Social Marketing Can Be Measured – Best Practices Applied to Four Concrete Examples

Lisa A. Skumatz, Skumatz Economic Research Associates, Inc.
Juri Freeman, Skumatz Economic Research Associates, Inc.

ABSTRACT

Communities, businesses, and institutions are embracing social marketing as a supplement to traditional marketing- and widget-based DSM approaches. Social marketing can be used to achieve not only increased measure uptake but also to change energy use behaviors, both of which result in savings. The authors will summarize "best practices" principles in evaluation of social marketing and behavioral programs, and present examples of four varied types of social marketing programs to illustrate elements of their evaluations, highlighting the successes / weaknesses of the design as it applies to evaluation, and also discuss portions of the program design that could be improved and serve as lessons for other programs and for future evaluations of the programs. The evaluation results indicate that different types of social marketing programs deliver different levels of savings, but the results indicate that social marketing and personal interventions have the potential to generate reductions in energy use and increases in measure saturation.

Introduction

Utilities and governments are constantly assessing and expanding the broad array of tools that can be used to reduce energy use and demand. Increased attention is being paid to social marketing as a strategy to reduce energy demand both with and without the installation of new measures. However, the question of how to achieve lasting savings cost-effectively through social marketing remains – and as an important component, the question arises of how to measure the savings achieved by these "soft" behavioral-type interventions.

We present the key steps to evaluating social marketing and behavioral programs, and then assess evaluations of four social marketing programs to illustrate the concepts and review the evaluation challenges and successes. We discuss the impacts of the programs, the aspects that made the program evaluations successful (and their shortcomings), and present information on what other programs can learn from the design that will support better programs, and more measureable behavioral evaluations.

Social Marketing and Behavioral Interventions – Great Potential, Difficult Measurement

Energy efficiency portfolios are heavily weighted toward measure-based programs. However, increasing amounts of literature suggest that social marketing and behavioral programs have the potential to provide energy savings; according to ACEEE, behavioral actions and choices (with just current technologies) could save perhaps 30 percent or more energy in the US alone (Earhardt-Martinez 2009). Energy-related behavioral programs have a long history (Skumatz and Freeman 2011b), including isolated work in the 1970s, the volume of Bonneville Power Administration Hood River initiatives in the 1980s, shorter shower (and other conservation) messages in California in the water crises in the 1990s, Canadian initiatives in the late 1990s, the energy challenges from the California energy shortages in the early 2000s, a decade of Energy Star™ initiatives, and most recently, the "***setsuden***" (energy saving) momentum in Japan after Fukushima (Kakuchi 2011, New York Times 2011), among many others.

Attention to behavioral options has increased, assisted by advancements in hardware and software. Examples include billing feedback programs, real-time pricing, thermostat feedback, smart meters, and other programs designed to affect behaviors and equipment use. Most of these programs have incorporated sophisticated outreach efforts that incorporate some or all of the principles of social marketing – techniques that have been applied widely to curb teen drug use, reduce drunken driving, shame litterers, and reduce obesity. Briefly, hallmarks of social marketing campaigns in energy include:

- Combining traditional marketing techniques with sociological and psychological tools to influence a target behavior.
- Going beyond the awareness focus of most traditional outreach by incorporating the identification of barriers and motivations, targeting a specific sector, and the use of tools such as social norms, prompts, and feedback.
- Working to incorporate multiple "touches" to try to change and habitualize behavior change (Community-Based Social Marketing / CBSM is seldom a one-off effort).
- Working to reach out through social networks (faith-based, neighbors, community partnerships, etc.) to make connections, aid in credibility and transfer, etc.

Elements of social marketing focus on influencing changes in how consumers behave to increase measure uptake, increasing overall program participation, changing the way consumers and customers think about and utilize energy, impacting maintenance and upkeep schedules, and helping build stronger customer relationships. Examples of CBSM tools used in energy efficiency programs include applications in billing information programs (using feedbacks, norms, prompts, and messaging); real time pricing (norms, prompts, and messaging), and energy audit / measure installation programs (incentives, door-to-door outreach, social networks, norms, and prompts).

Although energy-based social marketing and behavioral programs have been around since the 1970s, the real integration of social marketing and behavioral interventions into portfolios has been slowed by concerns that these programs are less reliable – in terms of savings and lifetimes – than measure-based interventions. This concern reflects two issues:

1) Behaviors lead to uptake of measure-based programs. The participation decision is behavioral in nature.
2) Behaviors influence measure performance. Behavior has an important influence on the performance of measure-based programs in terms of both savings and lifetimes. Although some interventions (e.g. building envelope upgrades), may have little-to-no interaction with behavior, most programs incorporate some behavioral component. Behaviors influence the performance of measures through the way customers use the measures and the behavioral decisions about upkeep and O&M. The measure-based program's performance is intimately related to the behaviors, which can be influenced with social marketing and behavioral-based strategies.[1]

The practice of drawing a clear line drawn between "widget" and behavioral programs misses the fact that the former are influenced by behavior, and most programs fall within the continuum between widget and behavioral extremes.[2] Traditional measure-based programs have established and tested

---

[1] Performance of energy efficient refrigerators (savings and lifetimes) is impacted by coil-cleaning behaviors; optimal performance of HVAC systems is only realized through proper heating and cooling settings; maintenance, and installation in locations with sufficient space. Purer behavioral efforts include large scale social marketing campaigns with little to no focus on measures, and of course, there is a wide array of programs in between.

[2] Classifying programs (or actions or savings) as one camp or the other is difficult, making it obvious that there is no clear line. Consider the case of someone deciding to install a CFL. Does the credit belong with the behavioral program that encouraged them to install it, or the measure? A really vivid example of this issue is provided by the OPower programs. OPower is claiming 95% of the savings are from behaviors (some argue it is in OPower's interest to claim this). Some who

evaluation protocols[3], but the measurement and tracking of the behavior change programs lags behind the techniques applied to measure based programs. Behavioral interventions can be inherently somewhat complex to measure and evaluate, and we found the literature had few strong energy-related examples. We reviewed evaluation protocols in a number of behavioral fields (including health, etc.), and also reviewed more than 200 case studies of social marketing programs. Using this review and analysis, we:

1) Identified three key weaknesses in existing program evaluations, and
2) Developed recommendations for evaluation "best practices" for CBSM and behavior programs.

Weaknesses in Social Marketing and Behavioral Evaluations

Our review (Green and Skumatz 2000, Skumatz, et.al., 2010, Skumatz and Freeman, 2011a) found impacts ranged from 0% reduction in electric or gas use (and in one case an increase of residential energy demand) to a high of 30% reduction in residential energy consumption (a multi-resource audit in Ontario using incentives, coupons, in-home demonstrations, and other tools). One commercial example reported a high of slightly over 30% reduction from a program. Bill information programs delivered about 1-3% savings, and traditional design CBSM programs delivered energy savings of about 5 - 15 %. Similar results, ranging from 4%-12% savings, were found for the case of residential real-time pricing (feedback) pilots (Foster, Mazur-Stommen, 2012). Most of the CBSM literature described the program, told of their message and described their attractive posters, and lacked serious evaluation attempts, citing the "pilot" nature of the program. The three critical weaknesses we identified include:

- *Weak impact evaluation information*: Most of the evaluations used small samples or omitted control groups or other methods addressing dynamic baseline issues.
- *No cost-effectiveness information*: The studies omitted costs, or costs plus impacts, and did not provide cost-effectiveness estimates (cost per impact).
- *No Retention Results*: None of the evaluations followed-up to estimate retention of savings and behavior changes (even well-funded statewide programs). Many utilities assign *ad hoc* "deemed" retention values of one year, and no higher than three years for behavior retention.[4]

The implications of these key omissions, and the reliance of self-report for many program impacts, underlie the mistrust of evaluations of behavioral change and social marketing programs. Without reliable methods, it is not possible to develop reliable estimates of the benefit-cost ratio (or real cost information), and the savings can't reliably be included in integrated plans.

---

are doing surveys are claiming more on the order of 25% of the savings are from behaviors, with the rest due to measures installed. One state is "deeming" that OPower's 1-2% savings are 67% behavioral and 33% measure-based, but without research to support this. One issue may be the categorization – where does the CFL belong? But some also argue a high percentage of measure-based savings is indicated by the shape of the uptake of the program and its savings, with larger savings appearing after a lag – reminiscent of the lag of purchasing equipment. These are all questions of interest in evaluations of these types of programs.

[3] Summarized in Skumatz et.al. 2010. We might argue that the existing protocols may need adjustment to take account of behavioral components and influences that may be important, even in measure-based programs, because of the influence of behavior on performance of the equipment (Skumatz et.al. 2010, Freeman and Skumatz 2012).

[4] Although it is unclear if a median EUL of 3 years for behavior modification can be justified or whether the 1 year assumption undervalues the impacts given that there is minimal research for this estimate. In many cases, the evaluations are based on results immediately after the program is delivered, so evidence for some programs doesn't even show months of retention, much less a year or multiple years of retention.

Recommended "Best Practices" Evaluation Methods for CBSM/ Behavioral Programs

We worked to develop the key steps for evaluating behavioral and CBSM programs. Measurement protocols for energy-related behavioral programs follow the same principles as many other types of behavior-related evaluation work[5]. The bottom line is that there are few substitutes for good up-front experimental design including control groups and random assignment.[6] An abbreviated summary of best practices principles follows. Although phrased in terms of residential programs, the principles extend to other programs and sectors. The five recommended steps are included in the following illustration, and each is described below.

1. **Identifying Goals and Conditions**: Identify the goals of the program (preferably measurable) and the effects of interest (including a definition of what constitutes "participation" or "adoption"), and ensure that the effects can be seen to be caused by the program's intervention(s) (and not spurious factors). Assure that the program is administered to a group of participants that can be seen to represent the (ultimate) population of interest.

2. **Experimental Design and Sampling**: Plan for a test and control group. Both the control and test groups should be large enough to support statistically valid and meaningful comparisons. Sample sizes supporting +/-5-10% at 90-95% confidence are preferred.[7] The control group should be as similar as possible to the test subjects (and both groups should be as similar as possible to the ultimate group that will be eligible for the program to maximize transferability of results). Pre-post measurement of the test group is not best practice. Pre-post alone is vulnerable to seasonal differences, and other factors; control groups allow easy and reliable netting out of these variations. The control group sorts out impacts from effects beyond the program (e.g. nationwide ads from EPA or others, etc.)– and serves as a dynamic baseline against which the effects can be measured to provide net impacts. The premier experimental design is random assignment of eligible customers into the test and control groups (Sebold, et.al. 2001). Random assignment also helps to eliminate self-selection bias. Other approaches that have been taken include use of "similar" counties or cities, neighboring / similar states, etc.[8] Controlling for other factors from these "similar" control groups can be attempted through corrections with statistical models, but random assignment is much more straightforward and reliable.

3. **Measurement Design**: Evaluation methods need to be clearly laid out before any data collection is conducted. When evaluation is concluded, all limitations of methods and results need to be clearly

---

[5] We reviewed dozens of documents, including Sebold et.al. 2001, GAO, 2009, Skumatz et.al. 2010, Sergici and Faruqui, 2011, Skumatz 2012 and protocols related to ARRA and other programs.
[6] This is the approach illustrated in the case study conducted by the authors that is described later in this paper.
[7] Assuming large populations, these requirements tend to require sample sizes of 68, 96, 270, or 380 observations, with higher numbers preferred. Greater specificity can be provided on sample size needs depending on the degree to which the measurement needs to address Type 1 error, Type 2 error, one or two sided hypothesis testing, single or repeated measures experiments, etc. However, a surprising number of social marketing programs have measured the impacts based on sample sizes substantially less than 100 (often 30), and only pre-post and not control group measurements, to the detriment of confidence in the results.
[8] As widespread education campaigns affecting both target and non-target audiences become more common, finding a baseline to measure against is more difficult - it is hard to uncover a population with a "zero" behavior baseline

identified. In addition, the evaluation should include an assessment of the associated uncertainty. Identify the way in which the impact(s) and costs will be measured. For energy behaviors and energy savings, there are several main approaches:

- *Metering*: If the project (and budget) allows, metering the equipment affected by the desired behaviors over the course of the experiment provide direct and reliable information on the behavior change and its energy impacts. With large budgets, metering may be installed in large[9], random (or representative) samples of the test and control groups; with more modest budgets, metering samples are small, strategic samples that can be generalized to larger samples.
- *Utility bills and impact evaluation:* Preferred data for this option includes monthly energy usage (and billing cycles/meter reading dates and possibly tariffs) for all treatment and control customers, or alternatively, for a significant and random sample of each.
- *Surveys and reported behaviors:* The researcher needs to identify the relative appropriateness of phone, in-person, web, or other types of surveys. Research should include well-crafted / tested question methods, for example, asking about behaviors undertaken in specific time frames, rather than "general" habits,[10] and other preferred survey approaches.[11] Again, control groups are highly recommended to provide "baseline" behaviors.
- *Demographic Information:* Gathering information on number of occupants, socio-demographics, appliance data, occupancy (move-ins, etc.), weather data, and other information can help in developing statistical models that control for these sources of variations in results when conducting impact evaluation work or other comparisons.

4. **Impact Analysis and Context**: The basic preferred analysis approach is a comparison of means between treatment and control group, using either one pre/post period, or periodic measurements. The appropriate tests for statistically significant differences are performed to identify impacts from the program. Multiple measurements over the course of the project / pilot provide advantages in efficiency and variance reduction (due to the correlation between measurements at different time points), and thus, provide greater confidence in the results. Analysis of the data up-front is valuable (monthly comparisons, plotting data, and conducting comparisons of "features" between test and control groups to ensure comparability). Impact evaluation work using statistical models and the energy data can provide reliable estimates of these means, and the results can be improved with the inclusion of appropriate causal factors. Present results with confidence intervals and other relevant performance statistics. Several methods are available for estimating impacts:
   - *Measurement and Verification (M&V):* using metering or estimating key parameters from a random sample (or all) of the participants and control group and applying to all members of the group.
   - *Statistical Analyses:* applying statistical regression models[12] to utility billing[13] or metering data of all program participants, including approaches like differences of means / ANOVA, difference

[9] Obviously the principle is that large samples reduce the variance and help detect significant differences between groups.
[10] For example, 'did you use the power strip yesterday', or 'how many of your last two laundry loads used cold water', rather than 'do you use power strips' or 'do you use cold water for your laundry'.
[11] To help increase confidence in the survey reports, the researcher might also research the "say / do" gap and run scenarios on the range, conduct a sample of on-sites to confirm some behaviors that might be easily observed (current laundry temperature settings, etc.), and benchmark against a sample of billing data, where possible.
[12] Econometric texts can address the issues involved in regression modeling including model misspecification, measurement error, correlations, etc.
[13] Billing analysis using weather-normalized consumption data provided by the utility commonly is used to estimate gross savings. Billing analysis requires consistent residency for two or more years, so one year of pre-program data can be compared with one year of post-program data. Billing analysis may be used to estimate gross savings of education programs

in differences[14] and panel data regression analysis[15], and other methods provide reliable estimates of impacts. Cross section and time series approaches are valid. There is an extensive literature on statistical, or statistical / engineering adjusted models.

- *Surveys and Self-Reporting*[16]: Surveying certain populations to gather information regarding knowledge or behavior to estimate the savings-related changes from behavioral / educational / social marketing programs, and analyzing for statistical differences in the adoption of the behavior. Assuming energy savings and energy are the key impacts of interest, an additional step identifying an estimated or deemed value for the savings "per adopted behavior" may be the best information on overall savings available from this method. Self-report is not the gold standard; however, good practices allow meaningful findings.
- *Other Effects and Attribution issues:* Note that we do not consider "impacts" in terms of energy or similar all the analysis that is needed. The analysis plan should include measurement / estimation needed for important non-energy benefits, and spillover effects[17], if relevant.

Costs, and Cost-Effectiveness:,*:* These evaluation efforts need to include strong cost tracking to support analysis of cost per impact, cost-effectiveness, and benefit-cost type-calculations that will support analysis of the impacts in context and in comparison to other programs and strategies for delivering energy and energy efficiency. This should be built in from the beginning,

Context: Numbers are meaningful in relation to other numbers. Context should be provided for the result, including, for example, comparisons to the program's past performance, to similar programs regionally or nationally, the "literature" or other context discussion. Putting the results in context better supports explanation, refinement, extrapolation, and better decision-making support.

5. **Retention Analysis and Re-Analysis**: Measurement of savings from behavioral programs over the course of a full year for behavioral programs is preferred, to help account for seasonal effects. However, we would recommend working to put in place a measurement protocol that follows beyond that period to test for retention of the effects – which is a very important special uncertainty component of behavioral programs.[18] The retention of an adopted behavior change lasts only as long as the behavior remains changed, but the impact on savings may be suffer because of total cessation of the behavior, occasional retention by an actor, or retention by only some occupants of the home. Conversely, some behaviors may form new habits and remain in place for a lifetime. All of these possible changes – and more – will have an effect on the lifetime (and level of) of the estimated savings from the program.

---

combining low-cost measures and behavior modification. However, as billing data are inherently too "noisy," gross savings less than 10% of pre-consumption levels are hard to detect. (Skumatz et.al. 2010)

[14] This involves netting out mean differences between treatment and control groups in the pre-treatment period from the mean differences between treatment and control groups in the post treatment period. If the difference in differences of the mean values is statistically significant, then the treatment is found to yield an observable effect in the usages of the treatment customers. (Sergici and Faruqui, 2011)

[15] Advantages of this approach include the possibility to increase the efficiency and precision of the estimate using repeated measures on each program participant and to account for time invariant unobservable variables that would otherwise lead to biased estimates. Modeling approaches include fixed effects or random effects models (Sergici and Faruqui, 2011)

[16] Self reported data are often augmented with site visits and selected metering (e.g., hours of use).There are many texts that address the issues in survey design, question development, bias reduction, etc. This is not addressed here.

[17] Represented by additional behaviors adopted beyond those explicitly encouraged by the program, or to others by word of mouth, etc. These may be powerful elements of behavioral and CBSM programs.

[18] The impact of behavioral and other influences on the retention for measure-based programs, or the combined measure and behavioral programs, is being recognized and addressed in some cases and by some agencies (Skumatz 2012).

Considerations and Alternatives:[19] Although random assignment is the "gold standard", the world (and budget) does not always allow for this design – particularly if large-scale broadcast media are used, and potential participants cannot feasibly be excluded. Other options include:

- *Quasi-experimental comparison groups,*[20]*, Statistical analysis of observational data,*[21] *or In-depth case studies or other approaches.*

Refinement / Re-Analysis: Once retention work is done, the time component and decay function relevant to the savings can be applied to get total savings associated with the investment. This allows refinement of the cost-effectiveness findings, and a chance to reassess the results in context.

The main differences between these steps and protocols for measure based programs are that random assignment and control groups are more critical for highly behavioral programs, because inspections and counts are impractical. In addition, the authors add several steps that have not been included in CBSM and behavioral evaluations in other fields:

- Including "contextualization" of the results explicitly provides a critical link in learning what makes a difference in programs, and how to improve them.
- Computing estimates of cost-effectiveness is critical.
- Measuring retention is ignored in most programs, and is an essential computation if savings (and realistic estimates of costs) are to be counted on in integrated plans. Budget should be saved for this effort.

Review of Sample Evaluations of CBSM / Behavioral Programs Against Recommended Methods

We selected four example evaluations from among four dozen we assembled from our literature review, to illustrate different types of programs and different sophistications of evaluation. The programs[22] included are:

Program 1: A program providing feedback on energy bills, using comparisons to other customer groups (nearby, high performers, etc.) to encourage customers to reduce energy.
Program 2: A smart thermostat program, allowing reactions to real-time pricing options / dynamic pricing rates designed to influence energy use. The program involved web portals, in-home displays, and/or programmable communicating thermostats.
Program 3: A program using college students as interns to design and deliver education, contests, events, and other social media tools to their peers to encourage reductions in energy use.
Program 4: A locally-based CBSM program providing mailers, handouts, pledge cards, contests, door-to-door, and other methods to encourage energy efficiency and recycling behaviors.

Table 2 provides a summary of the evaluation methods used for the four sample programs.

---

[19] Note that the GAO report also stated that improvements to any evaluation can be achieved by: collecting additional data, targeting comparisons, and gathering a diverse body of evidence (GAO 2009).
[20] These resemble randomized experimental design, but the groups are selected from un-served members. This might include groups denied participation when a program is full or those that will participate in the next period, etc. According to the GAO report (GAO, 2009), the approach requires statistical analysis to establish groups' equivalence at baseline, and potentially, specialized statistical modeling in some examples, regression discontinuity analysis (GAO 2009)
[21] This approach requires observing and collecting data prior to the intervention, and after the intervention
[22] We are avoiding naming the programs, because we recognize that there were factoring affecting the evaluations. Budgets may be limited, or timelines may have been rushed due to circumstances beyond the program manager's control, or extensive evaluation may not have been in the program's plan, etc..

Table 2: Assessment of Evaluation Methods Used for Four Sample Energy-Related Behavioral or CBSM Programs

	Program 1	Program 2	Program 3	Program 4
1.Goals & Conditions / Set Up	• Design: Residential; Provides periodic mailers (monthly vs. quarterly, depending on customer) comparing to neighbors / groups. • Goal: No quantitative goal – looking to estimate impacts	• Design: Residential and Small Commercial • Dynamic rate plans provided (and the associated technology) • Goal: No goal; looking to check load reduction impacts.	• Design: University students designing posters, events, contests, and other strategies to influence peers to use less energy. • Goal: no specific savings goal; checking intern-delivery model & potential.	• Design: Residential neighborhood-based CBSM mailers, calls, contests, door-to-door, web, and other outreach to decrease energy use and increase recycling. • Goals: increase knowledge on 4 energy behaviors; increase recycling 15% / 7 lbs.
2.Experi-mental Design	• Pre-post with control group • Large sample 84K HHs randomly assigned to test / control; • Baseline data – billing data for all, assessor data for HH characteristics	• Pre-post with control • Residential & commercial, 2400-2800 random test and control from same groups • Baseline data: billing data for each customer, hh characteristics available. • Experimental design required / advice by grant.	• Thousands of eligible students per campus; no control group (able to be) established. • Baseline: No metering of energy available pre - Limited demographics.	• Pre-post with control group, randomly divided neighborhood into three sub-neighborhoods. • Baseline: No energy metering available at all; Pre-data on weekly recycling; HH data from surveys.
3.Measure-ment	• Data collection: billing data for impact evaluation plus 1+ years baseline usage; characteristics data also available.	• Data collection: billing data collected over time; • Limited pre-data available • Gathered / assembled HH data.	• Post-surveys widely distributed, asking for changes in behavior from pre (recollection); Little ability to screen for distributing to high quality samples of participants and non-participants / poor response • No metered data available for direct impacts. • Surveys examined for attitudes (and changes) and non-energy benefits.	• Data collection: Pre-and post-surveys on self-report behavior, frequency, knowledge, HH characteristics, etc. • Gathered surveys from samples pre and compared to surveys post. Randomness of respondents not guaranteed. • Strong data on weekly recycling tonnages from all three routes, not household level, pre, during, post, and continuing.
4. Impacts & Context	• Strong statistical impact evaluation with influential factors using pre/post billing data and control groups. • Estimated 1.1-1.2% savings • Confidence intervals / statistics reported • Compared to one other similar program and discussed. • No costs or cost-effectiveness presented.	• Difference of differences used for impacts • Estimates of peak red'n 1.5-30%; off peak 1-10%, depending on time of day, group, etc. (some groups increased usage) • Confidence intervals given • No discussion of context. • No costs or cost-effectiveness computed.	• Statistical analysis of reported in (self-report) energy behaviors collected from post-surveys from mostly participants. (no billing data for impact evaluation) • Analysis of changes in attitudes / knowledge. • No confidence intervals given • No discussion of context (few similar programs) • No costs / cost-effectiveness computed.	• Computed impacts from self-report, using deemed values to derive kWh and GHG savings. • Recycling tonnages quantified / compared between test and control neighborhoods pre/post. • No confidence intervals given. • Cost-effectiveness computed • Context for savings provided, especially in terms of relative cost of reducing GHG from variety of energy & recycling & outreach program options.

	Program 1	Program 2	Program 3	Program 4
5.Retention & Re-Assessment	• No retention in design or conduct. • Assumes constant future impact, presuming constant / continuing delivery of program. • No revised / long-term cost-effectiveness provided	• No retention in design or conduct.[23] • No discussion beyond immediate impacts provided • No longer-term cost-effectiveness provided	• No retention in design or conduct. • No discussion beyond immediate impacts. • No cost-effectiveness (largely a process evaluation with preliminary attempts to summarize quantifiable effects)	• Retention for recycling tonnage tracked after program; strong retention from door-to-door treatment neighborhood after 1 year; weak for mailed households. • Cost effectiveness for MTCE from recycling recomputed accounting for retention and compared to other initiatives
Strengths	Great sample size, good design for control group (randomized) Billing and characteristic data, including pre-data supported strong impact evaluation with causal / controlling factors. Designed for internal "test" of impacts of mailer frequencies.	Great sample size, good design for control group (randomized) Billing and characteristic data, including pre-data supported strong impact evaluation with causal / controlling factors. Designed for internal "test" of impacts of different prices, groups, etc.	Assesses behavior changes and "other" effects (non-energy benefits); examines changes in attitudes and awareness; attempts to measure program with myriad / flexible / non-uniform activities on multiple campuses – an evaluation challenge.	Good sample size, good design for control group (randomized) Good recycling tracking, and analysis of recycling retention Cost-effectiveness computed for GHG impacts. Internal "test" of levels of CBSM intervention (door-to-door vs. mailers).
Weak-nesses and strategies	No feedback on which HHs / behaviors implemented; what accounted for savings (no surveys); Limited context, no spillover, no cost-effectiveness, no retention, no up-front goals.	No context, no spillover, no cost-effectiveness, no retention analysis, no up-front goals.	No metered / billing data is critical problem. No retention, no spillover. Sampling / no control groups, and only post-surveys (self-report) weakens reliability of results.	No billing data is critical problem / self-report energy behavior; no spillover.
Possible Remedies	Conduct follow-up impact evaluations. Assess retention by looking at billing data patterns / impact evaluations over time from cohorts starting at the same time, and should also consider stopping mailers to a group and reviewing savings pattern / impact / retention over time.[24] Should definitely conduct cost-effectiveness analysis to compare to other measure and behavioral programs for long term planning. Add spillover questions in survey.	Conduct follow-up impact evaluations. Assess retention by looking at impacts over time from cohorts on same plans, and by stopping feedback / pricing for a random group and review impact / retention over time. Should definitely conduct cost-effectiveness analysis to compare to other measure and behavioral programs for long term planning. Add spillover questions in a survey.	Because separating participants from non-participants within campuses is hard, consider control groups at other "similar" campuses, similar to control group approaches taken for Energy Star™, etc. Prioritize conducting "pre" surveys; perhaps establish random panels of students pre/post to improve measurement. Purchase portable metering equipment for use during key strategies like dorm challenges, etc. Self-report remains an issue, but well-worded questions can help.	Better if billing data / releases can be arranged, or even "community" billing data pre/post. Add spillover questions to surveys.

[23] AND no mention of retention in the multi-document detailed M&V procedures documents from funders.

[24] This occurred for one program we reviewed, as a problem arose in mailings to a group for a period of time; the interruption was found to provide an unintended, and relatively high quality (and inexpensive) method of examining retention.

Conclusion

Behavioral programs have the potential to deliver significant savings. ACEEE estimates 30%; however, although the bulk of the pilot studies have deliver 5-15% reductions in energy use, some examples save zero, and the bulk of the billing feedback studies we reviewed showed 1-3% savings. Credible evaluations are a key part of establishing behavioral programs as a part of a serious energy portfolio (able to deliver savings and behavior change), but credible evaluations also provide real and transferable lessons. There are legitimate concerns about behavioral programs – and social marketing – efforts.

- The savings have been well-measured (including control groups) in only a few cases, and programs are all distinct, potentially leading to different savings values. Attribution to specific program interventions is more complex than measure-based interventions.
- The programs have mostly been pilot in nature; full-scale implementation results may lead to different savings – and cost-- results.
- Costs are rarely measured, and neither is retention, hampering computation of cost per unit savings, and muddying the "place" in the portfolio that behavioral or CBSM programs should occupy.
- The retention from social marketing is a significant question.[25] Using "deemed" retention values does not inspire confidence in the programs.

However, behavioral programs have several major advantages when compared with traditional measure-based programs:

- They can have significant impacts on energy use, as mentioned above
- They can be implemented quickly --with widespread adoption in a matter of weeks to months.
- They do not require programmatic purchases, delivery, or installation of equipment, intrusions into homes, and other efforts.
- Preliminary results from a few evaluations that have put results in context indicate that cost effectiveness may be on the order of other measure-based programs.

To be consider in an integrated planning model, programs of all kinds need to provide data on savings, cost, and years the resource will last. Behavioral and social marketing programs have generally been ignored in plans – a result that arises because evaluations of these programs have generally been weak. The main weaknesses we identified (Green and Skumatz 2000) remain experimental design (largely issues of control groups and sample size), cost-effectiveness and retention. More research on these questions is essential if behavioral programs and social marketing / outreach programs are to be a more integrated and reliable part of the energy portfolio.

Best practices in evaluation design are available, and we provide information on the "gold standard" and a few alternatives that have been invoked for programs with limitations – programs that may have been faced with "ready/fire/aim" or timing situations, budget barriers or other evaluation challenges. While there is almost no really strong substitute for large sample, randomized control and test group designs, we should recognize that some level of evaluation should be conducted so the literature can expand and we can identify patterns of strong and weak programs, and so creative evaluation approaches can be tested and spread. Most importantly, some evaluation budget should be preserved so programs

[25] As an aside, retention results from one of the studies reviewed indicates that door-to-door methods have much stronger retention (and cost-effectiveness) than mail-type outreach of the same materials, making the results potentially stronger than those already known for traditional outreach programs.

can assess retention of the behavior change and/or associated savings (whichever is relevant and measureable). Without retention, we will not know the total kWh the program saves, hampering computation of the true cost-effectiveness of the program.

Program managers need to become comfortable with behavior-based programs. The truth of the matter is that most programs are a blend of measure and behavioral components, and the choice between them has been a false choice. Elements of behavior are critical determinants in real-world program performance and explain differences in carefully-measured savings reported for measure-based programs. The "measure" savings we've all been counting on are delivered from both measures and behaviors; even greater energy savings potential is possible if we embrace the full spectrum of program designs – from pure measure to pure behavioral and everything in-between, and get busy and apply good evaluation to the behavioral components. We all win if "setsuden", Japan's post-Tsunami-crisis ethic of energy savings, becomes part of our everyday lexicon as well.

REFERENCES

Allcott, Hunt, and Todd Rogers, 2012. "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation", NBER Working Paper Series, Working Paper 18492, Cambridge, MA, October 2012.

Earhardt-Martinez, Karen. 2009. *Testimony before the United States House Committee on Science and Technology*, Subcommittee on Energy and Environment. April 28, 2009.

Foster, Mazur-Stommen. 2012. *Results from Recent Real-Time Feedback Studies*. ACEEE Report Number B122, February, 2012.

Freeman, Juri and Lisa A. Skumatz, 2012. "Widgets versus Actions: Measuring the Role of Behavior Change in DSM Programs", Proceedings of the IEPEC Conference, Rome, June.

Global Energy Partners, 2012. "OG&E Smart Study Together Impact Results, Final Report, Summer 2011", Walnut Creek, CA, February 29, 2012

Green, John and Lisa A. Skumatz, Ph.D., 2000. "Evaluating the Impacts of Education/Outreach Programs – Lessons on Impacts, Methods, and Optimal Education". Proceedings from the ACEEE Summer Study on Buildings, Asilomar, CA, August, 2000.

Kakuchi, Suvendrini, 2011. "Energy-Saving 'Setsuden' campaign sweeps Japan after Fukusima", Guardian Environmental Network, guardian.co.uk, August 22, 2011.

New York Times, 2011. "In Japan, the Summer of Setsuden", Editorial, September 25, 2011.

Sebold, Fred, Alan Fields, Lisa A. Skumatz, Shel Feldman, Miriam Goldberg, Ken Keating, Jane Peters. 2001. *A Framework for Planning and Assessing Publically Funded Energy Efficiency*. 2001. Study ID PG&E-SW040

Sergici, Sanem, Ph.D. and Ahmad Faruqui, Ph.D., 2011. "Measurement and Verification Principles for Behavior-Based Efficiency Programs", The Brattle Group, Prepared for Opower, May.

Skumatz, Lisa A. 2009. *Do Energy Efficiency Strategies Outperform Recycling in GHG Mitigation and Job Creation?* International Energy Program Evaluation Conference. Portland, OR 2009.

Skumatz, Lisa A. 2010. *A kWh is not just a kWh: Comparing Energy Efficiency Programs in Terms of GHG, Job Impacts, and Policy Achievements*. Proceedings of the American Council on Energy Efficiency Summer Study on Buildings. Asilomar, CA. 2010.

Skumatz, Lisa A., 2012. *What Makes a Good EUL? New Protocols for Estimating Useful Lifetimes and an Analysis of Existing Estimates, Proceedings of the IEPEC Conference, Rome, June, 2012.*

Skumatz, Lisa A., and Juri Freeman, 2011a, *Getting the Most from Colorado's Recycling Programs and Infrastructure: Social Marketing Outreach and Education Toolkit – A Guidebook for Communities*, Skumatz Economic Research Associates, Superior, Colorado, May 2011.

Skumatz, Lisa A., and Juri Freeman, 2011b, *Getting the Most from Colorado's Recycling Programs and Infrastructure: Best Management Practices, Impacts, and Cost-Effectiveness of Social Marketing*, Skumatz Economic Research Associates, Superior, Colorado, May, 2011.

Skumatz, Lisa A., M. Sami Khawaja, Jane Colby. 2010. *Lessons Learned and Next Steps in Energy Efficiency Measurement and Attribution: Energy Savings, Net to Gross, Non-Energy Benefits, and Persistence on Energy Efficiency Behavior*. Prepared for California Institute for Energy and Environment. November 2010.

Skumatz, Lisa A., Lisa McClain, and Juri Freeman, 2006. "Alliance to Save Energy's Green Campus Program: Year 1 and 2 Evaluation Report", for CPUC, Superior, CO, October 31, 2006.

Skumatz, Lisa A., Carol Mulholland, and Sharyn Barata, 2009, "Social Marketing: Evaluating the Social Twist on Marketing Programs", AESP Strategies Article, Association of Energy Services Professionals, Spring 2009

United States Government Accountability Office. 2009. *Program Evaluation: A Variety of Rigorous Methods Can Help Identify Effective Interventions.* November 2009