From real-time to over-time: Developing a four-year perspective on an energy efficiency portfolio

Ryan Bliss, Mersiha McClaren, and Jordan Folks, Research Into Action, Inc., Portland, OR
Erika Kociolek, Energy Trust of Oregon, Portland, OR

ABSTRACT

Many program administrators and implementers have come to recognize the value of real-time (RT) data for monitoring program progress and making program adjustments to optimize energy savings going forward. But RT data has value that goes beyond use solely for contemporaneous monitoring of transitory information (e.g., program satisfaction in a given month). If properly collected, managed, and mined, RT data can provide information on trends in, as well as impacts of specific events on, assessed program indicators. Analyzing time-trend patterns can provide insights into program success and forecasts for key program indicators.

This paper describes one of the pioneering efforts in RT process data in the energy efficiency world, discusses the challenges faced in using large RT datasets for process evaluation, and presents results from analyses carried out nearly four years' worth of RT data from about 9,000 residential program participants. These analyses investigated the relationship between program success indicators and time, attributes of participation, and program and market changes. In addition, the analyses explored the consistency between qualitative and quantitative responses and provided Energy Trust with information to assess the level and consistency of performance of project contractors.

Background

In 2011, Energy Trust of Oregon initiated a process for collecting real-time feedback from participants in multiple residential and non-residential programs. Dubbed "Fast Feedback," this process involves a short phone survey with recent program participants (within two months of project completion) on an ongoing basis. The survey collects information on participant satisfaction, decision-making, and free ridership as well as suggestions for program improvements for approximately 20 distinct groups: 15 residential measures spanning three programs, and five non-residential programs. The survey assesses certain core topics (e.g., overall program satisfaction, satisfaction with the incentive application) consistently across all or nearly all program groups, but also tailors some questions to specific programs (e.g., satisfaction with equipment performance for appliances or satisfaction with home comfort for shell and sealing measures). The survey also includes some questions that are completely program-specific (e.g., satisfaction with scheduling and pickup for refrigerator recycling).

Program staff receive regular summaries of quantitative survey responses and verbatim responses specific to their respective programs, but these reports provide information about responses for a point in time and no systematic investigation of trends over time had previously been completed with these data. In 2014, Energy Trust selected Research Into Action, Inc., to analyze Fast Feedback data from Q1 2011 through Q2 2014 (42 months), to investigate the following questions:

- Are there time-related trends in survey responses that are independent of geography and/or utility territory, and prior program participation?
- Did certain identified program and market changes affect participant satisfaction, free ridership rates, or other key metrics?
- Are quantitative survey responses consistent with the verbatim comments provided?

- Do survey responses vary depending on what contractor or retailer did the work about which respondents were surveyed?

Answers to the above questions would provide several benefits. Identification of a general trend over time in satisfaction or other indices of program success – or of a single, event-related effect on such indicates – may allow Energy Trust and program implementers to make needed adjustments. Assessing the consistency of quantitative and verbatim responses will shed light on the validity of former and provide context for interpreting the latter. Finally, providing information on how satisfaction ratings vary depending on the contractor or retailer will help Energy Trust and the implementers provide more useful and actionable feedback to program-affiliated contractors and retailers.

This type of project poses certain challenges and risks, some of which are well known but still difficult to address. First, large data sets are beneficial but mean that small effects can be statistically significant even though they may not be practically significant. Analyzing effects across multiple dependent variables (DV) increases the risk that one or more of the analyses will be statistically significant by chance – a phenomenon that statisticians refer to as "Type I error." Analyzing changes over a 42-month period requires attention to any changes in survey methods during that time period. Finally, without a control group it can be difficult to determine whether a specific event produced a change that might otherwise be part of a general time-related trend. Below, we describe how we addressed those challenges together with the benefits that these analyses provide.

Methods

Energy Trust provided Research Into Action with datasets containing quantitative and qualitative survey responses from Q1 2011 through Q2 2014 for 13 residential and four non-residential program groups. As the residential dataset was much larger than the nonresidential dataset (about 9,000 vs. about 1,600 records), this paper focuses on the former. Below, we describe our data preparation and analysis approaches.

Data Preparation

We excluded responses for two residential respondent groups (gas fireplaces and ductless heat pumps) because they did not have enough data points to provide reliable analyses. This left a total of 8,763 records in the quantitative dataset. Each record included: program participation; satisfaction ratings; program services received; free-ridership; general feedback (open-ended); respondent demographics and electric and/or gas provider; information on participation before and after the project for which they were surveyed; and the contractor or retailer involved in the project. In addition, some records contained program-specific data, as described below. A dataset of qualitative information included 2,048 records from Q1 2012 to Q2 2014. Each of these records matched to a record in the quantitative dataset. Each record in the qualitative dataset included 25 binary variables that coded the open-ended feedback.

The evaluation team worked with Energy Trust staff to identify DVs of interest for each program group. These include variables that represent the two components of Energy Trust's free ridership assessment method – what would have happened absent the program ("project change") and the program's influence on the project ("influence") – as well as the combined free ridership rate. We excluded variables that had a high percent of missing data by program groups, either because of low response rate or because of conditional skip patterns within the survey instruments.

We identified a total of 25 DVs (DVs) of interest: the three free ridership variables described above, 14 satisfaction variables, and eight dichotomous variables relating to respondent behavior. Two dichotomous ("yes/no") variables assessed whether respondents visited the Energy Trust website or

received a brochure before participating ("consulted Energy Trust information") or considered the list of approved Energy Trust trade allies when selecting a contractor ("considered TA list"). Four assessed whether solar PV or solar water heating respondents paid for the system with cash, unsecured loan, home equity loan, or equipment loan /power purchase agreement. Finally, two variables assessed whether Home Energy Review participants had done any improvements since the review or plan any improvements in the next 12 months. Fifteen of the 25 variables pertained to only one or two respondent groups, while the other 10 each pertained to at least six groups.

Time Trend Analysis

We used regression analyses to assess whether the various DVs changed systematically over the study period, independently of other variables. For scaled DVs (e.g., satisfaction ratings), we used classical regression.[1] For dichotomous DVs, we used logistic regression. To reduce the number of multiple regression models, we first tested the bivariate correlations between each DV and time for each program group. If we did not find a statistically significant bivariate correlation, we did not proceed with a multiple regression analysis for that DV in that group.[2] For each significant bivariate correlation, we carried out a multiple regression that included all potential moderator variables (geography, utility, and respondent gender) that were related to both the DV and time, as well as their interaction terms.[3] We conducted a total of 93 bivariate correlations. To examine the risk of Type I error, we evaluated whether the pattern of *p*-values departed from what would be expected by chance.

Event Hypothesis Testing

At the beginning of 2013, two events occurred that affected the Existing Homes program: a new program management contractor (PMC) began implementing the program and new rules for "sorting" some projects between the Multifamily and Existing Homes programs took effect. One or both of those events could have affected survey responses for participants in the Existing Homes program.[4]

We first assessed whether those events were associated with significant differences in the various DVs for survey groups in that program, independently of region, utility, and gender. To do this, we restricted the analysis to the last three months of 2012 and the first three months of 2013. We created an independent "time" variable, with the two values representing those two time periods – i.e., a survey had a value of 0 if it was completed in the last three months of 2012, before the two events, and a value of 1 if it was completed in the first three months of 2013, after the events. Using observations only from the six months surrounding the event reduced the sample size, making lack of normality a greater concern. Therefore, we used logistic regression, recoding all scaled variables to dichotomous variables.

If the logistic regression analysis showed a statistically significant relationship between an event and a DV, we carried out a second analytic step to attempt to rule out general time-related trends. We attempted to determine whether the observed difference between the three month pre- and post-event

---

1 For a discussion of using scaled responses in regression, see, for example, Poole, M. and O'Farrell, P. (1971). "The assumptions of the linear regression model." *Transactions of the Institute of British Geographers* No. 52 (Mar., 1971), pp. 145-158. Available online at: *http://people.uleth.ca/~towni0/PooleOfarrell71.pdf*, accessed December 13, 2014.

2 Theoretically, it would be possible for a multiple regression analysis to reveal a significant effect of time on a given DV when the effect was not evident in the bivariate correlation, if one or more of the moderator variables were masking the effect of interest until controlled for in the regression model. We concluded that any effect not seen in the bivariate analysis likely would be small in the multiple regression and not of critical interest.

3 In all cases, the linear functional form explained at least the same amount of variance as the curvilinear functional forms, so we always used linear regression.

4 Since the two events co-occurred and both affected the same respondent groups, it would not be possible to determine which event caused an effect.

periods was substantially deviant from the variation in the program outcome across the rest of the survey period. For each three-month period, we calculated the percentage of respondents with a value of "1" for each given dichotomous DV. Then, for each pair of successive three-month time periods, we calculated the delta in the percentage of respondents with a value of "1" (e.g., the delta between Q1 2012 and Q2 2012). We then tested whether the delta between the three months before and after the event in question was significantly greater than the average of the deltas for the various successive periods. If the pre/post event delta exceeded the mean by at least 2.5 standard deviations, then we concluded that the event-related change was significantly greater than the mean and that the change in the DV was related to the event. In addition to helping rule out general time trends, by introducing an additional criterion for rejecting the null hypothesis of no effect, this second step helped reduce the chance of Type I error.

For each significant effect in a logistic regression analysis, we also examined the corresponding results of the general time trend analysis (described in the previous subsection) to determine whether any overall time trend existed and, if so, whether it was consistent or inconsistent with the pre/post change surrounding the event in question. We evaluated those comparisons on a case-by-case basis.

Qualitative Analysis

We carried out analyses on the 25 qualitative variables to determine whether, collapsing across groups and time periods, the coded comments were generally consistent with the quantitative ratings of program satisfaction (that is, we did not examine the time dimension for these analyses) for the approximately 2,000 respondents who provided verbatim comments. We first reviewed the 25 codes and identified each of them as indicating a positive comment, a negative comment, or a neutral comment.

We then carried out two sets of analyses. First, among respondents that provided any verbatim response, we created three dichotomous variables indicating whether or not they provided any positive, any negative, or any neutral comment and analyzed the relationship between each of those variables and overall program satisfaction. For example, we assessed whether overall program satisfaction ratings differed for those who did and did not provided positive comments.

We then analyzed the relationships between pairs of specific qualitative codes and quantitative survey responses that we hypothesized would provide consistent responses. For example, codes representing broad evaluations of the program experience (e.g., good experience with the program, expectations not met) should be related to most or all satisfaction indices. We identified several pairings that were more specific – for example, the qualitative variable "process or incentive took too long" should be related to the quantitative assessments of satisfaction with the incentive form and incentive turnaround time.

We assessed the relationships between the quantitative and qualitative variables with the Mann-Whitney nonparametric test.[5] To provide for the greatest generality of results, we carried out the above analyses only with quantitative variables that were applicable to most respondent groups.

Residential Contractors and Retailers Analysis

Survey respondents who completed an insulation, water heater, gas fireplace, heat pump, or window project; who did air sealing or duct sealing; or who participated in the Home Performance program rated their satisfaction with the contractor or retailer involved with the project on a variety of indices. The residential dataset identified the contractors and retailers that were associated with the surveyed respondents' projects. Using that dataset, we identified 39 contractors and retailers associated

[5] The Mann-Whitney test is a nonparametric alternative to a two-sample *t*-test, used when it cannot be assumed that the samples come from populations with normal distributions.

with at least 30 surveyed projects; we carried out analyses relating to customer satisfaction with those 39 contractors and retailers.

Like the analysis of consistency between qualitative and quantitative responses, this set of analyses does not examine the time dimension. The benefit of having a large number of survey responses is that it allowed us to provide information on the relative consistency of satisfaction ratings across contractors.

Results

General Time Trends

This section describes the results of our analyses of potential time trends across the study period, Q1 2011 through Q2 2014, in the various survey responses.

Of 93 bivariate correlations across the 11 respondent groups, and involving 25 DVs, we found 28 statistically significant correlations, involving 14 DVs and nine respondent groups. That number represents 30% of the tests run, which well exceeds the 5% expected by chance at the commonly accepted 0.05 significance level (confirmed by binomial test). Moreover, 16 of the 28 correlations were significant at the 0.005 level or beyond. Therefore, we conclude that the risk of Type I error is low.

For seven of the 28 significant bivariate correlations, one or more exogenous variables were related to both time and the DV and so were included in the resulting regression model. The effect of time on the DVs remained statistically significant in five of those seven models. Therefore, time predicted a DV, independently of any of the exogenous variables, in 26 of the 28 regression models.

Time accounted for a maximum of 4.4% of the DVs' variance, but it had a marked cumulative effect in many cases. The upper portion of Table 1 shows the regression results for the ten DVs where time had the greatest cumulative effect. In particular, satisfaction with the application form and incentive turnaround time increased over time in three respondent groups each, and satisfaction with the overall program experience and the scheduling process increased over time in one respondent group each. (The scheduling process applied to only two respondent groups.) Across these analyses, satisfaction showed average increases of about one-quarter to three-quarters of a point on the 5-point scale. The one exception to the trend for increased satisfaction is that respondents in the Home Energy Review group showed an average one-third-point decrease in satisfaction with information they had received on how to apply for incentives.

One or more of the free ridership variables had statistically significant relationships with time for several respondent groups. The bottom portion of Table 1 shows the analysis results for the free-ridership variables for which time had a statistically significant effect. The cumulative effects over the study period for the free ridership components (project change and influence) ranged from an increase or decrease of 0.02 points to an increase of 0.10 points. On a scale ranging from 0 to 0.5 points, this translates to a cumulative effect of 4% to 20% of the total scale range. For the total free ridership rate, the cumulative effect ranged from 0.09 to 0.11 – or 9% to 11% of the total scale range (0 to 1.0).

Effect of Change in the PMC and Sorting Rules

The change in the PMC and sorting rules for the Existing Homes program (the "events") appeared to have narrow impacts but showed no evidence for any broad impacts. The logistic regression analyses showed statistically significant pre-event/post-event differences for two satisfaction variables (incentive turnaround time and overall program experience) and with the "consulted information" variable (whether or not the respondent visited the program website or received a brochure before

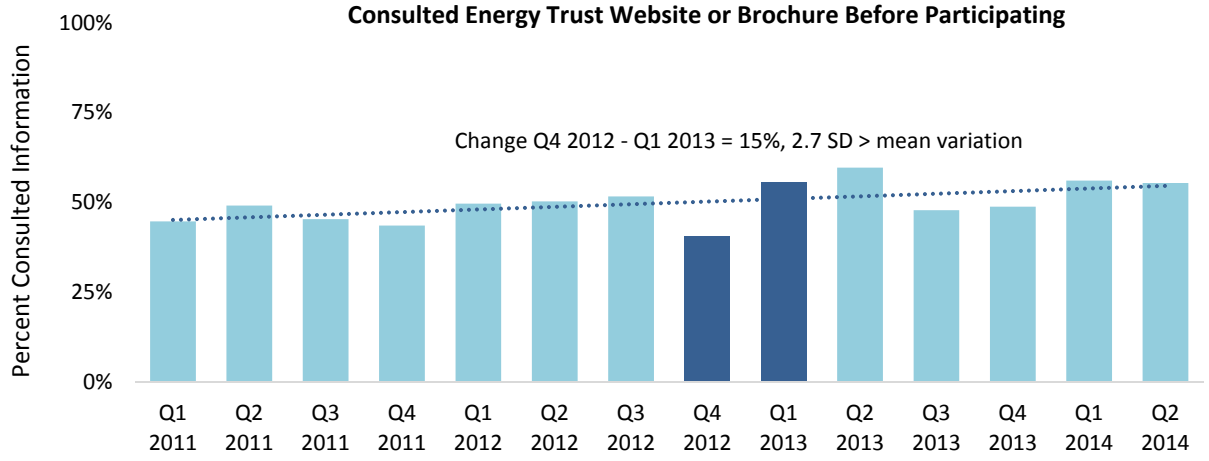
participating). For all three variables, the difference in pre-event and post-event responses occurred only among insulation respondent groups.

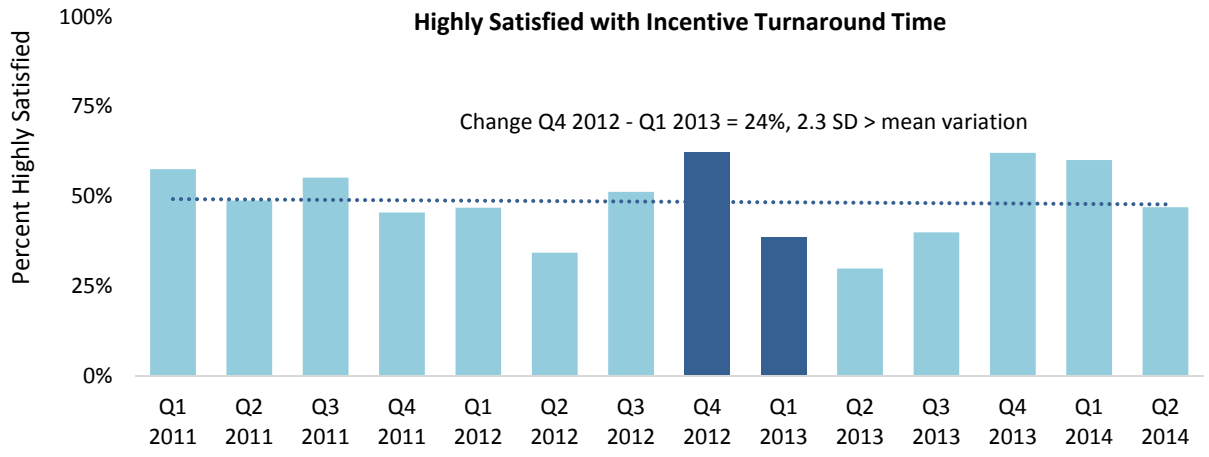
Table 1. Summary of Regression Models Showing Greatest Cumulative Effect of Time on Various DVs

Program Group	**Dependent Variable**	**Regression Type**	R^2	***p*-value**	**Cumulative Effect**
Dependent variables are satisfaction ratings					
Clothes Washer	Application form	Classical	0.009	0.002	0.25
	Incentive turnaround	Classical	0.04	< 0.00001	0.75
	Overall experience	Classical	0.01	0.0004	0.25
Windows	Application form	Classical	0.03	0.00001	0.50
	Incentive turnaround	Classical	0.01	0.001	0.50
Home Energy Review	Scheduling process	Classical	0.01	0.001	0.25
	Application information	Classical	0.01	0.002	-0.37
Water heater	Application form	Classical	0.02	0.0005	0.37
Refrigerator	Incentive turnaround	Classical	0.005	0.03	0.25
Dependent variable is dichotomous – respondent consulted website or brochure before participating					
Insulation	Consulted information	Logistic	0.003	.007	10%
Dependent variables are free-ridership component scores (range = 0-0.5) or total score (range = 0-1.0)					
Windows	Project Change	Classical	0.005	0.04	0.05
	Influence	Classical	0.004	0.04	0.04
	Total FR Rate	Classical	0.009	0.008	0.09
Clothes Washer	Project Change	Classical	0.007	0.009	0.04
	Influence	Classical	0.004	0.03	0.03
	Total FR Rate	Classical	0.02	0.00002	0.09
Insulation	Project Change	Classical	0.02	<0.00001	0.10
	Total FR Rate	Classical	0.01	<0.00001	0.11
Refrigerator	Project Change	Classical	0.004	0.04	0.03
Refrig. Recycling	Influence	Classical	0.005	0.01	-0.02

Note: For all analyses, the cumulative effect is the difference between the predicted value of the DV at the end of the 42-month period and the predicted value of the DV at the beginning of the period, expressed in terms of the DVs' units of measure. For the analyses of satisfaction ratings, the unit is one point in the 5-point satisfaction scale. For the "consulted website or brochure" variable, the unit is a percentage point reflecting the probability of a positive response. For the free-ridership variables, the units are continuous values along the respective ranges.

Only the change in the "consulted information" variable deviated by at least 2.5 standard deviations from the overall variation, our criterion for concluding a possible effect of the program changes, although the change in the two satisfaction variables approached the criterion (with 2.3 and 2.2 standard deviations, respectively). Again, one benefit of establishing this criterion was to reduce the risk that conducting multiple analyses would produce a Type I error. Therefore, although the satisfaction variables came close to meeting the criterion, we did not conclude that a possible effect existed. Figure 1 illustrates the trends in the three variables with statistically significant pre/post-event differences.





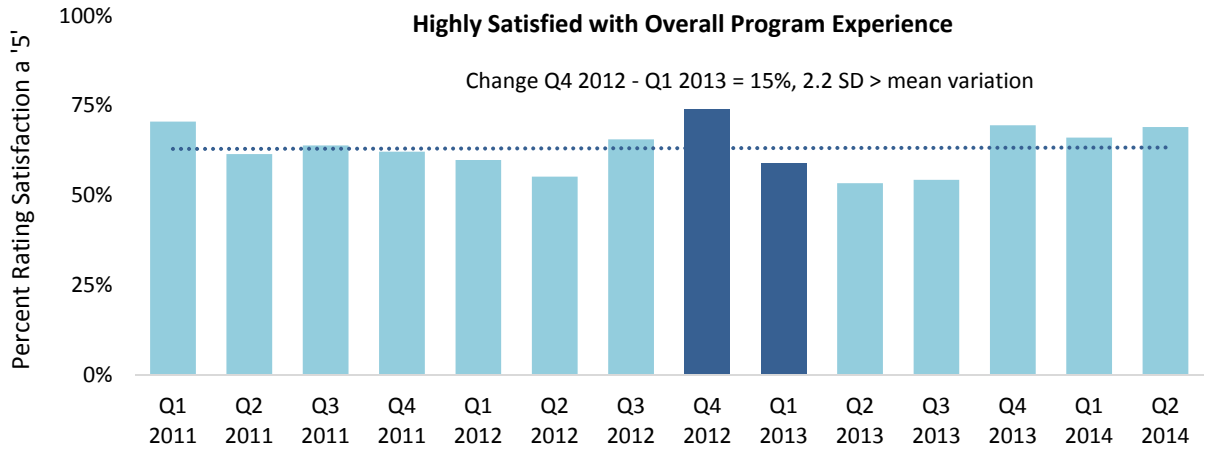


Figure 1. Change in Survey Responses, Q4 2012-Q1 2013, Compared to Overall Variation by Quarter and Linear Trend

We focused on the three-month periods before and after the event because we expected to see the most striking event-related change there, but as noted above, we also checked the overall trend across time for variables showing evidence of an event-related effect. In fact, there was a statistically significant linear trend for the "consulted information" variable in the same direction as the pre/post-event change (Table 1, above). Again, the fact that the pre/post-event change exceeded the overall variation suggests that the change was not a function of the overall time-related trend.

For the two "satisfaction" variables, in addition to the fact that the pre/post-event change did not meet our criterion for concluding a possible event effect, there was no overall linear trend in responses across the entire study period. The lack of an overall trend further suggests that the change in PMC and sorting rules did not bring about a change in satisfaction.

Consistency of Quantitative and Qualitative Responses

Among the approximately 2,000 respondents who provided verbatim comments (about 23% of the total residential sample), positive qualitative responses were associated with higher overall program satisfaction, and negative and neutral responses were associated with lower satisfaction (Mann-Whitney U, $p < 0.001$ in all cases; Figure 2). Although neutral responses were associated with lower satisfaction, this association was much weaker than the association between negative responses and satisfaction ratings. In addition, providing any verbatim response was associated with lower satisfaction; although the relationship was weak, it was statistically significant (Mann-Whitney U, $p < 0.001$), suggesting that a negative experience is more likely than a positive one to move someone to comment. Figure 2 summarizes the overall satisfaction ratings for respondents that did and did not offer positive, negative, neutral, and any comments.

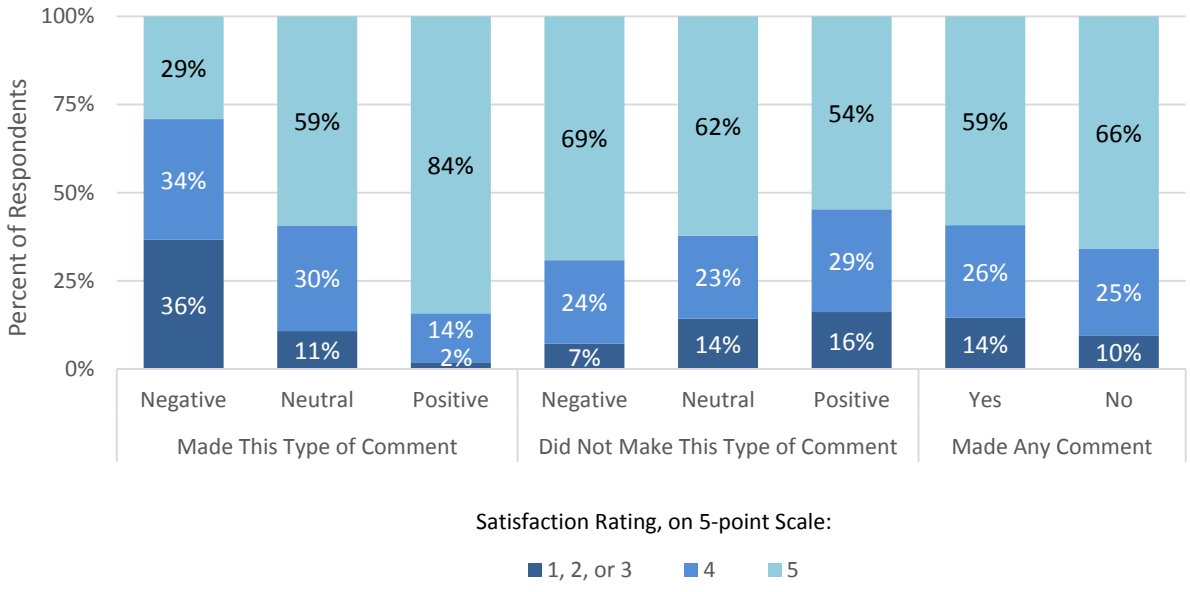


Figure 2. Relationships between Direction of Qualitative Responses and Overall Program Satisfaction

We further analyzed the consistency between pairings of the qualitative response codes and selected quantitative satisfaction responses. Table 2 summarizes the results of this analysis. Because the Mann-Whitney U statistic is not readily interpretable, Table 2 simply identifies all examined relationships that were statistically significant as either positive (↑), meaning that the qualitative response was associated with a higher satisfaction score, or inverse (↓), meaning that the qualitative response was associated with a lower satisfaction score. We identify statistically non-significant findings as "ns." A blank cell indicates that we did not hypothesize a relationship for that pairing and so we did not conduct an analysis.

The results show a general correspondence between qualitative and quantitative responses. In the case of codes indicating positive comments, 11 of 26 examined relationships showed a positive association with rated satisfaction, while the relationship was not statistically significant in the other 15. The codes indicating negative comments were more consistently related to rated satisfaction – with statistically significant inverse relationships in 18 of 23 pairings.

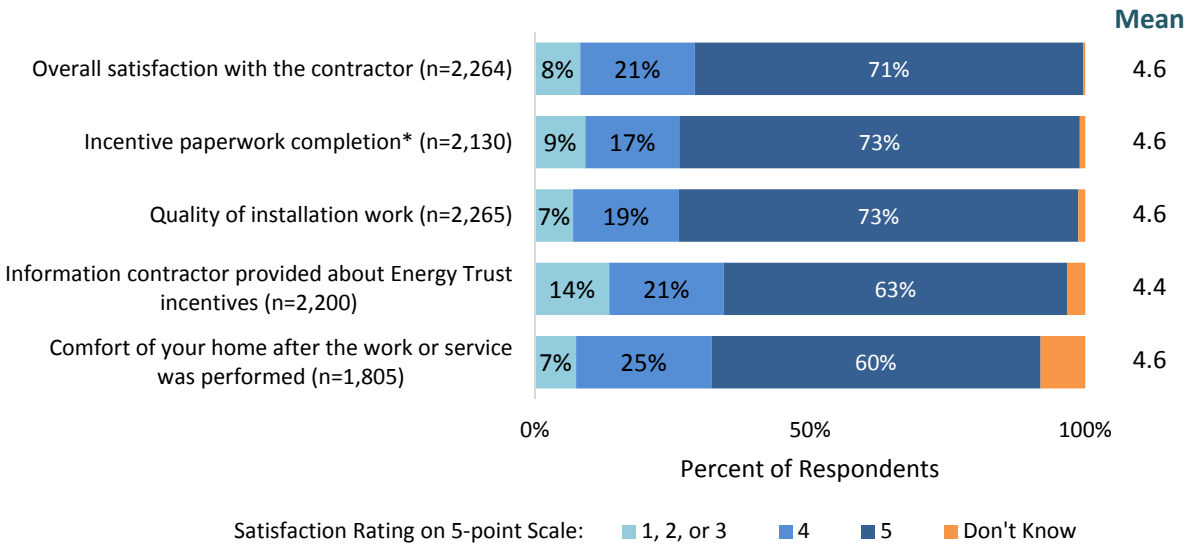
Table 2. Relationships between Qualitative and Quantitative Variables

Verbatim Code	**Quantitatively Assessed Satisfaction Variable**				
	Overall Experience	**Incentive Turn-around time**	**Application Form**	**Perform-ance of new Product or System**	**Contractor**
Good program experience	↑	ns	ns	ns	ns
Interested in re-participating	ns	ns	ns	ns	ns
Overall positive experience	↑	↑	↑	↑	↑
Recommend Energy Trust	↑	ns	ns	ns	ns
Process was overall easy	↑		↑		
Good contractor experience	↑				↑
Greater comfort				ns	
Reduction in bill				ns	
Expectations not met	↓	ns	ns	ns	ns
Overall negative experience	↓	↓	↓	↓	↓
Process took too long	↓	↓	↓		
Process too complicated	↓		↓		
Problem with contractor	↓				↓
Problems with program rep.	↓				
Inaccurate information*	↓	↓	↓	ns	↓

*Indicates information in print, website, or from a trade ally or program representative was inaccurate.

Satisfaction with Contractors and Retailers

The large sample allowed us to investigate satisfaction with contractors and retailers to a high level of precision. As Figure 3 shows, satisfaction with the contractor or retailer was high among those who hired a contractor for a home upgrade project.



* Respondents who did not have their applications completed by the contractor or retailer did not answer this question.

Figure 3. Contractor and Retailer Satisfaction Associated with Projects by High-Volume Contractors and Retailers

Of possibly greater interest, the large dataset allowed us to examine how consistently 39 "high activity" contractors and retailers received high satisfaction ratings from customers. By "high activity" contractors/retailers, we mean those who did at least 30 projects associated with survey responses. Contractors and retailers varied in how consistently they received the highest satisfaction rating of 5 ("very satisfied").

For each category of service, we categorized each contractor or retailer's quality of service based on the percentage of their customers that gave them the highest satisfaction rating for that service:

- "Highest quality" means that at least 80% of respondents rated satisfaction a '5'.
- "High quality" means that 70% to 79% of respondents rated satisfaction a '5'.
- "Moderate quality" means that 60% to 69% of respondents rated satisfaction a '5'.
- "Lowest quality" means that less than 60% of respondents rated satisfaction a '5'.

Figure 4 shows the percentage of contractors and retailers who have provided the above levels of service quality across five categories of service. The distribution of quality levels varies by service category. A casual comparison shows similarities between this figure and Figure 3, with relatively more contractors and retailers providing "highest quality" services, and relatively fewer providing "lowest quality" services, for incentive paperwork completion and quality of installation work than for comfort of home and incentive information.

But closer inspection shows a difference in the information shown by the two figures. While Figure 3 shows that satisfaction for information about incentives is close to the level of satisfaction for

paperwork and installations, Figure 4 shows a greater distinction among the service categories. Specifically, quality of installation work shows much more consistency regarding satisfaction – particularly, more "high quality" work – than comfort and providing incentive information, and even more than incentive paperwork. It also is apparent from this figure that the quality levels relating to installation work are closest to those for overall satisfaction, while this is not at all apparent from Figure 3.

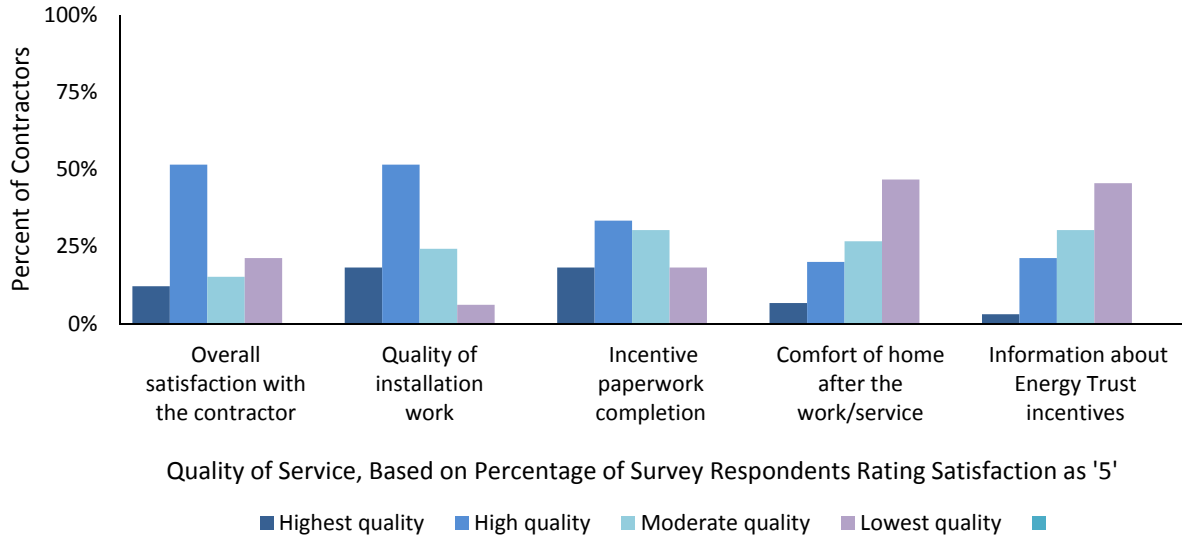


Figure 4. Distribution of Contractors and Retailers Providing Highest to Lowest Quality of Service

This information is valuable to program administrators who need to know which and how many of the contractors and retailers affiliated with the program are providing consistently high quality of various services. But more than this, the focus on consistency shows that those contractors and retailers who receive lower average satisfaction ratings do receive some high ratings as well as lower ones. This type of information may allow program administrators to provide more useful feedback to contractors, possibly allowing them to identify and focus on specific weaknesses to improve their service.

Discussion

This paper demonstrates that data collected to provide near-real-time feedback on program success, over time, can yield a large dataset that can be analyzed and identify findings going well beyond those that can be seen in periodic summaries. These analyses allowed us to: identify increases over time in several aspects of program satisfaction and free-ridership; determine that a change in a program management contractor and program rules likely did not have an effect on program satisfaction; provide internal consistency validation of satisfaction ratings by showing consistency with verbatim responses; and show that not only the overall levels of satisfaction but the consistency of high satisfaction varies across contractors and retailers and across categories of service.

These analyses are useful to Energy Trust in various ways:

- The comprehensive analysis of various research questions will help Energy Trust determine what analyses it want to continue doing on an ongoing basis, incorporating updated data.

- The analysis of a large sample of verbatim comments collected across time provides a larger context for interpreting relatively small number of verbatim comments provided to program staff in monthly reports.
- Showing how verbatim comments relate to quantitative satisfaction scores also provides context for interpretation, since tone is difficult to infer from the comments.
- Energy Trust will share contractor-specific results with the contractors to provide feedback on their performance.
- In the full report, Research Into Action developed a template for contractor results that Energy Trust can populate with updated data moving forward.

This paper also illustrates several actions taken to avoid some of the pitfalls of carrying out multiple analyses using large data sets. We examined the likelihood of Type I error by comparing the number of statistically significant outcomes with the number that would be expected by chance. We examined not just whether or not a trend was statistically significant, but also the cumulative size of the effect. In the "event hypothesis" analyses, we developed a method for determining whether a statistically significant pre/post-event difference was likely attributable to that event – a method that can be extended to analyze the potential effect of any event, including external, market events. Finally, in the analysis of consistency between qualitative and quantitative responses, we tested only relationships for which we hypothesized consistency, thereby reducing the total number of tests run; and, again, we examined the overall pattern of significance. Actions like these will become increasingly important as researchers in the energy world increasingly conduct multiple analyses from large datasets.