

How To Evaluate Programs

Energy Retrofits For Houses

PROG 4

October 28, 2009 4:00PM – 5:30PM

Toronto, ON

Presented by: Michael Blasnik, M. Blasnik & Associates

michael.blasnik@verizon.net

Why Evaluate?

- Assess impacts of program
 - compare to expectations, prior years, other programs
- Assess performance of specific retrofits
 - Find out what works (or when it works) and what doesn't
- Identify opportunities for improvement
 - based on energy savings analysis
 - based on technical review & field inspections
 - based on interviews and surveys of stakeholders
- Demonstrate value of program
 - to funders, regulators, and others

Energy Program Evaluation in the U.S.

- Many programs don't ever get evaluated
 - mostly done only when required by outside parties
- Many evaluations aren't very reliable
 - Many don't actually use any energy usage data!
 - rely on projections ("engineering" estimates) and surveys
 - Many evaluators know little about energy usage
 - and they're often not so great with statistics either
- Difficult to identify evaluation problems
 - especially when the econometric jargon starts flying

Types of Evaluation

- Impact Evaluation
 - Assess impact of program
 - Energy savings
 - explore patterns: by measure, house type, provider, etc
 - Non-energy benefits
 - Health/safety, environmental, economic, etc.
- Process Evaluation
 - Assess other aspects of program
 - technical review of procedures, field tools, training, QC
 - field visits using diagnostics
 - administrative/logistical systems of program
 - survey clients to get feedback on program implementation, marketing, education, etc.

Energy Savings Evaluation Methods: Projected Savings

- “Calculate” savings from energy audit software or engineering algorithms
 - Not reliable: Measured savings are often just 50%-70% of projected savings
 - Flaws in assumptions, inputs, and the engineering models themselves all tend to over-predict savings for virtually every measure
 - If you want to learn about actual savings you need actual usage data

Impact Evaluation Designs

- Experimental Design (rarely possible)
 - Random assignment of treatments
 - Untreated units = control group
 - Clear basis for cause and effect: treatment -> impact
 - Statistical tests assess whether “random chance” is a plausible explanation for observed differences
- Observational Study (typical)
 - Assess on-going program - treatments not randomly assigned
 - people choose to participate, retrofits based on needs
 - No control group, instead create a “comparison” group
 - Reliability depends on representative groups
 - Bias should be assumed, explored, and adjusted for
 - Statistical tests generally don't address this problem

Impact Evaluation Bias

- Observational studies need representative groups – expect problems....
 - Participants analyzed may not be typical of program
 - less likely to include renters, occupant turnover, intermittent utility service, supplemental heat
 - Comparison group may not be comparable
 - Participants usually differ from random customers
 - Homes: older, leakier, less insulated, etc.
 - Occupants: resources and interest in investing in home energy improvements
 - Future participants often used as comparison group
 - better than random, but participants may be changing or targeting may select households on usage trend
 - Some types of bias can be addressed (but some can't)
 - matching or stratification methods on key observables
 - propensity scoring may be useful if matching impractical
 - regression models (not as good)

Measuring Energy Savings

- Energy usage changes year to year due to
 - Program Treatments and ...
 - Weather, behavior, and other changes in the home
- Evaluation goal: measure change due to the program
 - Adjust usage for weather variation from average year
 - Degree day methods (e.g. PRISM) work well for heating, but big changes still cause problems.
 - Cooling loads harder to adjust, especially in mild climates
 - Use statistics as needed to adjust for other factors
 - Large groups average out the random non-program changes
 - Comparison group should reflect trends that don't average out
 - Trends could be related to naturally occurring load growth or efficiency changes or could be due to weather normalization bias
 - Usage typically changes only a little ($\pm 3\%$)
 - Matching methods and stratification can help with comparability

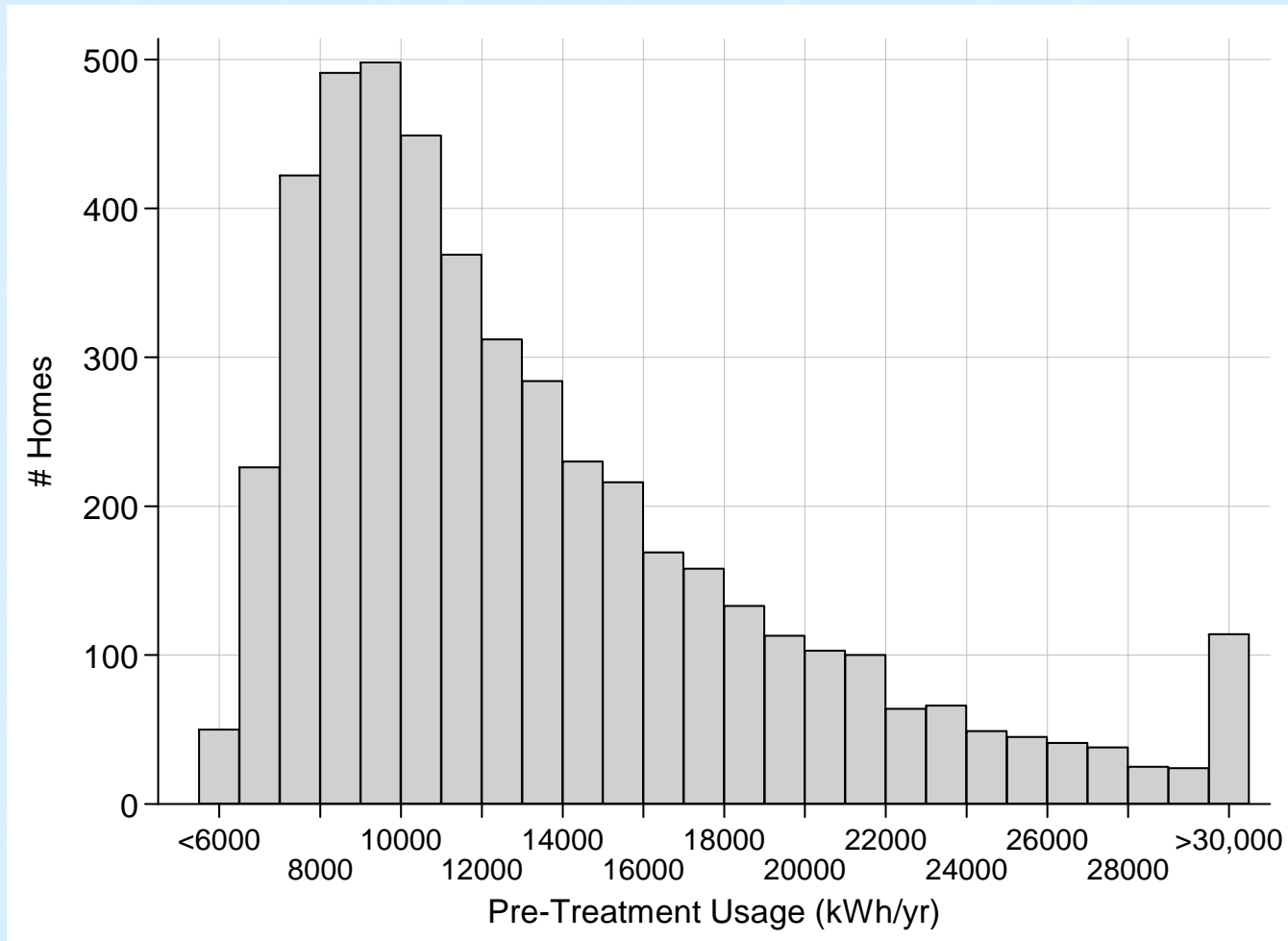
Billing Data Analysis

- Collect Meter Reading Data
 - typically 12 months before and year after treatment
 - include comparison group
 - need cooperation from utilities
- Billing Data Analysis Approaches
 - House level analysis (a.k.a. NAC method)
 - each home's usage is analyzed pre and post
 - gross savings = pre use – post use
 - "net" savings = participant savings – comparison savings
 - Pooled Econometric Models (e.g., TSCS regression)
 - data analyzed with single statistical model
 - savings derived from regression model coefficients
- Evaluation costs vary widely: \$20k - \$100k+
 - Depends on data collection and research questions
 - Field work and surveys more costly than billing analysis

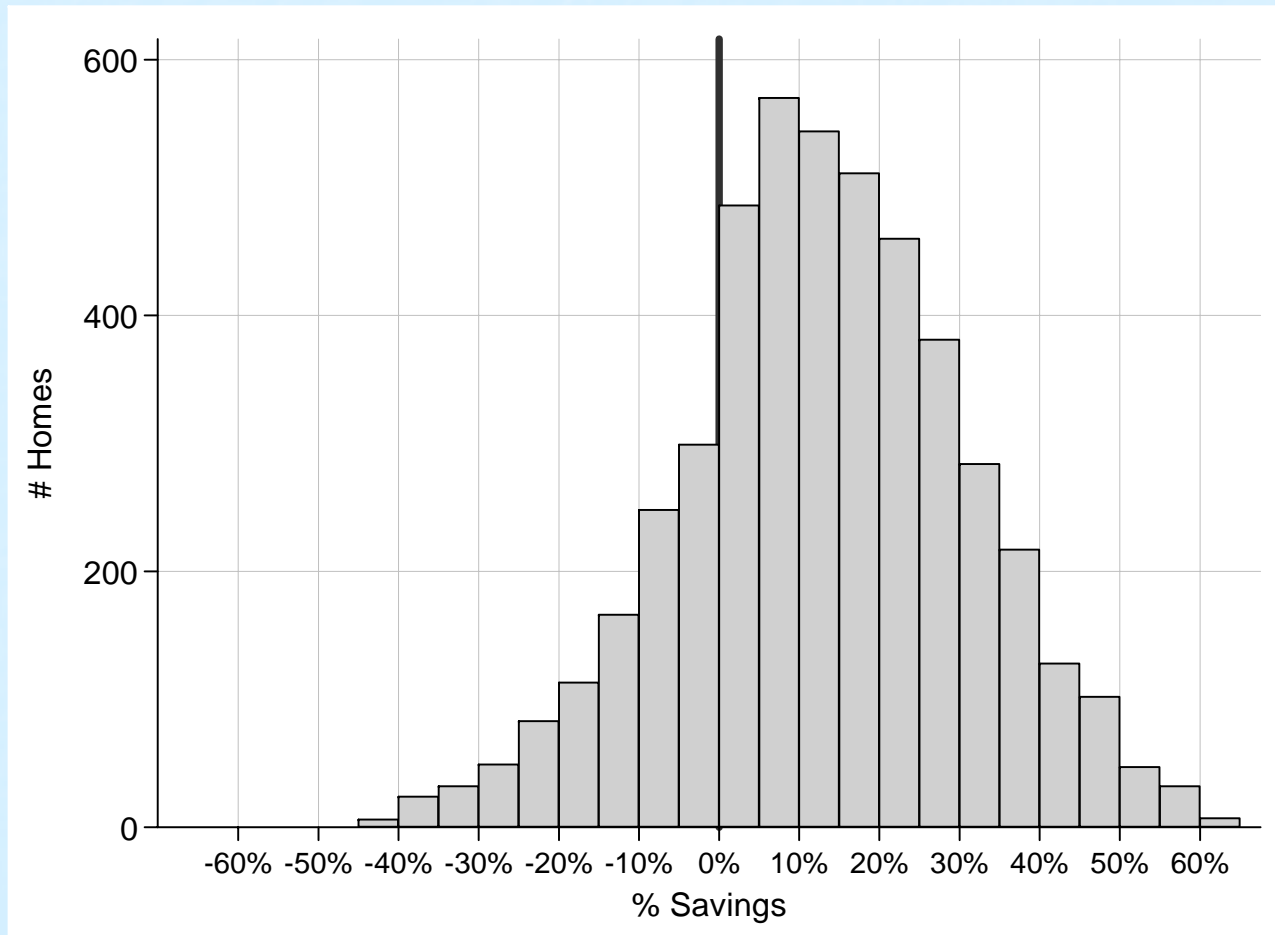
House Level vs. Pooled Models

- House Level Model Advantages
 - results can be “seen”, distributions examined
 - group savings easily calculated by treatments, housing characteristics, demographics, etc.
 - relationships can be explored with graphs and analyzed with statistical models of the usage or savings
 - can identify and explicitly deal with outliers, unoccupied homes, homes with wrong fuels, etc.
- Pooled Model Advantages
 - can utilize all data, even incomplete data
 - by assuming all homes are similar
 - easier and quicker than house level
- Other Pooled Model “Advantages”?
 - can play with models to get desired results
 - can lead to strange and even absurd conclusions
 - “Hey, I’ve just disproved the first law of thermodynamics!”
 - can use jargon to mystify clients (and themselves?)
 - heteroscedasticity, multicollinearity, endogeneity?
 - can charge more for things no one understands

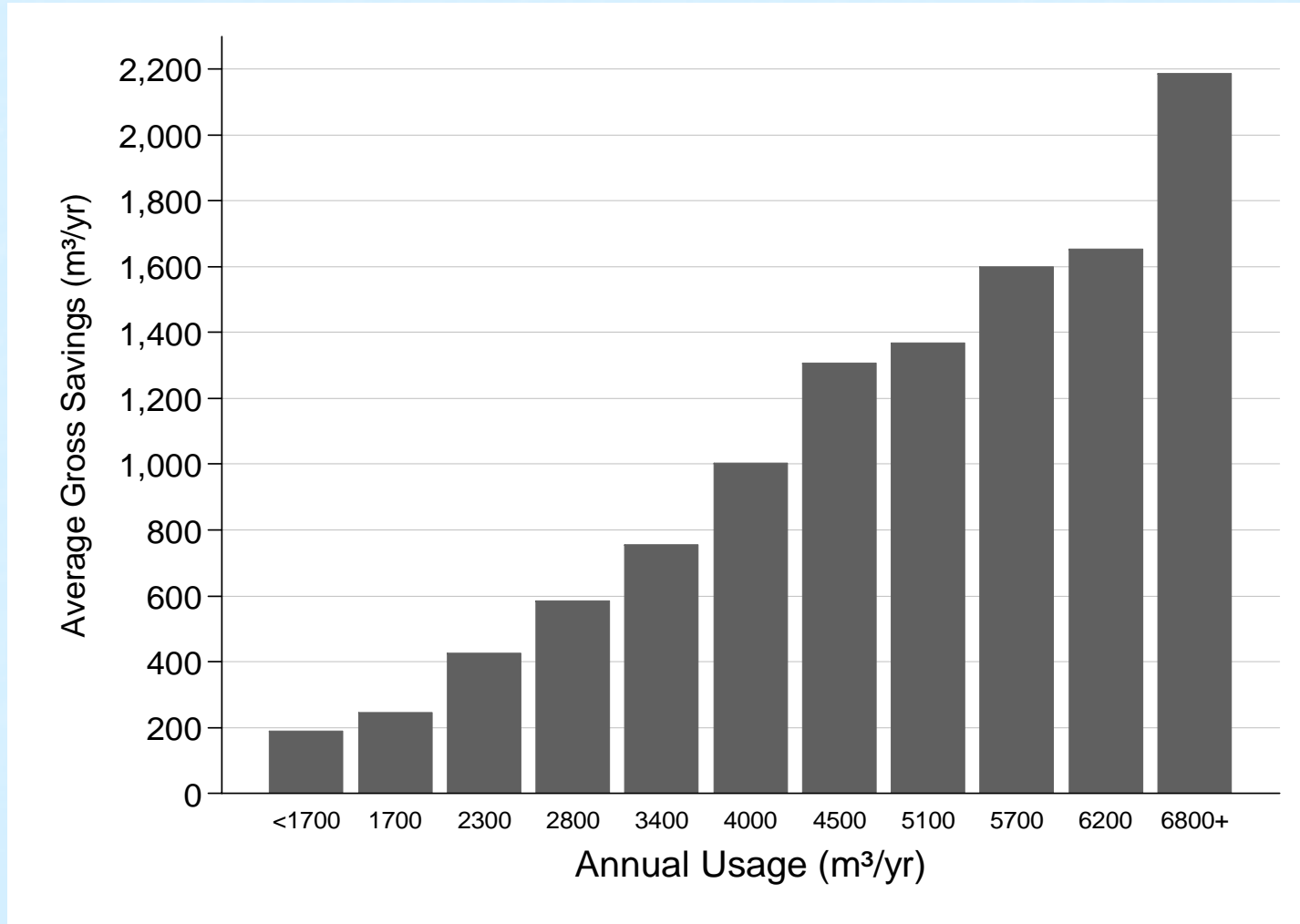
Things Pooled Models Don't Show You: Distribution of Pre-Use



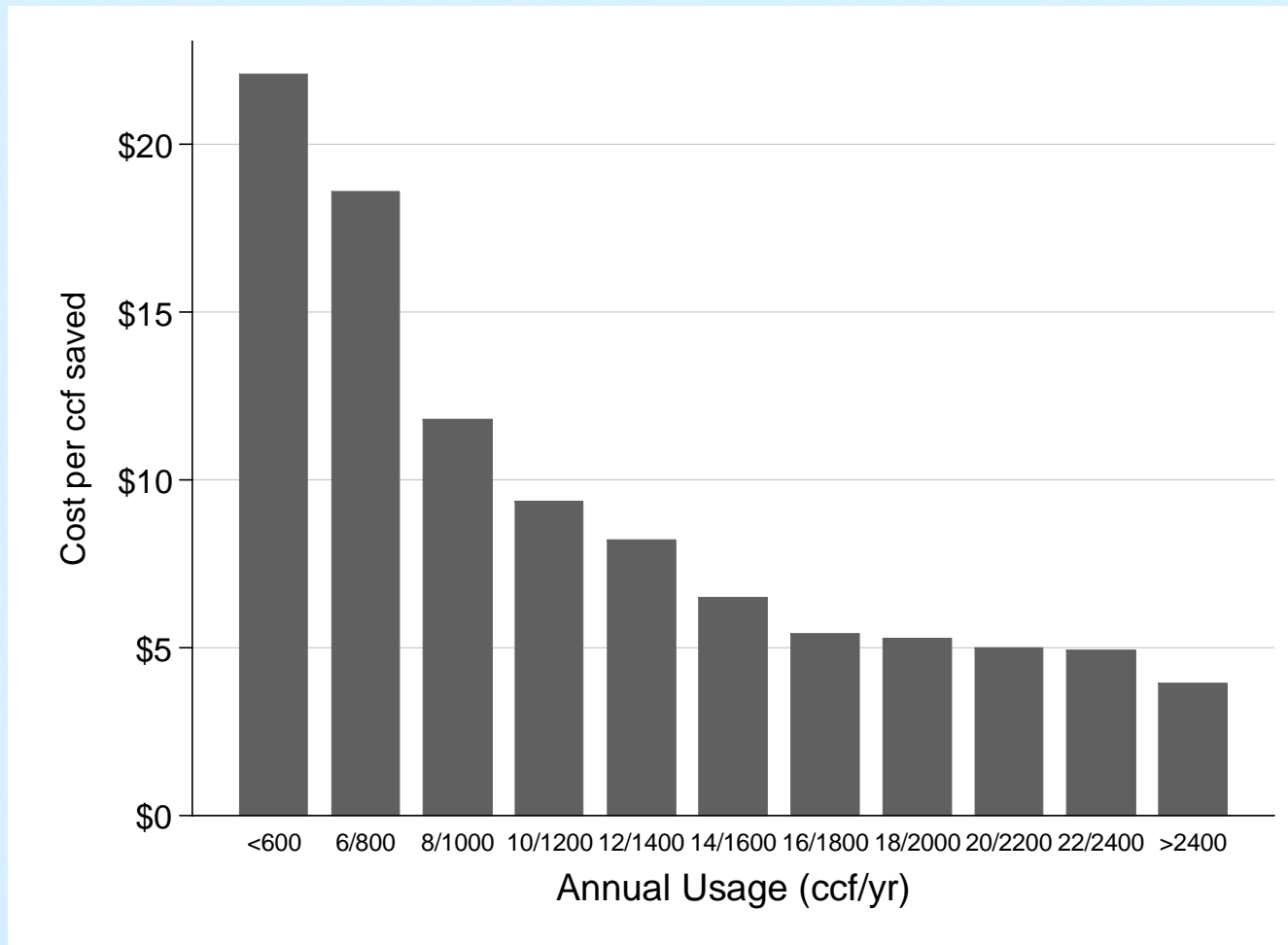
Things Pooled Models Don't Show You: Distribution of Savings



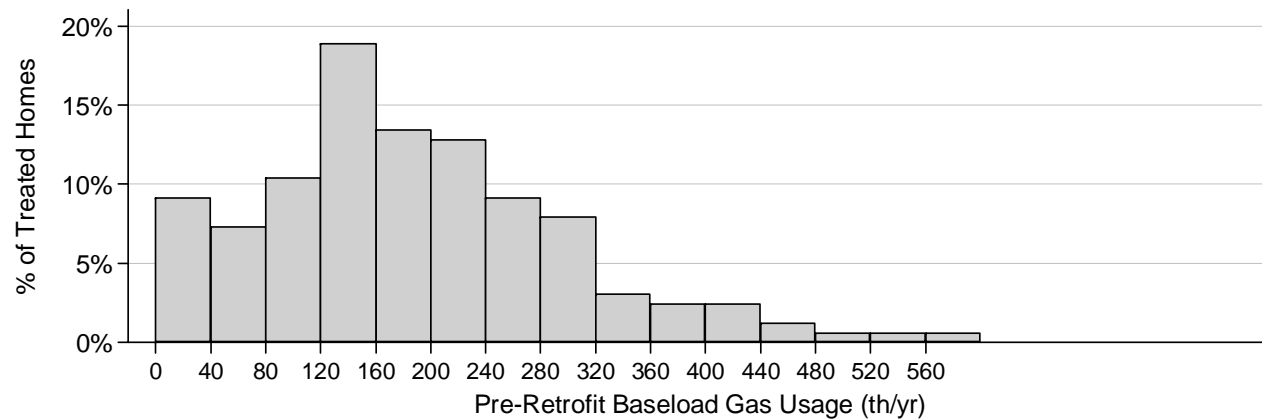
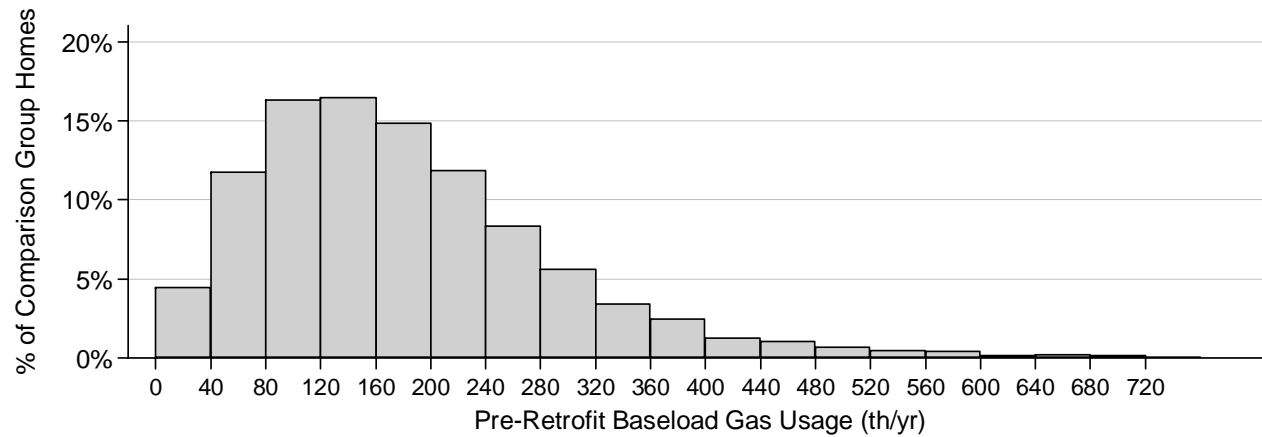
Things Pooled Models Don't Show You: Savings vs. Pre-Use



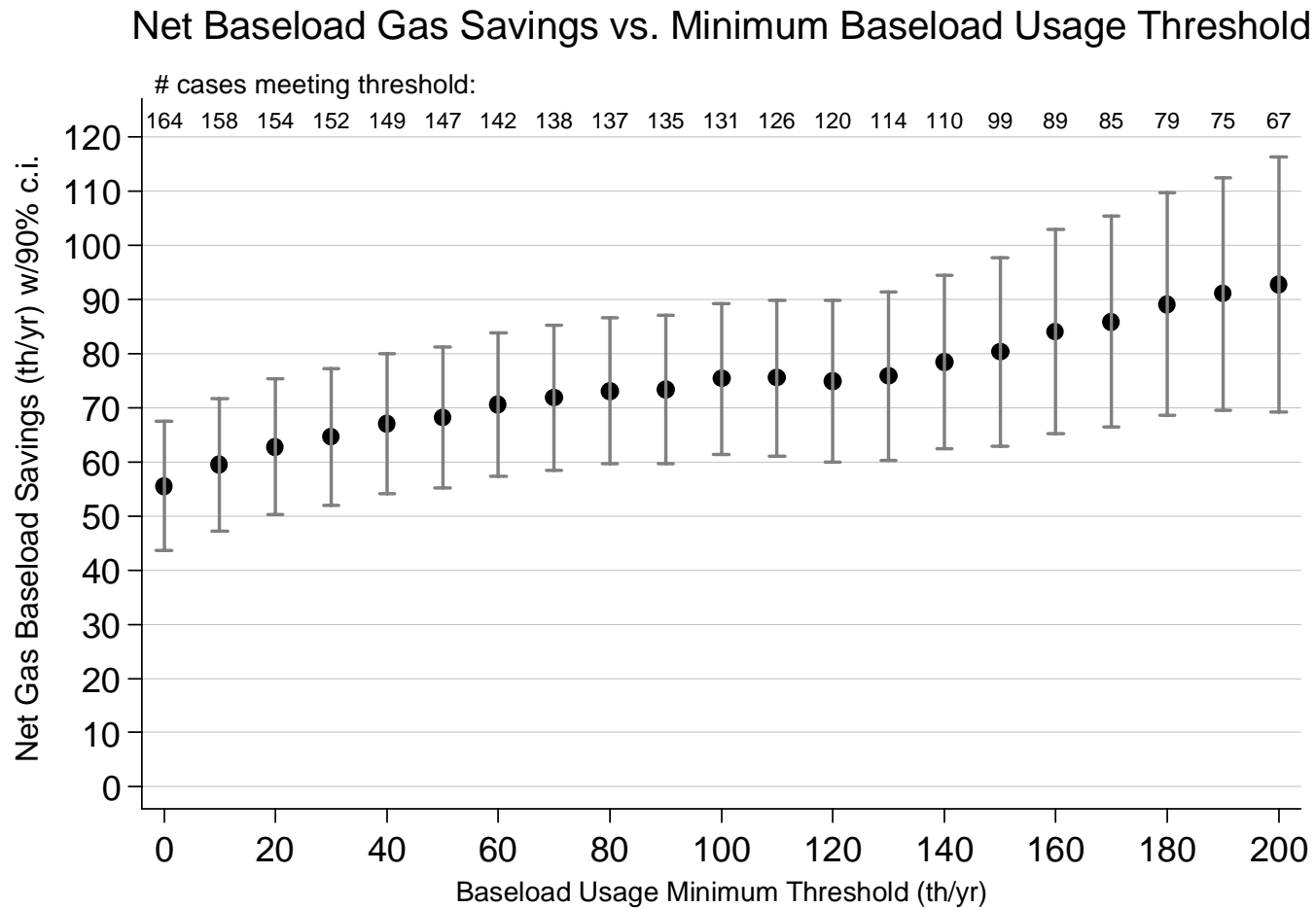
Things Pooled Models Don't Show You: Cost effectiveness vs. Pre-use



Things Pooled Models Don't Show You: Identify Problem



Things Pooled Models Don't Show You: Assess Problem Impact and Resolve



Things Pooled Models Don't Show You: Participant Characteristics by Savings Level

Characteristics	Savings Category		
	Low Savers (bottom quarter)	Mid Savers (middle half)	High Savers (upper quarter)
# Houses	390	778	389
Pre-Usage (th/yr.)	794	872	1,140
Net Savings (th/yr.)	-60	106	339
Net Savings % (of total)	-7.5%	12.1%	29.7%
Treatments:			
Attic Insulation	27%	55%	86%
Wall Insulation	13%	32%	56%
Floor Insulation	12%	19%	24%
Perimeter Insulation	16%	28%	35%
Heater Replacement	7%	8%	18%
Storm Windows	5%	9%	14%
Materials Cost (Energy)	\$181	\$329	\$545
Other:			
Area (living space sq.ft.)	1,392	1,297	1,282
Usage Intensity (therms/sq.ft)	0.57	0.67	0.89
Renter	16%	22%	28%
Mobile Home	18%	18%	18%

Things Pooled Models Don't Show You: Usage & Savings by Treatment Groupings

Measures	# Jobs	Pre-Use	Post-Use	Gross Save	Gross %Save	Net Save	Net %Save
No Insulation Measures	368	825	795	30	3.6%	13 (± 12)	1.6% ($\pm 1.5\%$)
Attic	249	956	810	145	15.2%	120 (± 16)	12.6% ($\pm 1.7\%$)
Wall	90	905	783	122	13.5%	101 (± 19)	11.2% ($\pm 2.1\%$)
Floor	61	804	712	92	11.5%	76 (± 21)	9.5% ($\pm 2.6\%$)
Perimeter	95	917	835	82	9.0%	62 (± 19)	6.8% ($\pm 2.1\%$)
Attic & Wall	162	982	742	240	24.5%	215 (± 17)	21.9% ($\pm 1.8\%$)
Attic & Floor	124	891	708	183	20.6%	161 (± 18)	18.1% ($\pm 2.0\%$)
Attic & Perimeter	133	999	807	192	19.2%	166 (± 18)	16.6% ($\pm 1.8\%$)
Wall & Floor	16	986	841	146	14.8%	118 (± 46)	12.0% ($\pm 4.7\%$)
Wall & Perimeter	54	919	754	165	18.0%	142 (± 25)	15.5% ($\pm 2.7\%$)
Attic & Wall & Floor	69	991	704	286	28.9%	260 (± 26)	26.3% ($\pm 2.6\%$)
Attic & Wall & Peri	120	1007	718	289	28.7%	261 (± 20)	25.9% ($\pm 2.0\%$)
No Heater Replace	1402	909	772	137	15.1%	116 (± 13)	12.7% ($\pm 1.4\%$)
Heater Replaced	155	1012	788	224	22.1%	198 (± 20)	19.5% ($\pm 2.0\%$)

Learning from the Results

- Average savings don't tell you much
 - Patterns of savings can provide insights
 - graphs can be useful and provide transparency
 - Group savings break-outs
 - average savings by measure (or combo of measures), house type, provider are interesting but don't show cause and effect
 - Statistical modeling (regression)
 - tries to account for multiple factors at once to provide better estimate of what drives savings
 - be cautious of "black box" fancy statistics
 - analysis must be guided by people who understand program and measures

Lies, Damned Lies & Statistics

- Simple Statistics
 - Compare average for one group to average for another group
 - Easy to calculate and explain (use t-test) but beware outliers
 - Relies on “other things being equal” to be cause and effect
- Some Problems with Statistical Tests
 - Tests assess likelihood of the difference being random, but not whether difference could be caused by another factor
 - Practical significance \neq statistical significance
 - Can't prove two things are equal, just test if they differ
 - The \pm uncertainty is itself an estimate
 - Bias and precision worse than the estimate itself
 - Std. errors are generally biased, look at alternatives such as bootstrapping

More Evaluation Problems

- Interpretation Issues
 - Results for any one house are unreliable - stuff happens
 - Savings vary between groups of homes for many reasons
 - Opportunities (housing or appliances), work selected, and work quality all play a role
 - Only field observations by experts can assess quality
- Evaluator Issues
 - Good evaluations recognize and try to deal with potential bias, don't over-state certainty
 - Many evaluators have little building science knowledge, so be wary of strange results that could be precisely wrong
 - Technical mistakes go unnoticed or unchallenged, hidden behind complex statistics

Good Evaluation as Storytelling

with numbers...

- Must be able to identify and assess potential sources of bias and factors that affect outcome
- Subject matter knowledge is more important than statistical knowledge
- Must develop a story that ties the data and analysis results to the real world
 - If a result seems unbelievable, don't just blindly believe it
- Use simple statistics to summarize and test, use more complicated statistics as needed
 - But be careful to understand the assumptions and tie it all into a narrative that makes sense

Impact Evaluation: an imperfect tool

- Impact evaluations are observational studies, not designed experiments, so bias should be expected
 - Results for any one house are unreliable - stuff happens
 - Houses evaluated may not be typical of program
 - Comparison group may not be comparable to participants
 - Savings vary between groups of houses for many reasons -
- housing stock opportunities, work selected, work quality.
Only field observations by people with expertise can assess quality.
 - Good evaluations recognize and try to deal with potential bias
 - Many evaluators have little technical building science knowledge, so program managers and operators should beware of strange results that could be due to evaluation errors or statistical problems

Wx Program Gas Savings Results

