ATTRITION BIAS IN FUEL SAVINGS EVALUATIONS OF LOW-INCOME ENERGY CONSERVATION PROGRAMS

Michael Blasnik
G.R.A.S.P.
Philadelphia, Pennsylvania

Common fuel savings evaluation methodologies require more consumption data than are available for many participants in low-income weatherization programs. These data requirements often lead to sample attrition rates greater than 50%. In the process of conducting a pilot weatherization program, the Grass Roots Alliance for a Solar Pennsylvania (GRASP) noticed substantial differences between houses which met the data requirements for evaluation (the evaluation sample) and those which did not (the attrition sample). GRASP compared the evaluation sample with the attrition sample and an unscreened sample in order to verify and quantify some of these differences and investigate the potential for bias caused by sample attrition.

GRASP discovered significant differences between the evaluation sample and the other groups in terms of initial measured air leakage rate, reduction in leakage rate due to program measures and pre-period fuel consumption. The evaluation sample houses had significantly tighter envelopes than the unscreened sample and had much smaller air leakage rate reductions from weatherization measures. A rough comparison of fuel usage showed significantly lower consumption for the evaluation sample than the attrition sample. There appears to be a correlation between the quality of billing data and the thermal integrity of the house. These results imply that low-income weatherization programs may be achieving greater savings than a standard high attrition evaluation would indicate.

GRASP's findings demonstrate attrition bias and call into question the generalizability of many low-income fuel savings evaluations which have comparable sample attrition. Further exploration is needed in evaluation methods which reduce this bias such as cruder billing data analysis techniques, statistical bias reduction techniques, and methods such as short-term submetering which create their own data.

Background

Energy savings evaluations of low-income weatherization programs are usually based upon available billing data. These evaluations are important as they are used to determine program cost-effectiveness and to select weatherization treatments. The representativeness of the evaluated sample is critical in demonstrating to policy makers the true value of these programs and ensuring the use of the most effective treatments.

Evaluation Methodologies

Most energy savings evaluation methods analyze utility billing data to estimate weather-normalized energy consumption before and after treatment for both the treated group and a control group. Two methods are a simple degree day analysis, sometimes called "Slash and Burn," and a computerized regression analysis called the PRInceton Scorekeeping Method, or PRISM (see Fels, 1986).

The "Slash and Burn" method has several variations, but typically estimates a baseload consumption based upon summer usage and subtracts this baseload from heating season usage to estimate a heating component. The heating component is adjusted to a "typical" weather year through multiplication by the ratio of long-term average to actual degree days at a chosen reference temperature.

PRISM is a linear regression model of usage per day against degree days per day. PRISM performs a mathematical search for the degree day reference temperature which provides the best fit for the linear model and estimates a Normalized Annual Consumption (NAC) based upon long-term average degree days at this reference temperature. PRISM provides standard errors for the estimates of NAC and all parameters and provides a measure of goodness-of-fit (R^2). Although the individual parameters are prone to bias (due to the seasonality of baseload), PRISM is generally accepted as the more accurate tool due to its stronger physical and statistical basis. The complexity of PRISM imposes more stringent data requirements for the houses analyzed.

Data Requirements of PRISM

Although PRISM requires only five meter readings to estimate NAC for a period, users often impose greater data requirements in order to produce more reliable estimates. These requirements have never been officially codified, but evaluators generally adhere to similar

rules (for discussion see Dunsworth and Hewett, 1985). Evaluators typically screen data for a minimum of 7-10 fairly evenly distributed meter readings for each period analyzed. This screen is not too stringent for households where utilities have monthly meter readings and no shut-offs. Unfortunately, such conditions are often not met for low-income weatherization program participants.

Sample attrition can occur at several points in data preparation and analysis. Billing data are screened before being run through PRISM to remove houses with obviously inadequate data. Cases are typically screened out if there are too few meter readings, very long gaps between readings, or shut-offs. Oil-heated houses are particularly problematic because billing data are rarely available and of questionable reliability. Most billing data evaluations are performed on utility heated houses only. Cases which meet the initial data screen are analyzed using PRISM. The PRISM results are then screened for reliability based on the quality of the linear fit (R^2) and the standard error of NAC [CV%(NAC)]. Evaluators have adopted various standards ranging from a minimum R^2 from 0.75 to 0.95 and a maximum CV%-(NAC) < 25% to > 5%. Some evaluators also screen out cases with parameters that are physically impossible or have large standard errors, or if there is no discernible heating load.

Sample Attrition: A Philadelphia Case Study

In Philadelphia, the data screening procedure examines 14 months of billing data and selects houses which have 7 or more actual meter readings, no data gaps greater than 3 months, and no shut-offs. The PRISM results are screened for cases with $R^2 > .90$ and CV%(NAC) < 10%. The experience of GRASP and the Office of Housing and Community Development Energy Unit is that less than 15% of the original sample generally survive this data screening. An examination of the available data and causes of attrition distilled from several low-income program evaluations in Philadelphia is instructive in understanding this process.

The Philadelphia Gas Works attempts bimonthly meter readings in the summer and monthly readings the remainder of the year. Meters are generally located inside and estimated bills are common, comprising approximately 50% of all billing points for program participants. Estimated readings resulting in insufficient data points has eliminated 45% of all households from evaluations. Additionally, complete billing data have been unavailable in about 10% of all cases due primarily to change of occupancy.

Temporary and permanent service disconnections are very common in Philadelphia: from 1981-1987 shut-offs averaged 30,000 per year. Approximately 3% of all billing points for program participants are at time of shut off, with about 40% of all participants having been shut off at some time during the previous 24 months. Approximately 20% of all houses have been screened out due to shut offs.

The remaining sample, only about 25% of the original, is then analyzed using PRISM. Approximately 50% of the resulting PRISM estimates fail to meet the reliability screen for R^2 or CV%(NAC) in either the pre or post period. The resulting sample is typically less than 15% of the initial cases analyzed. In essence, PRISM applied with standard screening criteria selects a sample for evaluation that is only a small fraction of the original group. The attrition rate of 85+% is an extreme case, but a brief review of other studies demonstrates high sample attrition is a widespread problem. Table 1 summarizes sample survival rates for several PRISM evaluations of low-income programs.

The vast majority of low-income houses are screened out of most PRISM analyses: in Philadelphia, New York, and Illinois attrition rates exceed 80%! If failing the data screen were a random event then the attrition would only be an annoyance; it would reduce sample size but not bias the results. The potential impact of non-random sample attrition has prompted a few evaluators to insert caveats about the generalizability of results, but most readers and policy makers just look at the "bottom line" energy savings. The phenomenon of large-scale sample attrition has been mostly ignored.

The GRASP Blower Door Pilot Program Experience

During the fall of 1987, GRASP initiated a blower door research and pilot program funded by the Office of Housing and Community Development and the Philadelphia Housing Development Corporation. The research phase included blower door testing and treatment of low-income houses. This research was used to design a program to be run by New Kensington Community Development Corporation (NKCDC). GRASP devised a performance-based payment system based upon the experiences in the research houses. NKCDC was to be paid according to the change in blower door reading.

Because fuel savings evaluation was a critical component of the pilot program and previous experience indicated large attrition, houses were selected for treatment which met a data screen for PRISM. Out of 309 gas-heated houses analyzed, the data screening yielded a pilot program sample of 66 houses. The pilot was

initially designed for 100 houses, so 34 unscreened houses would also be treated (the unscreened houses were randomly selected from new program applicants and data quality was unknown).

GRASP realized something was wrong when the three tightest houses we had ever measured were among the first five screened houses. The screened houses were in much better condition than the research houses. Because the payment schedule was based on experience in unscreened houses, NKCDC was losing money rapidly. The payment schedule was revised and it was agreed that tighter houses would not be treated. Eventually a total of 75 houses had pre-treatment tests performed—56 screened and 19 unscreened. NKCDC treated 61 of these houses—47 screened and 14 unscreened. Table 2 shows the results of the blower door testing and treatment for both groups of houses.

The screened houses were 30% tighter on average than the unscreened (prob < .05). But even more revealing, the weatherization work was more than twice as effective at reducing the leakage rate in the unscreened houses (prob < .05). This difference in leakage reductioncan be roughly estimated to equal 5%-8% greater energy savings for unscreened houses, a potentially significant bias. Figure 1 shows the distribution of the blower door pre-treatment readings for screened and unscreened houses.

The evaluation screen also appears to be an air-tightness screen including a disproportionate number of tighter houses. Because the distribution of screened houses appears non-normal and to minimize undue outlier influence, non-parametric tests were also used to analyze the two groups. The median pre-treatment blower door reading in screened houses is 23% less than that of unscreened houses (3791 vs. 4954). Contingency table analysis rejected the hypothesis that both samples are from populations which share the same median (Chi-square = 4.51, prob < .05) and confirmed that there are a greater proportion of tight houses (CFM @ 25Pa < 3000) in the screened sample (Chi-square = 5.1, prob < .025). A Wilcoxon Rank-Sum test offered the strongest

Table 1. Sample Survival Rates from Low-income Evaluations

Screen Used	Phil 1	Phil 2	Phil 3	NY	Minn	Mass	Wisc[a]	Illinois
Original sample N =	405	309	745	138	166	74	>460	220
Pre: raw data screen	47%	21%	35%					
Pre: RxR/CV screened		16%	26%					
Pre & post raw data screen	24%		25%	22%	72%		93%	
Pre & post RxR/CV screened	13		13%	17%	46%	39%	67%	19%
Criteria: Min RxR/Max CV	.9/.1	.9/.1	.9/.1	.9/.1	.9/.1	.75/.2	.9/—	.8/—

[a]This study used PRISM with a fixed reference temperature for all houses with <9 readings.

Sources: Phil 1 = Daspit et al., 1987; Phil 2 and Phil 3 = Daspit, personal communication, 1989; NY = Rodberg, 1986; Minn = Hewett et al., 1986; Mass = Nadel, 1987; Wisc = Goldberg, 1986; and Illinois = Hall, 1988.

Table 2. Blower Door Readings and Changes from Treatment (all Measurements in CFM @ 25 Ps.)

	Screened	Unscreened	% Difference	Sig. @ 95%
Pre-treatment Mean, All Cases (N = 56 screened; 19, unscreened)	3902	5630	31%	Yes
Treated Houses (N = 47, screened; 14, unscreened)				
Pre-treatment Mean	4041	5757	30%	Yes
Post-treatment Mean	2980	3536	16	No
Reduction Mean	1061	2221	52	Yes

evidence of sample bias, strongly rejecting the hypothesis that the two groups are random samples from the same population (prob < .001). There were no significant differences in house size or weatherization budget.

Fuel Usage Analysis

After finding the significant differences in blower door readings and reductions between screened and unscreened houses, GRASP decided to analyze the fuel usage of the original sample of 309 houses. Since the evaluation would have proceeded with an analysis of the 50 houses which met the reliability criteria for PRISM results ("selected" houses), these houses were compared with a sample of the houses failing the screen ("rejected" houses). Data was available on only 103 of the 243 rejected cases. These 103 cases are the first of two groups of rejected houses (the groups differed only in application date to the program).

GRASP ran into immediate difficulties in attempting to compare the fuel consumption between two groups of houses, one of which is defined by its lack of data. GRASP first tried using "Slash and Burn" but realized that this method was capable of analyzing only a few more cases than PRISM because of the local gas utility's bimonthly summer meter reading schedule. GRASP decided to utilize the least data-intensive method available, a linear regression of usage per day vs. degree days per day (base 60°F). This method could work with any house having three or more real usage points. Many houses still did not meet this requirement, so the analysis was extended to the previous 24 months. With no pure baseload data for many houses, the parameters are often poorly estimated, but normalized consumption is more robust. In an attempt to avoid obviously erroneous results, the data for each house was plotted for detection of outliers.

The regression method was tested using the 50 selected houses. The mean normalized consumption estimates were 1302 ccf/yr for PRISM and 1270 ccf/yr for the regression. This 2.3% difference is small but statistically discernible (prob < .05) and implies some bias, perhaps due to the 24 month analysis period. The relatively good agreement between the regression method and PRISM was encouraging, but for houses with little data the reliability is unknown. Nevertheless, with no other tools available the method was applied to the rejected houses. The regression method was capable of estimating normalized consumption on 92 of the 103 rejected houses, with 9 cases exhibiting no heating load and 2 cases with insufficient data.

The mean normalized consumption of the rejected houses using the regression method was 1478 ccf/yr, 16% greater than the consumption of the selected houses. This difference is statistically significant (prob < .05). Figure 2 shows the distribution of normalized consumption for the selected and rejected houses.

Because the regression method was assumed to be subject to substantial random errors and the selected group's distribution appeared non-normal, non-parametric methods were applied to analyze the differences between the two groups. The median usage of the rejected houses is 23% greater than the selected houses (1359 ccf/yr vs. 1108 ccf/yr). Similarly to the blower door analysis presented previously, contingency table analysis confirmed the difference in medians (Chi-Square = 3.98, prob < .05) and supported the hypothesis that the selected group had proportionally more low users (Chi-Square = 4.94, prob < .05), defined as the bottom third of the distribution. A Wilcoxon Rank-Sum test rejected the hypothesis that the two groups were drawn from the same population (prob < .03).

Implications and Explanations

The houses selected for evaluation due to the quality of their billing data records are different from the general program population. The differences discovered thus far involve factors which are strongly related to expected energy savings—pre-period energy consumption and blower door readings. Both of these factors indicate that the sample selected for evaluation is biased toward the tightest and most efficient houses treated in the program which have the smallest potential for savings. In one study which used submetered data for evaluation, the houses with blower door readings above the median had twice the reductions and saved 50% more energy than the houses below the median, and houses with pre-period fuel consumption above the median saved more than twice as much energy as those below the median (Synertech Systems Corp., 1987). A program evaluation based upon a data screened sample can therefore be expected to underestimate energy savings.

There are several explanations which could account for the relation between the adequacy of billing data and the thermal integrity of a house. One likely possibility is that low income people who live in very inefficient houses have higher utility bills and are therefore shut off more often. Conversely, low income people with the highest incomes may live in better, and more efficient houses and are more able to keep up with their bills. Utility meter readers may also introduce bias by not attempting to read meters in the lowest income neighborhoods as diligently as in higher income areas. Failure to pass the reliability criteria for PRISM results may indicate use of space heaters, more varied household behavior, or a non-linear dependence on degree days due to greater air leakage.

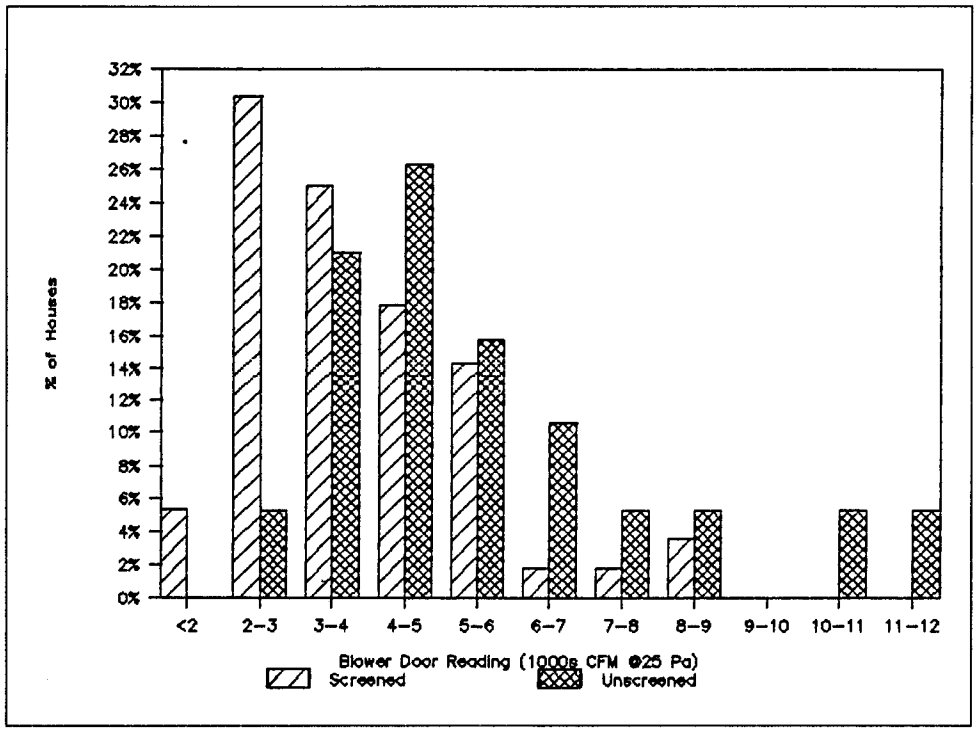

Figure 1. Blower Door Readings: Screened vs. Unscreened

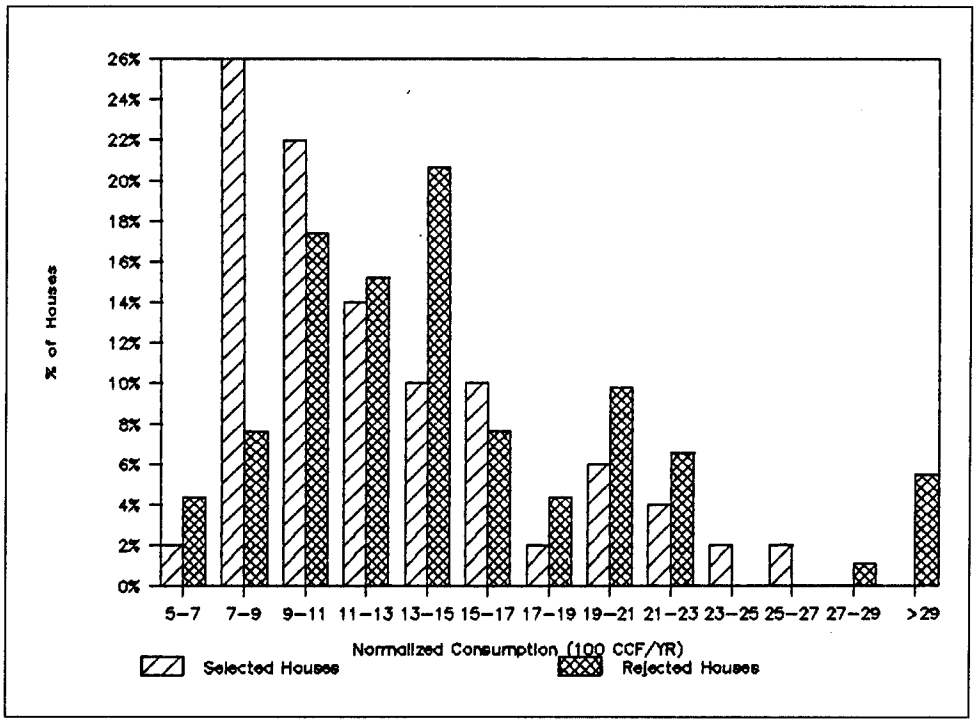

Figure 2. Normalized Consumption: Selected vs. Rejected Homes

Bias-reducing Alternatives

The identified attrition bias can be reduced through either evaluating a representative sample or adjusting the results from the biased sample to compensate for the bias. A representative sample can be evaluated through relaxed data screening, alternative billing data analysis methods, or independent data collection. Adjustment for bias can be accomplished through statistical techniques.

Evaluating a Representative Sample

The most obvious method to reduce sample attrition from screening procedures is to stop screening the data. PRISM can be used with as few as five meter readings for each period. "Slash and Burn" can be used with as few as three, if each reading is appropriately timed. "Slash and Burn" requires at least one pure baseload point, but PRISM can be run with a fixed reference temperature requiring no pure baseload data and then only four meter readings (this would be analogous to the regression method used to estimate rejected house consumption levels). The evaluation period can be extended to include houses which do not meet these minimal requirements. The impact of the presumably greater scatter in the resulting estimates can be reduced by using robust group savings estimates. If it can be shown that the result of relaxing data screens is a simple trade off between random error and sample bias, then the random error is preferable (especially since sample size is increased).

There are currently no validated alternative billing data analysis techniques which require substantially less data than PRISM or "Slash and Burn." At least two consumption points are needed—three if an estimate of reliability is desired—to estimate the two parameters used to normalize for weather. GSTEAM, a short-term method which accomplishes weather correction through weather-dependent period selection, has produced good results (Blasnik *et al.*, 1988). But GSTEAM still requires a pre-period PRISM analysis plus one well-timed consumption point in the post-period. GRASP has experimented with another version of GSTEAM that requires just one consumption point in each the pre and post periods. This new method has had promising results—performing weather normalization by choosing periods with "normal" weather instead of estimating parameters—but more research is needed. Another alternative technique is to compare usage levels without explicitly adjusting for weather, but instead relying upon a control group to reflect weather and non-weather related usage changes. This technique would increase random error by not accounting for a known co-variant (degree days), but may produce reliable group savings estimates if samples are large and data periods approximating a year are available for both groups.

The surest method for reducing attrition due to billing data is to not use billing data. Evaluators can collect their own data for analysis either through special meter readings or submetering the usage to be analyzed. Special meter readings have been used as a way to prevent sample attrition with some success in Philadelphia, but are labor intensive and therefore expensive. Special meter readings may be appropriate for smaller samples such as pilot programs. Submetering is an appealing method if reduction of heating consumption is the primary program goal because this usage can be measured directly. Submetering also provides results in a few weeks of winter as compared to the 12 months typically required by billing data analysis. The accuracy of submetering has been validated (Nadel, 1987) and it is the only viable evaluation method for oil heated households. The primary disadvantage of submetering is cost (approximately $100/house for labor and materials excluding any meter reading costs). Submetering studies usually involve small samples due to budget constraints and require more sophisticated planning and implementation than billing data analysis methods.

Statistical Bias Reduction

Stratified sampling techniques are commonly used to reduce bias when the nature of the bias is known. For example, if it is determined that the source of the bias between the evaluated sample and the general program population is initial air leakage rate, then the evaluated sample can be partitioned (stratified) into several bins representing different levels of air leakage. If the true distribution of air leakage rates for the program population is known (by measuring every house), then program savings can be estimated by taking a weighted average of the savings for each leakage bin, with the weight for each bin equal to its proportion in the general population. Two problems with this technique are applying a parallel method to the control group and properly identifying the true nature of the bias. Does the initial blower door reading capture the bias, or is the change in blower door reading or demographic characteristics more important?

Sophisticated statistical techniques have been developed for estimating participant self-selection bias (see Train, 1987). These techniques typically involve creating a model to estimate the probability of participating and then modeling energy savings through a multivariate regression which includes a term based upon the probability of participation. The energy savings are then simulated for participants and non-participants to estimate the true program impact. This method requires data on underlying characteristics which determine par-

ticipation for both participants and non-participants. By defining this method in terms of evaluator selection, a similar approach can be used for attrition bias. But the lack of data and small samples may limit the reliable adaptation of this method. A simpler approach may be possible by using a combined demographic and engineering regression model. For the selected houses, energy savings can be regressed against factors such as pre-period usage, blower door reduction, home repair expenditures, owner vs renter, and other household characteristics found to be important. The resulting model can be used to estimate savings for the attrition group using the measured values of the regressors. If pre-period fuel consumption is needed, a crude estimation method can be used. The more sophisticated approach taken for self-selection bias and this regression technique both warrant further examination but have many potential problems such as multicollinearity exacerbated by sample bias, model mis-specification, inadequate characteristic data quality and quantity, inadequate sample size, lack of data for inclusion of the control group, and a large potential for statistical abuse.

Conclusions

The high attrition rates especially common in low income weatherization program evaluations may be leading to biased estimates of program savings. In order to guard against such bias, careful examination of attrition groups should be an integral component of any evaluation. If bias is discovered, as was the case in Philadelphia, alternative evaluation techniques need to be used. Cruder billing data analysis methods need to be examined as the least cost option for representative sampling and statistical bias reduction methods need to be further developed and validated.

References

Daspit, C., and R. Roberts. ***High Efficiency Gas Furnace Modification in Low-Income Housing***, Energy Task Force of the Urban Consortium for Technology Initiatives, 1987.

Dunsworth, T.S., and M.J. Hewett. ***Data Quality Considerations in Using The "PRISM" Program***, Minneapolis Energy Office Technical Report No. 85-4-AM, 1985.

Fels, M. "PRISM: An Introduction," ***Energy and Buildings***, Vol. 9, Nos. 1&2, pp. 5-18. 1986.

Goldberg, M. "A Midwest Low-income Weatherization Program Seen through PRISM," ***Energy and Buildings***, Vol. 9, Nos. 1&2, pp. 37-44. 1986.

Hall, N.P., "A Residential High Efficiency Furnace Replacement Program In Illinois: An Examination of Energy Savings and Related Consumer Behavior," ACEEE 1988 Summer Study on Energy Efficiency in Buildings, (American Council for an Energy Efficient Economy) Vol. 9, pp. 22-33. 1988.

Hewett, M., T. Dunsworth, T. Miller, and M. Koehler. "Measured versus Predicted Savings from Single Retrofits: A Sample Study," ***Energy and Buildings***, pp. 65-73.1986.

Nadel, S. "Timely Evaluations: A Discussion Of Methods Used by the Massachusetts Audubon Society to Evaluate Fuel Savings in Under One Year," ***Energy Conservation Program Evaluation: Practical Methods, Useful Results***, Argonne National Laboratory: Argonne, Illinois; Vol. 2, pp. 73-83. 1987.

Rodberg, L. "Energy Conservation in Low-income Homes in New York City: The Effectiveness of House Doctoring," ***Energy and Buildings***, pp. 55-64. 1986.

Synertech Systems Corporation, ***Integrating Analytical Tactics into New York State's Weatherization Assistance Program: Project Findings***, New York State Energy Research and Development Authority Report No. 87-21, 1987.

Train, K. E. "Net Savings From A Rebate Program For Commercial And Industrial Customers," ***Energy Conservation Program Evaluation: Practical Methods, Useful Results***, Argonne National Laboratory: Argonne, Illinois; Vol. 1, pp. 258-268. 1987.