

# Integrated, Real Time (IRT), on-going data collection for evaluation – benefits and comparative results

Susan L. Gordon  
Skumatz Economic Research Associates, Inc.  
USA

Lisa A. Skumatz, Ph.D.,  
Skumatz Economic Research Associates, Inc.,  
USA  
skumatz@serainc.com

## Keywords

real time, on-going, data collection, evaluation, comparative results, building programme evaluation

## Abstract

Assessment of the energy and demand savings attributable to building energy efficiency programs are usually conducted using surveys of past program participants, and oftentimes past non-participants. However, retrospective surveys – calls back to participants and non-participants from past years – have drawbacks including potential recall issues, low response rates, and high cost. We propose and test one workable and effective approach that helps address these problems – integrating evaluation data collection as part of already-established points of contact with participants.

On-going data collection provides several advantages over traditional methods. Data are available at any point during the program's implementation – so evaluations can be conducted quickly. The data are collected at a point closer to when the program participation decision is made – potentially improving the quality of the data and consistency with actual decision-making. Finally, if data are collected as part of the ordinary course of the program, then it saves the expense of phone (or other) surveys, and can considerably increase the sample sizes available for the analysis. This approach has been used by several utilities, and the paper will describe the experience to date, benefits, costs, and lessons learned. The paper will compare results from retrospective attribution surveys and the real time, program-integrated surveys, and discuss how both types of surveys compliment each other to provide a more complete attribution evaluation.

## Introduction

Attribution of effects due to energy efficiency (EE) programs are usually based on surveys of past program participants. This approach, incorporating calls back to programs, have three key problems:

- recall issues,
- low response rates, and
- high cost.

This paper describes the methods and performance of an approach that helps address these problems. Incorporating evaluation data collection efforts into already-established points of contact with participants<sup>1</sup> provides several advantages:

- Evaluation data are always available, so evaluations can be conducted quickly and in enough time to support real-time evaluations for use as mid-course corrections in implementation. The questionnaires can be developed up-front, along with other program and/or evaluation materials, and the surveys are then sent with other documents that the programs require from participants (work completion forms, etc.). These can be gathered in a central place, and then keypunched quickly at any time an evaluation update is needed.

---

1. For example, established points of contact with participants include forms participants must already fill out and return to the utility; for example, they often have to return "work completion" forms, verifying that a contractor completed the work that is being rebated, other verification forms (for example, noting they received three bids), forms requesting rebates, and other participation-related documentation that varies according to the program.

- Evaluators would have a series of the same questions over time, allowing them to track and assess changes in values for free ridership and spillover. This enables changes to be tracked and program decisions / refinements to be made in real time.
- The data are collected at a point closer to when the program participation decision is made – potentially improving the quality of the data and consistency with actual decision-making.
- If data are collected as part of the ordinary course of the program, then it saves the expense of specially dedicated telephone (or other) surveys conducted at a later and separate time. The data have been gathered as one more form that participants fill out as part of their program participation – and minimal expense to the utility.
- Incorporating the data collection forms or surveys into standard forms associated with the program, the sample sizes will be larger, and this has been a critical problem in many evaluations.
- Even though forms may need to be short, different “modules” of questions can be asked of different samples of participants, supporting real time data collection on an array of useful information – process and satisfaction, decision-making and attribution, and other evaluation topics.

### Integrating Data Collection Into the Program Process

The rationale for what we have dubbed integrated real-time (IRT) data collection includes.

- **To be efficient and non-intrusive to the program and participants:** The real-time work collects data as part of the standard practices of the program, doesn't modify the program's design, and avoids extra interruptions to the program actor.<sup>2</sup>
- **To implement an on-going data collection system:** Establishing an on-going data collection system allows future evaluations to be conducted with short notice, reduces the need for other data collection efforts, and reduces costs.
- **To collect timely quality data:** The data collection effort gathers data at optimal points in time – specifically, nearer the point at which decisions are made than can be achieved with periodic large-scale surveys, improving accuracy and recall.
- **To collect higher quality data:** Adding several key questions keeps the data collection effort from becoming onerous, reducing the drop-out rate and respondent fatigue.

These efforts allow evaluators to gather data from all key program actors that interact directly with the program – potentially including owners, vendors, contractors, or others – in an on-going basis, piggy-backing on implementation and regular

program communication to provide data for current and future evaluation efforts. Specifically the integrated, real-time (IRT) data collection efforts allow:

- Information on the program process / administration and satisfaction;
- Updated information on the “market”, where improved information is available;
- Enhanced information on decision-making, program influence, spillover and free ridership with key market actors to address attribution questions;
- Incremental price or cost information;
- An array of market progress information from actors involved in the program;
- Priority program progress indicators; or other information useful for evaluation of the program.

Effective intervention points for these data collection efforts depend on the program's design. To be efficient, intervention points for collecting the data may include:

- Adding a separate survey form to accompany the forms submitted to obtain job-related program rebates,
- Adding questions or additional forms in association with existing forms participants must already complete and return (for example, forms requesting verification that work was completed, or other documentation, depending on the program).
- Adding a return postcard or other form after purchase of rebated equipment (e.g. CFLs, appliances, etc.)
- Adding a series of questions to possible follow-up telephone interviews that call center staff may conduct with participating and non-participating households / businesses / contractor or other actors; or other intervention points.

Incorporating data collection at these points assures good response – especially if the forms are required in order to get rebates or otherwise close out the program participation. However, evaluators may be concerned that respondents will “tell the researcher / program what they want to hear”.<sup>3</sup> We discuss differences in real time results compared to results from surveys conducted well after program participation later in this paper. It is important to promise that the responses to these evaluation questions are completely independent of getting the rebate (and potentially will be kept confidential). Whether responses differ depending on the stage of response (real time or after-the-fact) is a researchable question – and will be discussed below. However, it might be argued that, all other factors held constant, the information obtained from IRT data collection would be more closely related at point in time to the decision-making and influence of the program – if the participants can be convinced to answer honestly.

Integrated, real-time data collection may not be appropriate for all instances or for measuring all types of progress indica-

2. It will represent one additional form for participants, but has, to date, not negatively affected any standard aspects of the programs in which it has been used.

3. Which can be a concern no matter when surveys of this type (or any type) are asked.

tors. For example, it is more complex to obtain feedback from some types of actors in real time – particularly those influenced by the program, but for which there is limited direct contact. Situations in which IRT data collection may not be appropriate may include:

- Low priority data or indicators<sup>4</sup>: it may not be worth cluttering up the program forms with data collection on low priority or complex program progress indicators. Targeted phone or mail surveys may be more appropriate, or the information may be collected indirectly.
- Indirect actors: Feedback from lenders, building code inspectors, or regulators may not easily be obtained in “real time” as there may be few priority market progress indicators for these program actors and there may be limited opportunities for regular interaction points between these actors and the utility.
- Non-participants: Information from non-participating contractors is difficult to collect using IRT data collection because there are seldom regular interactions with this group at which to intervene and collect data. This is an important group, and separate data collection efforts should be devised to gather the necessary data, including market characterization / assessment, and attribution data, including information related to non-participant spillover. This is an important problem, and raises the issue that the two contexts -- IRT data collection from participants and off-side/ex-post data collection from non-participants -- occur in a different context. However, by conducting IRT for some elements, the evaluator can save money that they can apply to improved data collection for other aspects of the evaluation (e.g. non-participant interviews where there might not otherwise have been budget, or larger-sample surveys).

In our work for clients, our data collection focused on attribution and baseline information<sup>5</sup> and not on awareness or other indicators in order to minimize the number of questions, focus on priority indicators, and keep the instruments short. However, given the potential for integrating these data collection approaches into forms for all program participants, data to support evaluation of the wide range of program progress

indicators could be collected by randomly assigning different “modules” of data to each participant.<sup>6</sup>

### Description of data collection efforts and topics

Generally, the plan for IRT data collection for existing programs would not be to modify existing documents, which have gone through levels of approval. Instead, our suggested approach would be to prepare additional forms or surveys that will be forwarded (or asked) in addition to the process currently in place. However, for new programs, the questions can be integrated into forms being developed for the program. Depending on the program’s “progress indicators”, sample questions might include the following.

#### MARKET CHARACTERIZATION / ASSESSMENT

- How they found out about the program?
- The name of any contractors, vendors, etc. used
- Whether opinions of energy efficiency have changed since and because of participation
- Whether knowledge of energy efficiency has changed since and because of participation

#### ATTRIBUTION / CAUSALITY FOR OWNERS

- Which measures / features they installed, and which were different than they had planned before participating in the program – by measure or EE feature.
  1. Were they going to install it anyway?
  2. Would they have installed it anyway within a year?
  3. Did they upgrade efficiency beyond what they had planned?
  4. Did they add any other EE measures since participating in the program?
  5. Were any changes due to their participation in the program?
  6. If not, why did they make the change?
- Did they tell any friends or others about the program or about EE measures they installed?
- Did any of these persons or others participate in the program? Did they install EE measures or use EE design?

#### ATTRIBUTION / CAUSALITY / PRICING FOR CONTRACTORS

- Which measures / design features were installed in the home / facility?
- Which were installed because of the program, and what percent of savings that represents?

4. For example, net to gross elements (free ridership and spillover issues) tend to be fairly high priority data, as they relate directly to the cost-effectiveness of the program. Low priority indicators will depend on the program and the use of the evaluation and will be at the discretion of the evaluator’s judgement; however, lower priority indicators for some programs may be related to information channels (for programs that aren’t information-based), etc. In addition, depending on the program, there may be high priority data elements that do not lend themselves to collection via paper or email surveys, which are the basis of IRT data collection – they may need more exploration or “give and take” and require phone or other methods. In those cases, IRT may only help with lower priority data. The evaluator for each specific program will need to assess the priorities and data requirements to support evaluation needs overall; this IRT method provides an opportunity to collect a number of high priority elements in an efficient fashion.

5. Attribution referring to impacts that can be attributed to the program. One aspect is examining attributable savings, which is assessed through basic net-to-gross (NTG) work examining free ridership and spillover. There are myriad other aspects of attributable effects, a concept that will be familiar to all evaluators. Baseline is another basic evaluation concept, related to, as one example, the mark of “standard practices” from which the program may be trying to encourage more aggressive energy efficiency designs. In general terms, it is identifying what would have happened without the presence of the program, and can be a very challenging concept to measure. See many previous papers by the authors and others; Sebald et. al. 2001 is one paper addressing these concepts.

6. The content could include the modules described above, focusing on questions developed as part of the program logic and associated testable hypotheses and progress indicators.

- Did they add EE measures / features in this job beyond what they are claiming in the program? What additional EE measures they proposed were rejected by the owner?
- What is their estimate of the price differential for each measure / feature beyond a standard measure / feature, and for the job as a whole? What share of this difference is made up by the program?
  1. What percent of the non-program jobs they work on use each of these measures and others from a list of program-encouraged measures / features? What percent of similar jobs did they install these measures prior to program participation?
  2. Has EE equipment become more available since the program? What share of this change (if any) do they attribute to the program?
  3. Have they told other (non-participating) contractors about the program or about EE measures they installed? If yes, do they believe it affected measures installed by these contractors?
  4. Number of “qualified” jobs they do per year; number under the program.
  5. Whether there are bottlenecks in obtaining equipment; whether that has changed due to the program (and how)?

Depending on program goals, a host of different questions might be more relevant; however, these types of questions have been asked successfully in IRT data collection work we have conducted.

### Findings and Implications: Lessons for Evaluation Data Collection

We have conducted work for several clients that has allowed us to compare the results from IRT data collection vs. retrospective phone surveys with past participants. We collected data from residential and commercial new construction and retrofit programs, equipment rebates, renewables, and other programs. We compared results from attribution questions related to free ridership, spillover, and net-to-gross attribution questions.<sup>7</sup> These are the core questions allowing evaluators to estimate the share of program-rebated energy-efficient “widgets” that are due to the program, or that were installed above and beyond what would have happened without the presence of the program. This is a key element in the computation of program benefit-cost ratios.<sup>8</sup>

7. We have also collected data on non-energy benefits and other data using this approach.

8. Specifically, for context, brief definitions follow. Free ridership is the share of program participants or share of savings in program records (or rebated) that would have been implemented even without the program. Spillover is a factor that represents additional purchases of energy efficient equipment NOT counted in (rebates or) program records that were induced because of the influence of the program. This includes additional energy efficient measures included in a program job that weren't rebated, measures installed at other jobs by participants or non-participants because of the program or because the program led the market to stock higher percentages of energy efficient measures or other influences. The combination of these two effects – the directly attributable share (1-Free riders) times the spillover (1+all spillover factors) is called the net-to gross factor (NTG).

There are a number of reasons the results from the two sources of data could differ. This includes:

- Data administration effort: The IRT data were collected via mail / paper. The retrospective surveys were all collected via phone.
- Timing of answers: IRT data were collected at the time of participation; the retrospective surveys were conducted up to 2 years after program participation. Recollections could differ between that reported in “real-time” and that from surveys asking about decisions in the past..
- Sample sizes: In some cases, our sample sizes between the groups varied.<sup>9</sup>

In addition, the IRT data collection includes newer participants – and it may be that some of the participants from which IRT data were collected had been participating for a while, and their “baseline” has changed (for example, for builders). That is, you might argue you could expect free ridership reports from actors that have been participating in the program to be higher because they have internalized the program's qualified measures and think of them as their own plan without the need for the program.<sup>10</sup>

Happily, for evaluation purposes, the results were not dramatically different between the two sources. We compare results for a fairly high-priority and complex question in evaluation – the issue of free ridership. This is a key element in translating the program's installed gross savings into net savings attributable to the program's influence. The question asks variations around a question about the likelihood the respondent would have installed the energy efficient measures if the program had not been present. Table 1 provides the results for five programs for which data were collected from a sample of program participants. The retrospective surveys were conducted via phone, interviewing participants that had been involved in the program from six months to about two years prior. The IRT data were collected using written surveys incorporated with program forms, and had been administered to participants from the last 2-6 months, depending on the program. Program designs had not changed substantially across the time periods.

As shown in Table 1, we found there were few patterns in results between the two sources of information. In some cases, IRT results for free ridership were higher than results based on recall from participation in a program two year's prior to the phone survey (Programs B and C); in some cases IRT results were lower (Programs D and E), and in one they were very close (A). However, in no cases (where confidence intervals were available) were the differences significantly different. This variation bodes well for use of IRT in program evaluation. It may also suggest that there are no distinct patterns in people systematically giving higher (or lower) credit to the program depending on the time from decision-making and manner

This factor, multiplied times the program-recorded “gross” savings, provides the estimate of “net” program-attributable savings or program effects.

9. Note that in all cases, samples sizes are small; the client's goal was 30-60 responses and in some of the results in the evaluation report, results from IRT and retrospective were combined. These comparisons are indicative only.

10. And, as a reviewer points out, this would represent some spillover effect.

**Table 1. Results of Exploratory Research on Comparisons of IRT Data Collection and Standard Retrospective (Recall-based, "Retro") Surveys for Evaluation<sup>1</sup> (90% confidence interval reported when available)**

Estimates	Program A		Program B		Program C		Program D		Program E	
	Retrospective	IRT	Retro	IRT	Retro	IRT	Retro	IRT	Retro	IRT
Number of Observations	32	20	48	26	50	38	16	45	17	37
Free Ridership Factor	0.34 (.23-.46)	0.38 (.29-.47)	0.38 (.26-.44)	0.48 (.36-.60)	0.52 (.42-.61)	0.65 (.56-.75)	0.77	0.65	0.83	0.61

<sup>1</sup> Note that in no cases are these differences statistically significant. Small sample sizes are an issue; this was exploratory research. Differences in free ridership result because some programs were market transformation programs, others were equipment rebates, etc. However, the point for this paper is differences in IRT vs. "retrospective" data collection approaches.

(phone vs. mail) in which the survey is conducted.<sup>11</sup> Overall, we found that the results for this important factor were estimated to be about the same from both data sources.

### Summary and Conclusions

Our research indicates that IRT data collection provides results that are not dramatically different from the results obtained using retrospective surveys.<sup>12</sup> The approach provides a number of advantages that tend to make the results robust:

- It tends to increase sample sizes, especially if forms are provided to all participants. The budget savings can be used to 1) follow up on non-respondent participants, 2) data collection from more actors, and 3) gathering additional data from non-participants, which are critical to a real understanding of the program and which are often not addressed well in many evaluations.
- The data are collected near the point of decision-making, which may provide the best input on program influence on those decisions.

Based on our research, we recommend utilities and program delivery agents consider revising the data collection associated with their program evaluation work. We believe IRT data collection has distinct advantages.

- Integrated: The data collection requires no extra effort, and elicits higher sample sizes because it is 1) issued to all participants, and 2) gets higher response (if basically required with the forms that get them their rebates).
- Ongoing and timely: Responses reflect opinions at or close to the time of decision-making, rather than recall from years hence. Further, program managers gain a time series of the same questions over time that allow them to track and assess changes in values for free ridership and spillover (or other data) so changes can be tracked and program decisions and refinements can be made in real time. In addition, an evaluation

can be conducted at any time, even unexpectedly, as the data are collected and available continually. Program managers can (afford to) evaluate the program more frequently and use the results in the program refinements.

- Inexpensive: Relatively little extra effort needed to implement this process. The utility needs to design a survey they would need to design anyway, print up a few more forms, and conduct the analysis. Response rates are higher, which makes analysis easier, and the sample may be less biased because "callbacks" are reduced and recall is not necessary.
- Close to decision: The answers presumably provide more accurate recall of the program and decisions. This provides an opportunity to examine the role of the program (and rebates / initiatives) in energy efficient decision-making more or less at the time of the decision compared to their recall 1-2 years after the decision / installation work.

IRT data collection is a promising data collection method for evaluation research. It may require up-front assessment of program goals and their associated indicators, but this type of pro-planning is useful for programs in any case. The data collection method is very inexpensive. Phone surveys can cost anywhere from \$ 55-\$ 100 or more each, and on-site data collection routinely costs \$ 300 or more for each observation. IRT data are nearly zero cost for data collection.<sup>13</sup> Further, IRT data collection addresses an oftentimes key problem related to low response rates and sample sizes. The data are readily available for analysis, and potentially result in less biased samples of respondents,<sup>14</sup> Although our quantitative results are somewhat preliminary and not comprehensive, they indicate strong performance for data collected in this manner. Most importantly, collecting data in this manner can save significant evaluation dollars – on the order of \$70-\$300 or more per survey depending on respondent type – which can be responsible for savings of \$ 15,000-\$ 90,000 or more in data collection (per relevant actor) for many program evaluations. Again, the increment we are addressing is data collection only; all other steps

11. We compared means for several other variables, including estimates of energy savings from spillover (however, confidence intervals were not available). For inside spillover, we found three cases in which retrospective values were higher, and two in which IRT values were higher. For outside spillover, we found four for which the retrospective values were higher and one for which IRT values were higher.

12. At least for data that were high priority for the evaluations we were conducting (free ridership and spillover). Comparisons could still be made to examine whether there are differences in results and recollections regarding program satisfaction, etc. based on whether the recall is requested at the time of participation vs. two or more years after participation. We were not in a position to do so; others may wish to pursue it.

13. Both methods require development of survey instruments, data keypunching, and analysis; the data collection portion is the savings.

14. The level of bias from survey non-response can be helped especially well if the questions are appended to a required form – then all participants must respond, and a random sample can be selected for analysis if there is too much data (a rare occurrence in the evaluation world!). Regarding bias from respondents telling evaluators what they want to hear, there may be no difference in that tendency regardless of which point the surveys are administered. However, it is a testable hypothesis and one we were not able to incorporate into our work. Others may have the opportunity to investigate this question.

of analysis would remain nearly the same. If we assume the researcher would like to analyze data from 300 or more respondents (based on +/-5 % at 95 % confidence and better *a priori* confidence level computations), and that the program needs to gather responses from 1-3 groups (participants for one, two, or three stakeholder groups, depending on the program design), IRT can represent a fairly significant data collection savings. Of course, IRT doesn't work well for non-participants, and that data collection will remain to be conducted via traditional methods.

These savings are much better applied to improved survey work with non-participants to understand program participation decision-making, non-participant market spillover, and other important program and evaluation factors. IRT data collection should definitely be continued into the future – it provides an easy win for some of the range of efforts involved in evaluation data collection, and is a great value.