiency of available new equipment not subject to code or standards. Net impacts were then typically estimated as a fraction, the net-to-gross ratio (NTGR), of impacts attributable to the program and multiplicatively applied to the gross impact estimate. Gross impacts were usually calculated for an initial, first year. For cost effectiveness purposes, those same first-year impacts were usually held constant over the effective useful life (EUL) of the measure. This *first-year* equals *all years* approach to gross impact estimation is still commonplace in many jurisdictions.

Over the past five years or so, increasing attention has been paid by evaluation analysts, regulators, and program administrators to refining and expanding the range of gross baseline methods, in particular, by including the use of a dual baseline approach alongside the ROB and RET designations. Concomitantly, in applying the dual baseline approach, some jurisdictions have initiated efforts to enable changes in impacts over time, rather than continue the practice of holding first-year savings constant over the measure life. Another baseline-related change in some jurisdictions has been to move from a code, or minimum efficiency, baseline to a "standard" or "common" practice definition for setting the baseline efficiency level for estimating gross impacts. These changes require clarification of the process for selecting and applying each method based on a set of criteria, guidelines, and requirements. Application of dual baselines and adoption of common practice type baselines have also

engendered some re-thinking, clarification and debate over the definitional and market share relationships between gross and net impacts.

Traditional Gross Impact Baseline Approaches

With respect to the gross impact side of impact estimation, one of the most important issues, and one that has generated some debate of late, revolves around the determination of which baseline to use. This is important because energy efficiency impacts and cost effectiveness are usually estimated for the number of years associated with the life of the efficiency measures. Though there is often a tendency to focus on estimating "first year" savings (Collins and Loper, 2013), in practice such *first year* savings invariably find their way into *lifetime* impact and cost effectiveness estimates. For many years, dating back to the outset of voluntary programs in the US, baseline determination was considered to follow from the "event" associated with the installation or adoption of the energy efficiency measure. Initially, in the 1980s and into the early 1990s, "event" was typically characterized into one of three types: 1) replace-on-burnout (ROB), 2) retrofit (RET), and 3) new construction.

Replace-on-Burnout (ROB) – Code Baseline

In general, ROB measures were often, though not always, associated with capital-intensive equipment that end users typically only replaced at the end (or near the end) of the existing equipment's effective useful life (EUL).[1] Common examples of equipment in this category included room and packaged air conditioners, motors, furnaces, boilers, and chillers. Energy efficiency programs focused on moving end users from purchasing lower efficiency *new* equipment to higher efficiency *new* equipment. Incentives were often offered based on the savings estimated as the difference between the lower efficiency equipment's estimated energy usage and the higher efficiency equipment's estimated usage. Similarly, the cost of the efficiency upgrade was estimated as the cost difference between the lower and higher efficiency equipment.

Thus, under the ROB framework, savings, costs, and incentives were tied to the *incremental* difference between the lower and higher efficiency new equipment available in the market for those in the market due to the actual or imminent failure of their existing equipment. Importantly, savings, incentives, and costs were not compared to the existing equipment needing replacement since the circumstances dictated that the end user was in the market for a new piece of equipment regardless of the efficiency level. Estimating savings compared to the old equipment that had failed was irrelevant. What was relevant was the expected consumption of new equipment and it was from the availability and efficiency characteristics of *new* equipment that baselines were determined for gross savings estimation.

One of the first major accomplishments of energy efficiency policy in the US was the development and promulgation of mandatory codes and standards for energy efficiency equipment and building practices. These codes and standards were designed to significantly improve the *average* efficiency in the market and often did so by making the least efficient equipment *unavailable*. Consequently, by the late 1980s, for many types of capital equipment for which efficiency codes were in place, "code" was often used as the baseline for estimating savings for high efficiency equipment for ROB-related measures. In cases where code was not applicable, analysts tended to use minimum available efficiency levels as the baseline.

[1] "Effective" useful life is the term used to capture the fact that the real-world average lifetime of equipment is affected not just by the physical or technical life of the equipment, but also behaviors of end users and building owners that result in removal of functioning equipment for other reasons such as changes in space functionality and renovation (McRae, Rufo, & Baylon 1988).

Retrofit – Existing or "In Situ" Equipment Baseline

The next event type, *retrofit*, was first conceptualized to capture cases in which efficiency measures were either high efficiency equipment replacing existing *in situ* equipment that was *not* at the end of its life and cases in which an efficiency measure was essentially an *add-on* to existing equipment or equipment operation (e.g., an energy management control system). As it turned out, commercial lighting systems were an interesting case that analysts in the late 1980s and early 1990s often believed fell somewhere between relatively expensive, capital equipment turnover (i.e., ROB) and lower cost, add-on operational efficiency kinds of measures (i.e., RET). Because lighting equipment was relatively inexpensive and modular compared to other capital equipment like air conditioners, and because energy savings from early lighting efficiency improvements could rapidly pay back not just the incremental costs of the equipment but the *full* cost, end users and regulatory agencies (under the TRC test) were often able to financially justify replacement of fully functioning commercial lighting equipment with new efficient lighting under the RET case.

Partly as a result of these favorable economics, in both types of retrofit cases, the baseline from which gross savings were estimated was generally the consumption of the existing equipment. In such cases, savings were generally estimated over the life of the efficiency measure (e.g., over the life of a control system or other add-on measure). In the case of efficiency add-ons, this often seemed reasonable given that the existing equipment was continuing to operate and the life of the controls measures (often estimated to be on the order of 10 – 12 years) was considered to be close, on average, to the *remaining useful life* (RUL) of the key types of capital equipment being controlled.[2]

In the case of commercial lighting, while it was understood by many that the existing equipment would only be in place until the end of its estimated remaining useful life (RUL), several factors combined to result in the use of the existing equipment baseline (e.g., RET treatment) as the basis of estimated savings over the EUL of the replacement equipment even though that was longer than the RUL of the existing equipment. In theory, a dual baseline approach was arguably more accurate but favorable economics under the existing equipment baseline, and aversion to the perceived complexity of dual baseline calculations, may have contributed to most analysts defaulting to the existing equipment baseline and avoiding applying the more complex "dual baseline" approach (which is discussed below).

New Construction/Major Renovation – Code Baseline

The third event type defined in the early days of efficiency analyses was new construction and major renovation. For these cases, codes and standards (C&S) were almost always the basis for the baseline, whenever C&S was applicable and required by law. In many cases, C&S for new construction could be based on an overall efficiency performance target (e.g., lighting watts per square foot), and met through different combinations of efficiency measures and building characteristics, or based on equipment specific efficiency requirements (e.g., energy factor). "Major renovation" was associated with new construction in cases in which the extent of the renovation triggered C&S requirements like those required for new construction.

[2] This may be partly due to the fact that much of the capital equipment to which efficiency controls in the commercial sector were applied have relatively long EULs, e.g., on the order of 15-25 years for motors, air conditioners, and chillers. Assume that the average chiller has an EUL of 25 years, at any given time the average age of a chiller across the population of all chillers is likely around half of the EUL (if chillers were installed relatively evenly over time), thus the average remaining useful life (RUL) of the population of chillers would be about half of the EUL, or around 13 years.

Standard/Common Practice and Dual Baselines

"Industry Standard" or "Common" Practice Baseline

In the early years of efficiency programs, setting ROB baselines at code was often reasonable partly because market average efficiency levels tended to not be greatly above code, at least for the first few years following promulgation of a new code. In addition, it was believed that application of a net-to-gross ratio, an estimate of the portion of equipment adoptions that were program-induced, to the code-based gross savings, produced the correct ultimate outcome metric of net program impact.

Although some analysts always grappled with the issue of whether code was the most accurate basis for estimating gross impacts for ROB type measures, this aspect of baseline assessment has received increasing attention over the past five to ten years. This shift was due to the view that switching from a code baseline to a baseline that was more related to typical, rather than the minimum efficiency, would be more accurate. This type of baseline has been referred to in different jurisdictions and papers as "industry standard", "standard", "common", or "typical" practice (SEE Action 2012; UMP 2014; CPUC 2013). For this paper, we use the term - "common practice baseline" (CPB).

Regardless of the moniker, an important issue associated with CPM concerns application of a net-to-gross ratio. As summarized in UMP 2014, some analysts and policy makers have shifted to a common practice baseline with the belief that it obviates the need for estimating and applying a net-to-gross ratio, that is, that use of the common practice baseline produces results that are more akin to net than gross impacts and that this determination is best made on an ex ante basis. Other analysts and jurisdictions have a different view, believing that application of ex post methods and net-to-gross ratios are still warranted but that care must be taken in how each piece of the estimation is done to avoid under- or over-counting net impacts, which we address in the second half of this paper.

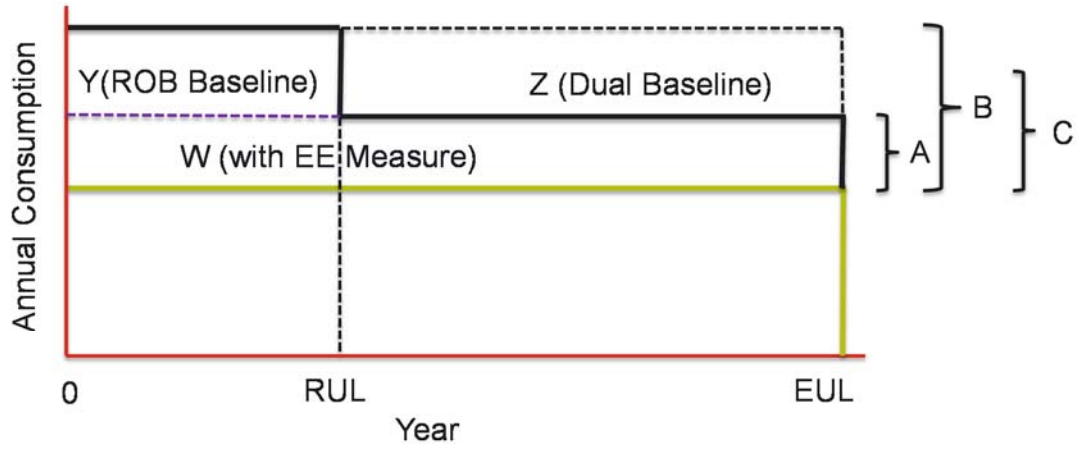
Another important issue regarding common practice baselines concerns spillover and market effects. If past programs have material spillover and market effects, these effects may contribute to higher common practice baselines than would otherwise be the case. Many jurisdictions do not include spillover and market effects in estimates of NTGR (Navigant 2013), meaning such NTGRs are more accurately characterized as net-of-free-ridership (NOFR) ratios. If these long-term market effects are not accounted for in attribution studies of past or current programs, than the cumulative long-term benefits of programs will be underestimated. This can occur with many different evaluation methods, including randomized control trials (RCT), not just two-step gross and net estimation processes. That said, it should also be noted that such *long-term* effects would not bias an estimate of the effect of a *current* program on *current* participants. That is, NOFR is a kind of "what have you done for me lately" metric which, though not the whole story, is important for ongoing improvement and assessment of a program (Rufo 2014). It is important that those responsible for funding and authorization decisions understand the distinctions between short-term and long-term assessments of net impacts.

"Program Induced" Early Replacement - Dual Baseline

As can be seen from the discussion above with respect to events treated as ROB or retrofit, there are cases in which neither of those approaches is adequately representative of the likely baseline of a particular efficiency measure or practice. For example, for measures that have an EUL that is longer than the average RUL of the existing equipment (almost always the case, on *average*), calculating lifetime impacts using the existing equipment as the baseline, over the number of years in the measure's EUL, is likely inaccurate given that the existing equipment would have to be replaced sometime during the length of time represented by the EUL; that is, at the end of the RUL period. To address such cases, some practitioners and policy makers have recently formalized and required use of a dual baseline

approach. In these cases, gross impacts are generally estimated using the existing equipment as the baseline for the RUL period of the existing equipment and code or common practice (see below) is used as the baseline for the remaining years. A central aspect of the dual baseline approach, and to triggering its application, is the question of whether the existing equipment was replaced before the end of its otherwise expected life due to the energy efficiency program intervention; that is, whether the early replacement was *program induced*.[3] Another condition that must be met is that the existing equipment had functional remaining useful life and, if so, how much, which may be difficult to estimate.

In Figure 1, we show a conceptualization of impact estimation using the dual baseline approach as compared to the ROB or retrofit approaches. In some jurisdictions, including California, the dual baseline approach was promulgated first for efficiency measures and projects found in calculated ("custom") incentive programs targeted at the nonresidential sector (CPUC, 2013). This was due to the fact that some capital intensive and complex projects associated with custom type programs were being claimed as pure retrofit projects. A retrofit treatment essentially equates to a claim that the program accelerated replacement of the existing equipment for a period equal to the EUL of the replacement equipment. While it is recognized that some pieces of industrial or commercial equipment may have significant RUL when replaced, the claim that most projects have RULs equal to EULs is difficult if not impossible to defend (one or the other must be wrong, on average, unless the program is inducing replacement of virtually brand new equipment). In the dual baseline approach, these become questions for analysis and evidence (see event treatment guidelines in CPUC 2013). The dual baseline approach has also been used for some lighting programs as they are more likely to have an early replacement influence (see Itron 2014b for an example of how the dual baseline approach can be successfully applied to a prescriptive commercial lighting program).



W = consumption with EE measure; X = consumption with in situ equipment; Y = consumption with standard practice new equipment; Z = consumption with in situ over RUL and ROB over post RUL period.

A = ROB savings, B = RET (in situ) savings; C = Dual baseline based savings (Z minus W)

If program induced early replacement, gross savings are based on Z (X over the RUL + Y over the post RUL period)

Figure 1. Conceptual Representation of Baseline Approaches

[3] We call "program-induced" the hidden adjective for "early replacement" baseline claims. This is a net-to-gross type of question having to do with whether the program changed the behavior of the end user to replace earlier then they would have absent the program. In theory, a two-part NTG may be required with each part fairly independent, that is, one part for the RUL period (did the program induce early equipment replacement) and one part for the post-RUL period (did the program induce a higher efficiency purchase than would have been the case at normal replacement).

Debate Over the Dual Baseline Approach

There have been several concerns and criticisms of the dual baseline approach. These generally fall into the following categories, to be discussed below: 1) results in lower lifetime savings, 2) is based on a hypothetical, 3) is too difficult to apply in program implementation and evaluation, and 4) is incompatible with tracking and cost effectiveness systems.

The first concern is that the dual baseline approach results in lower estimates of lifetime savings (MacCurdy et al. 2013). While this is generally true with respect to comparison with the retrofit baseline approach, it is not true with respect to comparison with ROB (the dual baseline approach has higher savings than ROB). The relative accuracy and defensibility of the baseline choice has to be determined on the merits, not based on what method produces greater impacts. For dual baseline, the merits are based on empirical information on measure life and end user decision making. Because of the empirical relationship between RULs and EULs, it is not mathematically possible to have a large, ongoing cohort of participants for whom the RUL is equal to the EUL. Even though estimation of the RUL is difficult, the alternative of using the retrofit approach of setting the baseline to existing equipment for the entire length of the EUL essentially says the RUL equals the EUL. The result of the (RET) approach is an estimate that is biased on the high side given that the program participants' RULs are likely to range from a few years to 1/3 or 1/2 of the EUL under relatively optimistic cases.

In addition, critics of the approach have argued that the dual baseline method is too "hypothetical" because it forecasts a baseline of code or common practice for the post-RUL period. Instead, these commenters suggest that the existing equipment baseline is "real" and appropriate because it *exists*. These arguments fail to recognize that what is being analyzed is not savings for the period *prior* to measure installation but savings for the period *after* measure installation for years into the future (as many years as the EUL). Using the existing conditions over the EUL is every bit as much of a *forecast* as using a dual or ROB baseline. In addition, this argument tends to ignore empirical evidence associated with the physical characteristics of the pre-existing equipment as well as information provided by end users indicating plans to replace equipment well before the end of the EUL period. Extrapolating existing equipment is more hypothetical than forecasting its replacement based on customer-specific evidence and empirical observations of average equipment turnover rates.

Lastly, critics have argued that the dual baseline method is too difficult to implement, whether through program implementation practices, evaluation studies, tracking systems, or cost effectiveness methods. While the difficulty argument may have had some traction in the early days of energy efficiency programs, for example, due to limitations of early computing systems and software, it is difficult to see how conducting net present value analysis by year or sub-period in today's computing and software environments is a barrier. That said, there are legacy tracking and cost effectiveness systems that were built on the premise that first-year savings from ROB or RET-based baseline assumptions can simply be multiplied over an EUL to obtain lifetime savings or copied for each year of the EUL to interact with avoided cost streams to produce a lifetime benefit-cost result. Frankly, such systems should have been modified years ago and, even in the worst case of their continuance, there are simple algebraic calculations that can produce approximate equivalences. For some program implementation and evaluation staff these can be difficult concepts that require additional work; however, in the long term, the training required to achieve successful field utilization of the dual baseline approach will likely pay off and lead to more sophisticated and influential efficiency analyses.

Cost Effectiveness and Dual Baseline Approach

Interestingly, there has been far less attention to the cost side of the dual baseline approach then the impact side. While concerns have been expressed about lower savings estimates, few have taken

note that the dual baseline approach almost always results in lower net present value costs than the retrofit approach (and, conversely, higher costs but higher savings as compared to ROB).

Managing the Gross and Net "Relationship" – Results from a Model

As discussed above, in the past, evaluators, analysts, and policy makers have often set gross baseline levels very low, at code or near minimum efficiency available, for the purpose of gross impact estimation prior to estimating a net-to-gross ratio, which was then applied to gross savings. In this approach, the belief was that the free ridership estimate imbedded in the NTGR would capture the portion of the market that would have adopted above code or above minimum efficiency in the absence of the program.[4]

Recently, some jurisdictions, including the California Public Utilities Commission (CPUC), have begun to set gross baselines at efficiency levels that reflect "standard" or "common" practice. For example, in the CPUC's evaluations, a NTG ratio is applied to gross impacts using a common practice baseline to produce overall net program impacts. At the same time, some other jurisdictions that have chosen to use a common practice-type baseline have considered forgoing application of an NTGR arguing that CPB already incorporate free riders and thus requires no further adjustment. Similarly, some analysts have questioned whether the application of a NTG ratio to an estimate of impact that uses a CPB over discounts or double counts for free ridership (UMP 2014).

Few authors have provided any calculations to support or refute positions in this debate over selection of CPB and application of NTGR. To address this gap, we have developed a relatively simply simulation. The question and issue driving the development of this simple model concerns the sensitivity of overall net program impact estimates to the level of efficiency chosen for estimating gross impacts via a common practice baseline. The purpose of the simple spreadsheet developed and summarized here is to try to simulate some simple hypothetical efficiency distribution levels, with and without programs, in order to see if the extent of any overlap or gap between gross and NTGR adjustments can be quantified or bounded under a set of reasonable conditions, or whether there is a clear or universally correct "answer" to the question of how optimally to set the gross baseline.

The problem is that any estimation of a program effect in the real world requires estimation, direct or implied, of a counterfactual, which, by definition, cannot be observed directly. However, in theory, it may be possible to simulate the distribution of efficiency levels, with and without the program (i.e., the counterfactual), for an entire market, in order to identify what percent of the market should be associated with the baseline used to estimate gross impacts. With this in mind, a simple spreadsheet model was developed to perform these market share and counterfactual simulations.

The analysis starts with a simple assertion: the bottom line, net impact of any program, is the difference in energy consumption and demand, for the entire eligible population, with and without the program intervention. (For the sake of simplicity, we exclude market effects and operational or rebound changes in this initial conceptual model.) Once a program is run one can no longer observe the counterfactual, in this case, what the distribution of efficiency levels would have been without the program: hence, the development of a simple simulation model to help illustrate the issue and possibly inform how to set CPBs when used with and without NTGRs.

We focus on energy consumption in explaining the model, following the example values and equations in Table 1. In the model, estimation of total market consumption, in the absence of the program, starts with a set of user-specified assumptions or estimates of the distribution of energy

[4] Some NTGR self-report methods included an adjustment for likely adoption of intermediate efficiency levels, in the absence of the program, that were higher than code or minimum but still lower than the program qualifying equipment.

efficiency levels in a hypothetical, no-program market. In the initial example, we divided the market into five efficiency bins and created market shares for each bin.

Although the entire analysis is conceptual, we use a commercial chiller market as the arbitrary example and express the efficiency metric as kW per ton of cooling capacity. The total number of units in the population, average size of the units, and average annual hours of operation are shown in fields (w), (x), and (z) in Table 1, but these are held constant across pre- and post- program scenarios and are not material to the overall analysis and results. Each efficiency level (b) is multiplied by the unit size, hours of operation, total population, and share of the market for that efficiency level (c), to produce the total consumption of the population by efficiency bin (l), the sum over which is the total consumption of the entire market. Consumption for the *participant* sub-population is calculated in (g). In this example, participants represent slightly more than half of the total annual ROB population, (y) as compared to (z). The results; however, described below, are not affected by the fraction of participants as a percent of the total market.

Table 1. Sample Inputs and Results from Gross Baseline, Net Impacts Model

Chiller Factors for Population Energy Calculations			
(w) Chiller Tons	(x) Chiller Hrs/Yr	(y) Participants	(z) Population
300	3000	105	200

(b') CPB Baseline =	**0.542**

		Counterfactual - No Program	**With Program**			
(a) Market Share Bins	(b) Efficiency Level (kW/ton)	(c) % of Market at Efficiency Level (Shares with *No Program*)	(d) % of Market at Efficiency Level (Shares *with Program*)	(e) % of Participants at Efficiency Level	(f) NTGR	(p) % of *No Program* that Participates in Program
(a)	(b)	(c)	(d)	(e)	(d - c)/(c*p+(d-c))	(p)
1	0.550	0.35	0.15	N/A	N/A	N/A
2	0.525	0.35	0.25	N/A	N/A	N/A
3	0.500	0.13	0.20	0.29	0.45	0.65
4	0.475	0.09	0.20	0.34	0.62	0.75
5	0.450	0.08	0.20	0.37	0.63	0.90
	Total or Average	1.00	1.00	1.00	0.57	

	Participant Population, "*Evaluation* -Based" Estimates					**Total Market Population, "Truth" in Simulation**		
	(g) Total Consumption for Participants	(h) Gross kWh Using Min. Eff. Baseline	(i) Net kWh Using Min. Eff Baseline	(j) Gross kWh Using CPB Baseline	(k) Net kWh Using CPB Baseline	(l) Total Market Consumption *No Program*	(m) Total Market Consumption *With Program*	(n) Net Impacts from Program - Delta Market Share
(a)	(w*x*d*e)	(See Notes)	(g*f)	(See Notes)	(j*f)	(w*x*z*b*c)	(w*x*z*b*d)	(m - l)
1	-	-	-	-	-	34,650,000	14,850,000	(19,800,000)
2	-	-	-	-	-	33,075,000	23,625,000	(9,450,000)
3	13,905,000	(1,390,500)	(630,000)	(1,168,020)	(529,200)	11,700,000	18,000,000	6,300,000
4	15,176,250	(2,396,250)	(1,485,000)	(2,140,650)	(1,326,600)	7,695,000	17,100,000	9,405,000
5	15,552,000	(3,456,000)	(2,160,000)	(3,179,520)	(1,987,200)	6,480,000	16,200,000	9,720,000
Total	44,633,250	(7,242,750)	(4,275,000)	(6,488,190)	(3,843,000)	93,600,000	89,775,000	(3,825,000)

Notes

(h) - This value is gross impacts for the entire participate population using the minimum efficiency as the baseline. First we calculate the consumption of each participant as (d*w*x) for each bin. Then we calculate gross impacts as the difference betweeen the consumption of the minimum efficiency baseline, in this case, the value associated with cell (b)1, which 0.55 in this example. Then we multiple these results by the participant population for each bin, that is, by (y*h). We calculate only positive values for efficiency levels of (b) that meet program-qualifying efficiency levels, in this example, all values above b2, that is, b3 through b5.

(j) - This value is gross impacts using the CPB baseline. Calculated the same as (i) except, instead of the efficiency level from cell a1, we use (b').

Similarly, the *with program* total market consumption can also be calculated. Market shares for each efficiency level under the *with program* case are set in (d) and the total market consumption

calculations are made in (m). The relationship between the *no program* and *with program* market shares are assumptions that the user can modify to create scenarios with greater or lesser degrees of program-induced shifts in the market. The total net impact of the program is then simply the difference between (m) and (l) and is shown in (n). In this hypothetical case, these calculations represent the "true" impact of the program to which the "estimated" impacts will be compared.

The next aspect of the model we describe focuses on the "estimated" net impacts, reflecting the two-step approach typical of evaluation in which one first calculates gross impacts and then estimates a NTGR and applies it to the gross impacts.[5] In our model, this process starts with estimation of the relative size of the program participants as a share of the total efficiency market (under the *with program* case). In (e), we input a set of assumptions about the share of the *no program* market that participates in the program by efficiency bin. Next these shares are used with the change in market shares to estimate the share of the *with program* market that participates in the program. This is made up of the sum of the share of the market that participated in the program that would have implemented the high efficiency measure anyway (p times c), plus the share of the market that is program induced (i.e., would not have implemented absent the program, namely, delta market share, which is d - c). Next we calculate the NTGR from these inputs by efficiency bin (f). NTGR equals delta market share (d - c), i.e., net participation, divided by gross participation, which is delta market share plus the portion of the counterfactual (c) that participates in the program (p), i.e., c times p (which represents "free riders").

As described in the notes of Table 1, we calculate gross consumption impacts relative to two baselines, minimum efficiency and "common" practice efficiency, respectively. Minimum efficiency is set at the lowest efficiency bin in the analysis, i.e., bin (1). Minimum efficiency could be associated with codes and standards or simply the least efficient new products available on the market. "Common practice" efficiency is user defined (b') and must be a value contained within the range of available efficiencies bounded by bins 1 through 5. This is the key feature of the simulation model: the ability to test the relative accuracy of different CPB baseline set points across a range of *no program* and *with program* efficiency market share scenarios and to solve for CPB values that produce the most accurate results when multiplied by an accurate NTGR and compared to the simulated "true" net impacts.

Gross program impacts are then calculated for each baseline, both the minimum and CPB, by efficiency bin for bins 3 through 5. Efficiency bins 1 and 2 are lower efficiency and are defined as non-program qualifying. Net impacts are then calculated as gross impacts times NTGR within each *program qualifying* efficiency bin. These simulated *evaluation*-based net kWh impacts, (h) through (k), can then be compared to the "true" value in the total row of (n). Note that this comparison is focused only on the accuracy of the gross impacts as related to the baseline choice in that the NTGR part of this hypothetical, model-based analysis is, by definition, perfectly accurate due to use of the simulation-derived value. Of course, in the real world, the NTGR also has measurement error, which could be random or systematically biased and cannot be directly estimated.

Now let's look at some results across several scenarios. One set of key metrics is shown in (r), (s), (t), and (u) in Table 2. Each of these compares the different evaluation-based estimates of impacts to the true net impacts (n, "Total"). Let's start with an example scenario, which we call Case 1. In Case 1, the market starts out weighted heavily to the lower efficiency shares, in the *no program* case (c), and is shifted out to the high efficiency side due to the program influence (d), with an associated NTGR (f) of 0.57. In row (r) for Case 1, we see that, if we use the ***minimum*** efficiency for the gross baseline and multiply by the correct NTGR, this introduces an error in the net result, in this case, an ***over-***

[5] We acknowledge that this two-step process, estimating gross and then estimating net and applying an NTGR, is used for many program evaluations but is, in theory, itself suboptimal in that one would often prefer an approach that goes directly to net impacts (e.g., via experimental design using random control trials (RCT)). Practical constraints on RCT result in the use of two-step estimation of net for many if not most program evaluations.

estimation of net impacts of 27%. In Case 1, we have solved for a common practice baseline that produces a gross impact result that, when multiplied by the correct NTGR, produces a net result that is equal to the "true" result. This is indicated by the result shown in (s) for Case 1 being equal to 100 percent. This CPB baseline value (b') is 0.534 kW/ton and falls between the minimum and second efficiency bin. For the *no program* case, roughly half of the cumulative market share (starting at the lowest efficiency bin and moving up), specifically, 0.54 in row (q), is less efficient than the CPB value and the remainder, 0.46, is more efficient. For the *with program* case, roughly 31 percent of the market is less efficient and 69 percent is more efficient than the CPB.

There could, of course, be a wide range of different input assumptions and scenarios across which the results from this simulation might vary. To test the sensitivity of results, we developed an initial set of 12 scenarios in which we varied the distribution of efficiency market shares. For each scenario, we solved for the CPB efficiency level that produced the correct gross and net savings when interacted with the "known" NTGR. In the remainder of Table 2, we present results from 4 of the 12 different scenarios (due to space constraints) we ran through our calculations.[6] The results are very sensitive to the initial *no program* distribution of efficiency levels and *slightly* sensitive to the *with program* efficiency distributions. Across the 12 scenarios tested, the average *no program* market share of the efficiency level that produced the correct results was roughly the *median*. This result varied widely for individual cases, from as low as the first quintile, to as high as 85 percent. It is important to note that the minimum efficiency level (a1) is rarely the correct answer for estimating gross impacts; in this model's results using the minimum baseline significantly overestimates "true" net impacts.

Table 2. Sample Results Across 4 Efficiency Distribution Scenarios

	Case 1				Case 2			Case 3			Case 4		
	(b') CPB Baseline =	0.534			0.525			0.542			0.526		
			No Program	With Program		No Program	With Program		No Program	With Program		No Program	With Program
	(a) Market Share Bins	(b) Efficiency Level (kW/ton)	(c) % of Market at Efficiency Level	(d) % of Market at Efficiency Level	Efficiency Level (kW/ton)	% of Market at Efficiency Level	% of Market at Efficiency Level	Efficiency Level (kW/ton)	% of Market at Efficiency Level	% of Market at Efficiency Level	Efficiency Level (kW/ton)	% of Market at Efficiency Level	% of Market at Efficiency Level
	1	0.550	0.25	0.15	0.550	0.16	0.15	0.550	0.35	0.15	0.550	0.05	0.02
	2	0.525	0.45	0.25	0.525	0.56	0.25	0.525	0.35	0.25	0.525	0.70	0.05
	3	0.500	0.13	0.20	0.500	0.13	0.20	0.500	0.13	0.20	0.500	0.05	0.05
	4	0.475	0.09	0.20	0.475	0.09	0.20	0.475	0.09	0.20	0.475	0.05	0.05
	5	0.450	0.08	0.20	0.450	0.06	0.20	0.450	0.08	0.20	0.450	0.15	0.83
		Total	1.00	1.00	Total	1.00	1.00	Total	1.00	1.00	Total	1.00	1.00
(q)	% of Market ≤ Baseline:		0.54	0.31		0.72	0.40		0.49	0.25		0.72	0.10
	NTGR =	0.57			0.61			0.57			0.80		
(r)	(Min Eff. Gross X NTGR)/"True" Net			127%			143%			112%			131%
(s)	(CPB Gross X NTGR)/"True" Net			100%			100%			100%			100%
(t)	Min Eff. Gross/"True" Net			215%			225%			189%			166%
(u)	CPB Gross/"True" Net			168%			152%			170%			125%

Some analysts may believe that use of a CPB baseline produces overlap with NTGR estimation. While picking a CPB that is too high, i.e., higher than the CPBs solved for in our simulation, would result in *underestimating* net impacts, using a CPB lower than those shown in Table 2 results in *overestimating* net impacts. Conversely, using the minimum efficiency rather than CPB *consistently overestimates* net impacts. Our conclusion is that use of CPB is more accurate generally than using a minimum, but that care must be taken to avoid using too high or low a CPB.

[6] Results for all scenarios, as well as the spreadsheet model, are available upon request from the author.

We also show in (t) and (u), how gross-only estimates of impacts compare with true net impacts. The results show that use of minimum baselines dramatically underestimates net impacts (as would be expected). The common practice baseline values in column (u) of Table 2 also significantly overestimate net impacts in the case where no NTGR is applied. However, in theory, the CPB can be set even higher such that it produces estimates equal to true net impacts, on average. To test this, we solved for CPBs to produce gross impacts equal to "true" net from the model. We found that the resulting CPBs came in at around three-quarters of the cumulative *no program* market share (from low to high).

Conclusions and Recommendations

For many years, two types of non-new construction baselines predominated in energy efficiency impact and cost effectiveness analyses: replace-on-burnout (ROB) and retrofit (RET). This approach was always overly simplistic and has needed expansion for some time. As argued in this paper, we believe the dual-baseline approach should be added to this choice set in that it is a useful, logical, and feasible method for estimating program-induced gross impacts for program-related market events that fall between the ROB and RET assumptions. Although the dual-baseline method does involve additional effort, the effort level is reasonable and appropriate given the increase in accuracy as compared to the binary ROB and RET approaches for certain measures and program intervention types.

Similarly, gross baselines have historically often been set at code or market minimums, while "common" or "industry standard" practice baselines are now becoming increasingly common. In theory, this emerging shift toward CPB is appropriate in that code and market minimums are often not the "right" baseline because, as shown in our results, they systematically overestimate net impacts when combined with NTGRs. That said, common practice baselines also can be chosen incorrectly and may over or underestimate net impacts, when combined with an NTGR, if the CPB selected is too low or too high. Moreover, where one selects common practice baselines depends on whether a NTGR ratio is subsequently estimated and applied at all. If no NTGR is going to be applied, then CPBs need to be set higher than would otherwise be the case to approach estimation of net impacts.

We found, using the model described herein, that the accuracy of two-step net impact estimation is sensitive to estimation of the baseline efficiency value. On average, over an initial range of simulations, we found that setting CPB at around the median of the *no program* market share produced a good estimate of net impacts when combined with a true estimate of the NTGR. Conversely, using the first quintile or market share associated with the minimum efficiency available, for the *no program* case, significantly overestimated true impacts when combined with an NTGR. Although the *no program* market cannot be directly and contemporaneously observed, pre-program or quasi-control group market share information can often serve as an informative and sometimes reliable proxy. In cases where only *with program* market share data is available, caution should be used as post-program market shares are generally not a reliable proxy, especially if the program represents a significant share of market activity.

From our review of baseline practices we offer the following recommendations:

- ***Expand use of the dual baseline approach both as part of evaluation and implementation practices***. There is a concomitant need for training of analysts to increase comfort with this approach and consistency in its application.
- ***Expand use of "common" (CPB) practice baseline definition*** as it is generally superior to defaulting to code or market minimums for many types of equipment installations.
- ***Jurisdictions should develop clear guidance on selection of appropriate gross baselines and associated analytical requirements*** and ground these in as much quantitative analysis of market share as possible.

Based on the results of our model, we offer the following *preliminary*[7] recommendations:

- When combining gross impact estimates with NTGR estimates to produce net impacts, in the absence of compelling information on the expected distribution of efficiency levels in the *no program* case, ***common practice baselines should be set at the median market share efficiency level (equivalent to the first 50 percent of the market*** in the ***no program*** case (from low to high)).
- If ***CPB is used to estimate net impacts without application of a NTGR***, then the estimated *no program* market share (via a proxy estimate) should be ***set even higher than the median***.

References

Collins, M. and J. Loper, 2013. "Kicking the Can: How First-Year Impact Evaluation Transfers Cost and Uncertainty," Proceedings of the 2013 International Energy Program Evaluation Conference, Chicago, IL

CPUC 2013. *Energy Efficiency Policy Manual, Appendix E: Custom Project Review Process*, Version 5, R.09-11-014, July.

Hall, N; Ladd, D.; Khawaja, M.S. (2013). "Setting Net Energy Impact Baselines: Building Reliable Evaluation Approaches." Proceedings of the 2013 International Energy Program Evaluation Conference, Chicago, IL.

Itron, Inc. (2014a). *2010-2012 WO33 Custom Net-to-Gross Final Report*. Prepared for the California Public Utilities Commission, Energy Division.

Itron, Inc. (2014b). *2010-2012 Nonresidential Downstream Lighting Impact Evaluation Report*. Prepared for the California Public Utilities Commission, Energy Division, September.

MacCurdy, Alex, et al., 2013. "Dual Baselines for Industrial Retrofits that Trigger Energy Codes," proceedings of the ACEEE Summer Study on Industry, August.

McRae, M., M. Rufo, and D. Bayon (1988). *Service Life of Energy Conservation Measures*, *ASHRAE Journal*, December 1988.

Navigant (2013). Custom Free Ridership and Participant Spillover Jurisdictional Review. Prepared for the Sub-Committee of the Ontario Technical Evaluation Committee, May.

Ridge, R.; Baker, M.; Hall, N.; Prahl, R.; Saxonis, W. (2013). "Gross Is Gross and Net Is Net: Simple, Right?" Proceedings of the 2013 International Energy Program Evaluation Conference, Chicago, IL.

Rufo, M. (2014). "Perspectives on Program Influence and Cost Effectiveness: Moving Forward from the Recent US Debates," Proceedings of the International Energy Policies and Programmes Evaluation Conference, Berlin, Germany. August.

SEE Action (2012). *Energy Efficiency Program Impact Evaluation Guide*. Prepared by Steven R. Schiller, Schiller Consulting, Inc.

UMP, 2014. *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. Chapter 23: Estimating Net Savings: Common Practices*, by Violette, D. and P. Rathbun. National Renewable Energy, Laboratory (NREL), NREL/SR-7A40-62678, September.

RTF (2012). *Guidelines for the Development and Maintenance of RTF Savings Estimation Methods*. NW Council, Released December 4. RTF 2012.

[7] Additional scenarios and vetting of the model are needed to further assess and develop these initial results.