Freeridership Borscht: Don’t Salt the Soup

Kenneth M. Keating, PhD
Portland, OR

ABSTRACT

Measuring and reporting freeridership through self-report mechanisms, which is one element of determining the net impacts of programs, is a difficult evaluation challenge. It is fraught with dangers of measurement error as well as potential response biases. In addition, the researchers themselves can add their own bias to the process. This paper isolates one particular problem that has cropped up even among experienced evaluators. The problem involves the use of a multiplicative algorithm to combine response scores in self-report methods. An exploratory effort to understand the potential frequency of the problem is included. The intention of the paper is not to cast blame on those who have inadvertently allowed the bias in the past, but to provide documentation of the issue, with the hope that future evaluations will avoid the problem.

Introduction

The observed changes that would not have occurred in the absence of the program are considered net effects of the program. Estimating net impacts in the energy efficiency field generally means accounting for alternative possible causes of the observed change, including economic and historical influences. Nevertheless much of the evaluation effort is focused by policymakers on freeridership and spillover from the program. These are logically off-setting, but not necessarily equal effects. Freeriders are program participants who would have implemented the program measure or practice in the absence of the program (TecMarket Works 2006). Spillover results from measures or behaviors adopted because of the presence of the program that would not have happened absent the program. Freeridership gets the most attention, because it is considered an indication of an unnecessary expenditure of public or ratepayer funds, and dilutes the benefit of running a program. Therefore, measuring freeridership is almost always an evaluation objective. Spillover, as real as it might be, draws less attention unless a program is specifically designed to change the market.

The most direct and transparent way to determine if someone would have taken the action without the program is to ask the participants a set of related questions to try to ferret out their motivation. This is called a self-report approach. It has become a very important methodology over the last 20 years (Ridge et al. 2009). The literature is replete with issues about the potential measurement error, respondent biases, and the nature of trying to discover what might have been done, but wasn’t done – the counterfactual (TecMarket Works 2004; McRae 2002; Peters 2008).

Many identified issues can be mitigated through careful research design, sample selection, survey wording, moving the survey closer in time to the decision, and identifying areas of potential bias. Over the years, evaluators have become sensitized to the nuances – “is influence a better term than importance?” or “is an eleven point scale better than a 5 point scale?” or “how long into the future is too long to actually believe a prediction of future intent?” How do we treat differing responses from different decision makers in the process? How are apparently conflicting statements by the same respondent to be handled? The assignment of probabilities to specific answers to questions have been developed into an art – “if they say somewhat likely to have done it, should that be a 0.5 probability or a 0.825 probability?” How do researchers handle “don’t know” responses? Most evaluators are trying hard to reduce measurement error, overcome the intrinsic conundrum of the counterfactual, and avoid bias. Sometimes, in the search for sophistication and nuance, what should be seen as obvious bias can be missed.

Inadvertent Bias

Many non-statisticians associate the term "bias" with an effort to favor one answer over another, and it is assumed to be deliberate act. It is possible for researchers to let personal judgments lead them to "shade" the analysis in one direction or another. For example, I have seen a case where the evaluator so distrusted the results of a free-rider analysis that he determined that it was only fair to subtract 10% off any freeridership finding above 25%. This is rare indeed, and technically, statistical bias is not a judgmental term. One of the more complete definitions is: "Bias (Epidemiology)

1. Any deviation of results or inferences from the truth, or processes leading to such deviation. Bias can result from several sources: one-sided or systematic variations in measurement from the true value (systematic error); flaws in study design; deviation of inferences, interpretations, or analyses based on flawed data or data collection; etc. There is no sense of prejudice or subjectivity implied in the assessment of bias under these conditions.[1]" (Sensagent 2008).

The distinction between bias and random measurement error is that measurement error will decrease with additional measurements, but bias won't improve no matter how many times you repeat the measurement. While never desirable, bias can be tolerable in research, because it may be a small impact relative to the standard error. (TecMarket Works 2004, 291) Nevertheless, if bias can be avoided, it should be; and if a technique that results in avoidable bias becomes widespread in a discipline, it needs to be eliminated before it undermines the credibility of the field.

Intrinsically Biasing Freeridership Algorithm

There is a biasing algorithm in the measurement of freeridership. It is a category of algorithm that has been possibly used for many years. The California Framework, for instance states that combining probability scores can be done additively or multiplicatively (TecMarket Works 2004, 137). The problem is simple to see once it is pointed out. Fundamentally it comes down to the mathematical properties of multiplying proper fractions or decimals. When they are multiplied they always produce a smaller number than either value being multiplied; for example, ¼ * ¼ = 1/16, or 0.5 * 0.5 = 0.25. It always results in a smaller number. How does this apply to freeridership estimates? As noted above, the field has focused a lot on assigning a probability of freeridership to an answer to specific questions; for example: "don't know = 0.5," and "'somewhat influenced' means they are 0.7 a freerider" for that measurement element." Other than being almost always arbitrary and always open to criticism on that basis, a large problem arises when the researcher decides how to combine the results from each score by multiplying them to create a summary score.

Multiple questions and scoring are thought necessary in order to avoid misunderstandings between surveyors and respondents that can result in blindly accepting a single answer with measurement error as reliable. A simple, single question that is obvious to the researcher may be interpreted differently by a respondent who is not aware of what the issue is. Any question can be answered, coded, or interpreted with error. Therefore multiple questions are needed to add hopefully to the reliability of the exchange with the respondent. The opportunity to deal with seemingly contradictory answers is one benefit of multiple questions, but the multiplicity of scores leaves the researcher with the need for an algorithm, or way to combine the scores provided.

[1] http://dictionary.sensagent.com/statistical+bias/en-en/

To provide a set of heuristic scoring choices to illustrate the problem from two perspectives, the following arbitrary, but recognizable, scores are postulated.
Assume the respondent answers:

- “somewhat important” to the question of ‘how important program technical advice was to the purchase decision?’ – score = 0.7 freerider;
- “very important” to the question of how important was the incentive offer – score = only 0.1 probability of freerider
- “don’t know” to the question of ‘how long the respondent had been planning to install the measure before hearing of the program?’ – score = 0.5 freerider;
- “within a year” to the question of “when the respondent would have taken the program action if the program had not been available?’ – score = 0.8 freerider.

While each score might be an accurate “measurement” of the underlying phenomenon, combining them multiplicatively can only result in a biased summary score that is lower than any single value. Hence, (0.7 * 0.1 * 0.5* 0.8) = 0.028 – lower freeridership than even the 0.1 for the “very important” incentive. A less dramatic result comes from saying that in the above example that there are two elements that can be scored separately: importance of program, and timing of planning. These two facets might be averaged together, but if they involve multiplication within each factor the problem remains. In this case, ((0.7*0.1) + (0.5* 0.8))/2 = 0.24, which is lower freeridership than would be indicated by three of the four variables. In fact, the downward biasing effect of multiplying decimal scores is easy to see if we ask “how we could get still lower freeridership.” The answer would be to simply add more questions that score with a decimal. “How important was the program marketing to your decision to purchase?” Answer “don’t know” = 0.5. Then the score becomes ((0.7*0.1*0.5) + (0.5*0.8)/2 = 0.22 freeridership.

The clearest example of why the bias can be very strong is to consider the case of four answers of “don’t know,” each scored at 0.5, for final multiplicative score of 0.07 freeridership or 93% net participant, even though the researchers haven’t actually elicited any information from the respondent. Add a fifth “don’t know” question/response, and follow the logical extension. This not only illustrates the problem of multiplicative algorithms, but of the treatment and scoring of “don’t know/refused/no answer” are handled. This is worthy of another brief research paper in and of itself, but suffice it to say here, that if the researcher chooses a fractional value for these non-responses, the researcher should be careful to avoid multiplying it with other factors.

A second perspective on the bias follows from changing one score in the above example to 0.0. Assume that the answer “very important” to the question on the importance of the incentive was the extreme answer, scored a zero probability of freeridership. In the first example of pure multiplication above, the one answer trumps all other indications from the respondent. Zero times anything is zero. This would be acceptable if the question were considered “critical,” but as noted above, why would we expect the “critical question” to be always answered without measurement error? That is why we have multiple questions, not simply a few critical ones. Why this is intrinsically biasing freeridership downward can be seen in the reverse case; if another score, for example the 0.8, was re-scored to be 1.0 – i.e. full freeridership – it could be reduced by multiplying it with other decimal scores, whereas the 0.0 can’t be increased by the other scores. Reversing the subject of the scoring as suggested below, would also indicate the directional bias, only this time, reducing net participation.

Assuming that each question stands alone as a separate indicator of free-ridership, each with its own score or probability, each score could be seen in terms of its complement. Generally net program effects can be expressed as (gross tracking system effects- freeridership) + spillover. In some places like California, net program impacts are expressed even more simply as (gross effects - freeridership).

Within the brackets net participation is captured, and is the complement of free-ridership, i.e., whatever impacts are not freeridership, are net impacts.

For a simplistic example of the issue, simply take the exact same questions and scoring used above and reverse the subject of the scoring. Now the score is for "net participation" or (1- the assigned freeridership score). This would create no problem if the researcher is averaging the scores, say 0.3 net participant (1- 0.7) and 0.9 net participant (1 - 0.1) and 0.5 net participation (1 - 0.5) and 0.2 net participant (1 – 0.8). The result would be 0.49 net participant. This would be the complement of the free-rider score compiled by averaging the complements – 0.51. In the case, however, of simply multiplying all scores, the result was 0.028 free-ridership. Using the complements of the scores, we would get ((((1-0.7)*(1- 0.1)*(1 - 0.5)*(1 – 0.8))) = 0.11 net savings[2]. Not only would program planners be outraged by such a net value, but it is likely that the source of the problem (multiplying decimals) would be quickly isolated.

Partial and Deferred Freeridership

While all scoring that doesn't result in a full net participant or a full free-rider could be considered a fractional free-rider or exhibiting a mix of intentional leanings, they generally aren't classified as partial free-riders. Partial freeridership is based on the concept that a program participant may have purchased fewer efficient measures without the program, or less efficient measures, but still more efficient than the baseline, standard for program measures without the program. Deferred freeridership is implied when the purchase decision may have only been advanced a short time sooner than the purchase of the program qualifying measure would have occurred without the program. Examples would be, respectively, that the respondent would have bought 20 efficient measures, but the program convinced the firm to buy 25; the respondent would have put in a SEER 14 air conditioner (above the federal standard of 13), but the program lead them to put in a SEER 15 (partial free-ridership); and that the program lead them to install the efficient measure now, instead of two years from now as they say would have happened absent the program (deferred free-ridership). While the third, deferred freeridership example, may be best seen as an indicator of a timing-related probability and considered as an indicator of the realism of the intention, the first two would appear to set themselves up to be measured as proportion, or a decimal value, of the full freeridership savings.

This would seem reasonable. Yet, how these values are combined with the intention/timing/importance probabilities remains important.

Conditional Probabilities

As with all simple appearing rules, there are exceptions. It would be proper to use a multiplicative algorithm if the phenomenon being measured is best represented by a contingent or conditional probability. In all of the above examples, questions, scoring, and interpretations, each question that is combined in the scoring is seen as a separate way to measure the same phenomenon or underlying construct – free-ridership or net participation. That is, each score assigns a likelihood or probability of the underlying construct.

It is theoretically possible, and even preferable, to view some of the questions as conditional facets of a single measurement. In this sense, conditional means that in order for the phenomenon to be valid, all facets must be true. The answer X is valid only if A AND B AND C are true. In that case the probabilities of A, B, and C, not only can be multiplied, but they <u>must</u> be multiplied together. In

[2] It would not make sense to multiply probabilities of the complements under any circumstances, but then the argument is that it doesn't make sense to multiply the related measured probabilities either.

the case of free-ridership, some aspects can be construed as only fully true if the respondent would have bought a product at the same time, would have bought the same quantity of the product, and with the exact same efficiency as they did with the program, except without the program. If any of these are not fully true, then the program should be given some credit for the observed savings. In the example above, if the respondent had said that he would have bought 20 lighting fixtures without the program, but bought 25 with the program, it is still possible that the 20 fixtures may have been purchased much later or would not have been as efficient as the program-supplied fixtures. In that case, the probability of free-ridership applies in a conditional fashion to the 20 purchases that were not clearly induced by the program. So the result would be 0.8 fixtures savings * 0.9 for timing[3] * 0.50 for efficiency or only 0.36 free-ridership. This result may be averaged or otherwise non-multiplicatively combined with other measures of free-ridership such "amount of program influence," "likelihood of paying the same amount as paid by the program incentives," or other non-conditional measures of the phenomenon.

There are two very important caveats on the use of conditional approach. The first is that the efficiency aspect has to be independent from the other aspects of timing and quantity. If they are not independent facets of the decision, the researcher and respondent are back to providing alternate assessments of the overall free-ridership phenomenon with separate questions. This occurs when the questions are asked in terms of "program measures" or "measures" instead of "products" because that mixes the efficiency of the product with the timing or quantity. This distinction is crucial, except when the product can only be the measure (Goldberg 2009). For this reason, instead of conditional probabilities, they become alternative measures of the probability.

The second caveat is that, even if the researcher is very careful about asking about the timing and quantity of a "product," and not a "measure," the respondent may not be keeping both concepts separate in their mind as they answer. This makes it difficult for the researcher to be confident that the respondent is answering the question being asked. This latter problem is one shared by most survey researchers, but can be troubling in the nuanced situation when a conditional approach is anticipated.

The confounding of the efficiency measure with the timing or size of the purchase decision is probably the main reason that free-ridership algorithms that appear to be conditional are suspect, but the question is left as to whether such powerful and theoretical approaches are the most interpretable and transparent alternative algorithms. Even if carefully done, are they worth it?

Alternatives

The fact that most evaluations are completed without the avoidable bias or the use of the conditional approach for at least part of the algorithm indicates that there are many ways to approach the combination of multiple measures. Some solutions are simple and are practiced by evaluators as standard operating procedures. Others have innovated quickly when this type of bias is pointed out. Certainly, additive algorithms do not create the same issue. Adding a 0.0 or a 1.0 probability does what it is supposed to – moves a combined score in the direction intended, and adding a zero does not negate all other indicators as does multiplying a zero. Usually additive measures are averaged to normalize the scale between 0.0 and 1.0, however some researchers add raw values and make a "preponderance of evidence" decision about freeridership at a particular value, such as 3.25 out of 6.0 (RLW 2007, 1-5). Perhaps the most frequent approaches that avoid the multiplicative bias are ones that (a) look at averages of probabilities obtained from similarly scaled responses to alternative ways

[3] The timing issue is often the hardest to use in a conditional sequence, because it can be asked in different ways: "how likely is it that you would have purchased the equipment 'at exactly the same time'"..."within a year"…."within two years?" etc. This makes the question and scoring very subjective, and can vary from researcher to researcher.

of getting at the construct; (b) averaging two macro indices, such as a four question sequence on influence of program, with a two question sequence on timing; or (c) weighting the scores based on *a priori* principles when averaging them – e.g. some respondents are more knowledgeable or some questions are considered more reliable. This latter approach is not frequently used, because of the subjectivity involved and the associated fear of criticism, but failure to weight is *de facto* a decision to weight all answers equally.

In terms of partial freeridership, rather than treat the level of freeridership as a decimal, as with SEER 14/SEER15, the researcher could use the self-reported non-program-induced efficiency level as the gross savings baseline for the calculation of savings. This is almost always necessary in complicated large commercial and industrial projects, where efficiency levels can not be distinguished easily in a scaled response category. Or, still yet, use the net savings – 5 fixtures -- as the net impacts and ignore further intentions, since the reported direct result is one alternative estimate of the quantified impact of the program.   Following on the concept of identifying the actual savings that are potentially not attributable to the program – by quantity and by efficiency level, there is an obvious multiplicative approach, which is not the multiplication of probabilities. If only half the measures installed by the program would not have been installed without the program, and they were only going to be half as efficient as the measures actually installed by the program, clearly the gross savings number should only be reduced by ¼. These are actual kWh, however, not probabilities of free-ridership, and are best expressed as an adjusted baseline for the savings.   Situations occur in which similar sounding responses are neither conditional probabilities nor adjustments to gross energy savings, and they are erroneously multiplied together.

For deferred free-ridership, researchers have frequently decided that one way to measure the likelihood of free-ridership would be to ask how far into the future the project would have been undertaken without the program. Although, as in footnote 3 above, the decision process can be very subjective and vary by sector and complexity of measure, most evaluators will assume that at some length of time, the probability of the measure actually being taken absent the program is zero. This determination may be at 1 year or 10 years into the future. This is a reasonable way to uncover uncertainty in the free-ridership, and provide one other gage of the counterfactual intent.

Evaluators have an enormous store of innovation and insight that they do bring and can bring to bear on the difficult issues of self-reported freeridership.

How Wide-Spread is the Problem?

The need for this paper was identified when the author, in reviewing dozens of evaluations every year, began to push back at some evaluators who were inappropriately using multiplicative algorithms. In a year, five drafts were returned to the authors to deal with this problem. Since it was a bias, it seemed important to point it out, and write this paper to help document why it should be avoided in the future. As noted above, depending on how frequently the algorithm actually was used within a sample, the impact of the bias could be quite small – 1 or 2 % of the estimated free-ridership in a field in which we can't claim such precision. Other times the changes were more in the range of 10%.

The question remains about how frequent is this bias among the evaluations produced over the years? This would be a vast undertaking that would serve little purpose – at least not as much as making sure that multiplicative algorithms are only used in the narrowly applicable circumstances. There are 20 years of evaluations in New England, 10 years or more in NYSERDA, Wisconsin, and the Pacific Northwest that would need to be examined. As an indicator, this author examined about 100 impact evaluations filed in CA since January 2003, where the evaluations are archived in a searchable database of the California Measurement Advisory Committee (CALMAC[4]). The result of the re-examination were reassuring on the issue of multiplicative bias, but not necessarily re-assuring to those who may think that free-ridership has been studied intensively in CA and that the results are well-established.

Three samples were selected: a systematic random sample of all evaluations published between 1/1/03 and 12/31/04 (sample = 17); a review of the universe of energy efficiency impact reports for 1/1/05 through 12/31/06 (73 reasonably valid matches to the search criteria); and all of those published from 1/1/07 through 6/30/07 (9). Surprisingly only 35 of the 99 impact evaluations actually made any effort to measure free-ridership, but of these only three had issues with the multiplicative bias. This is not evidence of a widespread problem. (That is not to say that all of the others were above criticism, but that is another research paper by itself.)

So what about the remaining evaluations without explicit free-ridership efforts? Thirty-four merely "deemed" the free-ridership as they were allowed by CPUC policy at the time. Another nine evaluations seemed to require a net savings assessment, but didn't have one. Three authors said the research included an estimate of free-ridership, but the methods and algorithms were not provided. The lack of efforts to measure free-ridership undermines the assumption that the field is constantly upgrading our knowledge and keeping up with changing markets. The remaining 17 studies were of non-resource programs, or other programs that did not require a net savings estimate – many of these could have been mis-characterized by the searchable database.

Conclusions

Bias can be inadvertent. Bias can be small in absolute terms, as when there are only a few decimals being multiplied among many respondents, e.g., when only relatively rare "don't knows" are given score other than a zero or 1.0. Bias can be small relative to the measurement and sampling error. But as long as the bias is recognized and is avoidable, evaluators have an obligation to take action to avoid it. The purpose of this paper is not to explain the obvious. It is not to blame evaluators who may be oblivious to the intrinsic bias in the algorithms they may have used. The purpose is to try to ensure that new evaluators are forewarned and forearmed, and to stimulate all evaluators to examine the implications of the details of their work.

The making of freeridership soup is a messy process. It is appropriate for evaluators to explain to users of their research that the results must always be taken with a grain of salt. It is not acceptable to inadvertently salt the soup in the making.

References

Goldberg, Miriam, 2009. Personal communication, April.

McRae, Marjorie R. 2002. "Sure You Do. Uh-Huh": Improving the Accuracy of Self-Reported Efficiency Actions," *Proceedings of the American Council for an Energy Efficient Economy Summer Study on Efficiency in Buildings*. Asilomar, Calif. 10.189-96.

Peters, Jane S. and Marjorie R. McRae 2008. "Free-Ridership Measurement Is Out of Sync with Program Logic…or, We've Got the Structure Built, but What's Its Foundation?" *Proceedings of the American Council for Energy Efficient Economy Summer Study on Efficiency in Buildings*. Asilomar, Calif. 5.219-34.

[4] <www.calmac.org>

Ridge, Richard, Willems, Phillipus, Fagan,Jennifer, and Randazzo, Katherine. 2009. "The Origins of the Misunderstood and Occasionally Maligned Self-Report Approach to Estimating the Net-to-Gross Ratio." *Proceedings of the International Energy Program Evaluation Conference,* Portland, Oregon.

RLW Analytics 2007. *An Evaluation of the 2004-2005 Savings By Design Program – Appendix.* Report # SCE0221.02, San Francisco, California.;

TecMarket Works 2004. *The California Evaluation Framework.* Oregon, Wisconsin.

TecMarket Works 2006. *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals*. Oregon, Wisconsin.

.