NAVIGANT

Memorandum: Discussion of Selected NTG Estimation Issues

Prepared for:

Enbridge Gas Distribution, Inc.

and

Union Gas Limited

Prepared by:
Navigant Consulting, Inc.
1375 Walnut Street
Suite 200
Boulder, CO 80302
303.728.2500

and

Apex Analytics, LLC
2500 30th Street, Suite 207
Boulder, CO 80302
303.590.9888

December 14, 2017

The Navigant team submitted the following memorandum to Deborah Bullock (Enbridge Gas) and Leslie Kulperger (Union Gas) discussing select NTG estimation issues. The memorandum covers the following:

NAVIGANT

Memorandum

To: Deborah Bullock (Enbridge Gas) and Leslie Kulperger (Union Gas)

From: Dan Violette (Navigant) and Scott Dimetrosky (Apex Analytics)

Date: December 13, 2017

Re: Discussion of Selected NTG Estimation Issues

This memorandum represents the Navigant team's perspective on a set of issues beyond the information gathered for the Jurisdictional Review report.[1] A discussion of six issues was requested:

1. Net-to-gross (NTG) scoring process used in the DNV *Custom Savings Verification and Free-ridership Evaluation* report prepared for the Ontario Energy Board (dated October 12, 2017)
2. Context around the use of NTG estimates given challenges in estimation
3. Best practices in the use of self-report survey methods
4. Addressing attribution when there are multiple programs operated concurrently by different entities (e.g., gas and electric utilities)
5. Baseline issues and their relationship with net savings estimations
6. Tailoring attribution studies to specific programs to appropriately address program objectives and delivery considerations

Each of these issues are addressed in the following sections.

Issue 1: NTG Scoring Process

The scoring process refers to the approaches used to translate survey responses into NTG values. Each survey respondent has an individual NTG value developed for them that is consistent with their responses to a set of questions. This makes it important to ask questions that respondents can understand and respond to with some degree of accuracy. Overly difficult questions or questions that require a level of precision in their answers that cannot reasonably be provided by program participants complicates the translation of survey responses into NTG estimates. Judgment is required in both developing the question batteries and the algorithms used to translate responses into NTG values.

The scoring algorithm is central to any resulting NTG estimates. As a result, it is important that the algorithms be as transparent as possible and undergo a stakeholder review process to build confidence in the approach. A process that allows for discussion of the scoring algorithms, includes sensitivity analyses to assess robustness, and is as transparent as possible is important for producing NTG values that will have buy-in from stakeholders.

[1] Navigant Consulting, Inc., *Net-to-Gross Policies: Cross-Cutting Jurisdictional Review*. Submitted to Enbridge Gas and Union Gas, December 14, 2017.

1.1 Ontario Free Ridership Evaluation Questions

DNV uses a methodology termed Life Cycle Net Savings (LCNS) to determine the free ridership component of NTG.[2] DNV indicates that "the treatment of timing is how LCNS differs from other estimation approaches for attribution." This overall approach uses three attribution parameters:

1. **Timing:** Did the program accelerate implementation of a measure or cause it to be implemented before it would have been without the program?
2. **Efficiency:** Did the program increase the efficiency of a measure above what would have been installed in the absence of the program?
3. **Quantity:** Did the program increase the quantity of a measure above what would have been installed in the absence of the program?

It is common for scoring algorithms to use more than one attribution parameter or influence factor. The method used by DNV focuses on program influence factors, but some algorithms will also include some non-program factors such as previous experience with the measure and organizational policy or guidelines.

This section focuses on the questions used to address the timing influence parameter in the DNV for brevity; however, a review of question wording and design for most any attribution parameter would have similar issues and considerations.

Timing – Questions Review

It is useful for reviewers to consider the questions asked in the telephone survey and their form. To assess the timing attribution parameter, DNV asked four questions of participants, as shown below:

LEAD IN -- Now I would like to get into some specifics of the <project_n>. I would first like to know about what effect, *if any*, that the <utility> <program> program had on your decision to perform the actions in that project *when you did.*

DTA1a Without the <utility> <program> program, would you have performed the <project_n> at the …

1	**Same time**	DAT1a
2	**Earlier**	
3	**Later**	
4	**Or Never?**	
98	Don't Know	
99	Refused	

[2] This is described in the *2015 Natural Gas Demand Side Management Custom Savings Verification and Free-ridership Evaluation* report prepared by DNV GL for the Ontario Energy Board. October 12, 2017.

DAT1a_O. Why do you say that? [RECORD VERBATIM]

77	Record Response	IF DAT1a = NEVER, SKIP TO DAT1c] [ELSE IF DAT1a ≠ LATER, SKIP TO DAT2a]
98	Don't Know	
99	Refused	

DAT1b. Approximately how many months later?

[Try to get a number. Try bracketing if necessary by beginning with more or less than four years later.]

1	Record Number of months	
98	Don't Know	
99	Refused	

DAT1c. How old was that equipment?

[Get age at time of replacement. If they cannot provide exact age, ask for year installed and calculate age.]

1	Record Age	DAT2a
98	Don't Know	
99	Refused	

These questions have a logical flow, and the question *DAT1a_O* about why a respondent selected the answer to the initial question to be "same time, earlier, later or never" provides additional insight into the participant's response. DNV provided the full set of verbatim responses, which helps in transparency and in assessing whether the questions were understood by the respondents.

This bank of questions is aligned with approaches used in attribution studies across the industry, and there is no single right way to ask these questions. Yet, there are some judgments made in the way questions are structured and how the responses are used in the scoring algorithm. As a result, it can be important for stakeholders to review the underlying judgments and assumptions to build confidence in the overall process.

Timing – Judgement in Questions and Scoring

One example of judgment in the survey question development and scoring is the DNV 48-month cutoff for partial free ridership. To illustrate, if the answer to question *DAT1a* is "later," then question *DAT2b* asks the respondent how many months later they would have performed the project. This is meant to get an estimate of when—if the program had not existed—the participant would have performed or installed the project. The person conducting the survey is directed to "try to get a number" and, if helpful, try bracketing by asking if the project would have been performed more than 4 years later (i.e., 48 months) or less than 4 years later. An approach like this can be used to help the respondent begin to think through their answer. Also, in the algorithm used by DNV, any project that would have been undertaken more than 4 years later is given an NTG value of 100% (i.e., zero free ridership). For estimates of less than 48 months, a partial attribution credit is developed, which produces an NTG less than 100%.

The use of 48 months as the cutoff is an informed judgment made by the DNV evaluation team. The importance of this assumption can be addressed through sensitivity analysis. For example, alternative cutoff values could be considered. For example, if respondents are saying the project would be delayed by more than 3 years (i.e., 36 months), would it be reasonable to assume that they are not free riders, as this may be a speculative response? Using 36 months as the cutoff in the scoring algorithm instead of 48 months would provide information to stakeholders about the importance of this selected cutoff value.

Timing – Question Design Review

A review of the timing questions involves thinking though how well participants are able to provide these values. DNV states that timing was the most important attribution factor for many of the programs evaluated. With respect to Union's Custom C&I program, DNV states, "Timing was the component most strongly affected by the program. The program affected the timing of projects that account for approximately half of the energy saving" (p.35). Given this context, it is important to consider whether respondents can provide accurate estimates of when they would have performed their recent project if the utility program had not existed and if this information can be provided at a monthly level. Is it reasonable to distinguish between 6 months, 9 months, or 12 months later?

Looking at some of the verbatim responses can give a reviewer a sense of how accurately respondents might be able to assign the month in which they would have performed the project undertaken as part of the utility-offered program. A few of the verbatim responses from participants that said they would have performed the project at a later date are shown below (from Table 8-9: Timing Verbatim Responses Union Custom C&I programs, p.C-2):

- At some point in time we would have learned the value of this and done it.
- It's one of those things that you put on a list and OK, we'll do it sometime, but it might be 5 years or 3 years. Hard to say.
- May have done it the next year without incentives. Hard to say if upper management would have approved.
- Probably would never have done it; if so, maybe a couple of years.
- They will wait to replace something until they really have to, unless it's a health and safety issue.
- It would have taken longer to get approval.
- Probably would never have done it; if so, maybe a couple of years.
- They would've had to do these eventually.
- Tough question - It's possible that we just would have done nothing at all. Maybe fewer if we did.

These verbatim responses would seem to imply that it might be difficult for these respondents to determine the month in which they would have performed the project implemented under the utility

program if the program had not been offered. This is not a criticism of the DNV approach, but it does show how stakeholder review may be helpful in assessing question form, the use of the monthly time period, and the resulting NTG value. For additional discussion of the effect of this response on the NTG result, see Section 1.2.

Timing – Potential for Response Bias

One concern when questioning a program participant about whether they would have undertaken the energy efficiency (EE) investment even if the program had not been offered is the social desirability bias: yes, they would have undertaken the EE investment in the absence of the program, as they may want to view themselves as a good citizen with an understanding of the benefits of EE investments. This translates into concern that respondents may be overly confident about how quickly they would have undertaken the project in the absence of the program.[3] As a result, statements about future respondent EE actions in the absence of the program may be biased high and the NTG from this set of timing questions may be biased low.

Given this concern, it might be useful to conduct sensitivity analysis to test the difference in the NTG values that would result if participant estimates were overly optimistic. One scenario might be that participants are overly confident in the actions they would have taken by 50%. In this scenario, a sensitivity analysis would be conducted by changing responses of 3 months later to 6 months later, and if the response was 6 months, the sensitivity would use 12 months in the NTG calculation. This analysis would provide information on how sensitive the overall NTG values are to response errors in the timing question. This would allow stakeholders to comment on whether is it reasonable to consider an outcome where the respondents were overly confident about their undertaking the project in the absence of a program.

Reviewing potential response bias and the role of judgment in the survey design and scoring algorithm is not meant to be a critique of DNV's approach, as developing and using counterfactual "what if" questions are difficult in all EE attribution studies. Additionally, this problem is not unique to EE evaluation. These counterfactual scenarios must be addressed in any evaluation performed across a wide number of fields including the evaluation of business management decisions (e.g., the benefits and costs of offering a benefits package to employees), health programs (e.g., a school lunch program), and any other evaluations of policies and investments.

Timing – Early Retirement

The fourth question in the battery of timing questions asks about the age of existing equipment, which may also be important in the NTG scoring as part of DNV's LCNS methodology. The question on equipment age is:

[3] The New York State Department of Public Service, *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approach* states, "Often a series of survey questions are asked of the participant about the actions they would have taken if there had been no program to derive a free ridership estimate. More specifically, this is asking the respondent to state their intentions with respect to purchasing the relevant equipment absent the program. Bias creeps in because people may intend many things that they do not eventually accomplish." Link:
https://www3.dps.ny.gov/W/PSCWeb.nsf/96f0fec0b45a3c6485257688006a701a/766a83dce56eca35852576da006d79a7/$FILE/NY_Eval_Guidance_Aug_2013.pdf

DAT1c. How old was that equipment?
[Get age at time of replacement. If they cannot provide exact age, ask for year installed and calculate age.]

The response to question *DAT1c* above is used in conjunction with an estimate of the remaining useful life (RUL) of equipment replaced to allow for the use of different efficiency baselines for equipment that is viewed as being replaced early versus equipment installed on failure. This is a component of DNV's LCNS methodology, which is presented in detail in Appendix J of the DNV report. In general terms, these baselines are:

- If it is an early replacement, the baseline for the years of RUL is the current efficiency levels of the replaced equipment. After that, the efficiency baseline is the industry standard efficiency for new equipment that is installed at that time.
- If the equipment is replaced on failure, then the industry standard efficiency is used for the expected life of the equipment.

These calculations require information on the age of the equipment replaced by the EE project (as provided by the respondents), estimates on the useful life of the equipment, and the expected life of the new equipment installed. Can the respondents provide the age of the equipment or the installation date of replaced equipment? This is probably true, but there may still be uncertainty and assumptions in this lifecycle savings estimate.[4]

1.2 Comparison of Scoring and Attribution between Ontario and Massachusetts

The questions used in the Ontario Free-ridership Evaluation are very similar to those used in Massachusetts.[5] The similarity of the questions allows for a comparison of calculated NTG values for the same responses to specific questions across the two jurisdictions.

DNV provides examples of attribution calculations that can be used in this comparison. Table 8-86 in the DNV report provides several examples of how survey responses are translated into an NTG ratio. The first row of Table 8-86 (reproduced below) states that the effect of project "acceleration only" with a participant response indicating that they would have undertaken the project 24 months later if the utility program had not been offered results in an NTG of 31%.
If the "months later" response was 48 months, DNV's algorithm assigns an NTG of 100% (i.e., free ridership is zero), as this is the cutoff value. In this example attribution calculation in Table 8-86, the "months later" is 24 months, a value that is one-half of the value that would produce an NTG of 100%. These results are summarized in the table below:

Months Later Response	NTG Score	Free Rider Score
24 months later	31%	69%

[4] Additionally, it is worth noting that the DNV approach applies early replacement credit only to those participants who cite the program for accelerating the timing of their investment. In some jurisdictions, such as California, this type of adjustment is handled in the gross savings realization rate estimation and does not require the program to influence project timing in order to claim early replacement credit[4]. Therefore, when net savings factors account for early replacement, it is necessary to ensure that double counting does not occur when estimating the gross savings realization rates. Although we assume this did not happen in the DNV study, Appendix J is not explicit in whether early replacement savings adjustments are accounted for in gross realization rates and if so, how DNV ensures that double counting of the gross realization rates and NTG adjustments is not occurring.

[5] This is not unexpected at DNV is working on the customer C&I NTG studies in both Ontario and Massachusetts.

48 months later	100%	0%

The table above shows that a “24 months later” response produces an NTG value of 31% (e.g., a free rider value of 69%). This is a non-linear relationship—the “24 months later” response is halfway to the “48 months later” response, but the difference in NTG scores is much greater (100% compared to 31%). Additionally, the free ridership scores differ quite a bit as well – 0% to 69%. The reason why this is not a linear relationship is likely due to the way DNV addresses early replacement within its LCNS methodology. However, there is not enough information presented in the report to replicate this value.[6]

[6] It could be the case that DNV presented these calculations in workshops in Ontario, as the Navigant team was not involved throughout the entire NTG stakeholder process. It is not uncommon for these calculations to be presented in greater detail in workshops, but the report does not contain the information necessary to track how this example attribution was made.

Table 8-86. Ontario Attribution Examples[7]

Example	Timing Response		Efficiency Response		Quantity Response		VGS_E	VGS_S	$Y_{V.RUL}$	$Y_{V.EUL}$	VGS_L	Y_A	E	Q	SPA	NS_L	NTG
Accl only	Later	Two Years	Same		Same		100	50	3	10	650	2	0%	0%	0%	200	**31%**
"Never" for timing	Never		Same		Same		100	50	3	10	650	3	0%	0%	100%	650	**100%**
No attribution	Same		Same		Same		100	50	3	10	650	0	0%	0%	0%	0	**0%**
Accl with partial efficiency	Later	Two Years	Less	Between	Same		100	50	3	10	650	2	50%	0%	50%	400	**62%**
"Never" with partial eff.	Never		Less	Between	Same		100	50	3	10	650	3	50%	0%	100%	650	**100%**
Partial eff. only	Same		Less	Between	Same		100	50	3	10	650	0	50%	0%	50%	250	**38%**
Accl with partial eff. and partial quantity	Later	Two Years	Less	Between	Less	Half	100	50	3	10	650	2	50%	50%	75%	500	**77%**
"Never" with partial eff. and partial quantity	Never		Less	Between	Less	Half	100	50	3	10	650	3	50%	50%	100%	650	**100%**
Partial efficiency and partial quantity	Same		Less	Between	Less	Half	100	50	3	10	650	0	50%	50%	75%	375	**58%**
"None" is equal to "Never"	Same		Same		None		100	50	3	10	650	3	100%	0%	100%	650	**100%**
Full eff. credit, no accel. or quantity (ER)	Same		Less	Standard	Same		100	50	3	10	650	0	0%	100%	100%	500	**77%**
Full eff. credit, no accel. or quantity (non-ER)	Same		Less	Standard	Same		0	50	0	10	500	0	0%	100%	100%	500	**100%**

[7] Appendix K, Table 8-86, *2015 Natural Gas Demand Side Management Custom Savings Verification and Free-ridership Evaluation*. Prepared by DNV GL for the Ontario Energy Board. October 12, 2017

Because the same questions are asked in Massachusetts, it is possible to compare NTG and free rider scores for the same responses. For the same 24 months later response, the resulting NTG value is higher in Massachusetts—i.e., a 50% NTG in Massachusetts compared to 31% in Ontario. This a 60% increase in the NTG value (i.e., 50% NTG/31% NTG) due to different scoring algorithms. The reasons for this difference cannot be diagnosed in this review due to incomplete information, and there are no judgments about one NTG score being more appropriate than another NTG score; however, it is something that stakeholders in both jurisdictions might want to better understand.

A second attribution example calculation from Table 8-86 using the efficiency attribution parameter also demonstrates the role of judgment in the NTG scoring algorithm. Table 8-86 indicates that a program participant that states they were influenced to:

1. Install the high efficiency unit through the utility program, but
2. In the absence of the program, they would have installed a unit that is:
    a. Higher than a standard efficiency unit; but,
    b. Lower than the efficiency of the program unit they installed.

These participants also state that they would have installed the equipment at the same time and in the same quantity. These responses put them in the partial free rider classification.[8]

Given the responses set out above, the NTG value assigned by the DNV algorithm in Ontario is 38% (from Table 8-86). The same responses in Massachusetts would produce an NTG value of 50%. In terms of free ridership, Massachusetts would have a free ridership score of 50% compared to a 62% free ridership score in Ontario. This is another example of how differences in the scoring algorithms can influence NTG values, even if the questions are very similar.

As a note, there are additional factors accounted for in Massachusetts that influence the NTG scores. The same question on efficiency partial free ridership is currently included in the Massachusetts algorithm; however, Massachusetts is considering removing this question because including an intermediate efficiency response in the NTG algorithm and also through the use of industry standard practice (ISP) baselines in estimating gross savings could result in double counting the efficiency penalty.[9]

Recent research in a forthcoming northeast US study has shown the responses to these partial free rider questions can be highly variable.[10] For example, when those respondents that selected the response option of “between standard efficiency and what you purchased” were further asked what they would have purchased, a number of respondents said they would have purchased equipment that met code; or, the least expensive option. This invalidates their prior response that they would have purchased intermediate efficiency levels. In addition, the most common response to these questions was “don’t know” or “we weren’t considering additional efficiency levels.” This shows a potential the lack of consistency in responses to this question on partial free ridership. It also demonstrates the value of using consistency check questions in the survey.

[8] As a note, there may be concerns about how respondents understand what the standard efficiency baseline represents in this question, particularly if they also respond that they would have taken this action 24 months later.

[9] See the *Massachusetts Commercial/Industrial Baseline Framework.* prepared by ERS and DNV GL for the Massachusetts Program Administrators, April 26, 2017 (http://ma-eeac.org/wordpress/wp-content/uploads/MA-Commercial-and-Industrial-Baseline-Framework-1.pdf)

[10] Tetra Tech. *2016 Commercial and Industrial Programs Free-Ridership and Spillover Study* (Draft), Prepared for National Grid Rhode Island, September 2017.

Another example of where judgment plays a role is in the assignment of NTG values to question responses concerning partial free ridership. When a participant responds with the answer that they would have purchased "between standard and efficient equipment" in absence of the program, the Ontario algorithm decreases the efficiency attribution by 50%.[11] However, the actual decrease in efficiency might be higher (e.g., 90%) or lower (e.g., 10%) depending on equipment specifications and the resulting savings relative to the efficiency of the program-installed unit. The effect of this 50% assumption could be examined through sensitivity analyses. In addition, surveys on a test sub-sample of participants can try to get more refined estimates that can be used to inform the scoring algorithm.

It should be noted that while there is evaluator judgement embedded in the Ontario methodology, this is also true for most all other methodologies for NTG scoring. The question for stakeholders is whether the expert evaluator judgment seem reasonable, and whether the NTG values are stable and robust across a reasonable range of assumed values.

1.3 California and Illinois NTG Scoring Examples

The California and Illinois NTG survey methods use similar NTG questions, but they are based on an alternative approach to the one used in Ontario and Massachusetts. These NTG values are based on three program attribution components, or indices:

1. **Program attribution index 1 score (PAI-1)** reflects the influence of the most important of the various program and program-related elements in the customer's decision to select the specific program measure at the time.
2. **Program attribution index 2 score (PAI-2)** captures the perceived importance of the program (whether rebate, recommendation, training, or other program intervention) relative to non-program factors in the decision to implement the specific measure that was eventually adopted or installed. The program influence score is adjusted (i.e., divided by 2) if respondents said that they had already made their decision to install the specific program-qualifying measure before they learned about the program.
3. **Program attribution index 3 score (PAI–3)** captures the likelihood of various actions the customer might have taken at the time and in the future if the program had not been available (the counterfactual).

The algorithms in California and Illinois are not very comparable to the one used in Ontario. However, in contrast to Ontario, the California and Illinois scoring algorithms take the maximum of the timing or efficiency scores for the PAI–3 score, meaning the program only has to influence either timing or efficiency to receive the full value of that score. This tends to produce higher NTG scores and lower free ridership scores. This free ridership score is also tempered by the PAI–1 and PAI–2 scores that address a wider range of influence factors (both program and non-program factors) that were influential in their decision-making process.

Issue 2: Context Around the Use of NTG Estimates

This discussion examines the statistical error in NTG estimates and its relationship to overall uncertainty in the NTG estimates. Statistical error addresses one aspect of the uncertainty in NTG estimates, but there are a number of other important contributing factors. A study can produce a point estimate of NTG with a high level of confidence and precision, but the overall uncertainty in the estimate may be far

[11] DNV GL. *2015 Natural Gas Demand Side Management Custom Savings Verification and Free-ridership Evaluation*. Prepared for the Ontario Energy Board. October 12, 2017. Appendix K, Table 8-84.

greater. Stakeholders should understand the uncertainty in the NTG estimates when they are considering how to these estimates for planning and potentially when calculating shareholder incentives.

2.1 Statistical Error in Survey Methods

The context of NTG estimates involves understanding the representations of confidence and precision accompanying these estimates. DNV has a good explanation in a section titled “Understanding Statistical Error” on page 18 of the report. DNV designed the samples used in the studies to target 10% relative precision with 90% confidence (90/10) based on the best available assumptions at the start of the evaluation.

These statistical criteria may seem too precise given the discussion of issues in scoring algorithms presented in Issue 1 above. Scoring algorithms may be based on questions that can be difficult for program participants to answer, and evaluator judgment is used in the construction of NTG scoring algorithms. The crux of the issue is that there is a difference between uncertainty[12] and statistical error.

Statistical error in the context of self-report survey findings represents a somewhat narrow concept. Usually a sample of participants is selected for the survey as it would be too expensive to survey the entire population of participants. Statistical error in this context refers to the relationship between the estimates and findings from the survey, and what would have been obtained if the entire population had participated in the survey instead of only a sample. A sample NTG estimate of 80% obtained from the survey with a confidence of 90% and a relative precision of 10% indicates that if the entire population was surveyed (to eliminate sample error), there would be a 90% probability that the value obtained from the entire population would fall between 72% and 88%, i.e., it would fall within an interval of 80% +/- 8%. In this case, the confidence and precision only represents sampling error and does not capture any issues related to response bias, or judgments applied by evaluators (even expert evaluators) in the construction of the NTG scoring algorithm.

The result is that attribution studies can produce NTG values that have 90% confidence and 10% precision, but addressing statistical error only may not appropriately dimension the overall uncertainty in the NTG values. This can be due to (real or perceived) biases in the inputs to the NTG scoring algorithm—due to biases in the responses to NTG questions and in accuracies in evaluator judgment used in the scoring algorithms. The effect of these factors can be assessed using sensitivity analyses (as discussed above) or other simulation methods that show the potential impacts of alternative inputs to the NTG algorithms. This is an important component of any assessment of an NTG scoring algorithm and should be part of a stakeholder review of the NTG estimation process.

2.2 Assessing Reliability and Consistency of NTG Estimates

Assessing the overall robustness and consistency of the NTG scoring algorithm across a range of reasonable assumptions and scenarios should be a part of any overall evaluation process. The result will typically provide a range for the NTG estimates. The case studies in the Jurisdictional Review report indicated that all three states used a collaborative, transparent stakeholder process to finalize NTG estimates.

This review process can highlight issues with the survey questions and format. One issue that often comes up is whether the survey provided enough context and explanation of the program to remind the

[12] Definitions of uncertainty in the sciences can be a complex concept; however, a definition for uncertainty that is useful in this context can be found on Wikipedia: “The lack of certainty, a state of limited knowledge where it is impossible to exactly describe the existing state, a future outcome, or more than one possible outcome.”

participants responding to the survey of the program features, information, and potential influences. If a survey is conducted 1 year or more after participation in a program, the respondent may not recall all the features of the program and all the assistance provided. Instead, respondents may focus narrowly on the influence of the rebate or incentive payment.[13] It is also difficult for respondents to isolate the influence of one utility's programs from other influences if multiple programs are offered or if incentives are provided by more than one entity (e.g., incentives by a gas and an electric utility).

Sometimes a triangulation process is used where trade allies are surveyed to get their views on the influence of a program on selected participants or on the market as a whole. On occasion, trade allies can have different opinions on whether program participants would have really undertaken the actions offered through the program even if the program had not existed. Past industry experience has indicated that some participants may be overly optimistic regarding what they would have done on their own—i.e., in the absence of the program. Trade allies may be able to offer responses that are less prone to this bias and provide a means for adjustment.

The New York State DPS EM&V Guidelines states that "when multiple questions, weights, and complex algorithms are involved in calculating the NTG ratio, evaluators should also consider conducting a sensitivity analysis (e.g., changing weights, changing the questions used in estimating the NTG ratio, changing the probabilities assigned to different response categories, etc.) to assess the stability and possible bias of the estimated NTG ratio."[14] The NY DPS guidelines go on the state that "the onus is on the evaluator to demonstrate that the algorithm is not biased." However, the team would amend this to be the responsibilities of an overall stakeholder or evaluator review group in a jurisdiction. All parties have an interest in robust, stable results.

2.3 Promoting a Collaborative Process

An important component of NTG context depends on how the estimates are to be used. If the NTG estimates are to be used to calculate shareholder incentives, then the review and estimation process can become much more contentious. The potential for bias in responses to survey questions and from the survey frameworks (e.g., inadequate context on the overall program for respondents) can result in disagreements about exact NTG values or the fairness of the application of NTG estimates in an incentives calculation. The fact that some evaluator judgment is a part of every NTG scoring algorithm only compounds the discord.

In the Massachusetts case study,[15] the use of NTG retrospectively in incentive calculations was described as causing tension in the system, with significant disagreements over the NTG estimates. It was

---

[13] The New York State Department of Public Service, "*Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approach"* states, "This makes it essential that the interviewer guide the respondent through a process of establishing benchmarks against which to remember the events of interest. Failure to do so could well result in, among other things, the respondent 'telescoping' some events of interest to him into the period of interest to the evaluator. Set-up questions that set the mind of the respondent into the train of events that led to the installation, and that establish benchmarks, can minimize these problems." Link: https://www3.dps.ny.gov/W/PSCWeb.nsf/96f0fec0b45a3c6485257688006a701a/766a83dce56eca35852576da006d79a7/$FILE/NY_Eval_Guidance_Aug_2013.pdf

[14] The New York State Department of Public Service, *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approach* p. 10 Link: https://www3.dps.ny.gov/W/PSCWeb.nsf/96f0fec0b45a3c6485257688006a701a/766a83dce56eca35852576da006d79a7/$FILE/NY_Eval_Guidance_Aug_2013.pdf

[15] See Section 2 in Navigant Consulting, Inc., *Net-to-Gross Policies: Cross-Cutting Jurisdictional Review.* Submitted to Enbridge Gas and Union Gas, December 14, 2017.

described by one expert as “really, really bad in the past.” Another expert described NTG results to be “extremely negative as retrospective tool.” A third expert noted, with respect to incentives, that when program administrators (PAs)—typically utilities—are “losing money based on subjective studies, it gets ugly.” Therefore, in the most current cycle, Massachusetts adjusted its policies to apply agreed-upon NTG factors on a prospective basis. One best practice is to develop net savings and NTG approaches that are transparent, understood by the parties, and seek consensus where possible.

The case studies in the Jurisdictional Review contain examples where stakeholder groups have enhanced collaboration throughout the program planning, evaluation, and incentive calculation process. In all three states, stakeholders had the opportunity to question, challenge, and suggest modifications to the initial estimates produced by an EM&V study. One comment that was made was that because NTG “answers are not easy” there needs to be a process with “enough room for reasonable people to disagree.”

Another best practice is to use agreed-upon, pre-defined C&I NTG survey questions and algorithms. These are often viewed as common practice for that jurisdiction, but they are not strictly required or followed. This pre-defined approach allows NTG results to be compared over programs and time, and allows the algorithm to be updated, tweaked, and improved. This leads to a more sophisticated approach and increases stakeholder confidence, as it has been incrementally changing over time as part of a transparent review process.

Issue 3: Discussion of Best Practices

This section discusses some components of best practices in applying self-report surveys to C&I custom EE programs. This is a broad topic and too expansive to fully address in this memorandum. The Jurisdictional Review report, through the three state case studies, discusses a number of best practices. Given concerns with self-report methods, experts interviewed for that report noted multiple approaches that can be used improve the accuracy of self-report studies and thus promote confidence in the study findings. These same points are set out in the Jurisdictional Review report, but to provide the information in this document, key points are presented below:

1. **Fast feedback:** Fast feedback refers to survey methods where the respondents are asked about factors influencing their participation in a program at a time near to when they participated—e.g., within 3 months of completing participation. Experts noted the value in using fast feedback to gain the most accurate responses for free ridership, but it is not required in any state. A number of Illinois utilities use a parallel path evaluation approach for selected custom projects that allows for real-time NTG. In California, they are pre-screen custom projects with respect to an estimating initial NTG value to reduce risks of surprise NTG values when the full impact evaluation is performed. This two-step approach helps produce a “no surprise” approach that builds confidence in the NTG estimates.
2. **Sensitivity analysis:** Sensitivity analysis (with full transparency regarding the scoring) has been used in all three case study states, primarily when the pre-defined batteries are first developed and tested, but the algorithms are also periodically revisited. This can be important as different but reasonable assumptions used in translating question responses into NTG values can result in different NTG values.
3. **Triangulation:** The perspective of vendors is collected in all states for custom projects on a project-by-project basis (e.g., if the customer states the trade ally recommendation was important) and can increase the NTG result. Triangulation with vendors/trade ally surveys is also used to address the influence of factors that program participants may not be well positioned to address—e.g., the relative influence of multiple program influences on program impacts. As noted below, multiple experts noted the difficulty of participants understanding attribution of any

individual influence on their decision-making, as there are many potential influences in the EE marketplace.

4. **Other best practices*:*** Other best practices mentioned by experts include: including multiple factors in the NTG scoring (program influence and other non-program influences), ensuring the questions and weighting are fully vetted, consistency checking, and gaining insight into the project story by spending additional time with the participant to understand.

Information on best practices are contained in the references to documents for the three case study states (Massachusetts, California, and Illinois). Three other documents providing guidance on the use of self-report surveys for estimating net savings are:

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

2. New York State Department of Public Service, N*ew York Evaluation Plan Guidance for EEPS Program Administrators -- Appendix G- Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approach,* 2013. Updated August 2014. https://www3.dps.ny.gov/W/PSCWeb.nsf/96f0fec0b45a3c6485257688006a701a/766a83dce56eca35852576da006d79a7/$FILE/NY_Eval_Guidance_Aug_2013.pdf

3. Research Into Action. *Review and Analysis of Net-to-Gross Assessment Issues for Natural Gas Demand Side Management Custom C&I Programs.* Prepared for Enbridge Gas Distribution, Inc. August 25, 2017.

Issue 4: Addressing Attribution with Multiple Programs

The landscape of utility program evaluation is made more complex when there are multiple programs targeting the same customers. The Jurisdictional Review report addressed the issue of overlapping programs. In that report, experts reported that it is difficult for program participants to disentangle the influence of multiple programs (e.g., when more than one entity is providing incentives or information to encourage program participation), and they recommend viewing simultaneous programs as a single offering for free ridership purposes. However, this may be less than satisfactory to stakeholders when a utility has shareholder incentives and the goal is to have those incentives tied directly net savings attributable only to their programs.[16]

If utility-specific net savings estimates are viewed as necessary then one approach might be the use of a stakeholder review process to finalize NTG estimates parsed out for each utility. In the Jurisdictional Review report, all three case study states indicated that stakeholders had the opportunity to question, challenge, and suggest modifications to the initial NTG estimates produced by the evaluations. There was a recognition that all NTG estimation methods face challenges in application, and working toward agreed-upon NTG values informed by an NTG evaluation was worth the effort.

Another issue when dealing with the influence of multiple programs is the stacking of incentives. This occurs when a utility can help a customer not only obtain incentives from its program but also can help

[16] Problems in parsing out the influence of multiple programs is also discussed in: Research Into Action (2017). *Review and Analysis of Net-to-Gross Assessment Issues for Natural Gas Demand Side Management Custom C&I Programs.* Prepared for Enbridge Gas Distribution, Inc. August.

the customer qualify for incentives offered by other entities. When this occurs, it is difficult to parse out the influence of the utility that may have first worked with the customer. For example, one verbatim response to the timing question on Enbridge's C&I program indicated that Enbridge made the customer aware of other incentives. Estimating the partial contribution of one incentive over another may be difficult. The respondent may simply put the greatest weight on the largest incentive. However, the utility that initially interacted with the customer may have provided the impetus for program participation. There is a concern that some customers, when responding to influence questions, remember only the financial incentives and not the engagement, informational, and process elements derived from working with a utility to assess EE investments.

The issue of multiple programs and influences was addressed in a recent update to the New York State DPS EM&V Guidance, which stated that:

> *… some level of net savings assessment, or examination of program influence, can serve as an effective tool for program design and implementation. However, given the variety of activities occurring in the marketplace, including Commission direction for NYSERDA and utility offerings to become more complementary in nature, it will be increasingly more difficult to parse out the effects of any one specific program action.*[17]

Multiple programs and influences is an issue for other jurisdictions and regions as well. One regional study estimated total net savings from all the programs in a C&I sector, then a trade ally/market expert survey was conducted to allocate these savings to specific programs. In this case, the trade allies were judged as having a better perspective on the relative influence of individual programs than the customers would, as those customers might not even know how all of the programs interacted.[18]

Trying to parse out the influence of a single program in a market with multiple programs, incentives, and market influences is increasingly becoming a challenge. As was recognized in the New York updated guidance, it remains important to perform net savings assessments, but the limitations of the methods need to be considered. Good decisions do not require perfect information; rather, they require information that can be used to make good policy decisions regarding EE investments.

Issue 5: Baseline Issues in NTG

The determination of the baseline against which to measure energy savings is one of the most important aspects of an EE evaluation. One of the stronger aspects of the DNV report is its discussion of baselines in Appendix B and specifically which baseline is appropriate for various situations. This is particularly important in the verification of gross savings, which was conducted as part of the custom program savings verification (CPSV) analyses. It is also important for self-report survey design and implementation. In the efficiency attribution parameter NTG assignment (discussed in Section 1.2), the survey tries to get at partial free ridership by asking the following question:

> *Without <the program>, would you have installed <measure> that was "standard efficiency on the market at that time," or "between standard efficiency and the efficiency that*

[17] New York State DPS/Office of Clean Energy, *Evaluation, Measurement & Verification Guidance*, November 2016, p. 10. http://www3.dps.ny.gov/W/PSCWeb.nsf/96f0fec0b45a3c6485257688006a701a/255ea3546df802b585257e38005460f9/$FILE/CE-05-EMV%20Guidance%20Final%20%2011-1-2016.pdf

[18] A version of this approach was taken in Violette, D.; Ozog M.; Cooney, K. (2003), Retrospective Assessment of the Northwest Energy Efficiency Alliance -- Findings and Report. Prepared for: Northwest Energy Efficiency Alliance, Ad Hoc Retrospective Committee, December 8. See: http://www.theboc.info/pdf/EvalBOC_SummittBlue_NEEA_2003.pdf

you installed?

The participant can only answer this question if they know what the standard efficiency is for alternative equipment to compare to the higher efficiency equipment installed by the program, and if they need to know what the efficiency levels might be if they make the investment several years into the future (i.e., responded that they would have made the investment at a later date). There is sometimes a concern that customers are responding based on their view that any new equipment simply exceeds the efficiency of the existing equipment. In this case, they may be just moving from the efficiency of their existing equipment to standard efficiency and are not exceeding standard efficiency levels. This can be a complex question, as industry standard efficiencies may be higher than the efficiency of the lowest cost replacement equipment. Respondents may be overly confident about the likelihood that they would have installed higher than standard efficiency equipment in the absence of the program. This may result in an inaccurate estimate of partial free riders.

Making sure that the respondents are able to accurately project what investments they might have undertaken in the absence of the program often requires the use of setup questions that help respondents recall the sequence of past events, how these events affected their decision to adopt the measure, and awareness of equipment options that represent standard efficiency as opposed to the lowest cost replacement equipment.

Additionally, as noted in Issue 1 above, there are baseline considerations in the calculation of early replacement and whether that is approached through gross or net savings. Again, the only way to really address concerns about potential biases in the response to NTG questions that have a baseline assumed (e.g., installing equipment above standard practice) is to perform sensitivity analyses.

Issue 6: Tailoring Attribution Studies

It is important to tailor attribution studies to the specific programs being evaluated. Programs may have unique elements in terms of program design and delivery, which can affect net savings. Addressing these program design and implementation features may require tailored evaluation approaches.

Self-report survey methods have been most often used for incentive-driven C&I programs. However, there are a growing number of programs that are trying to move away from paying out large incentives. Instead, there is greater emphasis on engagement, information, and business case development—all of which support a more favorable environment for investments in EE. For these programs, it is often important that the survey introduces the ways support was provided through the program. This would include making sure that program training, analysis, and support are described to the participant. These can be particularly difficult for the respondent to recall if the survey takes place 1 year or more after participation. A program driven by financial incentives to induce participation has one major event (i.e., the payment of incentives) that the respondent can easily recall. More sophisticated programs that work to engage and support customers in making EE investments can require different survey designs to capture these non-incentive influence factors.

It can be important to work with program implementation managers and experts in the market (e.g., utility account managers and trade allies) to develop hypotheses that can be explored through the survey effort. These would include hypotheses regarding the different influence pathways used by a program to reach customers—particularly if there is a goal to move away from or better support customers and reduce the need for high cost incentives and rebates.

Finally, one of the best practices identified in the Jurisdictional Review report involved gaining insight into the project story by working through events that led to the installation of equipment. The goal is to have

participants talk about their project-related decision-making and what factors went into that process. This may be conducted for a smaller set of customers than the overall NTG sample, but these insights can be important when making judgments about attribution.