Opinion Dynamics

MEMORANDUM

TO: IL Stakeholders

FROM: Opinion Dynamics Corporation

DATE: August 8, 2014

RE: Alignment of NTG Evaluation within Illinois

This memo outlines first steps taken in an effort to align net-to-gross evaluation approaches within Illinois. As agreed in our conference call in mid-May, on behalf of the evaluators in Illinois, Opinion Dynamics agreed to perform the following tasks:

1) Systematically compare the questions and possible closed-ended responses between the DCEO and ComEd/Ameren survey battery (for those questions relating to setting up for the NTG battery as well as the questions used within the algorithms) to have a side-by-side comparison of the questions.

2) Obtain redacted information from a ComEd, Ameren, and DCEO survey with customer responses to specific questions from the first step as being part of the NTG "story". Analyze across the three sets of respondents and look at how both approaches deal with the various concepts of attribution.

Our analysis showed that the ComEd and Ameren surveys are virtually identical and use the same conceptual framework. Many of the questions in the DCEO survey are similar to those in the ComEd/Ameren surveys, but response categories are generally less nuanced and the framework to calculate a free rider (FR) value is different. Overall:

- For DCEO, use of a threshold question causes 6% of custom respondents and 9% of prescriptive respondents to have a FR value of 0 when they would have a higher value (0.33) based on their other responses. While a small percentage, we believe that the use of a threshold question should be reconsidered given the fact that there is contradictory information within a single survey. Alternatively, adding consistency checks (the DCEO survey does not contain any) could help eliminate/reduce inconsistencies.
- For ComEd/Ameren, the approach of averaging the three FR concepts may be creating FR values that are too high and should be reconsidered. For example, timing is considered in one of three components. If we learn that the customer would have installed the same efficiency and quantity, but they indicate this would have occurred in four years or more, it seems that this response has sufficient uncertainty around what would have occurred to mean that the customer at this point is not a free rider. The current algorithm does not take that into account.
- Due to the scalar choices made within the surveys, the distribution of final FR values for DCEO is "chunkier" (less smooth) than the ComEd/Ameren distribution. The group should

discuss the choice of numeric versus verbal scales,[1] including the range numeric scales (is 0 to 10 subject to the charge of false precision?) and how to assign FR values if verbal scales are chosen.

The remainder of this memo presents findings of the two tasks. The last section includes a few questions to start the discussion, among IL stakeholders, of standardizing the IL NTG approach. Additionally, our team has access to Dr. Richard Ridge, an expert in self-report methodology. Within the course of all of our evaluation activities, he serves as a quality assurance check. We sent him our original memo for comment and addressed many of them in this current document. He also provided us with a companion document around self-report that we include as an attachment herein.

Our analysis focused on FR questions for the Standard/Prescriptive and Custom programs.[2] For each utility/Program administrator, we analyzed data for PY5, which was the most recent survey data available.

The accompanying spreadsheet contains the full comparison of FR questions and additional comparative analyses, as summarized below. The spreadsheet contains the following tabs:

- **Algorithm Questions:** This tab provides a comprehensive list of FR-related questions and response categories for both surveys, aligning them side-by-side to facilitate comparison. It also provides a few observations that emerged during the review/comparison process. Questions are color-coded as follows:
  - Orange = Used in FR algorithm
  - Black = Set up question
  - Blue = FR-related question but not used in algorithm[3]
  - Purple = Question designed to be used in algorithm on case-by-case basis
- **Comparison of FR Ranges:** This tab compares three key FR questions (likelihood to install without program, timing of installation without the program, importance of program factors) and possible FR ranges for each response option. Based on the comparison of these questions and their impact on the overall FR score for a project, we identified a few questions for discussion (included in the last section of this memo).
- **DCEO Analysis:** This tab analyzes the effect of the threshold question about financial ability on project-level and program-level FR values. It also explores one aspect of potential inconsistency in the DCEO algorithm by cross-tabulating responses to the financial ability question with the question about the likelihood that the respondent would have installed the equipment without the program.

---

[1] The DCEO survey uses 4-point scales with each point assigned a label to aid the respondent. We call them verbal scales here to differentiate them from the ComEd/Ameren scales which asked the respondent to provide a numerical value.

[2] The original discussions included performing this analysis on the Retro-commissioning program as well. However, after considering the surveys performed recently, there were insufficient numbers of respondents for a rigorous analysis, and we excluded this program.

[3] In some cases, the information is not directly applied in the algorithm, but is considered in any adjustment of the NTGR based on this further information. In other cases, the questions are around consistency checks and play a role in the overall values that are included in the algorithm.

- **ComEd Analysis:** This tab looks at the distribution of FR scores for the three FR components, comparing their ranges and averages for different levels of overall FR. This analysis was done for the PY5 ComEd Prescriptive survey.
- **Ameren Analysis:** This tab contains the same analysis as the ComEd Analysis tab, only for the PY5 Ameren Custom survey.
- **DCEO Response Key:** This tab summarizes how the DCEO algorithm assigns overall FR scores based on the three components and the questions that go into each component.
- **ComEd Calculator:** This tab implements the project-level ComEd algorithm.
- **Ameren Calculator:** This tab implements the project-level Ameren algorithm.

Overview of FR Algorithms

Ameren/ComEd

The survey questions and FR algorithms for Ameren and ComEd's standard/prescriptive and custom programs are virtually identical. There are minor differences in question phrasing, and a few questions are specific to either Ameren or ComEd, or to either the Standard/Prescriptive survey or the Custom survey.

The FR algorithm consists of three FR components, which are averaged:

$$\text{FR} = \frac{\text{Program Components Score} \ + \ \text{Program Influence Score} \ + \ \text{No Program Score}}{3}$$

The three scores represent the following:

1. **Program Components Score**[4] reflects the importance of various program and program-related elements – e.g., availability of the incentive, recommendation from program staff, or program marketing – in the customer's decision to implement specific program measures. Greater importance of the program components means lower level of FR.

2. **Program Influence Score** reflects the degree of influence the program had on the customer's decision to install the specified measures, relative to other non-program factors. Greater importance of the program means lower level of FR.

3. **No Program Score** captures the likelihood of installing the exact same equipment if the program had not been available as well as the likely timing of the installation. A greater likelihood of installation without the program means a higher level of FR. Later implementation without the program means a lower level of FR.

Each of these three scores can take values between 0 and 10 where a higher score indicates a higher level of program attribution. These scores are then converted to FR values, which can range from 0 to 1.

DCEO

[4] Called "Timing and Selection Score" in the ComEd Prescriptive algorithm files.

The DCEO algorithm consists of four considerations:

1. **Financial ability** to install the equipment or measures without the financial incentive from the program;
2. **Plans and intentions** of firm to install a measure even without support from the program, including the level of efficiency and timing;
3. **Influence the program** had on the decision to install the measure;
4. **A firm's experience** with similar equipment and previous installation of energy efficient equipment without a program incentive.

If a participant indicated they would have been financially unable to install the equipment without the financial measure, a FR score of 0 is assigned to the project.

For decision makers who indicated that they were able to undertake energy efficiency projects without financial assistance from the program, rules are applied for each of the three latter factors to develop four binary indicator variables (yes/no) indicating whether or not a participant's behavior showed FR. For each participant, a FR value of 0%, 33%, 67%, or 100% is assigned based on the combination of the four indicator variables.

Comparison of Freeridership Questions

The ComEd/Ameren (considered as a single instrument here, except where noted) and DCEO survey instruments overlap on a number of key FR concepts. However, they often differ in the response options (e.g., 11-point scale versus a verbal scale such as very/somewhat/slightly/not at all). In addition, both instruments do not use all questions in the FR algorithm.

- Both instruments inquire about **when the respondent first learned about the program**. The DCEO survey does not use this information in the algorithm.
- Both inquire into **plans and intentions to install** the measure without support from the program.
- The DCEO survey uses a "threshold" question about the respondent's **financial ability** to install the measure without the incentive. Respondents who answer "no" to this question automatically get a FR score of 0, independent of other responses they provide (they are still asked all of the other questions).
- Both instruments ask about the impact of the program on the level of **efficiency**. The DCEO survey asks directly if the respondent chose a more energy efficient model because of the program; the ComEd/Ameren survey asks about the likelihood that the respondent would have installed the exact same equipment without the program.
- Both ask if the program influenced the **timing** of the installations. If the installation would have happened later without the program, both surveys ask how much later.

- The DCEO and the ComEd surveys ask about the impact of the program on the **quantity** of measures installed. However, neither survey uses the response in the FR algorithm. The Ameren survey does not include this question.
- Both ask about the importance of **program factors** in deciding to install efficient equipment, but there is little overlap in which factors are included in the algorithm.[5] The ComEd/Ameren questions are asked on a scale from 0 to 10; DCEO questions ask about the level of importance (very, somewhat, slightly, not at all).
- The ComEd/Ameren survey asks about a variety of **"other factors"** that might have influenced decision making.[6] The DCEO survey asks only about previous experience with the measure and energy efficiency purchases without a program incentive. Again, the ComEd/Ameren questions take on responses ranging from 0 to 10 while the DCEO questions tend to be binary (yes/no). The DCEO "other factors" questions are included in the FR algorithm. The "other factors" in the ComEd/Ameren survey are designed to be taken into account for projects with larger savings (i.e., standard rigor projects); however, in the PY5 analyses, none of these responses were used in determining the final FR score.
- The ComEd/Ameren survey contains **consistency checks** that are triggered when respondents give contradictory responses to key questions. The DCEO battery does not contain consistency checks.

For a detailed side by side comparison and related comments of the DCEO and ComEd/Ameren surveys, please see the "Algorithm Questions" tab in the accompanying spreadsheet.

Comparison of Freeridership Ranges

The "Comparison of FR Ranges" tab in the accompanying spreadsheet shows a side-by-side comparison of FR ranges for two key questions that were similar enough between survey instruments to enable a comparison. We show the low and high range the overall FR value

[5] The ComEd/Ameren algorithm includes: The availability of the program financial incentive, Recommendation from a program staff person, Information provided through a program-sponsored feasibility study (Ameren custom only), Information from program or utility marketing materials, Endorsement or recommendation by a utility key account executive, Information provided by program through technical assistance (ComEd custom only). The DCEO algorithm includes: Previous experience with the DCEO programs, Recommendation from a Public Sector Energy Efficiency Program representative, Recommendation from a SEDAC representative. The DCEO survey also asks about the importance of Past experience with energy efficient equipment, Incentive or grant payments from DCEO, and Recommendations received from DCEO but does not include these in the algorithm.

[6] These other factors are only asked for larger (standard rigor) projects, i.e., projects with relatively high savings. They include: Condition of the existing equipment (prescriptive/standard surveys only), Previous experience with this type of equipment, Recommendation from an equipment vendor or contractor, Recommendation from a design or consulting engineer, Industry standard practice, Corporate policy or guidelines, Payback on the investment, and Other (open end).

can take on, given responses to these key questions. The following are observations based on this comparison.

Likelihood to install equipment without the program

The questions are very similar. The ComEd/Ameren survey asks about the likelihood of installing "exactly the same equipment," i.e., it incorporates the concept of efficiency into this question; the DCEO question asks about the measure in general. As with other questions, the ComEd/Ameren survey asks on a likelihood scale of 0 to 10 while the DCEO uses four response options.

In both surveys, responses to this likelihood question can affect the range of overall FR values. There is the potential for the overall FR score to be a low value of 0% for all response options because this question gets combined with the timing question (and a separate efficiency question, in the DCEO algorithm). However, the possible high overall FR value is greater in the ComEd/Ameren algorithm, compared to the DCEO algorithm, because of the averaging of the three FR components within the ComEd/Ameren algorithm.

Timing of installation without the program

Both surveys ask fairly similar questions about the timing of the installations in the absence of the program, although the response options for later installations vary slightly. In the DCEO algorithm, a respondent who would have installed the equipment "less than 6 months" later can get a FR score of between 0 and 0.67, i.e., the program gets attribution of 0.33 for a slight acceleration of the project, if the program had no other influence on the project. In contrast, a ComEd/Ameren project that would have happened "less than 6 months" later without the program would get a FR score of 1 (zero attribution), if there was no other program influence. In the ComEd/Ameren algorithm, a respondent who says they would have installed the equipment "4 or more years later" can get a FR score of between 0 and 0.67, i.e., the program gets the same attribution of 0.33 for a substantial acceleration of the project (with no other program influence) as a DCEO project gets for a slight acceleration.

Analysis of Key Questions/FR Components

We looked at key questions and FR components in the DCEO and ComEd/Ameren survey to identify possible areas of internal inconsistency, i.e., contradictory responses, and their effect on the overall FR score. Spreadsheet tabs "DCEO Analysis," "ComEd Analysis," and "Ameren Analysis" contain the data underlying these analysis.

DCEO

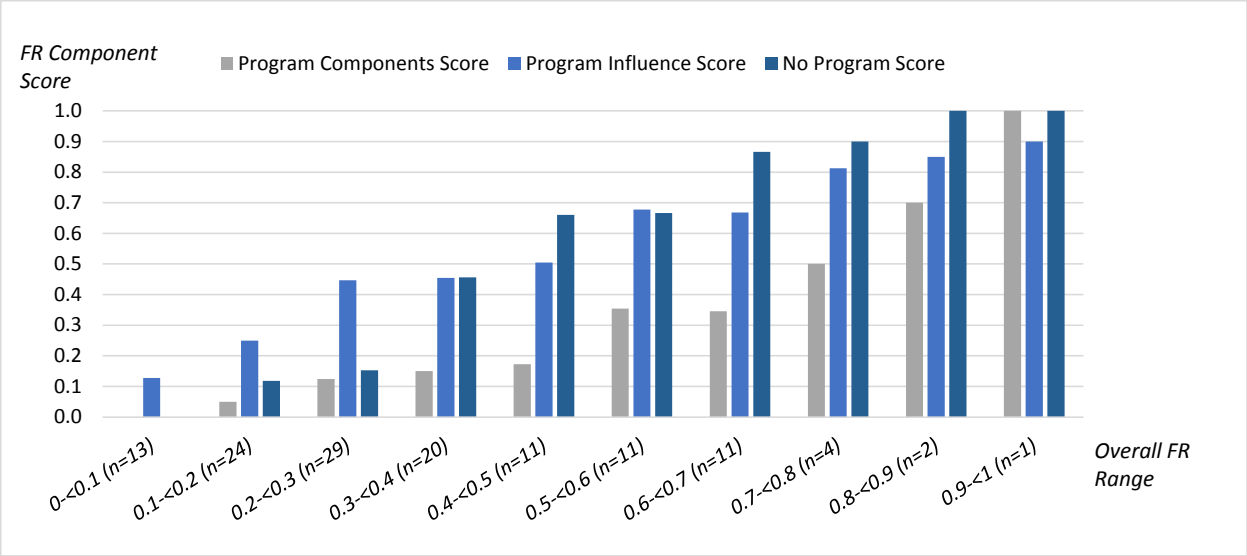
In the DCEO algorithm, anyone who says "no" to the question "Would you have been financially able to install the equipment or measures without the financial incentive from the program?" is automatically assigned a FR score of 0 for the measure. These respondents are still asked all of the other FR questions, but they are not considered in the

FR score and any inconsistent responses are not explored. Overall, 6 out of 109 custom measures (6%) and 19 out of 218 standard measures (9%) received a FR score of 0% due to their financial ability response. Based on other responses, these measures would have received a FR score of 33% if the threshold question had been ignored. If this revised score were used, the average Custom Program FR would increase from 6.4% to 8.3% (straight average; no weighting) and the average Standard Program FR would increase from 5.5% to 8.4% (straight average; no weighting). Inclusion of this threshold question therefore appears to have a fairly small impact on program level FR results for the PY5 population of respondents. However, while this particular case may not have a large impact, we are uncertain of the level of difference possible in any future surveys.

ComEd / Ameren

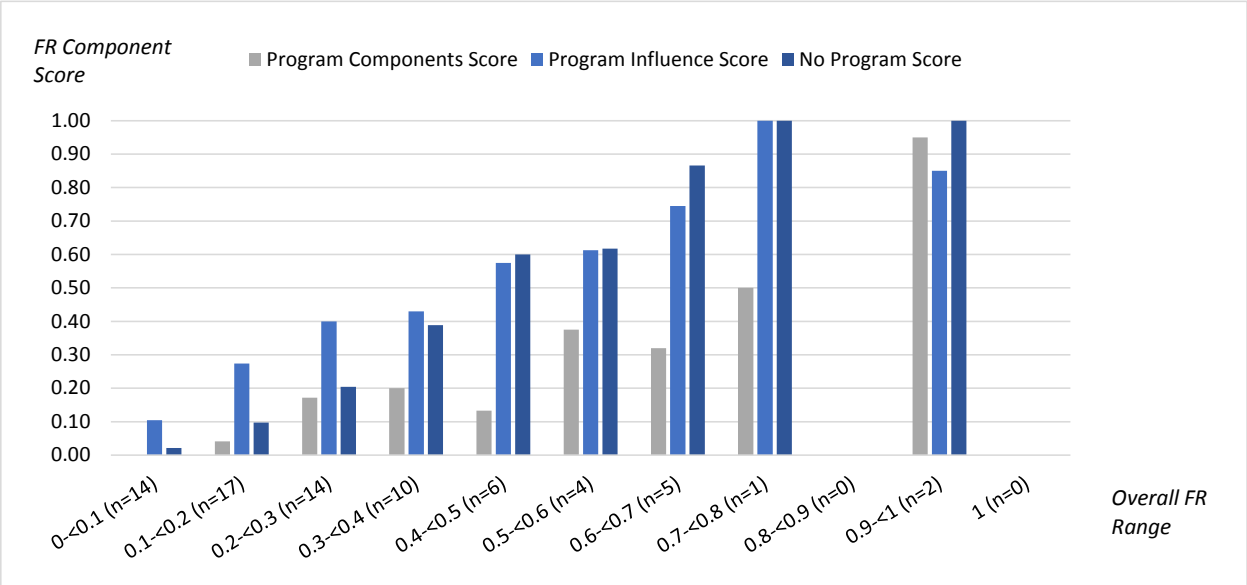
We looked at the distribution of average component scores from redacted ComEd Standard and Ameren Custom surveys. As shown in the figures below, in both cases, the Program Components Scores were lower than the other two components which contribute to the final FR score. For example, there were 20 ComEd respondents whose overall FR score was between 0.3 and 0.4. As shown in Figure 1 below, the average Program Influence Score and No Program Score are substantially higher than the average Program Components Score.[7]

Figure 1. Average FR Component Scores – ComEd Prescriptive



[7] This is just one way to look at the different components. Another method, which we did not employ at this point, is to calculate a Cronbach's alpha.

Figure 2. Average FR Component Scores – Ameren Standard



Questions for Consideration / Discussion

Based on our comparison and review of the survey questions and FR algorithm used in the DCEO and ComEd/Ameren evaluations, we identified a few questions for consideration/discussion among the group. These questions are not intended to be exhaustive but rather provide a starting point for discussion.

Question 1: What is the best way of combining the different FR components?

Averaging of three FR components in ComEd/Ameren algorithm leads to high possible FR score for all responses. While all three components measure an aspect of FR, they do not all include the same concepts (e.g., timing is only part of the No Program Score). On the residential side, some ComEd and Ameren algorithms multiply certain FR components. This choice can be considered by the group.

Question 2: What level of granularity is desirable?

The DCEO algorithm develops four binary indicator variables (yes/no), resulting in four possible measure-level FR scores. The ComEd/Ameren algorithm use 0-10 scales for many key questions and often produces a less "chunky" distribution. However, an 11-point scale with the accompanying nuance in FR scores might give a sense of false precision.

Question 3: How much credit should the program get if the installation would have been identical except for the timing? How much credit for slightly earlier (6 months, 1 yr) installations? How much for much later (3-4 yrs, 4+ yrs) installations?

When assessing timing of installation without the program, DCEO algorithm assigns an attribution score of 0.33 attribution (i.e., 1 – FR) for marginal acceleration of the project (when no other program influence is present). The ComEd/Ameren algorithms assign low attribution for substantial acceleration. It seems that the choice of high or low FR based on time should be the same throughout the Illinois application of self-report data.

Question 4: Is a threshold question desirable? If so, should there be consistency checks within the survey when other questions appear to support an overall result other than a 0 FR?

The DCEO algorithm uses a threshold question, the ComEd/Ameren algorithms do not. For PY5 responses, application of the DCEO algorithm ignoring the threshold question affected the measure-level FR score for few respondents and changed the program-level FR score only slightly.

Question 5: Regardless of a threshold question, should there be consistency checks? If so, on what questions? How should they be used to modify inconsistent responses?

The ComEd/Ameren survey contains consistency checks and, in some cases, gives the respondent an opportunity to change their previous answers. The DCEO battery does not include consistency checks.

Question 6: Should previous experience with the program count to reduce FR?

One of the program factors in the DCEO algorithm is the importance of previous experience with the program in the decision to install the measure. In the DCEO algorithm, high importance of previous program experience leads to a binary score for program influence that reduces FR (it has the same effect as high influence of a recommendation from program staff). The ComEd/Ameren surveys do not ask if previous participation was important in the installation decision. In the energy efficiency program evaluation community, there is debate over how previous program experience should be considered in the FR determination for a given program year. It seems that consistent treatment across all IL evaluations would be desirable, making this a good topic for discussion among stakeholders.

Question 7: How should non-program factors be included in the survey and considered in the FR algorithm?

The ComEd/Ameren survey asks several questions about the importance of non-program factors, but only if projects have high savings (standard rigor projects). For most projects, no non-program factors are asked about, placing the subsequent question about dividing 100 points between program and non-program factors somewhat out of context. For larger projects (standard rigor), these questions are asked but do not appear to be used in the algorithm. The DCEO survey only asks about two non-program factors, prior experience with the measure and energy efficiency purchases without a program incentive.

Attachment 1 – NTGR Comments by Dr. Ridge

As stated earlier, Dr. Richard Ridge is an expert in self-report methodology, having been instrumental in the creation of the method over the past decade or more. When we requested he review our memo, besides providing comments that we have incorporated to the extent practicable, unasked by our team he also provided us with a companion document around self-report. That document, in its entirety and unchanged by us except for formatting, is included next.

Comments on NTGR Methods in Illinois

Introduction

Any approach that is eventually adopted in Illinois should be based on methods that are supported in the social science literature. This would include textbooks, journal articles, evaluation reports and best practices protocols drawn from the first three. Appendix A contains the literature to which I'll refer plus others. These references address such topics as the reliability and validity of measurement, potential sources of error present in any research, the types of response scales, scale development, and problems with the self-report. In my comments, I'll reference the relevant literature in support of one approach versus another. Note that my intent is not to make invidious comparisons but to point out the extent to which each approach has support in the current literature.

Matrix Approach

The DCEO estimates the NTGR in two steps. The first step identifies free riders based on the answer to a single question. Those who are not classified as free riders in Step 1 proceed to Step 2 where they are asked three additional questions. For those who are asked these three additional questions, their final NTGR is based on particular response patterns for all four questions. Unfortunately, this approach, which have been referred in other jurisdictions as the "matrix" approach, introduces unnecessary measurement error.

DCEO appears to have assigned the FR scores to a particular response pattern based on expert judgment, which is essentially an exercise in the coding of each particular pattern of responses to these questions into a particular FR score. However, when we are addressing the reliability of these FR assignments, we are asking whether these FR assignments are repeatable. If twenty groups of 20 evaluators each provided their freerider assignment to each question response pattern, would they all agree (i.e., would their inter-rater reliability[8] be high or low)? While we do not expect contractors to form twenty groups of evaluators in the development of their FR assignments, we do expect them to pay special attention to

[8] Inter-rater reliability, inter-observer reliability, and inter-judge agreement are some of the words that have been used very often in the literature to designate a wide variety of concepts. All of these terms, however, refer to the extent of agreement among raters, judges, and observers (Gwet, Kilem L. (2012). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters*. Gaithersburg, MD: Advanced Analytics, LLC.

inter-rater reliability among those evaluators who do contribute their assignments. Simple statements that one has relied on industry experience and staff expertise to assign FR scores to specific question response patterns should not be sufficient since this, by itself, doesn't insure that the estimated NTGR is sufficiently reliable.

Also, when using this approach, one should have to demonstrate why having evaluators assign FR scores to particular response patterns is more reliable than more traditional approaches which eliminate the need to code responses by simply asking respondents to provide their own estimates of program influence and freeridership along a continuum (e.g. a 11-point scale). Note that I am not arguing that achieving reliable estimates of freeridership using the more traditional approaches is easy but that their use when guided by best practice guidelines is more likely to produce reliable estimates.

Reliance on a Single Question

Program influence, as represented in the net-to-gross ratio (NTGR) or its converse, free ridership (FR), cannot be directly observed but rather measured indirectly. As a general principle, whenever one is attempting to measure an underlying construct like the NTGR or FR, using multiple questions is considered more reliable than using only one question. Regardless of the magnitude of the savings or the complexity of the decision-making process, one should assume that using multiple questionnaire items (both quantitative and qualitative) to measure a construct such as freeridership is preferable to using only one item since reliability is increased by the use of multiple items (Blalock, 1970; Crocker & Algina; 1986; Duncan, 1984). In other words, no matter how straightforward any question seems, it always contains some potential for measurement error.

Temporal Sequence

The ComEd/Ameren questionnaire asks: "Did you learn about the program BEFORE or AFTER the decision was made to implement the measure that was installed?" If they respond "After," the overall influence score is reduced by half. I agree with ODC that it might be useful to ask if equipment was INSTALLED before they learned about the program. Some evaluators have used a series of yes/no questions to help the respondent get at the temporal sequence such as:

a. When did you first learn about the [Utility's] Program?
    i. Was if BEFORE you first began thinking about installing the new equipment or was it AFTER you first began to think about installing the new equipment?
    ii. Was it BEFORE or was it AFTER you began to look at or collect information about the new equipment?
    iii. Was it BEFORE or was it AFTER you selected or picked out the particular equipment that you were going to buy?
    iv. Was it BEFORE or was it AFTER the new equipment was installed.

Only if the response to Question a.iv is "After" is an adjustment made to the NTGR.

Questions Asked But Not Used

The questions asked in the DCEO approach are reasonable. However, it is not clear why they omitted some questions in their calculations. For example, they ask:

a. How did the availability of information and financial incentives or grants through the Public Sector Energy Efficiency Program affect the quantity (or number of units) of [MEASURE 1] that you purchased and installed? Did you purchase and install more [MEASURE 1] than you otherwise would have without the program?
b. How important are incentive or grant payments from DCEO for your decision making regarding energy efficiency improvements (Q11)
c. Did a representative of the Smart Energy Design Assistance Center (SEDAC) recommend that you install the [MEASURE 1]? (Q22/Q32/Q42)
d. If the SEDAC representative had not recommended installing the equipment, how likely is it that you would have installed it anyway? (Q22a/Q32a/Q42a)
e. How important is advice and/or recommendations received from DCEO for your decision making regarding energy efficiency improvements? (Q13)

The ComEd/Ameren approach also asks some questions that appear to be potentially useful but play no role in the calculation of the NTGR. For example, they ask:

a. What specific corporate policy influenced your decision to adopt or install the <ENDUSE> through the program? (QN12)
b. If I understand you correctly, you said that your company's corporate policy has caused you to install energy efficient <ENDUSE> previously at this and/or other facilities. I want to make sure I fully understand how this corporate policy influenced your decision versus the program. Can you please clarify that? (QN17)

It would be useful to agree on which questions are key to understanding the influence of the program as well as other non-program factors, agree on appropriate response categories, and agree on how the results can be analyzed to produce estimates of NTGRs or free ridership.

Cognitive Interviews

For questions such as those in these complex and challenging surveys, there are many opportunities for misunderstanding the meaning of each question. Prior to launching a pre-test of the questionnaire(s), it would be useful to conduct cognitive interviews with a small group of potential respondents. Biemer and Lyberg (2003) define cognitive interviewing as ". . . a set of methods for interviewing respondents so that the errors arising from specific stages of the response process (i.e., encoding, comprehension, information retrieval, response formatting, and communication) can be identified by the survey designer." Cognitive interviews are usually conducted in somewhat controlled settings such as a survey methodologist's office or a cognitive laboratory which is a room specially equipped for tape recording or video taping the interview. During the interview, the interviewer probes for information about the respondents thought processes immediately following the

response to a particular question. Problems discovered using this technique can then be addressed by questionnaire revisions, modification of data collection methods, interviewer training, and so on. This process is followed by a more traditional pre-testing of the questionnaire on a larger sample of potential respondents under normal interviewing conditions.

Measures of Reliability

The validity and reliability of *each question* used in estimating the NTGR must be assessed (Lyberg, et al., 1997). In addition, the internal consistency (reliability) of multiple-item NTGR *scales* should not be assumed and should be tested. Testing the reliability of scales includes such techniques as split-half correlations, Kuder-Richardson, and Cronbach's alpha (Netemeyer, Bearden, and Sharma, 2003; Crocker & Algina, 1986; Cronbach, 1951; DeVellis, 1991). An evaluation using self-report methods should employ and document some or all of these tests or other suitable tests to evaluate reliability, including a description of why particular tests were used and others were considered inappropriate.

Neither ComEd nor Ameren appear to have calculated some measure of internal reliability such as Cronback's alpha. Such a measure of internal consistency might be useful to calculate. A rule of thumb for Cronback's alpha is 0.70. For DCEO, perhaps a series of Chi-Square tests could be conducted that could address internal consistency?

Sensitivity Analysis

Ridge et al. (2013) noted that when multiple questions, weights, and complex algorithms are involved in calculating the NTGR, evaluators should also consider conducting a sensitivity analysis (*e.g.*, changing weights, changing the questions used in estimating the NTGR, changing the probabilities assigned to different response categories, etc.) to assess the stability and possible bias of the estimated NTGR. However, ComEd/Ameren and DCEO do not appear to have calculated any sensitivity analyses.

Ruling Out Rival Hypotheses

Ridge et al. (2013) note that most significant events in the social world are not mono-causal, but instead are the result of a nexus of causal influences. Both in social science and in everyday life, when we say that Factor A is strongly influential in helping to cause Event B, it is rarely the case that we believe factor A is the sole determinant of Event B. Much more commonly, what we mean to say is that Factor A is among the leading determinants of Event B. Thus, an evaluator should attempt to rule out rival hypotheses regarding the reasons for installing the efficient equipment (Scriven, 1976). For example, to reduce the possibility of socially desirable responses, one could ask an *open-ended question* (*i.e.*, a list of possible reasons is **not** read to the respondent) regarding other possible reasons for installing the efficient equipment. A listing by the interviewer of such reasons such as global warming, Energy Star, other utility programs, the price of electricity, concern for future generations, and the need for the US to reduce oil dependency might elicit socially desirable responses which would have the effect of artificially reducing the NTGR. The answers to such questions about other possible influences can be factored into

the estimation of the NTGR. The DCEO asks about only two non-program-related factors (prior installation of similar equipment and the purchase of energy efficiency equipment without an incentive). This list could have included more factors such as those in the Ameren and ComEd surveys. Note that the Ameren and ComEd surveys also ask about previous experience of a similar type of equipment.

Response Categories

In the self-report batteries of questions, what we are attempting to measure, among other things, is a participant's perception of the influence of the utility program on their decision to implement the energy-efficient measure. Because this is not something that is directly observable and measureable, we must rely on answers to a series of questions regarding the reasons for the installation. To assess the strength of any reason, many have chosen response categories along a 0-10 scale since the strength of the reasons cannot be adequately captured by a "yes" or "no" response. This is the reason why scales are used to measure such underlying constructs as personality and attitudes where responses are scored along a continuum. In this respect, program influence is no different.

The question is whether a typical respondent can accurately assess the strength of program and non-program influence factors along an 11-point scale. I don't think there is any empirical evidence that they cannot. Historically, surveys have used a variety of scales (e.g., 4-point, 5-point, 7-point, 10-point, 11-point). While methodologists continue to argue over scale length, the reason that a consensus has not emerged is because there is, to date, no definitive proof one way or the other. An additional advantage of using an 11-point scale (0 to 1) is that the responses are consistent with probabilistic statements that range from 0 to 1 that people make every day. I think there is no compelling evidence that using a 0 to 10 scale is a case of false precision. Finally, using a scale of whatever length is far more preferable to evaluators assigning FR values to specific response patterns, a process that can be arbitrary or at least appear to be arbitrary.

Comparably of NTGR across Different Programs

On page 5 of your memo, you state: "In both surveys, responses to this likelihood question can affect the range of overall FR values. There is the potential for the overall FR score to be a low value of 0% for all response options because this question gets combined with the timing question (and a separate efficiency question, in the DCEO algorithm). However, the possible high overall FR value is greater in the ComEd/Ameren algorithm, compared to the DCEO algorithm, because of the averaging of the three FR components within the ComEd/Ameren algorithm." Are such comparisons informative? In order for such comparisons to be informative, customers mix, technology mix, incentive levels, program years and regional economies would have to be reasonably similar. I suspect that, as yet, you do not know which of these conditions, if any, has been met.

Or, are you are suggesting that the timing question, rather than being averaged with the other responses to produce a NTGR or FR score, override the responses to the other questions because, on its face, it is more compelling and error free? If that is the case, see

my comments above in the section *Reliance on a Single Question*. In my view, I think that the timing question is not just counterfactual question about what you would have done in an alternate universe but what you would have done in the future in an alternate universe. Because the errors in such questions are compounded, they are inherently less reliable.

Methods Used in Other Jurisdictions

You might want to point out than in other jurisdictions (e.g., California, New York, and Wisconsin) with relatively long histories of evaluating energy efficiency programs, the ComEd/Ameren approach is preferred.

Appendix A

Methodological Literature

Biemer, Paul P. and Lars E. Lybers. (2003). *Introduction to Survey Quality*. Hoboken, New Jersey: John Wiley & sons.

Blalock, H. (1970), "Estimating Measurement Error Using Multiple Indicators and Several Points in Time," *American Sociological Review*, 35, pp. 101-111.

Bradburn, Norman, Seymour Sudman, and Brian Wansink. 2004. *Asking Questions: The Definitive Guide to Questionnaire Design: For Market Research, Political Polls and Social and Health Questionnaires.* San Francisco: Jossey-Bass.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart & Winston.

Cronbach, L. J. 1982. *Designing Evaluation and Social Action Programs*. San Francisco: Jossey-Bass.

Duncan, O.D. (1984). *Notes on Social Measurement: Historical and Critical*. New York: Russell Sage.

Groves, Robert M., Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2004. *Survey Methodology*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Lyberg, Lars, Paul Biemer, Martin Collins, Edith de Leeuw, Cathryn Dippo, Norbert Schwarz,and Dennis Trewin. 1997. Survey Measurement and Process Quality. New York: John Wiley & Sons, Inc.

Gwet, Kilem L. (2012). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters*. Gaithersburg, MD: Advanced Analytics, LLC.

PA Government Services. 2003. *Standardized Methods for Free-Ridership and Spillover Evaluation – Task 5 Final Report (Revised)*. Prepared for National Grid, NSTAR Electric, Northeast Utilities, Unitil, and Cape Light Compact.

Ridge, Richard, Ken Keating, Lori Megdal, and Nick Hall. 2007a. *Guidelines for Estimating Net-To-Gross Ratios Using the Self Report Approach*. Prepared for the California Public Utilities Commission.

Ridge, Richard, Nick Hall, Ralph Prahl, Gil Peach, and Paul Horowitz. (2013b). "Guidelines for Estimating Net-To-Gross Ratios Using the Self Report Approach." Prepared for the New York Department of Public Service.

Scriven, Michael. 1976. Maximizing the Power of Causal Explanations: The Modus Operandi Method. In G.V. Glass (Ed.), *Evaluation Studies Review Annual* (Vol. 1, pp.101-118). Beverly Hills, CA: Sage Publications.

Scriven, Michael. 2009. Demythologizing Causation and Evidence. In Stuart I. Donaldson, Christina A. Christie, and Melvin Mark (Eds.). *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* Los Angeles, CA: SAGE Publications.

Stone, Arthur A., Jaylan S. Turkkan, Christine A. Bachrach, Jared B. Jobe, Howard S. Kurtzman, and Virginia S. Cain. 2000. *The Science of the Self-Report: Implications for Research and Practice*. Mahwah, New Jersey: Lawrence Erlbaum Associates.