Feasting at the Ultimate Enhanced Free-Ridership Salad Bar[1]

Dr. Lori Megdal, Megdal & Associates, LLC, Acton, MA
Yogesh Patil, Energy & Resource Solutions, Inc., Haverhill, MA
Cherie Gregoire and Jennifer Meissner, New York State Energy Research and Development Authority (NYSERDA), Albany, NY
Kathryn Parlin, West Hill Energy & Computing, Inc., Chelsea, VT

ABSTRACT

Public investment in energy efficiency has policy makers demanding reliable information on what the investments are actually purchasing. This requires comparing what occurred to the counter-factual, what would have occurred in the absence of the program. Using the highest quality science and ensuring reliable and valid measurement of free-ridership is quite difficult. It is much easier, however, to pick-up the fast-food of evaluation: throwing survey questions, response scoring, and algorithms into a mindless meal on the run. This approach can produce questionable estimates of free-ridership and undermine confidence in energy efficiency estimates resulting in the potential to undermine the decisions to invest in energy efficiency. (These problem evaluations are unhealthy for evaluation and its consumers.)

The New York State Energy Research & Development Authority (NYSERDA) and its impact evaluation team created an in-depth mixed methods (quantitative and qualitative) approach to estimate free-ridership in a cross-program study of the largest expected savers from the commercial/industrial programs. Referred to as their "salad bar approach," where all respondents received a core set of questionnaires ("the lettuce"), then select instruments were applied per project ("the toppings"). The components included first learning how decision-making is made for that particular firm and project. Then this information is used to combine components.

The mixed-methods approach was used for both the inquiries and the process, providing in-depth information for a customized assessment. Several robust outcomes were obtained from this approach and process: construct validity, consistency and reliability within the free-ridership estimates (healthy evaluation outcomes.)

Free-Ridership Estimates – Important Ingredient to Assess Investment Value

Public investment in energy efficiency has policy makers demanding reliable information on what the investments are actually purchasing. Since the purpose of an energy saving program is to induce changes that would not otherwise occur, program evaluation includes an assessment of which energy saving initiatives were taken as a result of the program, and how much an organization would have done even if there were no program. This requires comparing what occurred to the counter-factual, what would have occurred in the absence of the program. This can be done through a variety of methods: market analysis, statistical comparison methods on decision-making and self-report approaches through surveys and

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the views of the New York State Energy Research and Development Authority (NYSERDA).

interviews. It is difficult to do any of these methods while ensuring the highest quality science is used, and that they are able to measure with validity the program-induced impacts.

A common long-standing approach is to use direct query, self-reports, surveys or interviews as to what action participants think they would have undertaken (or at least intended to do) in the absence of the program. For program participants, assessing the savings that are "net" of what would have occurred involves estimating the program measures (or the proportion of the savings) they would have adopted within the same time frame but absent the program's existence (free-ridership). The latter, free-ridership (FR), essentially dilutes the efficacy of the program. Free-ridership is expressed as a percentage reduction in the total gross energy and demand impacts of a program.

An off-setting effect, spillover (SO), includes those energy saving actions an organization takes as a result of the program, but for which it does not receive any program incentives. Program participants can also take additional efficiency actions due to what they learned or experienced through the program even when these actions are not explicitly recognized or directly supported by the program (spillover). There are two types of participant spillover:

- "Inside" spillover occurs when, due to the project, additional actions are taken to reduce energy use at the same project site, but these actions are not included as program savings.
- "Outside" project spillover occurs when an actor participating in the program initiates additional actions that reduce energy use at other sites that are not participating in the program.

In addition, non-participants can also be influenced by the program. A simple example, reflecting a common phenomenon measured within NYSERDA evaluations (Megdal & Associates 2008; Summit Blue 2006; Siems, Meissner and Megdal 2009), is where formerly participating design firms or contractors promote energy efficiency or green building practices to their customers due to what they learned when they participated with the program. Non-participant spillover can also occur when non-participating design firms/contractors are exposed to program information from marketing and outreach efforts (brochures, mailings, web sites, etc.) as well as from current and past program participants and incorporate these practices into their business in order to maintain competitiveness.

Both participant and non-participant spillover increases the effectiveness of the program, and is also expressed as a percentage increase in the total energy and demand impacts in a program. External, or non-participant, spillover was not included in the study reported in this paper, which focused only on the specific participants themselves. However, NYSERDA does include estimates of non-participant spillover in all their reported savings estimates. For many of their commercial and industrial (C&I) programs, the non-participant spillover rates comes from periodic evaluation studies that estimate this impact rate across NYSERDA C&I programs, which can have overlapping effects across markets.

To determine the net effect of the program, then, the free-ridership and the spillover need to be determined and taken into account. The ratio of net savings to program-estimated gross savings is presented by the net-to-gross ratio (NTGR). The NTGR can be calculated as follows: NTGR = 1 – FR + SO where FR and SO are expressed as percentages compared to the program-estimated gross savings.

Self-reported Free-Ridership – Making Poor Selections or Healthy Ones?

In many jurisdictions the need to understand program-induced savings to assess appropriate investment levels has been linked to utility incentive mechanisms to encourage effective energy efficiency programs. This linkage makes a target out of the science of evaluation, creates "an incentive" to undermine the use of accepted scientific methods for assessing causality. (See Ridge, Willems and Fagan (2009) and Ridge, Willems, Fagan, and Randazzo (2009) for a review of the broader literature on the realist view of causality assessment.) Just as "de-coupling" of utility incentives for increasing sales can potentially

encourage efficiency, de-coupling utility incentives from the art and science of estimating net impacts could allow healthy review and improvement for the art and science of measuring the counter-factual and net impacts.[2]

The simple presentation of the self-report approach and calculation of NTGR hides the great complexities that need to be addressed within the approach, the inquiries and the analyses. Its appearance of simplicity is deceiving. All evaluations should require well-trained and experienced evaluators leading or reviewing the evaluations. Other fields often include critical review or independent assessment to minimize the bias and measurement error within their scientific investigation or tools. Licensing and regulations have often been developed as a result of horrific consequences from poor applications by untrained personnel.[3] It is now standard medical practice to require trained individuals to use the current tools in medical diagnoses and to require double-reading before diagnoses are provided to the patient (HIV, mammograms, cancer tests, etc.). Critical review and independent assessment are also important for the application of the self-report approach to estimate free-ridership.

Estimating free-ridership in an environment where there can be many influences for an efficiency investment has been claimed by some critics to be an unrealistic expectation. Critics of self-report free-ridership measurement also cite that the self-report measures are likely upwardly biased given an increase in social desirability to be "green" and environmentally sensitive. There are many other fields of evaluation that face challenges at least as great and still depend upon experienced evaluators to use the highest quality science and art to separate program-induced impacts from impacts due to other factors or potential biases. The range of programs with evaluation challenges for assessing causality is quite diverse. It would seem logical that self-report bias due to social desirability in energy efficiency program free-ridership ought to be lower than that for drug abuse or sexual behavioral issues. Validation studies on self-reported drug abuse compared to urinalysis testing have shown a relatively low level of self-report bias. "Overall, there are indications of a small underreporting bias in the National Household Survey on Drug Abuse and the Monitoring the Future Survey, but its overall effects are relatively small" (Harrison 1995). Another example, this one including measuring an underlying construct, that of sexuality acceptance and its relationship to self-efficacy, can be found in evaluations of condom programs for young women (Bryan, Aiken and West 1996). Most evaluations of condom programs rely upon self-reported condom use, with a few studies that tested the validity of self-reported condom use data, such as Shew, et. al., 1997.

Other fields can also offer suggestions on ways to improve the validity of self-report data based upon survey technique and incorporating other related measurement indicators. Again, there is a broad range of fields using self-reports surveys to create valid measurements. These range from reporting of pubertal status among adolescents (Petersen, Crockett, Richardson and Boxer 1988), the validity of self-report strengths and difficulties scores to distinguish youth with mental illness versus the general population as compared to teacher and parent scoring (Goodman, Meltzer and Bailey 2003), and youth self-reports of alcohol and drug abuse (Winters, Stinchfield, Henly, and Schwartz 1990). We can not simply accept any self-report measurement of free-ridership. At the same time, we should not throw out all measurements of free-ridership just because they are self-reports.

There are evaluations and analyses that have been done that make self-reported free-ridership studies an easy target for criticism. These *problem evaluations* are often those performed without experienced and skilled evaluators or those not taking the time to ensure that both art and science is closely monitored with

[2] Incentive mechanisms could be designed to "re-couple" utility incentives to specific measurements that more directly measure utility efficacy and performance. This could allow utilities to be measured on what they can manage and allow the science and art of evaluation to stay focused upon obtaining the most reliable net savings estimates.

[3] An interesting example that few are aware of is that licensing and regulations for professional engineers began as a consequence of the death and destruction caused by a similarly little known unusual disaster – the great molasses flood of 1919 in Boston (Puleo 2003).

experience and critical thinking. It is much easier (and cheaper) to pick-up the fast-food of evaluation: throwing survey questions, scoring of responses and algorithms into a mindless meal on the run. The willy-nilly variation introduced by this approach can be "unhealthy" for evaluation and its consumers.

Background on the Largest Savers Evaluation Project

The New York State Energy and Research Development Authority (NYSERDA) operates **New York Energy $mart[SM]**, a portfolio of System Benefit Charge (SBC)-funded energy efficiency and demand management programs in New York, as well as energy efficiency programs sponsored with other funding sources. NYSERDA receives approximately $175 million per year for the SBC-funded commercial, industrial, residential, low-income and research and development efforts, which are expected to increase energy efficiency, lower energy demand, and develop renewable energy technologies in New York. NYSERDA's Energy Analysis group manages work by independent evaluators to conduct impact, market, and process evaluations of these efforts.

NYSERDA conducted a risk analysis in 2007 (Meissner, Gregoire, Meyers, et. al. 2008; Megdal & Associates 2007) in order to help evaluators target its limited impact evaluation resources for its next round of evaluations. The risk analysis (RA) used a comprehensive method of quantifying risk by describing full, probabilistic distributions of values, uncertainty, and the underlying drivers of such uncertainty to identify and quantify uncertainty within NYSERDA's portfolio of efficiency program evaluation realization factors (for gross savings and net-to-gross ratios (NTGR).

Net-to-gross realization rates for the larger NYSERDA programs were major contributors to the remaining evaluation uncertainties. The top two uncertainty contributors were found to be the NTGR for the Commercial/Industrial Performance Program (CIPP) and the Technical Assistance (TA) program. Three-quarters of the uncertainty surrounding the impact evaluation estimates for the Peak Load Reduction Program (PLRP) was with its NTGR estimates. The two programs that are the largest contributors to energy and demand savings, CIPP and PLRP respectively, also contain rigorous post-installation verification or data collection activities. These program-level efforts helped minimize the uncertainty in their gross savings realization rates, even with the limited sample sizes used in previous evaluations due to budget constraints. Nevertheless, the substantial contribution of NYSERDA's largest programs to the entire portfolio in combination with the uncertainty associated with the NTGR estimates means that they represent a large component of the overall uncertainty for the portfolio of efficiency programs.

The two primary evaluation actions that can be undertaken to increase reliability is to use methods that better capture the underlying construct and directly address the areas of potential bias (and/or test validity) and increase sample size (increasing sampling precision). Methods with better measurement of the causes of different decisions within a firm's decision-making process can greatly reduce any unknown risks of potential bias that can go unobserved within less comprehensive methods.

NYSERDA and its impact evaluation contractor designed a project to reduce overall portfolio uncertainty with cost-efficient impact evaluation by conducting a cross-program evaluation of the portfolio's largest expected savers. Almost all energy efficiency evaluation is at the program level or evaluating specific measures. The largest customers/savers may be part of a census stratum but the research design and reporting seldom focuses only on these customers, making the Largest Savers evaluation unique. The census stratum may receive customized gross savings evaluation plans (site-specific M&V plans). Yet, it is unusual for the census stratum to receive customized site-specific in-depth NTGR exams, as was used in the Large Savers Project NTGR method. The sophistication of the methods and the research approaches explored also provides a rich learning experience to indicate where further evaluation method enhancements should be considered across future program evaluations.

The impact evaluation contractor, a Megdal & Associates team, is in the process of evaluating the energy savings for a group of large savers that completed participation in NYSERDA's programs in 2005 – 2007. The study targeted 25 participants with expected savings exceeding 1.5 GWh (the Large Savers Project). The following programs were represented among these participants:

- CIPP (Commercial/Industrial Performance Program)
- DG-CHP (Distributed Generation – Combined Heat and Power)
- NCP (New Construction Program).
- PLMP (Peak Load Management Program)
- TA (Technical Assistance)

The Large Savers Project is designed to evaluate a census of the largest expected savers. All 25 projects represent 128 GWh in savings, or 18% of the 2007 incremental savings reported for the whole NYSERDA portfolio. In Phase I, the impact evaluation was completed for fourteen of the participants. The Phase I project results presented here represent 79 GWh/yr or 11% of the 2007 incremental savings reported for the NYSERDA portfolio.

Phase I Large Savers included a wide range of businesses and institutions, including communications, utilities, manufacturing, higher education, recreation, and retail. Phase I included approximately three of the largest savers per program. This is a census of the largest expected savers and the results are not expected to apply to any other groups of participants in any of these programs or to provide any general conclusions about the programs overall. When the Large Savers study is completed, increasing reliability of the savings estimates for these specific large projects will contribute to increased reliability NYSERDA's overall program and portfolio estimates.[4]

A Nourishing Salad Bar Approach

Historically, the free-ridership and spillover effects have been assessed through telephone surveys with the primary program contact at the subject companies, for large savers and smaller participants alike (as is done with almost all self-report survey-based free-ridership studies around the country). The creation of site-specific measurement and verification plans for the highest savings projects is a common expectation for high quality gross savings evaluations. NYSERDA's Largest Savers evaluation created a systematic approach to do likewise for free-ridership and participant spillover estimation. NYSERDA and its impact evaluation team created a customized and in-depth mixed methods (quantitative and qualitative methods) approach to estimate free-ridership for this cross-program study of its largest saving participants.

Referred to as their "salad bar approach," where all respondents received a core set of questions ("the lettuce"), then select instruments were applied per project ("the toppings"). In order to understand the decision-making process in the participating organization, the assessment team sought to identify and conduct in-person interviews, where possible, with all the participants' staff who have influence in making decisions relating to energy equipment or energy usage related products/operations. The initial step involves learning how decision-making is made for that particular firm, the project being discussed and for decision-making criteria in general for different categories of end-use equipment. Then this information is used to combine components based upon the type of decision-maker (facility managers, financial personnel, operations personnel) and the type of application (renewables, new construction, DG standard technology and DG emerging technology).

The survey instruments include the same set of questions used in the previous NYSERDA NTGR evaluations plus other inquiries to assess construct validity. This strategy allows the Impact Evaluation

[4] Additional information on the Large Savers Project evaluation study and the methods and results for the gross savings evaluation can be found in Maxwell, Gregoire and Megdal 2009.

Team to compare the results using the previous method with alternative approaches. All interviewees received a common set of inquiries so responses could be compared across decision-makers and that firm's decision-making process. Open-ended questions invited decision-makers to comment at length.

Batteries of questions were developed to be specific to each of the different types of sites and programs, and were customized for each type of decision-maker. The lead engineer for each site conducted the project review (including discussions with the project implementation manager at NYSERDA) and the initial decision-making survey. From these, they pulled together the instrument(s) for each of the NTGR interviews to be conducted on-site. A diagram depicting this "salad" construction is provided in Figure 1. Initial training on the salad bar approach and the subsequent expected interviews was conducted through structured training provided via web and teleconference.

Figure 1. The Largest Savers Evaluation Salad Bar Approach for Assembling the NTGR Instruments

The analysis combined quantitative and qualitative information from these interviews to ensure consistency and in-depth, firm-specific assessment. The NTGR team for each project consisted of that project's lead engineer, another senior engineer from the Large Savers evaluation study, and a senior NTGR

expert (with a social science background supporting this behavioral assessment and the mixed use of quantitative and qualitative information).[5] The process itself also included a mixed-methods approach.

The process for estimating the NTG factors for each site was conducted as described below.

- The three senior professionals first independently assessed and estimated free-ridership and participant spillover for that project based upon the interview data collected for that project (across all interviewees). These independent assessments allowed the process to test and enhance inter-rater reliability.
- Once the independent assessments were made, a teleconference was scheduled. Each member of the project-specific NTGR team presented his or her initial estimates (prior to any other questions or discussion). The lead engineer answered questions from the other two reviewers, described perceptions, intonations, and other context. This was followed by an open discussion covering various issues arising from the interviews and the challenges in interpreting the responses across the various interviewees.

The discussions varied considerably depending upon the responses received, the decision-making process, the type of project, the type of customer, whether the organization is a non-profit, a government entity and their funding circumstances. Some of the discussion questions asked during the NTG teleconferences can not be incorporated into standard telephone surveying, or even most enhanced self-report interview approaches. The topics covered by the Impact Evaluation Team during the deliberations for determining a project's free-ridership included the following:

- Had the participant previously enrolled in a NYSERDA program? If so, did that seem to influence their decision-making? Their responses?
- Are there indicators the responses from the participants may have had embedded bias?
- Where did each influential person fit into the decision-making process? Could differing roles have influenced the type of information they were provided by the others (i.e., did the organization process for deciding upon this project's investment create filters in the information provided to the different parties? How do we think that these factors affected the reliability or direction of their responses?
- What percent of the vendor's business is based on NYSERDA programs? Then the deliberation discussion included how that might be influencing their responses or the interactions they had with the customer (and their responses regarding vendor influence)?
- How did specific responses compare among the decision-makers and vendors?
- What can be deduced from the body language? (There was one case where the personnel refused to schedule individual interviews and would only schedule a joint meeting with management present. The review of this information included significant discussion on body language and the ways in which the different parties interacted as questions were asked.)

This process produced a consensus estimate of free-ridership and participant spillover, and an estimated upper and lower bound for these values.

[5] Most of NYSERDA's commercial and industrial programs include pre and post-retrofit M&V, and for the larger projects interaction and review by both program staff and an additional quality assurance contractor. Besides impact evaluation, there are also separate market and process evaluations conducted. Conducting multiple interviews of participants for the impact evaluation was not reasonable without potentially creating significant bias in evaluation participation. Obtaining cooperation after the many interactions already required of participants was still quite difficult. Since a senior engineer was required to conduct the site visits and on-site measurement, a process had to be developed, and training provided, for them to do the NTGR interviews while conducting the gross savings interviews. Designing interview instruments and a process for joint analyses with social scientists brings together the recommended use of engineers for complex projects recommended by Goldberg and Scheuermann (1997) with social science expertise in analyzing organizational psychology and decision-making.

The Meal Consumed: Overall Phase I Free-Ridership Results

The overall free-ridership level seen in Phase I is fairly high at almost 53%. This result is not completely unexpected given that a higher proportion of the largest customers can be expected to have well trained engineering staff and internal resources to search, consider and finance efficiency improvements.

Free-ridership is based upon customer knowledge concerning the equipment and building options, the vendor's depth of experience and knowledge of efficient equipment, the customer's decision-making process and financial situation, and the circumstances surrounding the equipment purchase or building construction. There can be significant variation across customers and projects. For the Phase I projects, free-ridership estimates range from 10% for CIPP to 88% for NCP. The free-ridership for most of the large C/I programs were around the average with DG-CHP at 52% and 59% for PLMP.

Table 1. Phase I Project Free-Ridership Estimates from the Mixed Quantitative/Qualitative Depth Method

Program[1]	Free-Ridership Estimate[2]
PLMP (n=3)	59%
CIPP (n=3)	10%
NCP (n=3)	88%
DG-CHP (n=4)	52%

[1] These results are for the projects studied only and will not be applied to the entire program. The program name is used, and the TA program is not within these program-by-program free-ridership results, so as to maintain the confidentiality of the sole participant in the study.

[2] All estimates are weighted by the *ex ante* gross savings estimates.

Since the Large Savers evaluation was based on selecting a census of the projects in NYSERDA's portfolio with the highest savings, it is not necessary to calculate sampling statistics. The upper and lower bounds for FR in this study are not confidence intervals, but rather are derived through the mixed quantitative-qualitative method described above and setting the intervals based upon uncertainty and potential measurement error. The upper and lower bounds for each project were discussed during the project-specific NTGR teleconferences and decided upon based upon the variation seen within the participants' responses and the variability in responses among the different players interviewed for the same project. Some projects were assigned a narrow range between the upper and lower bound while others had very wide ranges, reflecting the uncertainty in their FR estimates on a project-specific basis. Similarly, the upper and lower bounds were not constrained to be equidistant from the consensus point estimate. In some cases, the lower bound was the same as the consensus, best estimate, and for others the upper bound was the same as the consensus estimate.

Healthy Outcomes: Construct Validity, Consistency, In-Depth Understanding and Strong Reliability

Ridge, Willems and Fagan (2009) discuss estimating free-ridership as an influence index given the measurement is of a latent or underlying construct. Many other areas of scientific research depend upon measurement of an underlying construct. These include assessments of intelligence, deprivation or mental retardation versus human functioning, or, in the earlier example of young women's condom use, sexuality

acceptance and self-efficacy. One of the principals for undertaking quality science when measuring an underlying construct is to include an assessment of construct validity.

Construct validity refers to the extent to which an operating variable/instrument accurately captures an underlying concept/hypothesis, properly measuring an abstract quality or idea. In program evaluations, construct validity enables the researcher to "generalize from the variables and their observed relationships in a program evaluation back to the constructs and their relationships in the program logic" (McDavid & Hawthorne, 109). When conducting an evaluation, one must be sure that what is being measured and how it is measured relates to the construct on which it is based. An evaluation that is based on an unexplored concept can lead to faulty assumptions and invalid evaluation outcomes.

The construction of the customized (salad bar approach) in-depth mixed method deliberately included exact replication of the prior NYSERDA method within its inquiries and calculations to allow comparison of the results for each project, by program and overall between this evaluation's method and NYSERDA's standard self-reported free-ridership method used for all its program evaluations.

The results from the Large Savers evaluation provide support for NYSERDA's prior method. The prior method replication produced a weighted average FR of 53% compared to an estimate of 53% from the more in-depth mixed method used in this evaluation (less than 1% FR difference that disappears with rounding of each estimate). There is more variation between NYSERDA's prior method and the Large Saver's method on a program level, where the prior method is not within the lower and upper bound of this evaluation's estimate for PLMP and DG-CHP. The in-depth mixed method produced a lower estimate of FR for both of these programs. There were, however, only three projects assessed in Phase 1 for most of the individual programs, *i.e.*, the differences between the prior method's and the in-depth mixed method FR estimates on a per-program basis is only based upon three cases each.[6]

The comparison of NYSERDA's prior method with the much more in-depth and comprehensive method used in this evaluation significantly supports the construct validity of NYSERDA's standard self-report survey method. The in-depth method also allows for direct comparison of responses among a range of the parties instrumental in the installation of the efficiency measures, and thus allows for the potential to mitigate self-reporting bias from a particular decision-maker by combining the self-reports from the various perspectives. Through the depth process, it was also possible to weight the NTGR impacts based on the relative importance of the decision-maker at each site.[7] The consensus process also included explicit discussions concerning potential biases, direction and evidence for any biases within a set of interviews.

Many evaluators and implementers would like to minimize the number of types and variety of FR questions asked of the participants. Minimizing the number of these questions could also allow other important inquiries to be made of participants in the same instruments without fatiguing the interviewee. The in-depth mixed method used in this evaluation included a substantially larger variety of self-report FR questions than is typically employed, and many of these questions could be viewed as providing independent FR estimates. This approach allows for a more detailed comparison of the range of possible questions and testing the consistency of the responses, which is useful for more fully understanding the consequences of restricting the number of FR questions.

From this vantage point, the individual responses to the FR-related questions were analyzed as independent implied FR estimates for each participant interview. This analysis was conducted on a case-by-

[6] Due to changes in classifications of two projects one program ended up with only one project in the Phase 1 evaluation. That program is omitted from the program-level reporting to ensure confidentiality for that project's interviews.

[7] In previous NTG studies, NYSERDA has interviewed some of the different decision-makers, but there was no clear method to weight the results to account for the relative influence of the each party in the decision-making process.

case basis. This procedure resulted in some participants having implied FR estimates for the variation assessment based on only a few questions while others had estimates based on 10 or more questions.[8]

The full range in indicators of implied free-ridership for each project was examined (the minimum and maximum indicators for each participant's responses). Some participants showed little variability in all responses, contributing to a tight upper and lower bound estimate of their FR through the in-depth mixed method process. Figure 2 shows all the implied FR estimates for the Phase 1 participants. (The solid line in Figure 2 presents the mean for these participant-implied FR estimates.) A few projects gave highly consistent responses. One participant answered all questions in a manner that would imply very high free-ridership (Participant 8 in Figure 2), while two answered all questions with indications of very low free-ridership (Participants 4 and 7). However, most participants provided at least some variation their responses. Over half of the participants provided responses that could vary these implied NTGR estimates by more than a 50% FR score for that participant.

Figure 2. Within Project Implied Customer FR Estimates if Viewed Independently

Large differences between any respondent's minimum and maximum FR equivalent does not, however, necessarily mean that using these questions together provide an unreliable measurement of the underlying construct of free-ridership. The differences between the minimum and maximum FR equivalents are substantial. Yet, these differences only point to the difficulty of measuring the underlying construct –

[8] These were only constructed for the FR variation assessment. The project's free-ridership estimate, and lower and upper bounds were derived via the three professional quantitative and qualitative data reviews and discussions described earlier and including all the information in its entirety.

different respondents interpret the same questions differently. Those with a range in responses are also those where it is important to use the responses together to derive the underlying "story". This is best illustrated through a few examples of the different scenarios seen. These include the following cases:

- A participating decision-maker (DM) says they planned/intended on installing the exact same efficiency equipment at the same time but then other inquiries show that they could not get financing to adopt these plans;
- One DM said the firm definitely would have made the installation but that individual only received information by another DM in that firm's decision process that included the incentive in the project's payback information, a case of filtered information, and the DM that provided the filtered information said the firm would not have installed that equipment without the incentive; and
- A DM said they had plans to make the changes (but in context this response was assessed as them having plans to do something but not the exact measure implemented) because they would not have funded the necessary technical assistance, had no knowledge of the measure recommended and implemented, and the entity had difficulty obtaining funds for anything more than basic necessities.

The variation assessment examined the variation and the clustering of each project's responses. Figure 2 shows significant clustering of responses. While free-ridership questions appear to have been interpreted differently by different respondents, utilizing several inquiries to measure the underlying construct would appear to allow the FR construct to be measured where the clustering occurs. Varying interpretations of question wording or in the context of firm circumstances may be the cause of the range seen in a respondent's responses.

The average weighted standard deviation and the coefficient of variation (the standard deviation divided by the mean) of customer responses and of vendor responses within this variation assessment are presented in Table 2. Comparing the results in Table 2 to those seen in the project-specific range estimates (Figure 2) demonstrates that while the minimum and maximum FR equivalents may be distant for a respondent, using several measurements creates clustering of responses around the likely measurement of the construct of free-ridership for that respondent. If this clustering did not occur then the large differences between minimum and maximums discussed earlier would be replicated as large coefficient of variations in Table 2. That does not happen, which corroborates the simple examination of data that showed this clustering. This comparison provides strong evidence supporting the need to have multiple questions to measure the underlying construct of free-ridership and that doing so can provide defensible estimates of free-ridership.

Table 2. Within Project Customer and Vendor Free-Ridership Variation Assessment

	Customer FR Responses	**Vendor FR Responses**
Average weighted standard deviation	16.8%	10.2%
Coefficient of Variation (*c.v.* = sd/ mn)	0.37	0.35

The variations in responses and the intertwined inquiries needed to understand the decision strongly support the need for self-reported FR inquiries to include a battery of questions best suited to measure the underlying construct, the need for experienced evaluators that understand both the science and art behind self-report FR measurement, and the need for self-report FR survey methods to recognize and deal with inconsistency and variation in DM responses. The intertwined nature of inquiries needed to understand a

firm's context and decision-making may also indicate that it might be useful to replicate this case by case mixed quantitative and qualitative in-depth method for a portfolio's largest projects to gain greater reliability in those FR estimates.

Conclusion

The customized site-specific inquiries that were possible due to the salad bar approach allowed the gathering of important information that could only be obtained through in-depth interviews with qualitative and quantitative investigations. Both the interviews and the estimation process relied upon a mixed-methods approach, using qualitative and quantitative data and processes, to derive the consensus free-ridership estimates. These estimates provided in-depth, participant-specific assessment for any inconsistencies, effects of common social psychology concerns, and assessment of interactions between the flow of information between the different players and that site's decision to invest in efficiency. The method and process produced free-ridership estimates that demonstrate construct validity, consistency, low variation (as an indicator of estimate reliability). The method and process also provided a deeper understanding of these decisions at these particular firms, enabling the evaluators to strongly support the reliability of the free-ridership estimates produced.

References

Bryan, Angela D., Leona S. Aiken, and Stephen G. West 1996. "Increasing Condom Use: Evaluation of a Theory-Based Intervention to Prevent Sexually-Transmitted Diseases in Young Women," *Health Psychology* 15 (5): 371-382.

Goldberg, Miriam L. and Kurt Scheuermann. 1997. "Gross and Net Savings for Unique Projects." *In Proceedings of the 1997 International Energy Program Evaluation Conference,* Chicago, IL: International Energy Program Evaluation Conference.

Goodman, Robert, Howard Meltzer and Veira Bailey. 2003. "The Strengths and Difficulties Questionnaire: a pilot study on the validity of the self-report version," International Review of Psychiatry 15 (1 & 2), February: 173 - 177

Harrison, Lana D. 1995. "Validity of Self-Reported Data on Drug Use" *Journal of Drug Issues* 25 (1): 91-111.

Maxwell, Jonathan B., Cherie Gregoire, and Lori Megdal. 2009. "Large Lessons Learned: Impact Evaluation of Projects That Reported Over 1,500,000 kWh/yr Savings" *Proceeding of the 2009 International Energy Evaluation Conference*.

McDavid, James C., and Laura R. L. Hawthorne 2006. *Program Evaluation & Performance Measurement*. Thousand Oaks, California: SAGE.

Meissner, Jennifer, Cherie Gregoire, Steven Meyers, Lori Megdal, and Kathryn Parlin 2008. "Allocating Impact Evaluation Resources: Using Risk Analysis to get the Biggest Bang for your Buck," *Proceedings of the 18th National Energy Services Conference*.

Megdal & Associates 2008. *Prospective Benefits Impact Evaluation of the New Construction Program*, Prepared for the New York State Energy Research and Development Authority (NYSERDA), Principal investigators: Antje Siems of Opinion Dynamics Corporation and Dr. Lori Megdal, Megdal & Associates, November.

_____ 2007. *Risk Analysis for NYSERDA's System Benefit Charge (SBC)-Funded Impact Evaluation Estimates*, Prepared for the New York State Energy Research and Development Authority (NYSERDA), Principal investigators: Steven Meyers of Rational Energy Network, Dr. Lori Megdal, Megdal & Associates, and Kathryn Parlin, West Hill Energy & Computing, November.

Petersen, Anne C., Lisa Crockett, Aryse Richards, and Andrew Boxer. 1988. “A self-report measure of pubertal status: Reliability, validity, and initial norms”, *Journal of Youth and Adolescence* 17 (2): 117-133.

Puleo, Stephen 2003. *Dark Tide: The Great Boston Molasses Flood of 1919*. Beacon Press.

Ridge, Richard, Phillipus Willems, Jennifer Fagan and Katherine Randazzo. 2009. “The Origins of the Misunderstood and Occasionally Maligned Self-Report Approach to Estimating the Net-To-Gross Ratio,” *Proceedings of the International Energy Program Evaluation Conference*.

Ridge, Rick, Philippus Willems, and Jennifer Fagan 2009. “Self-Report Methods for Estimating Net-to-Gross Ratios in California: Honest!”, *Proceedings of the 19th National Energy Services Conference*.

Shew, Marcia L., Gary J. Remafedi, Linda H. Bearinger, Patricia L. Faulkner, Barbara A. Taylor, Sandra J. Potthoff, and Michael D. Resnick 1997. “The Validity of Self-Reported Condom Use Among Adolescents,” *Sexually Transmitted Diseases* 24 (9): 503-510.

Siems, Antje, Jennifer Meissner and Lori Megdal 2009. “Prospective Benefits Analysis for NYSERDA's Commercial New Construction Program”, Poster at the 2009 International Energy Program Evaluation Conference, Portland, Oregon, August 12-14.

Summit Blue Consulting 2006. *New Construction Program (NCP) – Market Characterization, Market Assessment and Causality Evaluation, Final Report*, Prepared for the New York State Energy Research and Development Authority (NYSERDA), May.

Winters, Ken. C., Randy D. Stinchfield, George A. Henly, and Richard H. Schwartz. 1990. “Validity of Adolescent Self-Report of Alcohol and Other Drug Involvement,” *Substance Use & Misuse* 25, (S11): 1379 - 1395