Evaluator as Fool

Tricking the Evaluator into Confirming Phantom Energy Savings as Real

Dr. H. Gil Peach, H. Gil Peach & Associates LLC, Beaverton, OR
Dr. Pam Brandis, Bonneville Power Administration, Portland, OR
Dr. Michael J. Maranda, University of Maryland, Baltimore, MD
Ryan Miller, RNM Research, Philadelphia, PA
Anne West, West & Company Research LLC, Eugene OR
Howard Reichmuth, Stellar Processes, Hood River, OR

Abstract

This study provides a list of improper devices, to assist an evaluator confronted with a residential performance evaluation or verification assignment *to recognize and overcome illusion in the construction of program savings estimates*.

Several improper devices can boost payments to performance contractors in residential performance contracting arrangements. These include: (A) Omitting negative savings, (B) Exerting unusual control over the record of energy use in the baseline year; (C) Bait and switch of energy efficiency enhancements, (D) Assertion of proprietary estimates of "per installation" savings and/or of installed material and equipment lives; (E) Capitalizing on external treatments; (F) Rigging tail-trimming; (G) Fuzzing selection of homes; (H) Use of regression to the mean; and (I) Capitalizing on a hidden auxiliary variable. Most (all but "B" and "C") are particular forms of selection bias. All nine devices play on the good faith of the program manager and of the evaluator to introduce illusion into measurement of energy savings.

The evaluation approach for this study is squarely within the tradition of modern science; it is both materialist and realist. The basic understanding is Darwinist – the search for truth is the key to survival of both people and organizations. This is the philosophical (evolutionary epistemology), evaluation design and statistical analysis tradition of the Campbell school. It emphasizes doing our best to learn about material reality, given that all human activity is inherently fallible. The goal of the study is to aid in attaining truth in measurement.

Introduction

In a verification study, the central evaluation research question is: "What percentage of the claim for project savings is due to the material energy savings caused by the contractor's installation of energy efficiency enhancements, and what percentage is due to illusion?" In performance contracting, the savings claim put forward by the contractor (or its agent) is also a claim for payment.[1] The lure of performance contracting is in the concept of "*pay for performance*." In this approach to securing energy efficiency, the buyer (agency) contracting for residential energy savings *theoretically* pays only for results. Rather than pay directly on a time and materials basis for the installation of retrofit treatments to homes, the buyer attempts to

[1] This is an essential difference from the case in which a utility or government agency contracts directly for work on a time and materials basis from the private sector or from a Community Based Organization (CBO) or Community Action Agency (CAA). For the private sector, the CBO, or the CAA energy savings is a technical result rather than a claim for payment.

purchase conserved kWh or conserved therms. The legitimation of performance contracting rests on the representation that the agency will pay only for demonstrated savings, measured according to a measurement protocol that is given legal status as a part of a performance contract. *Theoretically* this arrangement transfers many categories of risk from the agency to the contractor.[2]

The appearance of the performance contracting approach appeals to the buyer's interest in reducing risk. Also, the approach is typically pushed by "free market" economists who believe that markets are inherently superior to planning, that markets police themselves, and that the "free market" model of entry level economic texts explains how actual markets work. Further, performance contracting is generally accepted to work well in areas of engagement such as large government facilities.[3] It is not so much the emergence of governmentally approved generic measurement protocols that has made performance contracting acceptable within the government sector as – more importantly – that there is sufficient per facility energy savings available to produce a win-win situation for the buyer and the performance contractor. That is, there is a substantial margin for the performance contractor, without the need to game.

Residential Performance Contracting comes and goes in cycles. At the beginning of each cycle it is promoted based on appeal to idealized "free market" economic perspectives. Near the end, it goes into decline due to perceived problems in actual practice. It is then "reinvented" and the cycle repeats.

The first cycle of residential performance contracting began with the New Jersey demonstration program in the mid-1980's evaluated by the Oak Ridge National Laboratory. By the early 1990s the approach had been largely discredited within the utility industry as a result of the strict inspection and evaluation regimens that were integral to the context of least-cost planning and demand-side management.

The second cycle coincided with the pro-capitalist ("markets are the answer"), anti-planning ideology that was a part of the campaign to destabilize traditional integrated electric utilities in the early 1990s. Beginning about 1992, this general movement led to the destruction of integrated planning and control systems and to a host of defective market applications (as well as some valid market applications in some regions). During this period, the inspection and evaluation regimens associated with integrated least cost planning were largely destroyed and generally weakened as "planning and control" approaches were replaced by "market" approaches to provision of energy. Most utilities lost institutional memory of necessary control functions as they downsized or outsourced DSM and DSM evaluation, and lost the planning context that provided the underlying rationale for strict measurement in program evaluation.

The third cycle occurred with the collapse of deregulation in California, the exposure of the corruption of broad energy markets by ENRON and other power supply entities and the resulting material need to produce direct and immediate demand reductions and energy savings. Another stimulus to the third cycle

[2] There are several risk components. For a presentation of the theory that risks transfer to the performance contractor, see Hansen & Weisman (1997) and earlier work by Hansen; for a critical treatment showing that the key risk components do not actually transfer, see Peach (1992; 1995).

[3] At the same time, in global manufacturing where industries maintain fully competent corporate and plant energy engineering staff, though performance contracting can work, performance contractors are sometimes viewed as pests. In global industry the appeal of performance contracting is redundant to the financial system of the global corporation and the internal expertise of energy engineers. Workable exceptions can occur in contexts of sustained downsizing in which internal competence is lost and surviving engineers are carrying double or triple workloads and subject to critical "time window" constraints to accommodate plant production cycles. Also, higher management may wish to focus technical capability on core competencies and simplify the outsourcing of energy and energy efficiency so that industrial performance contracting is seen as a workable alternative. As to the government sector, the attraction of performance contracting lies primarily in the way it has been grafted within the project finance system to make projects possible, and in the general tendency to require government to outsource technical work rather than maintain the necessary staff capabilities to do work internally.

ramp-up was the political need to *appear* to produce energy savings that followed the deregulation crisis and dramatic rise in energy costs (gas and electric), which continues.[4]

The alternative to residential performance contracting is to contract for residential weatherization services with the local Community Action Agencies (CAAs),[5] supplemented where necessary by local private contractors. The Community Action Agencies are interested in doing quality work as a public service because they are local, community service minded, and in a service role for the long-term. In this model, the agency contracting for energy efficiency work takes responsibility for the energy savings. The contract is for the installation of certain equipment and materials using defined standards of work and defined rules of precedence as to what is installed in each type of dwelling, and in what order.

In contrast, the contractor in performance contracting is paid not for installations or for time and materials expended in the work, but only for resultant units of energy savings as calculated by methods specified in the contract. Since (in contrast to the commercial and industrial sectors) the margin in residential work is quite low, there is a built-in incentive for the contractor to try to control contract language to shape program measurement and evaluation of results. Examples are ordered from the simple to the more sophisticated improper devices.

Omitting Negative Savings

An early device (Peach 1992) was omitting negative savings.[6] The pitch goes like this:

(1) As a performance contractor, let me say that we are *partners* looking for energy savings, and I will be installing measures in each home to produce energy savings at your (the buyer's) request. All of these installations are positive actions and none will reduce energy savings.

(2) But, it is common knowledge that some homes, for whatever reason, will show increased energy use in the post-weatherization period as compared to the baseline period before we do our work. Logically, this kind of change does not occur due to our work but for some other reason.

[4] The current emerging consensus is that retail energy markets do not begin to resemble the "free market" model and, further, that the natural end-state of evolution of institutions in the energy area following deregulation is the polar opposite of the appearance created by the deregulation rhetoric. That is, absent government control, it will tend towards a handful of multi-region merged integrated companies, similar to the pre-deregulation utilities but much larger. That is, huge, multi-state and multi-region utility holding companies, returning to the institutional situation prior to the Public Utility Holding Company Act of 1935. A further outcome of the attack on planning was the breakdown of system diversity planning and the emergence of a large set of gas-fired electricity generating plants in direct price competition with residential and commercial gas customers, causing gas and electricity price increases to be mutually reinforcing.

[5] Coordinated programs that combine utility or state dollars with federal dollars are particularly effective. For design of coordinated programs see Hill & Brown, 1995.

[6] Utility data for the "no program" example was provided by Portland General Electric. Note that omitting negative savings is one variant of regression to the mean, also discussed in its more general form in this study. In omitting negative savings, moderate savings cases and cases with high savings are kept. When the cases that go negative are eliminated the mean for the overall distribution of cases shifts up, sometimes dramatically. But even small upward bias to the mean is important because the mean is multiplied by the number of case to arrive at total energy savings. Sometimes the mean is calculated in this manner on a small subset of cases that meet specified criteria for inclusion in the analysis, and then the mean is multiplied by the much large number of cases of treated homes. Plugging with the mean in this manner can create a large upward difference in energy savings and so in the dollar value of payment to the performance contractor.

(3) So, *to be fair,* let’s structure the contract so that only the homes that show positive savings will be measured to determine the amount of energy savings my work has produced (and so determine what you pay me). It is true that I will not get credit for what I install in homes that show zero or negative savings, but for the good of our joint effort I’ll offer to donate that equipment, materials, and labor to the overall cause, and not ask for payment for that gift to the community and to conservation.

(4) On second thought, I guess to be more fair all-around, let’s put a line in the contract that will pay me for every home I treat based on the mean net savings for the homes that have positive savings.

The result is shown in Table 1. The evaluator, given a contract that specifies a measurement agreement that eliminates homes that increase in energy use from measurement would report savings of 137 kWh per home as the contractual savings, while the savings actually produced and available in an analog of a dispatch model is 50 kWh per home.

The power of *payment amplification* through the omission of negative savings is further illustrated in Table 2. In this table, actual utility data is used and the table is laid out as if there were a performance contracting program, however, there was no program. Here, by securing a contract clause to eliminate homes with negative savings from the calculation of mean (per home) energy savings, the contractor would have built in a payment for 2,200 kWh per home before starting any work. The “savings” by contract (prior to the effect of any improvements actually performed by the contractor – since for this example we use actual utility data incorporating normal year to year variations in energy use but there is not program) would be 2,200 kWh per home. In fact, as shown in the first column, the average home in this group of homes increased energy use by 1,200 kWh. If there had been a program, there should be no payment to the performance contractor under the theory of transfer of risk from the utility (or unit of government) to the performance contractor. Yet, by securing two contractual clauses, one that calls for plugging all cases with the calculated mean, and the other that calls for elimination from the analysis of homes with negative savings, the calculation using actual utility

Table 1: The Consequence of Omitting Negative Savings.

Simulation: Omitting “Negative Savings” Tips the Balance

Home	**Measured Savings**	**Measured Savings without Negative Savings**
	(kWh)	**(kWh)**
1	110	110
2	110	110
3	-200	
4	150	150
5	200	200
6	-150	
7	-110	
8	100	100
9	100	100
10	200	200
Total Project Savings	500	960
Savings per Home	50	137

Note: Data in this table is simulated to serve as an example.

energy use data results is the illusion of a mean savings of 2,200 kWh per home. The mean savings is then multiplied by the full number of treated homes to determine payment. Suppose 20% of 5,000 homes in a project have negative savings. For this case, the 2,200 kWh would be calculated based on 4,000 of these homes. The mean result of 2,200 kWh per home would then be multiplied by 5,000 homes producing a claim for payment for 11,000,000 kWh first year energy savings, with additional payments for subsequent years based on assumed lifetimes of energy efficiency enhancements reportedly installed in the dwellings.

A more sophisticated variation employs a comparison group. Suppose a valid comparison group is used. Suppose, further, that the comparison group shows that energy increased (even when data is weather normalized), and the result happens to be the same as shown in the first column of Table 2. Now the net mean difference between the treated homes used in the calculation (2,200 kWh) and the comparison group (-1,200 kWh) is 3,400 kWh. Then, as above, with 5,000 treated homes in the project, a claim for 5,000 times 3,400 kWh or a total of 17,000,000 first year kWh savings is submitted for payment.

Table 2: Omitting Negative Savings (No Program; Real Data).

Omitting "Negative Savings" in the "No Program" Situation

Home	Mean (per Home) Normalized Annual Consumption	
	All Households (kWh)	Omitting Homes with Negative Savings (kWh)
Baseline Year	23,300	24,400
Post Year	24,500	22,200
Savings	-1,200	2,200

Note: Data provided by Portland General Electric.

This version is more sophisticated since using the comparison group creates an illusion of objective measurement. This version mimics the development of net savings in the standard non-equivalent control group design. This colors the process with an *image of technical objectivity*. Hidden manipulation while introducing elements to make the situation appear objective is called an *"illicit dialectical shift*" (Walton, 1995:120-123). This is a technique of persuasion in which one participant in a discussion is intentionally manipulating the other by using a format that creates an illusion of objectivity or fairness. Here the format, language and some of the basic mechanics of a standard evaluation design create the image of objective measurement, and the comparison group (in itself) is sound, but treated cases with negative savings are still omitted. The mechanics of objectivity are mis-used to cast a spell of illusion, and the use of the (valid) comparison group does not provide protection.

The Adjustable Base Year

To simplify analysis and to make results more open to critical review, it can be a good practice to define a common baseline year for a project, rather than to develop a separate baseline year for each house. However, in a set of such data, there would be some missing months for some subset of homes. For example out of 1,000 homes, maybe 25 would be missing a January usage. This situation requires rules for adjustment of data. For example, such cases could be eliminated from analysis, data could be inspected to back-allocate based on the next meter reading, or a value for usage (and temperature) from a prior January could be substituted.[7]

The pitch for the "adjustable base year" goes like this:

(1) As the performance contractor I believe we should take certain realities into account so we can have the same expectations through the project.

(2) Since it is common knowledge that there will be cases that we have to plug, and rather than throw them out of the analysis, why don't you give me – say 60 months of back data for each home and I will use the best 12 months to figure out the baseline energy use.

(2) Or, if some individual case has a problem, I might use more months and average them out to get the best baseline measurement.

(3) So, let's get that language into the contract.

Here, the buyer thinks the discussion has been about truthful technical measurement as it would be with an in-house measurement and evaluation staff. The performance contractor has been careful to maintain that illusion while instrumentally guiding the discussion.

If possible, the performance contractor may have a few other people at the table as team members who appear to be unaffiliated or to have a relative degree of technical independence. The value of these agents is greatly leveraged if they appear to be independent representatives of groups unaffiliated with the performance contractor, or "neutral" technical experts (that, for example, the performance contractor has previously infiltrated onto technical boards or committees). For this to work, participants "at the table" do not realize that there are pseudo-independent participants at the table who are working by script to coordinate the thought and discussion of a working group towards particular results.[8]

Through this kind of process, the buyer agrees (without realizing it yet) to legalize by contract the performance contractor's power to control the reading for the baseline year for each home. That is, the contractor secures the ability to use 60 months for one home, 15 for a second, and 12 for a third. Also, he may pick the "best" January or the "best" December for each home, and the like.

This creates the ability to pad actual savings with potentially large amounts of additional pseudo-savings. For example, given that energy efficiency may gradually improve in untreated homes, using such a participant home with 60 months of baseline measurement is likely to overstate the baseline energy use for that home. Tables 3 and 4 illustrate inflation of baseline energy use.

[7] To keep the example simple, appropriate weather adjustment is not discussed.

[8] Classically, these are often classed as "Communist" or "Labor Radical" infiltration and front group tactics (Selznick, 1952). In realty, however, these are old tricks used by groups on the far right as well as the far left. Today they are primarily used by major corporations in corporate issue campaigns.

Table 3: Adjusting the Baseline by Plugging with the "Best" Year.

Adjustable Base Year						
Home	Average Normalized Annual Consumption					
	kWh					
	2000	2001	2002	2003	2004	Selected for Baseline
(1)	23,300	48,000	24,000	22,000	21,000	27,660
(2)	25,000	32,000	24,500	23,500	22,500	25,450
(3)	27,000	35,000	26,000	25,000	25,000	35,000

Table 4: Adjusting the Baseline by Plugging with the "Best" Month.

Adjustable Base Year (by Month)						
Home	Average Normalized Annual Consumption					
	kWh					
	Jan 2000	Jan 2001	Jan 2002	Jan 2003	Jan 2004	Selected for Baseline
(1)	3,000	4,500	2,500	3,000	Missing	4,500
(2)	2,400	3,000	2,400	2,200	Missing	2,400
(3)	2,800	3,300	2,700	2,600	Missing	2,850

The "Bait and Switch"

It is generally a good practice to establish and follow a clear protocol in which there is no ambiguity about which kinds of energy efficiency enhancing materials and equipment will go into which homes. Also, it is good practice to define the order in which different kinds of energy efficiency remedies will be installed and to define these in standard "bundles." In contrast, it is of advantage to a performance contractor to control assignment and to keep the rules for assignment fuzzy. Here is a "bait and switch" pitch:

Step (1) [The Bait] As the performance contractor, I really want you to look at this wonderful list of materials and equipment that we have worked out and then modified along with the buyer and the working group in arriving at a final list to include as approved in the contract.

Step (2) [The buyer focuses on the materials and equipment and other remedies while the performance contractor is massaging the meeting so the buyer feels good about the many "major installations" on the list]. If artfully presented with wit, charm, and a personable style, the buyer may be lulled into allowing contract language that will permit complete ambiguity on assignment of materials, equipment, and other remedies. It will all seem 'logical' and like everything is being covered.

Step (3) Since I (as the performance contractor) will be at each house, or, that is, my crews will be at each house, it will be best if we let the crew determine the best package of things to install for each house. After all, if we were not good at this, we would not show the high savings. So, *trust me*, we really have *market motivation* to do this in the most t*echnically correct* (though, of course, secret, proprietary and undisclosed) fashion. Trust me; we are the experts on this.

Step (4) Don't worry about it. You stay in the office and administer, we'll take care of the customers and their homes.

Step (5) [The Switch] When the project is complete you (the buyer) will see very strange patterns in allocation of items on the approved list to the homes. For example, if you have allowed a formula with an initial payment of $500 for each home treated, and the balance to be "measured," you will likely find hundreds or thousands of homes "banged out" with quick "low-cost/no-cost" items in order to create a claim for the per home bounty fee.[9]

This device works best where the buyer (agency) has devolved its purchasing function since experienced purchasing staff are professionally aware to "bait and switch" as a generic problem (Kitching, 2001:1-2). In organizations in which responsibility is devolved, for example, an officer with a function marginally related to energy conservation may unknowingly sign a contract with "bait and switch" language, and then hand it to the conservation group.

It is not unusual for performance contracting to be instituted in utilities by means of a Commission order. In some cases, Commissions are moved by a theory of the free market to order utilities to do residential performance contracting. The utility may be uninterested in this business approach and unprepared for a negotiation that will displace ongoing working relationships.

"Bait and switch" is also facilitated by other kinds of "stress" conditions, in which, for example, an agency is losing its competency and institutional knowledge through deregulation, downsizing, and repeated reorganizations. "Stress" conditions occur in agencies that have over-downsized and not yet recovered.[10] With experienced staff in early retirement or otherwise downsized, institutional memory is lost. Then conservation and energy efficiency may appear to be a simple purchasing problem, like buying pencils. With no one on staff with direct knowledge of weatherization work, quality control, inspections, or – especially – evaluation, an agency that contracts out all of its programs and has no senior in-house evaluation capability will be a sitting duck for this device.

Measure Lives and *Ex Ante* Savings

Another improper device may occur when the performance contractor is to be paid *per item of equipment or other remedy installed*, based on *ex post* energy savings and assumptions regarding the life of materials, equipment, and other remedies. In some accepted measurement protocols contractors are paid a fixed increment per item installed in a performance contract, based on an asserted or calculated savings per item.

Early problems associated with this approach were countered by application of strong evaluation designs in the context of integrated utilities using least-cost planning. The non-equivalent control group design and the adoption of forms of regression analysis (including conditional demand analysis and statistically adjusted engineering analysis) made simple assertions for inflating savings impossible.

Then utility deregulation replaced "planning/inspection/evaluation" with reliance on markets. During the deregulation era both program knowledge and evaluation capabilities were often lost to institutional memory. Less than a handful of states maintained systems like CALMAC, NYSERDA, or the Oregon Energy Trust where knowledge was retained over the great dip in material focus on integrated resource

[9] It gets much more sophisticated; there are many variants of this device.

[10] "Stress" is one of six technical elements of classic cons (Swierczynski, 2003: 9-10).

planning, inspection, and materially oriented evaluation of energy efficiency results. Now, as energy efficiency effort has been rebuilding, vulnerability in this area has again occurred, due to this gap. Here is the pitch:

(1) As the performance contractor, I have here a list of savings per measure for energy efficiency equipment, materials, and practices developed from our (secret, proprietary, and un-disclosed) experience. I also have a similarly constructed list of lives for these items. Plus we have updated them with some of the *latest research* from studies at *prestigious national laboratories*. Let's put these into the contract.

(2) So, we will do some measurement to determine payment but let's reserve out a set of measures for which you will pay per bulb, or per outlet gasket, or per widget, etc.

(3) Also, evaluation is such a hassle, and you no longer have an in-house evaluation capability -- and hiring an external evaluator would just mean there would be less money to make homes more energy efficient and fewer households would be treated. So, to make this work for everybody we could just throw in the evaluation and have our person do it – maybe I will offer to split the cost since we are partners in this joint effort. But, in any event, whether we do the evaluation for you ourselves as a gesture of good will, or we spilt the cost and have our person do it in partnership with you, or – if you really think you have to – you bring in an independent evaluator mutually agreeable to us…lets agree up front on the savings for each listed item and the lives for each listed item and put them in as an appendix to be referenced in the payment formulas.

The problem here is that the claim of secret proprietary knowledge of residential weatherization does not pass a simple straight face test, but rather there is a broad engineering and retrofit specialist community that shares this knowledge.

What about the incorporation of the "latest research from prestigious national laboratories?" Well, that work is being done at Oak Ridge, LBL, and PNL, but without the specifics, all you know is that the contractor claims the latest knowledge. If you go forward on the unwarranted assertion of secret knowledge and prestigious authority you are likely legally surrendering buyer (agency) property in return for phantom savings.

Tail Trimming

Sometimes a contract will specify that annual savings will be calculated using "tail trimming." Tail trimming is a crude but acceptable data cleaning method that can be used to remove outliers and/or extreme values from a set of data. Utility energy use data is typically quite dirty because the system is designed for billing rather than for analysis. When errors creep into energy use records as they do to every inherently fallible human activity or system, corrections are applied. However, the corrections are applied to correct bills but not to correct the use values in the records. Also, nearly every residential study with a large statistical sample of homes will contain some cases with absurd readings, sometimes from businesses or multiple family dwellings that have been improperly classified as homes, or occasionally from apartments that have been improperly billed for common area lighting, and similar errors. Tail trimming lops off these absurd cases.

In a correct implementation of tail trimming, the same number of cases is deleted from the upper and lower tails of the treatment group and from the upper and lower tail of the comparison group. For example, the ten highest and ten lowest cases might be deleted prior to calculating the mean gross energy savings of the treated homes. Then (in parallel) the ten highest and lowest cases might be deleted prior to calculating the mean gross energy savings of the comparison group. The group means calculated after removing the outliers and/or extreme cases are called “trimmed means.” But, if the outlier exclusion is not correctly performed, it is easy to create hundreds of thousands of dollars of payment to the performance contractor for phantom savings that really are not there.

This is the “two-sample problem” (Kruskal & Tanur, 1979:237). If the treatment group and the comparison group are of the same size, equal *numbers* of cases are deleted from the upper and lower tails of each distribution. If the two groups are of different size, equal *percentages* of cases are deleted (Yuen & Dixon, 1973).[11] The treatment is applied independently to both the treatment group and the comparison group, that is, the same treatment is applied separately to each of these groups.

This is in keeping with the *principle of symmetry of treatment* that runs throughout the non-equivalent control group design and the calculation of difference of means between baseline and post-year for the treatment group and for the comparison group. Symmetry of treatment is a general rule applied in measurement and evaluation.

A tail trimming problem is pictured in Figure 1. Think of this as generic diagram of energy savings showing a distribution curve with five percent of cases trimmed off each end of the distribution. Note that this particular generic distribution happens to have a long upper tail.

Suppose the lower tail for the treatment group looks the same as the lower tail of the comparison group (we are making these the same to simplify so we can concentrate on only the upper tail). For example, both lower tails might look like the generic lower tail in Figure 1.

Suppose, further, that the upper tail of the *comparison* group selected by the performance contactor is long (the shape shown in the right half of Figure 1), while the upper tail of the treatment group is short. When trimming is correctly carried out by trimming the upper 5% of cases in each distribution, separately, the basic relationship of the means of the two distributions is preserved and the extremes of both distributions are eliminated prior to calculation and comparison of the trimmed means.

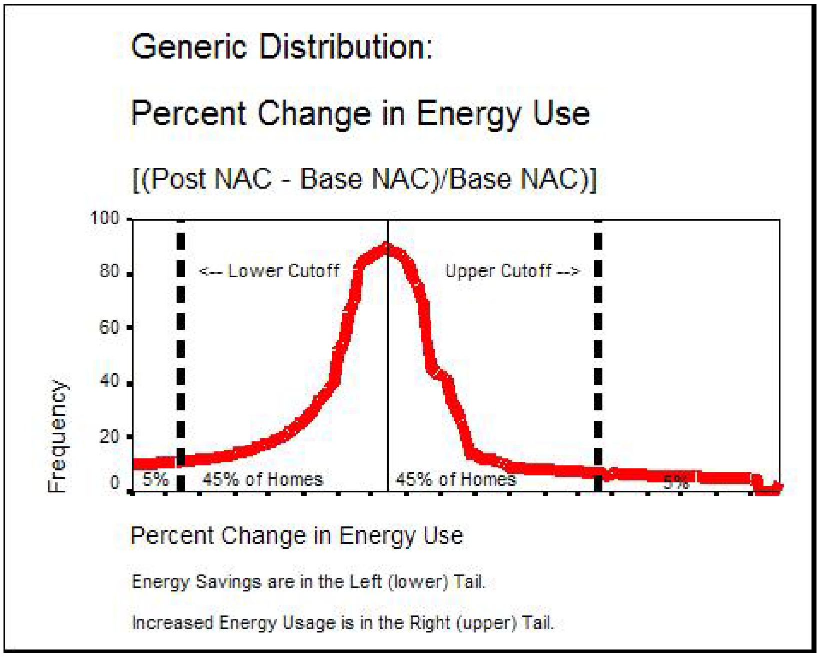

Figure 1: Tail Trimming.

[11] A further challenge is that the statistical models for this type of analysis assume random assignment with a true experimental control group which is almost never the case in this field of application.

However, if, improperly, the two distributions are first combined – the data points in what were previously the upper tails of two separate distributions become intermingled. Now, if tail trimming is performed on the combined distribution, the cases excluded from the distribution with the longer tail (the extreme high use cases from the comparison group) are lopped off resulting in a lower mean gross energy savings for the comparison group. However, none of the extreme high use points in the treatment group's upper tail are excluded which means that extreme cases are included to pull up the mean gross energy savings of the treatment group. When the net savings is calculated by taking the difference of these figures, the energy savings of the comparison group has been artificially lowered and the energy savings extremes of the treatment group have all been retained, resulting in an amplification of apparent net energy savings.

Here is the "tail trimming" pitch:

(1) [Contract Language] Sort homes in both treatment and comparison groups by the percentage change in Normalized Annual Consumption. Delete 5% of homes with the lowest percentage change and 5% with the highest percentage change. Calculate the baseline year and post-year Normalized Annual Consumption using the homes left in the analysis.

(2) The performance contractor manages to control the selection of either treatment or comparison group homes, or both, so that the 5% cut off hits at very different points in the two distributions.

(2) The performance contractor then insists that the contract language means that the treatment and comparison groups should be combined into a single distribution and the extremes removed from that merged distribution since, he asserts, "homes in both treatment and comparison group" means the pooled homes from "both" groups rather than the symmetric application of tail trimming to the separate distribution of each group.

This can amplify returns to the performance contractor for phantom savings.

Capitalizing on External Treatments

When developing contract language, the performance contractor may (reasonably) ask for contract language specifying payment based on "measured" savings aggregated across all treated homes. The buyer will (reasonably) insist on contract language that specifies a comparison group and may be successful in securing a valid comparison group. But a valid comparison group, though important, is not sufficient to protect against capitalization on external treatments. External treatments can enter through a contractual loophole in the language defining the program participants.

For example, if the performance contractor independently secures a list of homes treated by Community Action Agencies during the project period he may focus low-cost/no-cost efforts in certain homes with full weatherization (including major energy saving installations) by the Community Action Agency. If the contract does not anticipate this problem by ruling out of the calculation of results homes with external treatments and these results are loaded on the treatment homes but not on the comparison homes, the profit potential inherent in the amplified savings can be quite large.

In this case, the comparison group performs as it should. The buyer can test the comparison group and be sure that it is working. But the savings due to treatment is artificially inflated by the work of the federal Weatherization Assistance Program effort to yield a high multiple of the results actually produced by the performance contractor's efforts. Even if the comparison group contains some homes with external treatment, if the treatment group is systematically loaded with these homes, there is a swamping effect that overcomes the protection of the comparison group.

Here is the pitch:

(1) [Performance Contractor] Let us set the calculation of savings to reflect the change in energy use between the baseline and the post period, defining the baseline and the post period as a year each, and normalizing for weather.

(2) Language to this effect is then placed in the performance contract. Note that the buyer assumes that the energy savings measured will be due to the contractor's work.

(3) The contractor then piggybacks a subset of treatment group homes on the comparable federal program homes while adding at least one minor energy savings enhancement to each of these homes and almost zero cost. This legally fulfills the contract in that his crews have at least entered each of homes and have claimed to have done something in each home.

(4) The buyer's program manager looks at what was installed in the homes (as claimed on the contractor's invoices) and realizes that there is no way that these installations could produce more than a tiny fraction of the apparent energy savings reflected in the energy use data taken from utility records.

(5) The buyer knows something major is wrong, but never figures out the overlap with the federal effort, and pays the contractor for all apparent energy savings; or the buyer happens to get a list the federal homes and realizes where the savings come from. In that case, the performance contractor insists the legal language of the contact be followed and the contract is silent on this point.

Fuzzy Selection: Which Residences will be treated?

More generally, the performance contractor will try to fuzz selection of homes. In contrast to a Community Action Agency or a local time and materials contractor, a performance contractor usually tries to hide the exact criteria they will use to select residences to be treated. Also, as a follow on to a claim to *mysteriously secret supposedly special and proprietary knowledge* in selecting homes to be weatherized, the exact criteria for selection of a comparison group may also be fuzzed. If the performance contractor is able to con the buyer (agency) into putting this fuzziness into contract language, the treatment and comparison groups can be made systematically dissimilar through selection bias as shown in Table 5.

In this example, it is easy to see the bias. The treatment and comparison group may be similar on many criteria but they are not equivalent on usage in the baseline year.

Table 5: Fuzzing Participant and/or Comparison Case Selection.

Fuzzy Selection (Achieving Systematic Bias)		
Measurement Period	Average Normalized Annual Consumption	
	Treatment Group	Comparison Group
	(kWh)	(kWh)
Baseline Year	23,300	15,000

When energy efficiency work has a socially responsible or community oriented focus on improving the housing stock of a region, this pattern is referred to pejoratively as "cream skimming." When a wider social perspective informs energy efficiency efforts, "cream skimming," "cherry picking," or "hi-grading" is a serious technical sin, a "crime against the people" from a socialist perspective. The contractor is taking the "low hanging fruit" but damaging the vineyard and to get the rest of homes in the area later will substantially raise the overall cumulative cost of conservation. The cost of reaching the neighborhoods to serve the remaining homes will be repeated but the attainable savings will be less as the rest of the community is served.

When energy efficiency work has a narrow "Chicago School" profit maximizing motivation, "cream skimming" is encouraged. Under this model "cream skimming" is a virtue and the contractor who bags the most money from utilities and government while spoiling the market becomes not only rich but is socially regarded as smart. However, for the evaluator, the comparison group is still technically in error because it will inflate results by introducing substantial phantom savings for the performance contractor. This can greatly amplify apparent energy savings.

Regression to the Mean

Omitting negative savings, capitalizing on external treatments, tail trimming and fuzzy selection are different ways of selecting cases. Each introduces a form of selection bias. Another kind is selection of participants to introduce regression to the mean.

Figure 2 illustrates how residential weatherization is supposed to work. The performance contractor applies treatment at a particular point in time. Normalized energy use is measured for the year before and the year following treatment. The true benefit is the difference (weather normalized energy savings) between the post-year and the baseline year, adjusted for the change in the comparison group.

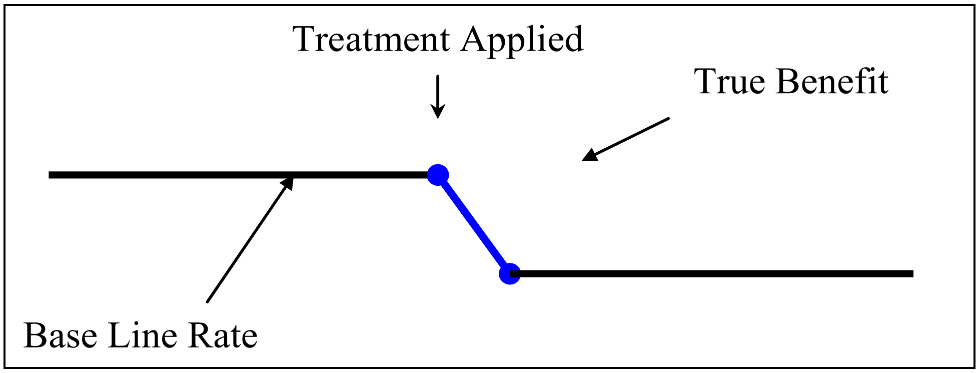

Figure 2: Real Conserved Energy - No Regression to the Mean.

Figure 3 shows what happens in regression to the mean. Treatment cases selected are biased towards cases with an unusual blip in energy use in the baseline year. The blip can look like randomness or noise in the data. But this is not measurement error – it is the result of multitudinous actual non-proximate research factors acting at a distance (only some of which may truly be random).

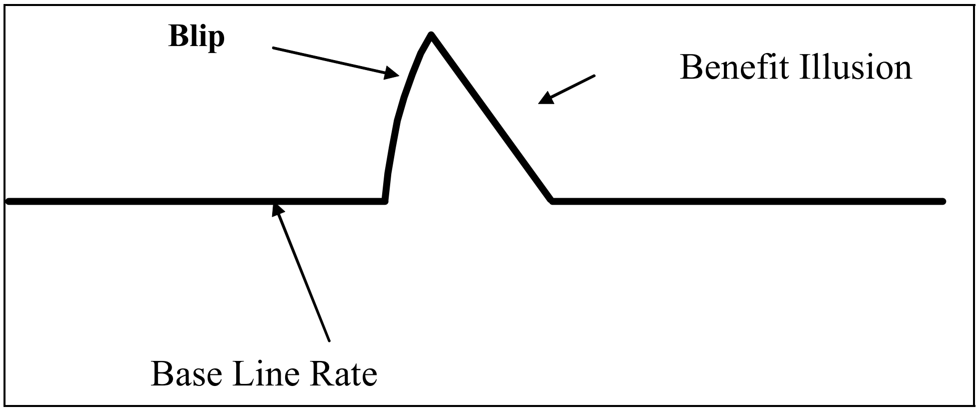

Figure 3: Blips in the Baseline Year - Regression to the Mean.

In regression to the mean extreme scores on the predictor variable or variables for individual cases (pictured as the blip in Figure 3) tend to be associated with less extreme scores on the predicted variable. This is shown in Figure 4 where the scores on the left vertical axis are *individual* baseline year normalized energy use scores for individual homes. The lines across the graph tie these points to their *mean* scores in terms of normalized energy use in the post year.[12]

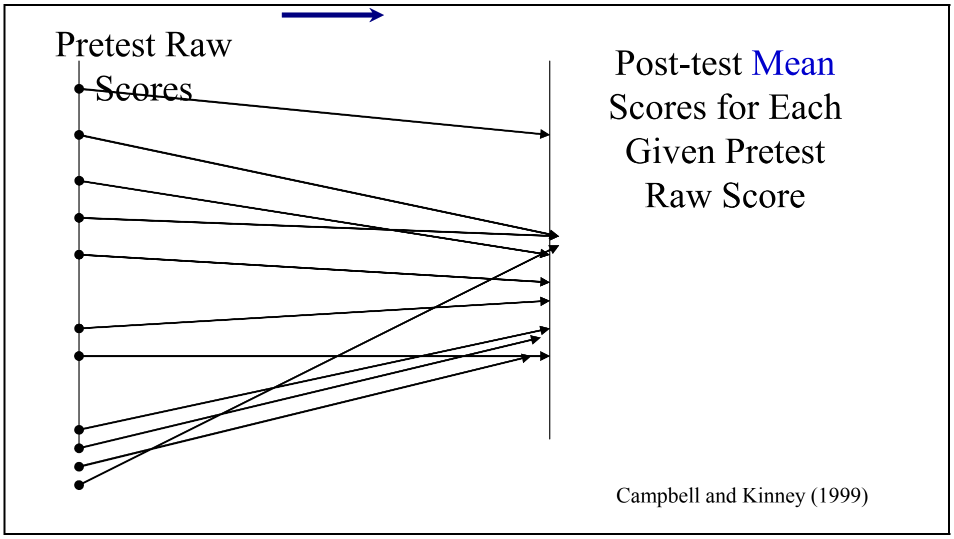
Pretest Raw Scores
Post-test Mean Scores for Each Given Pretest Raw Score
Campbell and Kinney (1999)

Figure 4: Galton Squeeze Diagram showing Regression to the Mean.

[12] Figure 3 is a "Galton Squeeze Diagram" as developed by Campbell & Kenny. In the Galton Squeeze Diagram, each score on the left is conceptualized as one of many scores with a common mean on the right. Campbell, Donald T. & David A. Kenny, *A Primer on Regression Artifacts*. New York & London: The Guilford Press, 1999.

As shown in the figure, scores toward the middle of the distribution show little regression to the mean (no regression to the mean would be represented by a horizontal line across the graph, connecting all of a set of identical baseline scores with the same post year mean score). At the same time, individual scores that are very high in the baseline year link to lower mean values in the post year; while the individual scores that are quite low in the baseline year tend towards a higher mean value in the post year.

With five years of data, it is possible for the performance contractor to select for homes that are particularly (but apparently randomly) low compared to their normal level and homes that happen to have an anomalously high value for the baseline year in comparison to their normal level. This knowledge can be used to separately create both a comparison group that will show low energy savings and a treatment group that will show comparatively high energy savings even in a "non program" situation. (Or, the knowledge may be used to create the effect for only the treatment group or only the comparison group.)

Note that though this device operates through the mathematics of the statistical methods used in standard evaluations, it is specifically a *statistical device* rather than a mathematical device. That is, it is not a consistent mathematical adjustment to all cases. It works by affecting to different degrees a preponderance of cases (for a large effect) or a concentration of cases (for a small effect).

In a program situation, the combination can greatly inflate material savings due to the program with substantial phantom savings. A partial protection against this device is to provide the contractor with only one calendar year of baseline data.[13]

The Hidden Auxiliary Variable

This improper device combines selection bias with use of a hidden auxiliary variable. This method can defeat all current measurement protocols and two of the evaluator's major tools for assessment of residential energy savings. It can defeat the non-equivalent control group design with difference of means testing, and it can defeat the regression approach (both conditional demand analysis and statistically adjusted engineering analysis).[14]

In the simple version of "fuzzy selection" ("G" above) systematic bias is evident in the usage numbers for the baseline and post-retrofit years. In the more sophisticated "regression to the mean" version ("H" above) the Galton Squeeze Diagram and the Backwards Galton Squeeze Diagram (not shown) will tip off what is happening. Using a hidden auxiliary variable is not initially obvious.

Here is the pitch:

(1) As a performance contractor, I am taking the risks on this project, so to mitigate that risk, I am sure you will agree to supply me with three to five years of data for all residential customer homes that might be chosen for treatment or for the comparison group. After all, if you were doing this work internally or with a Community Action Agency or with local independent contractors it would only be prudent to have this kind of depth of data available. To be fair all around, let me have equal access to

[13] Of the nine devices, this is the only one we have not yet found in practice. We have encountered all of the others at least once, and some many times. We include it because regression to the mean is a classic problem in evaluation and it should always be checked in any case.

[14] The non-equivalent control group difference of differences in means approach is much less sensitive to the inflation effect than are the regression approach, however both types of approaches are affected.

all the data I need to have a level playing field with the utility and the community agencies and the independent evaluators.

(2) The performance contractor then permits the development of an adequate comparison group because that is not where the action is in this method. A great show of open-handedness may be made in helping the buyer (agency) focus on developing a very good comparison group. This diverts possible attention to what is happening in the treatment group. The protection offered by the comparison group can be defeated in a way that does not show in the analysis.

(3) For the treatment group, the contractor analyzes data patterns over the three or five year period. He may have exact knowledge of an auxiliary variable or simply look for apparently non-random patterns in the data. Over this period, some homes are decreasing in normalized energy use. Others are showing an increase in normalized energy use, sometimes a substantial increase from year to year. In contrast to the "regression to the mean" approach which involves a search for positive and negative blips, here the search is for homes in which change is non-random and systematic.

(4) The crews then report *many* energy efficiency enhancements and *major* energy efficiency enhancements to the homes decreasing in energy use. They assign *few* and *minor* energy efficiency enhancements to homes showing marked increase in energy use. At the same time, some cases are treated normally because this method does not depend on mathematically uniform treatment of each case; it works by *statistical preponderance*. This means the device is used for a substantial subset of cases while other cases are left unaffected. For this reason, it is harder to detect. If the buyer asks to do field inspections, there will always be many normal cases available for office review or field inspection.

(5) The evaluator conducts a conditional demand analysis (CDA) or statistically adjusted engineering estimation analysis (SAE). Or, the evaluator conducts a difference of differences of means test referenced to a non-equivalent control group design.

(6) For the sample of homes analyzed, the regression (conditional demand or statistically adjusted engineering estimate) automatically picks up the association of the projected savings of the enhancements with utility data on savings per home and converts the *ex ante* savings for each type of energy efficiency enhancement to a high *ex post* savings. The high *ex post* values are then applied to the full set of energy efficiency enhancements in the full population of treated homes.

(7) The difference of means test automatically functions on the difference of differences in pre *vs*. post energy use between the treatment group (which has been rigged) and the comparison group (which is sound). The protection usually offered by the comparison group is thereby defeated.

In a variation of this device, the energy efficiency enhancements are installed without reference to the patterns of systematically decreasing and systematically increasing energy use in the homes over the baseline years, but the contractor "takes credit" for the full *ex ante* value of installed enhancements in the homes systematically decreasing in energy use and "contributes at no cost to the project" by failing to claim the *ex ante* values for the energy efficiency enhancements in homes that are systematically increasing in energy use over the baseline year. The most likely pattern is a mix of these two, along with a normal cases since the mixture makes it harder for the evaluator to spot the device.

This method can produce multiple millions of dollars of return to the performance contractor for phantom savings and it slips by current evaluation protocols which do not take it into account. The performance contractor does treat the homes and does produce some material savings, but may leverage a deserved payment of a few hundred thousand dollars into an actual payment of multiple millions of dollars.

There are two ways to control this device. The method can be partially blocked by never giving the performance contractor more that one calendar year of baseline data. However, sometimes Commissions require utilities to provide information similar to the extent of their own internal information to alternative vendors, for example, to promote the development of competitive markets. Or, the utility management may have agreed to provide several years of data at the request of the performance contractor.

If the performance contractor is given three or five years of baseline data, the apparent energy savings can be corrected by the evaluator by requiring the assessment of energy savings to take place in two steps:

(1) An initial regression is run using the dependent variable (energy savings) against the change in energy use over the baseline period.

(2) The residual savings from this regression is used as the dependent variable for the full subsequent analysis. This will separate the residual material savings actually due to the contractor's work from the bubble of phantom savings.

Conclusions

There are many improper devices to create phantom savings; the list above provides some of the most interesting, and can be used as a check list by evaluators working on residential performance contracting evaluation or verification studies.

However, in a quasi-Darwinian competition, residential performance contractors continue to evolve improper devices to fool our best evaluation methods and techniques. Many of these methods can be combined in a typical residential performance contracting project; some of those most easy to spot can be seeded as "giveaways" to satisfy the evaluator. These can be "trade offs" to make the evaluator go away without finding the millions of dollars of phantom savings in the deep structure.

The best evaluation protocols can be fooled where there is a will to do so. The saving grace when evaluation protocols are defeated may be the program manager who will insist, based on practical experience with work in a particular region or service territory that "these claimed savings cannot be coming from these reported energy efficiency enhancements." Sometimes evaluation engineers can look at data patterns and realize that the claimed savings are not possible in the context of the physics of buildings. It is these kinds of practical understanding of projects by experienced program managers and engineers that can cue the evaluation team to go to the roots of the project data and method and find improper devices that cause the *illicit amplification* of (apparent) energy savings far beyond that actually created by the contractor's performance.

References

Campbell, Donald T. 1988. "Evolutionary Epistemology." In *Methodology and Epistemology for the Social Sciences: Selected Papers*, edited by E. Samuel Overman, Pp. 393-434. Chicago: University of Chicago Press, 1988.

Campbell, Donald T. & David A. Kenny, *A Primer on Regression Artifacts*. New York & London: The Guilford Press, 1999.

Campbell, Donald T. and J. Stanley (1966), *Experimental and Quasi-Experimental Design for Research*. Chicago, Rand McNally Publishing Company.

Cook, T. D. and Donald T. Campbell (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston, Houghton-Mifflin.

Hansen, Shirley J. & Jeannie C. Weisman (1997). *Performance Contracting: Expanding Horizons*. Liburn Georgia, Fairmont Press; Upper Saddle River, New Jersey: Prentice-Hall PTR.

Hill, Lawrence J. & Marilyn A. Brown, "Estimating the Cost-Effectiveness of Coordinated DSM Programs," *Evaluation Review*, 19(2):181-196, 1995.

Kitching, Trevor, *Purchasing Scams and How to Avoid Them*. Aldershot, Hampshire, England & Burlington, Vermont: Grower Publishing Company, 2001.

Khawaja, M. Sami & Connie Colter (2001), *Pay-for-Measured Savings: Review of Programs, Final Report*. Portland, Oregon: Quantec LLC, prepared for the California LIEE Pay for Measured Savings Pilot Project Planning Team.

Kruskal, William H. & Judith M. Tanur, eds., *International Encyclopedia of Statistics*, Volume 1. NY, NY: The Free Press; London: Collier Macmillan Publishers, 1978: 237.

Marriott, F.H.C., "Trimmed Mean," *A Dictionary of Statistical Terms*, fifth edition. Essex, UK: Longman Scientific & Technical; NY, NY: John Wiley & Sons, 1960:209.

Peach, Gil, "Performance Contracting: Advice to Utilities," *Home Energy*, Nov/Dec 1992: 19-21.

Peach, H. Gil, "Perspective on Performance Contracting, What is Real and What is Not in Residential Sector Performance Contracting. Beaverton, Oregon: H. Gil Peach & Associates LLC (Monograph), 1995.

Selznick, Philip, *The Organizational Weapon, A Study of Bolshevik Strategy and Tactics*. Glencoe, Illinois: The Free Press of Glencoe, 1960; originally published by the Rand Corporation, 1952.

Swierczynski, Duane, *The Complete Idiot's Guide to Frauds, Scams, and Cons.* Indianapolis: Alpha Books, 2003.

Walton, Douglas, *A Pragmatic Theory of Fallacy*. Studies in Rhetoric and Communication, Tuscaloosa, Alabama & London: The University of Alabama Press, 1995.

Yuen, Karen K., "The Two-Sample Trimmed t for Unequal Population Variances," *Biometrika* (1974), 61, 1: 165-170.

Yuen, Karen K. & W.J. Dixon, "The Approximate Behavior and Performance of Two-Sample Trimmed t," *Biometrika* (1973): 369-374.