Energy Efficiency (2024) 17:27
https://doi.org/10.1007/s12053-024-10207-y

ORIGINAL ARTICLE

The inscrutable baseline and the problem of attribution

Stephen P. Waite

Received: 16 May 2023 / Accepted: 28 February 2024 / Published online: 20 March 2024

Abstract This paper examines the problem of attribution in the evaluation of energy efficiency program impact. The methodological problem concerns the observability of consumer behavior under the baseline condition of no program intervention. The statistical solution to the problem, which entails randomized exposure of targeted individuals to program influence, is not a viable alternative in most applications. Randomized opt-in and randomized encouragement designs do not conform to this requirement because all targeted individuals are encouraged to participate in the program, resulting in negative exposure bias. Quasi-experimental methods which utilize non-targeted individuals or targeted nonparticipants as baseline surrogates are further subject to selection bias of unknown magnitude and direction. Valid attribution in the general case of unrestricted eligibility depends on prior knowledge of the determinants of measure adoption and program participation. In default of such knowledge, evaluators must rely upon structural assumptions that have no foundation in empirical science. On the other hand, established measurement and verification methods which exploit scientific knowledge of the determinants of end-use energy consumption should be utilized to obtain unbiased estimates of individual measure and gross program energy savings.

S. P. Waite (✉)
West Haven, USA
e-mail: spwaite3@comcast.net

Keywords Energy efficiency · Program evaluation · Causal inference · Net savings · Experimental and non-experimental methods

Introduction

The methods employed to quantify energy efficiency program impact can be broadly categorized as either experimental, quasi-experimental, or survey methods. Experimental methods incorporate randomized eligibility into program design. Quasi-experimental methods are based on comparisons of program participants and eligible nonparticipants or comparisons of targeted and non-targeted populations. Survey methods solicit responses from participants, nonparticipants, and trade allies to hypothetical questions regarding market behavior under the counterfactual condition of no program implementation. Whereas experimental methods seek to selectively control program influence on targeted individuals, non-experimental methods are applied to programs in which eligibility to participate within the target population is unrestricted, which are the norm given the policy objective to make program benefits available to all individuals within the market segments addressed by each program.

All programs are designed to promote energy efficiency measure adoption. Promotional activities utilize various communication channels to inform consumers of the benefits of improved end-use efficiency

as well as the availability of program incentives to purchase products that conform to qualifying standards of energy efficiency and performance. A primary policy objective is market transformation, i.e., the reduction of market barriers to adoption via the dissemination of information pertaining to the availability, performance, and economic benefits of high-efficiency alternatives to less efficient energy-consuming equipment. Marketing channels include mass media, utility web sites, retail outlets, contractors, wholesale equipment vendors, utility bill inserts, and direct mail. Marketing and outreach activities are typically designed to achieve maximum exposure to program influence, in order to promote energy savings, participation, and program cost-effectiveness. Indeed many states have adopted the requirement that program portfolios be designed and implemented to achieve all cost-effective potential savings to maximize the net benefits to consumers.

Energy efficiency program impact is defined in terms of the difference between the energy consumption (net savings) and measure adoption (net adoption) of targeted individuals under the mutually exclusive conditions of program implementation and non-implementation (Violette & Rathbun, 2017). Because impact evaluations are retrospective, outcomes under the former (program) condition are observable whereas under the latter condition, termed the "baseline" by convention, they are not. In his seminal paper on causal inference, Holland (1986) referred to the impossibility of observing both program and baseline behavior of the same individual as "the fundamental problem of causal inference." The focus of Holland's discussion is on the "statistical solution" to this problem formulated in terms of the potential outcomes model of causal inference developed by Donald Rubin.

Rubin's model provides a formal treatment of the problem of attribution which is generally applicable to empirical studies that seek to quantify the causal effect of an intervention on a population of targeted individuals and consequently has been widely adopted by investigators and evaluators working in diverse disciplines including statistics, psychology, education, sociology, political science, economics, epidemiology, and clinical research (Rubin, 1990; Winship & Morgan, 1999; Greenland & Robins, 2009; Sobel, 2009; Stuart, 2010; Yamamoto, 2012; Imbens & Rubin, 2015; Abadie & Cattaneo, 2018). The generic formulation of causal effects in terms of potential outcomes allows for a coherent and transparent framework of analysis that is directly applicable to the diverse array of methods and programs which comprise the main body of current practice of impact evaluation. Nevertheless, it has received scant attention among program evaluation practitioners in the energy efficiency field of application, suggesting that insights into the problem of attribution and the transparency of model assumptions afforded by this framework have not been fully internalized in current practice.

Since the inception of energy efficiency programs in the late 1970s, program evaluators and utility regulators in the USA have struggled to achieve consensus on a standard of practice to establish confidence in the inference of attribution from impact evaluation findings. In their review of emerging issues in the evaluation of energy-efficiency programs, Vine et al. (2012) observe that, after decades of program evaluation experience, a number of issues are still unresolved and, in certain cases, "are highly contested," noting that the authors themselves are not in agreement on the resolution of some of the issues discussed in the paper. The review identifies net savings as one of the most important issues that are yet to be resolved, both in terms of its technical definition and problems with estimation, noting the implications of jurisdictional inconsistencies in definition for the evaluation methods employed and the resulting values derived therefrom.

The American Council for an Energy-Efficient Economy (ACEEE) conducted a national survey of current practice and expert opinion on this subject. The survey found that there is remarkable variation among states in the approach taken to the net savings issue. The authors were particularly struck by the divergence of opinion among the surveyed evaluation experts: "Even among evaluation professionals, there is no consensus on whether net savings is the metric that should be used, much less on what specific methodologies should be utilized to determine net savings" (Kushler et al., 2014, p. 19). The lack of consensus on methodology is underscored by the introductory comments to Chapter 21 of the Department of Energy/National Renewable Energy Laboratory (DOE/NREL) Uniform Methods Project (UMP), which is devoted to the estimation of net savings:

> The chapter provides a definition of net savings, which remains an unsettled topic both within the EE evaluation community and across the broader public policy evaluation community, particularly in the context of attribution of savings to a program. The chapter differs from the measure-specific Uniform Methods Project (UMP) chapters in both its approach and work product. Unlike other UMP resources that provide recommended protocols for determining gross energy savings, this chapter describes and compares the current industry practices for determining net energy savings but does not prescribe methods. (Violette & Rathbun, 2017, p. 1)

The restriction of the UMP net savings guidelines to a description of current practice and the deliberate exclusion of specific recommendations concerning methodological approach mirrors the reluctance of the expert respondents to the ACEEE survey to characterize any of the methods employed in current practice as superior to any others, finding them all to be acceptable alternatives and commenting further that "best practice would be to use multiple methods and triangulation to develop defensible estimates" (Kushler et al., 2014, p. 18). Regulators, for their part, must rely on evaluated savings to assess program performance and cost-effectiveness, but, in default of a working consensus among evaluators on a normative methodological standard for valid attribution, they are placed at a distinct disadvantage in their review and interpretation of reported evaluation findings because there are no established criteria to assess the credibility of the results.

This paper presents a critical examination of the methods utilized by evaluators of utility ratepayer-funded energy efficiency programs to quantify the net impact of program implementation. The substance of this critique is an analysis of the sources of bias that undermine a valid inference of attribution and the credibility of the implicit assumptions required to sustain such an inference. The analysis is structured in terms of policy-relevant impact parameters defined by program and baseline potential outcomes and the processes employed by different methods to generate observations on the potential outcomes. The objective of this approach is to translate the general conditions for valid attribution formalized by Rubin and others into context-specific assumptions that energy efficiency program evaluators must make about the market behavior of targeted individuals in order to justify their interpretation of reported estimates of program impact. The articulation of these assumptions in terms familiar to energy efficiency program evaluators is intended to clarify the summative and formative interpretations which may or may not be warranted by study findings.

This formulation yields new insights into the comparative validity of alternative methods. The analysis reveals that accepted methods, as indicated by common practice or EM&V guidance documents, exhibit categorical differences when seen through the lens of the potential outcomes model. Foremost is the distinction between experimental and non-experimental methods, which depend on radically different assumptions to support valid attribution. There are, moreover, methods that may be categorized as experimental, because they employ some form of randomized selection, which cannot realistically be characterized as equivalent or even close substitutes for methods that selectively control program eligibility, as discussed below. These core methodological differences directly contradict the view that, since all methods are subject to some form of bias, the choice of a particular alternative is reducible to practical considerations such as applicability to the type of program being evaluated, data availability, evaluation cost, or other resource requirements. On the contrary, methods which rest on fundamentally different identifying assumptions or different baseline surrogates are not "complementary." They are intrinsically inconsistent because they are not quantifying equivalent parameters of the target population. The implications of the findings of the analysis for EM&V practice and policy are explored in the discussion. The discussion concludes with some thoughts on changes to current practice that could be undertaken by program evaluators to address some of the issues raised in this critique.

The statistical solution

The two essential elements in Rubin's exposition are (1) definition of the causal estimand (impact parameter) as a comparison of the individual program and baseline potential outcomes, termed the "unit-level

causal effects," or the population average of the unit-level effects, termed the "summary causal effect," and (2) a "posited assignment mechanism" which specifies a probability model of the selection of targeted individuals for exposure or non-exposure to the program intervention (Rubin, 2005). Rubin (1990, 2005) credits the statistician Jerzy Neyman with the origination of the potential outcomes formulation for application to the design of agricultural field experiments. In addition to introducing the potential outcomes notation, Neyman (1923) proved that the difference between the average of the observed outcomes of individuals randomly selected for exposure to an experimental treatment and the average of the observed outcomes of individuals randomly selected for non-exposure to the treatment is an unbiased estimator of the population average of the unit-level causal effects, i.e., the average difference between the program and baseline potential outcomes for every individual, only one of which is observable. This finding establishes the formal validity of causal inference in randomized experiments. The designation "Neyman-Rubin Model" (NRM) employed in this paper follows Pearl (1996).

In Neyman's formulation, the statistical solution to the problem of attribution requires that a subset of the target population of N energy consumers be randomly selected for exposure to the influence of program implementation, a design known as a completely randomized experiment. The subsets of n exposed and $N–n$ non-exposed individuals thus represent random samples selected from the target population. Under this design, the average of the observed outcomes (e.g., measure adoption, energy consumption) for each subset is an unbiased estimator of the corresponding target population parameter and the difference between the average outcomes of the exposed and non-exposed subsets is an unbiased estimator of net impact (adoption/savings), which, in Rubin's parlance, is the summary causal effect.

Randomized *eligibility* is intended to selectively control *exposure* to program influence. Successful implementation of this program design requires that targeted individuals excluded from eligibility to participate be effectively insulated from all program influence, including marketing content pertaining to the benefits of measure adoption, in order to simulate the baseline condition of non-implementation. The program is thus designed to be a controlled experiment that yields an unbiased estimate of program impact, calculated as the difference between the average outcomes of the groups of eligible and ineligible individuals randomly selected from the target population. Effective exclusion from program influence is a prerequisite for the validity of any program design which purports to establish an observable condition that can serve as a valid surrogate for the unobservable baseline condition of non-implementation. However, selective exposure requires restricted program marketing as well as restricted eligibility because, for an unknown segment of the population, the threshold of measure adoption may be purely informational, not transactional or financial.

The feasibility of controlled exposure is further complicated by the effects of market transformation — in economic terms, "general equilibrium" effects — on market adoption of energy efficiency measures. Even when consumers can be selectively insulated from the direct influence of program promotional activities, depending on the scope and scale of program implementation, they will be subject to the potential impact of program-induced changes in the production, availability, and price of efficient alternative products and services.

Failure to control program influence on targeted individuals whose eligibility is restricted results in *exposure bias*. If it is assumed that program implementation has the potential to induce some ineligible individuals to adopt energy efficiency measures, then the estimates of baseline energy consumption and net energy savings will be negatively biased. While elimination of exposure bias is a necessary condition for valid attributional inference, it is not sufficient unless restriction of eligibility is *randomize*d. Even though the ineligible individuals are not subject to program influence, their observed energy consumption may not be representative of the baseline consumption of the target population. This is the problem of *selection bias*. Thus, experimental methods which effectively randomize exposure to program influence, via randomized selection for eligibility, can eliminate both sources of bias and yield valid estimates of attributable program savings.

In energy efficiency applications, the prototypical example of randomized eligibility is the randomized opt-out (ROO) design which entails the random selection of targeted individuals ("treatment group") to receive home energy reports containing site-specific

information, including historical energy consumption data and recommended actions to reduce consumption. In this design, selected individuals are sent monthly, bimonthly, or quarterly reports, unless they inform the program administrator that they have elected to opt out of the program. The treatment group is thus comprised of all individuals who are selected, informed about the program, and received an initial report, regardless of their decision to participate or not to participate, i.e., to opt out. Under the assumption of no exposure bias, that is, no influence of the intervention on individuals who are not selected to receive the reports ("control group"), the difference between the measured average energy consumption of the control and treatment groups yields an unbiased estimate of average net savings. The net savings estimand is the difference between the average potential response of targeted individuals if all were sent the initial mailing of home energy reports and the average potential response if no targeted individuals were sent reports.

Net savings thus accounts for the impact of program implementation on the energy consumption of both program participants and nonparticipants. Depending on the program design, participants are defined as eligible individuals who either voluntarily opt in to receive excludable program benefits or voluntarily decline to opt out. In certain programs, e.g., "upstream" interventions, participants may be unaware that they are the recipients of program benefits, so there is no conscious decision to participate or not to participate. In any case, the defining characteristic of net impact, whether the potential outcome is energy consumption or measure adoption, is the comparison of the program and baseline potential outcomes of all targeted individuals. The population average impact is directly scalable to the population total impact via multiplication by the size of the target population. Program net savings is implicitly a weighted average of participant and nonparticipant net savings, so its value reflects the rate of participation, differences between the participant and nonparticipant rates of measure adoption attributable to the intervention as well as the respective differences in the average measure savings per targeted participant or nonparticipant.

The net impact of randomized eligibility is sometimes referred to as an "intention-to-treat" (ITT) parameter (Imbens & Rubin, 2015), because it captures the average causal effect of treatment "assignment" on potential outcomes rather than the average effect on individuals who "received" the treatment. In the context of energy efficiency program evaluation, this terminology is ambiguous, if not misleading, because the intervention under evaluation is program implementation. The primary focus is the impact on all individuals who are eligible to participate, which includes the effectiveness of marketing efforts to make individuals aware of the program and to inform them of the potential benefits of measure adoption and the incremental excludable benefits of program participation. The scope of potential influence is not restricted to program participants because targeted individuals may be induced to adopt measures but decline to participate, and program effectiveness depends critically on the overall rate of net adoption by all individuals exposed to the influence of program implementation. The ROO program design provides a good illustration of the fact that program participation is never under the direct control of the program administrator, but is rather a potential outcome of the intervention that may not be realized by assignment to receive the home energy report. Stewart and Todd (2017, p. 9) make the point: "For example, some households may opt out of an energy reports program, or they may fail to notice or simply ignore the energy reports. Thus, the effect is ITT, and the evaluator should base the results on the initial assignment of subjects to the treatment group, whether or not subjects actually complied with the treatment."

Allowing for the possibility of program impact on nonparticipants rules out the separate identification of participant and nonparticipant net savings and net adoption impact parameters, meaning that the observable data cannot differentiate between alternative values of the impact on the respective subpopulations of the target population. Of course, under the assumption of no program influence on nonparticipants, the net savings and net adoption ITT parameters are exclusively attributable to measure adoption by program participants, in which case the average net savings of participants is identified by the ratio of the population average net savings and rate of participation, sometimes referred to as the average treatment effect on the treated (ATT). While the estimates of the population average net savings and participation rate

parameters are unbiased, the ratio of the two unbiased estimates is consistent but biased.[1]

However, most programs are not designed to limit eligibility within the target population; every targeted individual decides whether or not to participate. The consequence of unrestricted eligibility is exposure of the entire target population to the influence of program marketing and promotional activities that would not occur under the defined baseline condition of no intervention; consequently, *there are no source data to reveal the baseline behavior of targeted individuals*. Quasi-experimental methods that rely upon the market behavior of eligible nonparticipants to serve as a valid baseline surrogate are fundamentally flawed because valid attribution requires the unrealistic assumption that the program has had no influence on nonparticipant energy consumption decisions, i.e., that there is no nonparticipant spillover. Furthermore, even under the assumption of no exposure bias due to nonparticipant spillover, there remains the problem of selection bias, because nonparticipants and participants are not randomly selected from the target population. As noted above, some quasi-experimental methods, e.g., cross-sectional market sales analysis of upstream programs, utilize a non-target population as a surrogate for the target population under the baseline condition of program non-implementation. For this method, the ideal surrogate population consists of individuals who are representative of the target population but are not subject to the influence of the evaluated program or a similar program that would bias the estimate of baseline energy consumption.

It is important to note that the definition of experimental methods stated above is limited to programs that randomize eligibility. There are two program designs typically classified as experimental which do not conform to this definition (Stewart & Todd, 2017; Violette & Rathbun, 2017). In randomized opt-in (ROI) designs, also known as randomized recruit deny/delay, the program is marketed to all targeted individuals. A subset of the pool of intended participants, i.e., those who opt in to the program, is then randomly disqualified from participation. Those who are disqualified serve as the baseline surrogate for all eligible individuals who intended to participate. Every targeted individual, including the baseline surrogate group, is therefore exposed to program influence, resulting in exposure bias.

Failure to recognize the distinction between randomized exposure and randomized denial or deferral of eligibility to a subset of all targeted individuals who have been recruited to participate creates the misconception that randomization automatically confers internal validity on program impact estimates, which is not the case. In their discussion of residential behavior-based (BB) programs, Stewart and Todd (2017) assert, without qualification, that the ROI program design produces an unbiased estimate of the average net energy savings of targeted individuals who opt in to the program, but that the estimated program impact lacks external validity because it does not apply to targeted individuals who do not opt in. What randomization does in this program design is eliminate selection bias from the comparison of the participants and the customers who wanted to participate but were randomly disqualified. However, unless the consumption behavior of all individuals in the latter group is assumed to be uninfluenced by the program marketing and recruitment process, the estimate of net savings of the opt-in subpopulation is subject to exposure bias. The validity of inference cannot be correctly characterized in terms of a limitation on the scope of application of an unbiased estimate of program impact, which is to say that it is only a matter of external validity; internal validity is impaired as well. Moreover, the negative exposure bias generated by the effect of nonparticipant spillover on the baseline surrogate group is compounded by the failure to account for the net adoption and energy savings attributable to the potential impact of program implementation on *all* targeted individuals who did not participate, i.e., those who did not opt in as well as those who opted in and were subsequently denied eligibility.

Stewart and Todd (2017) recommend a different approach in applications where selective exclusion of targeted individuals from participation is unacceptable. In randomized encouragement designs (RED), evaluators randomly select targeted individuals to receive supplemental encouragement to participate, in addition to the normal promotional and recruitment program process. The potential outcomes and

[1] This ratio is an instrumental variable (IV) estimator of the ATT. The critical distinction between the IV estimands under randomized eligibility and randomized encouragement designs is discussed below.

corresponding impact parameters are accordingly defined by the presence or absence of encouragement to participate, as opposed to the presence or absence of the program as designed. The ITT estimate of net savings is the average consumption difference between the encouraged and non-encouraged individuals. In this approach, the implicit baseline is the condition of program implementation as designed and the intervention is supplemental encouragement. The RED method exemplifies the problem of exposure bias in the extreme because the net impact of program implementation is embedded in the estimate of baseline energy consumption. The RED baseline parameter thus understates the true value of non-implementation baseline consumption by the amount of program net savings and the RED intervention parameter understates the potential program consumption as designed by the incremental amount of net savings attributed to supplemental encouragement.

The nature of the randomized intervention fundamentally changes the definition of the baseline and program conditions and hence the definition of the impact parameter. While the RED impact parameter may have formative policy relevance as an experimental exploration of the efficacy of a specific enhanced recruitment design element, it cannot serve as a valid metric to quantify the net savings of all eligible individuals under the condition of program implementation in the absence of supplemental encouragement, which is the summative objective of program impact evaluation. When multiplied by the number of individuals in the target population, it is equal to the total savings of individuals who were induced to participate and adopt measures by the supplemental encouragement, which is *additional* to the total savings attributable to the program intervention. These net participants are commonly referred to as "compliers," because they represent the subset of participants with encouragement who would not have participated without encouragement, whereas the remaining participants who would participate independently of encouragement are termed "always takers." The targeted individuals who do not participate with encouragement are referred to as "never takers." Compliers are induced to participate by the encouragement intervention whereas always takers are induced to participate by the program intervention.

The incremental nature of an encouragement intervention is clearly illustrated in the situation where the source of encouragement is a behavior-based (BB) program that encourages program participants to enroll in an existing program that offers different benefits such as equipment rebates or other financial incentives to purchase and install qualified energy efficiency measures. Stewart and Todd (2017, p. 36) provide recommendations regarding the quantification of the BB program savings attributable to "program uplift," which is the term commonly used for the additional participation generated by the encouragement intervention which would not have occurred without implementation of the BB program. As the authors state, quantification of this component of BB program savings is important because (1) it is "an important effect of BB programs and a potential additional source of program energy savings" and (2) in order to avoid double counting the savings from uplift, it is necessary that the amount of savings be subtracted from the evaluated energy efficiency program savings. This recommendation clearly differentiates between the savings attributable to the uplift generated by BB program encouragement to participate and the evaluated savings attributable to the existing program in the absence of encouragement. The BB program thus functions in part as an RED intervention to increase uptake in another program; however, the savings from uplift conveys no information about the impact of the existing program on energy consumption.

In this context, it is clear that the estimated net savings from program uplift are directly attributable to the BB RED intervention and are not comparable to the estimated net savings attributable to the incentive program intervention because, by definition, without supplemental encouragement, the compliers would not have participated in the incentive program and the always takers would have participated. This obvious distinction illustrates the problem of latent heterogeneity of program impact within a target population. The program with and without RED interventions represents mutually exclusive implementation scenarios. The success of any particular encouragement intervention depends on the extent to which the inducements devised by the evaluator address the diverse barriers to participation confronting targeted individuals within the program nonparticipant population. Every different encouragement intervention will selectively attract different segments of this population with correspondingly different rates

of baseline measure adoption and potential measure savings. Under randomized assignment, comparisons of different interventions relative to each other and to the existing program can accordingly yield valuable insight into the potential impact of program enhancements, but such studies cannot realistically serve as a sound empirical basis for valid attribution of existing program impact.

Stewart and Todd (2017) are careful to point out that the RED ITT parameter quantifies the savings attributable to the encouragement, not to the program intervention. The authors then summarize the technical approach to estimation of an alternative impact parameter, the Local Average Treatment Effect (LATE), formulated by Angrist et al. (1996). Angrist et al. (1996) first define the unit-level ITT effect of the encouragement intervention in terms of the difference between the potential outcome of interest — in this context energy consumption — of each targeted individual under the encouragement and no encouragement conditions. Given certain assumptions, the individual ITT effect is shown to be equal to the product of a binary indicator of the individual effect of encouragement on participation and the individual effect of participation on energy consumption, i.e., the difference between an individual's potential energy consumption under the two participation conditions. Therefore, under the maintained assumptions, the target population average causal effect of encouragement on energy consumption is equal to the product of the population rate of net participation, i.e., the proportion of compliers, and the average net energy savings of the complier subpopulation. The characterization as "net" savings is essential because, depending on program design, some compliers may be baseline measure adopters, in which case participation has no effect on energy consumption.

The average net energy savings of compliers is termed the LATE by Angrist et al. (1996), who show that the LATE is equal to the ratio of the RED ITT net savings parameter to the population rate of net participation, which can also be characterized as an ITT parameter that quantifies the impact of encouragement on program participation. Angrist et al. (1996) demonstrate that under randomized encouragement, the ratio of the unbiased estimators for the two ITT estimands is equal to the standard instrumental variables (IV) estimator for binary instruments. The formulation of the LATE was motivated by the practical objective to establish conditions under which the instrumental variable (IV) estimand warrants a causal interpretation. However, this interpretation of the LATE is not valid if the encouragement intervention has a direct influence on energy consumption apart from the indirect effect via induced participation. This critical assumption is termed an "exclusion restriction." Stewart and Todd (2017) accordingly advise against the use of encouragement materials that could affect energy consumption, the feasibility of which is certainly open to question. Another critical assumption, termed "monotonicity," requires that the intervention has a non-negative effect on participation; in other words, it rules out the possibility that the intervention may discourage participation.[2]

In their discussion of the formalization of the IV estimand and its interpretation as a causal parameter, Angrist et al. (1996) emphasize the point that the impact on compliers is the only estimable causal effect of the intervention because under the RED design, the data are not informative about non-compliers, for whom only one potential participation outcome is observed. They further comment that the only way to identify the impact on non-compliers is to make the implausible assumption that the average net savings of those subpopulations are equal to the net savings of compliers. This point is crucial to a valid interpretation of the LATE parameter under alternative program designs.[3] As noted by Angrist et al. (1996), given the exclusion restriction and monotonicity assumption, under randomized *eligibility*, the LATE is equal to the ATT, defined above, because in that context, there are no always takers and the "compliers" consist of all targeted individuals who participate when eligible in the absence of supplemental encouragement. However, in the RED context, these same individuals are defined as always takers whose net savings count nothing toward the RED impact on

[2] The monotonicity assumption should not be taken lightly. Discouragement of measure adoption can be an explicit objective of program design in order to reduce implementation of measures which will in certain applications increase the total cost of end-use service.

[3] In this context, the RED design is characterized as "two-sided noncompliance" to distinguish it from the "one-sided noncompliance" randomized eligibility design. See Imbens and Rubin (2015) for a thorough discussion of both designs and Freedman (2006), who uses the respective characterizations of "double crossover" and "single crossover" designs.

participants. So, as discussed above, the true value of average net savings attributable to participation in the program as designed has no relation to the value of the RED LATE which, when imputed to program participants in the evaluated program, reflects the implicit assumption that the causal effect on always takers is the same as that for compliers, who in the absence of encouragement contribute nothing to the ATT.

Consider an idealized upstream program designed to randomize a discounted price of high-efficiency equipment among all targeted individuals who are purchasing a new central air conditioner. HVAC contractors and equipment vendors agree to randomly select purchasers to be offered either the current price or a discounted price for the more efficient product, while maintaining the current price for standard efficiency products. Suppose further that the purchasers who must pay the undiscounted price have no knowledge of the discount being offered to others. Selection bias and exposure bias are thus eliminated, assuming as well that there are no market effects. Consequently, the differences between the observed average consumption and market adoption of the purchasers at the current and discounted prices yield unbiased estimates of the corresponding ITT net savings and net adoption impact parameters. The ATT net savings of program participants is identified by the ratio of the ITT net savings and participation rate parameters, assuming monotonicity and recognizing that there is no nonparticipant spillover bias in upstream programs that discount all qualified products. All measure adopters are program participants because they no longer have the option to purchase the efficient product at the undiscounted price.

But suppose that, given the infeasibility of randomized selection for eligibility and selective exposure of targeted individuals to an upstream intervention, all purchasers of the high-efficiency products were charged the discounted price and an impact evaluation was undertaken to randomly provide a subset of purchasers with information designed to encourage the selection of the high-efficiency product. What conclusions could be drawn from the findings of such a study? What is the correct interpretation of the estimated values of the ITT and LATE impact parameters? If the encouragement intervention generates a substantial increase in efficient product sales, then the ratio of the difference between the average energy consumption of non-encouraged and encouraged individuals to the corresponding difference in the proportion of efficient product sales provides an estimate of the average net savings of the complier subpopulation. But what can be inferred about the impact of the program price discount from the estimated impact of encouragement, that is to say, from a comparison of the market response of two separate populations of consumers to two fundamentally different interventions? While the findings indicate the positive effect of encouragement on some consumers for whom the price discount was not sufficient to induce measure adoption, there are no data to support any conclusion about the effectiveness of the discount on the always takers who required no encouragement to purchase the efficient product. Therefore, the data generated by the RED cannot differentiate between alternative values of the average net savings of the targeted individuals who participated in the program being evaluated. Indeed it is quite plausible that randomized selection for the price discount would produce results indicating no impact of the upstream program on the target population. Likewise, a finding of no effect of encouragement could plausibly be consistent with a substantial impact of the program intervention on measure adoption and energy consumption. Clearly, conclusions regarding program effectiveness that are based on a spurious interpretation of the RED causal estimand, especially if they were to inform policy decisions pertaining to program design and implementation, could seriously undermine the realization of energy efficiency policy objectives with obvious adverse economic consequences for the population of energy consumers targeted by the evaluated programs.

In summary, estimates of the net impact of programs that are not designed to randomize exposure to program influence are subject to negative exposure bias and selection bias of unknown magnitude and direction, both of which undermine the internal validity of the estimator. The source of both types of bias is a defective baseline surrogate that is not representative of the target population under the condition of non-implementation. In what follows, I will consider two aspects of the methodological problem. First is the question of whether certain non-experimental methods can in some way overcome or to some extent compensate for the limitations on valid attribution of program impact imposed by program design. The

answer to this question is formulated in terms of the choice between two alternative paradigms that rely on radically different assumptions to justify the proposed solutions. The final section undertakes a close examination of the latent structure of program impact which identifies two separate questions of attribution that entail fundamentally different solutions.

Alternative paradigms

The experimental paradigm

The methods of data analysis developed by Rubin, Rosenbaum, Imbens, Angrist, and others to estimate causal effects are routinely employed in energy efficiency program evaluations to quantify net program impact. In these evaluations, the implications of nonrandom selection for attributional inference are generally ignored or glossed over. This is unfortunate because the implications are quite serious: the results of methods that produce unbiased estimates when applied to experimental data are subject to bias of unknown magnitude and direction in the absence of randomized exposure. Rubin's innovation was the extension of Neyman's experimental model to observational studies, which he variously refers to as "nonexperimental," "nonrandomized," or studies "with unknown assignment mechanisms" (Imbens & Rubin, 2015).

The cardinal principle of the NRM perspective is the essential role of the assignment mechanism: if the investigator cannot "posit" a plausible assignment mechanism that generated the observed data, then causal inference is not possible. The task for the researcher "when trying to estimate causal effects from an observational dataset is to conceptualize the observational dataset as having arisen from a complex randomized experiment, where the rules used to assign the treatment conditions have been lost and must be reconstructed" (Rubin, 2008, p. 815). Of first importance is the identification of "key covariates," i.e., measured background variables that were available for use in the selection process. Given a set of key covariates:

> The next step is to try to find subgroups (subclasses, or matched pairs) of treated and control units such that within a subgroup, the treated and control units appear to be balanced with respect to their distributions of key covariates. That is, within such a subgroup, the treated and control units should look as if they could have been randomly divided (usually not with equal probability) into treatment and control conditions. (Rubin, 2008, p. 817)

In Rubin's conception, adherence to this methodological approach constitutes:

> [an] objective observational study design in the sense that the resultant designed study can be conceptualized as a hypothetical, approximating randomized block (or paired comparison) experiment, whose blocks (or matched pairs) are our balancing groups, and where the probabilities of treatment versus control assignment may vary relatively dramatically across the blocks. (Rubin, 2008, p. 818)

If the posited assignment mechanism approximates the actual process that was employed to select targeted individuals for exposure to the intervention, then the Neyman unbiased estimator can be employed to calculate the average net savings within every subgroup (block or matched pair). Rubin's conception of observational data as having been generated by a randomized block design is formalized in the assumption of "strong ignorability," defined as an assignment mechanism that combines the properties of unconfoundedness and positivity (Rosenbaum & Rubin, 1983):

$$\Pr(PI|X, EC(1), EC(0)) = \Pr(PI|X)$$

$$1 > \Pr(PI|X) > 0 \tag{1}$$

In the NRM formulation, the energy consumption variables $EC(0)$ and $EC(1)$ represent the potential outcomes for each individual i in the target population consisting of N individuals under the respective states ($PI=0$, $PI=1$), only one of which is observable depending on the alternative conditions of program implementation or non-implementation. The binary selection variable PI indicates selection for exposure or non-exposure of each targeted individual to the program intervention. Strong ignorability sets the unconfoundedness condition of *conditional independence* of selection and potential outcomes, given the observed values of key covariates X, and the additional condition that the probability of

selection is positive for both selection conditions. The conditional probability of selection $\Pr(PI_i|X_i)$ is known as the "propensity score."

Under the NRM assumption of a strongly ignorable assignment mechanism, i.e., a hypothetical randomized block design, selection bias can be eliminated or mitigated by conditioning via stratification, matching, or parametric regression on the key covariates, which adjust for variation among targeted individuals in the propensity score, $\Pr(PI_i|X)$. The randomized design balances, in expectation, the distribution of *unobserved* as well as observed covariates within each stratum. The paradigm is "experimental" because it is based on the assumption that individuals within each target population stratum, defined by measured key covariates, were randomly selected to participate or for eligibility to participate. This is simply an extension of the completely randomized program design to one in which selection is randomized within each population stratum, that is, the probability of selection is uniform within each stratum but can vary among strata.

The NRM addresses exposure bias via the "consistency assumption":

$$EC_i = EC_i(1)PI_i + EC_i(0)(1 - PI_i) \quad (2)$$

This assumption implies that knowledge of the values of the observed outcome and exposure of each individual are sufficient to empirically determine the corresponding value of the potential outcome. Rubin's term for this condition is the "stable unit treatment value assumption" (SUTVA). Consistency implies no exposure bias.

Most energy efficiency programs are not conformable to the experimental paradigm because the assumption of random selection is not credible; the selection variable used in the evaluation is either voluntary program participation or program eligibility. In the latter approach, observations of non-targeted consumers are utilized to represent the unobserved baseline condition for the target population. An example is the method of market sales analysis, in which market adoption data are collected for a baseline surrogate population of consumers in a separate geographical area outside of the target population area. There is no sense in which the participant and nonparticipant subpopulations or eligible target and ineligible surrogate populations can be conceptualized as the outcome of a random selection process.

In their discussion of statistical inference for non-random samples, Copas and Li (1997, p. 55) cite the seminal contribution of R.A. Fisher regarding the critical role of randomization "as the logical underpinning of methods of analysis" in experimental research, noting further the logical flaw in the assumption that the same underpinnings of valid inference are sustained by application of these methods in the non-experimental context:

> However, methods designed for analysing experimental data are also routinely applied to observational data, sometimes (often?) with little or no recognition of the fact that the absence of randomization has, in Fisher's sense, removed the grounds for the validity of these methods. Essentially, randomization becomes a model for the data rather than a factual statement of how the data were obtained.
> Modern statistics places great emphasis on the testing of assumptions. But the argument that randomization underpins the standard model assumptions is not reversible — the empirical verification of these assumptions does not imply that the hidden assumption of randomization is necessarily justified so that standard inference statements can safely be made.

Imbens (2010, p. 407) elaborates on this confusion of fact and assumption which glosses over the categorical distinction between experimental and quasi-experimental methods. To contentions that randomized methods do not merit special priority over non-randomized methods, because all such methods rely on the validity of certain assumptions to justify causal conclusions, Imbens responds that what sets randomized experiments apart "is not the *assumption* of randomization but the actual *act* of randomization that allows for precise quantifications of uncertainty, and this is what gives randomization a unique status among study designs."

The structural paradigm

The statistical solution is not a viable option in the absence of random selection. Consequently, program evaluators must invoke an alternative "structural"

paradigm to justify the inference of attribution or accept the possibility that such inference is not possible in most energy efficiency program applications. In the alternative paradigm, elimination of selection bias that arises in the comparison of targeted and non-targeted individuals, or targeted participants and nonparticipants, requires adjustment for population differences in the *determinants* of baseline consumption. Put differently, what is needed in the structural paradigm is an accurate model of the baseline potential outcomes as opposed to a model of the assignment mechanism. The *rationale* for covariate adjustment marks a fundamental point of divergence between the structural and experimental paradigms. In the experimental paradigm, the "key covariates" are the set of observed variables used to determine the target population strata employed in the posited assignment mechanism. In the structural paradigm, the "relevant covariates" function instead as posited determinants of the baseline potential outcome. In Rubin's terms, a "model on the science" is now required to identify the causal effect of the intervention (Rubin, 2005).

In most energy efficiency applications, quasi-experimental methods employ a comparison of the metered energy consumption of participant and nonparticipant target subpopulations. In the absence of exposure bias, i.e., nonparticipant spillover, an unbiased estimate of participant net savings requires the additional assumption of conditional mean independence of baseline energy consumption and program participation (*PP*):

$$\left(\overline{EC}(0)|X, PP = 0\right) = \left(\overline{EC}(0)|X, PP = 1\right) \tag{3}$$

Equation (3) states that conditioning on the observed covariates X is sufficient to eliminate selection bias, because the observed average consumption of nonparticipants is equal to the unobserved average baseline consumption of participants who have the same values of X. Therefore, the difference between nonparticipant and participant conditional mean consumption is an unbiased estimate of conditional participant net savings. Under this assumption, an unbiased estimate of unconditional net savings is calculated as the average of the conditional savings estimates weighted by the corresponding population frequencies of the covariates. The estimator is formally equivalent to the unbiased stratified estimator under the NRM assumption of strong ignorability. The critical difference is that the latter is a design-based assumption of random selection whereas the former is a model-based assumption concerning "the science" of baseline energy consumption.

A valid structural model must therefore account for all determinants of baseline energy consumption (or measure adoption) in order to eliminate selection bias; all relevant covariates must accordingly be known and measurable. This is the crux of the problem: it is not possible to condition on unobserved causal factors. There are two reasons for failure to observe the critical determinants of baseline potential outcomes. First, most quasi-experimental methods that purport to control for selection bias do not collect the essential site-specific data that enable conditioning on *known* causal factors. Second, it is not possible to measure *unknown* determinants of potential outcomes. Both of these problems — the data problem and the problem of prior knowledge — present formidable barriers to valid inference.

The data required to properly account for variation in known determinants are typically not collected because the cost of on-site measurement and verification (M&V) of the relevant covariates for both program and baseline surrogate population samples is generally considered to be prohibitive. Nevertheless, the feasibility of the data collection process is demonstrated by standard practice of site verification and monitoring of the installation and operation of program measures installed by a sample of participants, routinely employed to provide accurate estimates of individual measure, and gross program energy savings required to evaluate measure and program cost-effectiveness and program administrator performance and to support program and resource planning. Established standards of data collection and analysis are documented in various M&V protocols and guidelines, including the International Performance Measurement and Verification Protocol (IPMVP), the Uniform Methods Project (UMP), and the US Department of Energy Federal Energy Management Program (FEMP) M&V guidelines. Various elements of established M&V standards are incorporated into Technical Reference Manuals (TRMs) utilized by program administrators as approved by state regulatory authorities.

EM&V budget constraints limit the range of practical options for large study populations to methods which rely upon available data. Impact evaluations of residential retrofit and opt-in behavioral programs

often employ regression analysis or matching methods that condition post-implementation metered consumption of participants and nonparticipants on historical metered consumption and local weather data. Impact evaluations of equipment rebate and upstream incentive programs condition efficient product sales to target and surrogate baseline populations on demographic and product price data.

However, even if the requisite funds were allocated to conduct primary site-specific data collection, evaluators would be left with the insoluble problem of controlling for population differences in unknown determinants of consumer behavior, because there is no scientific basis to establish equality of the average baseline outcomes of distinct populations. Equation (3) is an invalid assumption because it omits unknown relevant covariates that are by definition unobservable. Heckman and Robb (1985) characterized the implicit structural assumption of conditional independence as "selection on the observables" (SOO).[4] Heckman (2005) has argued that SOO is not credible because it is inconsistent with the widely accepted hypothesis that the program participation decision is based in part on individual unobserved expectations pertaining to the benefits of participation, i.e., the potential outcomes. He proposes an alternative approach which allows for selection on unobservable causal factors (SOU) as well as on measured covariates. The proposed solution is to model the selection bias in terms of observed covariates and hypothetical latent (unobserved) variables in the form of a discrete choice model of the individual participation decision. The modeled probability of selection is used to adjust for differences between participant and nonparticipant baseline outcomes. Heckman's method of "control functions" explicitly models the dependence between potential outcomes and the decision to participate. The model specifies the unknown potential outcomes as a function of observed and unobservable determinants of the participation decision which is consistent with economic theory (Heckman, 2010a).

[4] Parallel but technically distinct identifying assumptions are routinely invoked by investigators working in diverse fields of application under various terminology: admissibility, comparability, exchangeability, exogeneity, sufficiency, no omitted variables, no unmeasured confounders, etc.

Heckman (2010b) has written extensively concerning the dichotomy between the structural and "program evaluation" approaches to causal analysis. The explicit formulation of structural assumptions pertaining to consumer decision-making provides a level of methodological transparency that is wanting in impact evaluations that employ standard methods of data analysis that rely on implicit assumptions which, at best, are only vaguely articulated, let alone critically examined. For example, Heckman (2010a) characterizes the conditional independence assumption invoked by matching methods as the result of some undefined natural process of randomization that functions as a surrogate for an actual experimental manipulation.

At the same time, Heckman (2005, p. 65) allows that: "Offsetting these disadvantages, the method of matching … does not require separability of outcome or choice equations, exogeneity of conditioning variables, exclusion restrictions, or adoption of specific functional forms of outcome equations." Nevertheless, Heckman and Navarro-Lozano (2004, p. 30) contend that: "Because the method of control functions explicitly models omitted relevant conditioning variables rather than assuming that there are none, it is more robust to omitted conditioning variables." Heckman (2005, p. 5) also stresses the importance of incorporating unobserved heterogeneity in models of program impact: "Another reason why epidemiological and statistical models are incomplete is that they do not specify the sources of randomness generating the unobservables in the models — i.e., they do not explain why observationally identical people make different choices and have different outcomes given the same choice." Sobel (2005, pp. 121, 122) takes Heckman to task on his assertion that the control functions method is more general than matching (Heckman, 2005, p. 73), observing that the validity of the two approaches rests on alternative sets of untestable assumptions, neither of which is implied by the other. In Sobel's view, Heckman "is far less critical" of the assumptions required to justify the control functions method than those required in the use of matching and instrumental variables methods.

Heckman (2010b, p. 367), as noted above, emphasizes that the structural model specification of observed determinants of potential outcomes and the participation decision must be derived from economic theory because "there is no 'objective'

way to choose these conditioning variables. Any argument for inclusion or exclusion of variables has to be made by an appeal to theory — implicit or explicit." Unfortunately, in this context, economic theory cannot provide any meaningful guidance because there is no body of knowledge that can be drawn upon to favor one set of structural assumptions over another. If critical determinants of baseline consumption are unknown, then the magnitude and direction of the selection bias of estimated net impact are unknown and indeterminate in any non-experimental application. Like SOO methods, SOU methods based on models of program participation and potential outcomes employ untested assumptions about the relevant measured covariates that determine the probability of participation and its relationship to potential outcomes.

The dependence of the validity of attribution on the credibility of assumptions regarding unknown determinants of the behavior of targeted individuals under unobserved conditions raises the question: "What is the threshold of credibility of such assumptions?" Regarding the credibility of SOO, Heckman (2005) asks why some otherwise observationally identical individuals decide to participate in a program while others do not. There must be unobserved factors that account for heterogeneity of response among targeted individuals who are homogeneous in the measured covariates. Equation (3) implies that the unobserved factors that generate individual differences in baseline consumption have no effect on the average rate of participation within each homogeneous (on X) segment of the target population. How does one come to such a conclusion?

Provencher et al., (2013, p. 6) present a different perspective on this question in their discussion of the rationale for matching participants and nonparticipants on historical monthly consumption to estimate the energy savings attributable to an opt-in behavioral-based program. The authors reject the hypothesis "that even though the participants and their matches behave the same on average for 24 months before the start of the program, in the absence of the program their energy use would not continue to be the same on average because unobservable factors cause the development of systematic differences in the energy use between the two groups." Their explanation for this assumption is based on the following rationale:

> Suppose an underlying set of unobservable variables **Z** reflect a household's behavioral propensity to save energy, and these variables are correlated with participation in the program. One can reasonably expect that close matching on the energy use history will, on average, generate the same distribution of **Z** among the matched households as among the participant households.

Why is this supposition a reasonable expectation? With matching on energy use history, the covariates X in Eq. (3) consist of observed monthly metered energy consumption during the pre-implementation period. These covariates are not *determinants* of *baseline* consumption, which is the unobserved energy consumption of program participants during the post-implementation period under the counterfactual condition of non-implementation. They are the observed historical outcomes which, like the baseline outcomes, are complex functions of unobserved physical and behavioral factors, some of which are known and measurable. Balance between participants and nonparticipants in the distribution of historical whole-building energy consumption does not imply balance in the corresponding distributions of the determinants of baseline end-use consumption, i.e., that they "behave the same."

There are two intrinsic difficulties with this methodological approach. First is the problem of measurement boundary: metered consumption confounds the targeted end-use consumption, which is the object of program influence, with the other loads at each site; matching on monthly, or even hourly, billing data cannot identify the baseline consumption value that defines the net savings parameter. Second is the problem noted that matching on energy consumption, even if it were measured at the relevant end use, does not match on the critical determinants of baseline consumption, viz., the magnitude and utilization of equipment capacity that vary according to individual differences in the demand for end-use services and the end-use efficiency of the targeted equipment. Moreover, the comparability of matched participants and nonparticipants is further compromised by the lack of data necessary to cull from the latter sample individuals who are not eligible to participate for various reasons, such as the end-use technology or fuel type, building type or size, etc.

Agnew and Goldberg (2017) recommend an alternative approach designed to mitigate selection bias in estimates of energy savings attributable to residential whole-building retrofit programs. The approach entails the use of a comparison group of targeted customers who participate in a time period following or preceding the implementation year under evaluation. The critical assumptions supporting this recommendation are that future and past participants: (1) "are similar to the participants being evaluated with respect to energy consumption characteristics" and (2) "are unlikely to install the program measures on their own during their non-participating years." Under these assumptions, the authors conclude that the savings estimate is properly interpreted as gross savings and that the estimate is likely to be "less biased with respect to self-selection." The implied comparison is: less biased than estimates derived from a matched comparison group of nonparticipants.

The recommended method of estimation is a mean difference-of-differences (DiD) regression analysis. The authors stipulate that the validity of this estimate depends on a critical assumption, that the determinants of the baseline consumption time trends are the same on average for the participant and comparison groups. This condition expresses the "parallel trends" assumption of DiD analysis, which is undermined by the strong possibility that the timing of the decision to participate is related to unobserved differences in customer or physical site characteristics. For example, eligible customers who anticipate a near-term increase in the demand for end-use services due, for example, to a change in occupancy, have a stronger incentive to participate than customers who do not face an impending increase in consumption and energy costs, in which case the observed pre-post difference in energy consumption of future participants will underestimate the current participant baseline trend and produce negative bias in the savings estimate.

The SOU approach takes selection bias as a foregone conclusion; there is no assumption of balance in the participant and nonparticipant distributions, conditional on observed covariates, of unobserved determinants of the potential outcomes. Rather, the assumption of mean independence conditional on observed covariates in Eq. (3) is replaced with an "all causes" model of the dependence between unobserved determinants of the participation decision and unobserved determinants of the potential outcomes, conditional on observed covariates. Heckman (2005) articulates the basic rationale: "Knowledge of the relationship between choices and counterfactuals suggests appropriate methods for solving selection problems. By analyzing the relationship of the unobservables in the outcome equation, and the unobservables in the treatment choice equation, the analyst can use a priori theory to devise appropriate estimators to identify causal effects."

The thrust of Heckman's argument is that SOO methods are based on implicit assumptions which, when clearly articulated, are not plausible. The hypothesis of selection bias carries a presumptive validity that cannot be assumed away, whereas SOU methods are based on assumptions about the sources of the bias that can yield bias-corrected estimators of program impact. In other words, evaluators should not make inferences of attribution that rely upon assumptions that most, if not all, evaluators believe to be false. The structural modeler's predicament, on the other hand, is not a question of the plausibility of the maintained assumptions so much as the availability of an unlimited set of equally plausible alternative models from which to choose. The posited underlying determinants of measure adoption and program participation represent one out of many untested, but intuitively plausible, alternative hypotheses that can be invoked to fill the void in our scientific understanding of the consumer decision process. The less we know, the greater the scope of plausibility, which undermines the credibility of any one of the hypothetical alternatives as a valid basis for attribution. As Heckman himself acknowledges, the distinction between SOO and SOU methods is not reducible to the presence or absence of the assumption of conditional independence. The modeled allowance for selection bias must be traded off against the validity of other structural assumptions which are imposed in order to identify the impact parameter of interest.

SOU methods model potential outcomes as a specified mathematical function of measured covariates and unobserved random variables defined by certain distributional assumptions. Unfortunately, Heckman's appeal to economic theory to guide model specifications does not provide a solution to the evaluation problem in energy efficiency applications. In the preceding quotation, he blurs the critical distinction between a priori theory and *knowledge* of the

relationship between the unobservable determinants of participation and potential outcomes. Thus, both SOO and SOU approaches claim to account for selection bias via conditioning on measured covariates that are assumed to determine baseline energy consumption and/or measure adoption; where they differ is in the nature of the assumptions about unobservables.

Valid application of the formal solutions to the problem of attribution ultimately comes down to questions of subject-matter knowledge. Failure to confront these questions leads to methodological abstraction from the substantive interpretation, that is, the semantics of the identifying assumptions in a specific application. For example, the formal equivalence of the assumptions of strong ignorability and conditional independence can blur the categorical distinction between the experimental and structural paradigms of attribution. Within the experimental paradigm, application of the statistical solution requires knowledge of the randomized design employed in program implementation. Quasi-experimental methods, on the other hand, require knowledge of the determinants of the potential outcomes. In the concluding section, I examine the problem of knowledge in energy efficiency applications and the methodological implications of a clear separation of the known and unknown determinants of program impact.

The latent structure of program impact

It is instructive to interpret Heckman's question regarding observational equivalence within the context of market demand for end-use efficiency: Why is it that consumers who exhibit the same level of demand for end-use services and face identical prices of commodity energy and energy-consuming equipment adopt different levels of equipment efficiency? The question reveals the principal dimensions of heterogeneity of the known and unknown determinants of net energy savings within the target population. The individual demand for end-use services and the market prices of energy and equipment efficiency are presumed to be observable determinants of measure adoption. The presumption is that they are determinants. That they are observable is not in question. It is therefore possible to identify subsets of targeted individuals who are homogeneous in these variables and to identify individuals within such subsets who do not purchase and install equipment with the same rated efficiency.

The question calls attention to two observable sources of heterogeneity: end-use service demand and measure adoption among individuals who are homogeneous in demand for the same end-use service. The question posed by the latter source of heterogeneity embodies two distinct hypotheses: first is the hypothesis that a consumer's demand for energy efficiency is informed (in part) by the magnitude of the potential energy cost savings and the incremental cost of higher efficiency; second is the hypothesis that there are other unobserved determinants of consumer choice which account for differences in adoption between individuals who would realize the same return on the investment in improved efficiency. End-use demand thus has dual significance as both a known and unknown determinant of net energy savings. It is a known determinant of net savings because the magnitude of savings produced by measure adoption is a function of the magnitude and utilization of the installed capacity of the equipment serving the relevant end use. It is an unknown determinant of net energy savings because its role as a causal factor in the adoption decision process is purely hypothetical.

The following exposition formalizes the dichotomy between the known and unknown determinants of the net savings impact parameter. The causal structure is composed of two functions that determine the values of the two factors that combine to produce program impact. Each factor is a policy-relevant impact parameter in its own right with independent significance for program planning and evaluation. However, the two questions of attribution associated with each factor correspond to the two extremes of the problem of knowledge, i.e., the known and unknown determinants of program impact, which, as such, require fundamentally different solutions.

Two questions of attribution

The derived demand for delivered energy is a function of the site-specific end-use demand for useful energy (EU) and the equipment-specific efficiency of conversion (η) from energy input to the useful energy output required by the end-use application:

$$EC = EU/\eta = (HO \bullet C)(CL/C) \tag{4}$$

Energy utilization at the end use is expressed as the product of rated equipment capacity (C), assumed equal to the maximum demand for useful energy (e.g., lumens, BTU/h), and equivalent full load hours of equipment operation (HO). Conversion efficiency (η) is the ratio of useful energy output per unit of energy input (e.g., lumens/W, BTU/h/W). The connected load (CL) is the input energy demand at full capacity.

An energy efficiency measure is a consumer action that reduces the amount of energy input required to produce the same level of end-use service as a specified (less efficient) baseline alternative. Energy efficiency measures can be classified as one of two types: actions that result in the installation of equipment with a higher conversion efficiency and actions that reduce the hours of equipment operation. The former reduces the amount of delivered energy required to produce a given amount of useful energy while the latter reduces the amount of useful energy that must be produced to maintain the desired level of end-use service, e.g., illumination and thermal comfort. Examples of conversion efficiency measures include installation of lighting equipment with a specified minimum rated lumens per watt and air conditioners with a specified minimum rated BTU/h per watt. Actions that reduce hours of operation include occupancy sensors, programmable thermostats, building weatherization, and "behavioral" measures whereby end users manually adjust the operation of energy-consuming equipment. For clarity of exposition, the following formal treatment of the problem formulates the market adoption decision as a choice between two alternative levels of efficiency.

Individual potential energy consumption can be expressed in terms of the potential adoption (MA_i) or non-adoption of a specific measure under the corresponding conditions of program implementation or non-implementation (PI). For each targeted individual, there are two potential values of energy consumption (EC_{Mi}, EC_{Bi}) associated with the realized value of the binary measure adoption variable:

$$EC_i\,(0) = MA_i(0) \bullet EC_{Mi} + \left(1 - MA_i(0)\right) \bullet EC_{Bi}$$

$$EC_i\,(1) = MA_i(1) \bullet EC_{Mi} + \left(1 - MA_i(1)\right) \bullet EC_{Bi}$$

Net savings (NS_i) is accordingly defined as the difference between the baseline and program values of energy consumption:

$$NS_i \equiv EC_i(0) - EC_i(1)$$

Substituting the previous equations for potential energy consumption yields an equivalent expression for net savings as the product of measure savings (MS_i) and net adoption (NA_i):

$$NS_i = MS_i \bullet NA_i \tag{5}$$

$$MS_i \equiv EC_{Bi} - EC_{Mi}$$
$$NA_i \equiv MA_i(1) - MA_i(0)$$
$$EC_{Bi} = (HO_i \bullet C_i)/\eta_{Bi}) \text{ or } (HO_{Bi} \bullet C_i)/\eta_i$$
$$EC_{Mi} = (HO_i \bullet C_i)/\eta_{Mi}) \text{ or } (HO_{Mi} \bullet C_i)/\eta_i$$

Equation (5) makes transparent the composite nature of the individual net savings impact parameter. Net savings embeds two separate and intrinsically different questions of attribution: net adoption is the impact of program implementation on measure adoption and measure savings is the impact of measure adoption on energy consumption. Structural separation of these two impact parameters is central to the assessment of methodological validity, because it reveals the dichotomy between the known and unknown determinants of net savings. Whereas the determinants of measure savings are known and measurable, the determinants of net adoption are unknown.

Measure savings quantifies the site-specific impact of independent variation in the relevant measure-specific determinant of energy consumption, i.e., conversion efficiency or operating hours, holding the other determinants at fixed values. By definition, it represents the *potential* savings that can be realized by the specified efficiency improvement, regardless of the factors that induce measure adoption. Measure savings is the source of the economic and environmental benefits of measure adoption and, as such, a principal determinant of the cost-effectiveness of measure adoption and program implementation. Accurate quantification of measure savings is therefore fundamental to effective program and portfolio design.

Program planners and regulators rely heavily upon site measurement and verification of the determinants of energy savings attributable to measure adoption by program participants. The principal source

of uncertainty associated with estimates of measure savings and measure cost-effectiveness is the annual cycle of equipment operation. The rated capacity and efficiency of installed equipment can be verified by onsite inspection at the time of program implementation or as part of a subsequent evaluation. Quantification of annual hours of operation is a much more involved process that entails direct or indirect measurement during a representative range of annual operating conditions. These measurement and verification activities are required by regulation to conform to rigorous protocols designed to minimize the bias and control the statistical precision of estimates of the total energy savings of the participant or target population.

On the other hand, net adoption quantifies the site-specific impact of exposure to program influence on measure adoption. There is no *known* equation, parallel to Eq. (4), which specifies the functional relationship between measure adoption and observable determinants. Measure savings is the difference between two physical outcomes generated by two different values of one of the variables in Eq. (4). The quantitative relationship between the physical variables is governed by physical laws of energy transformation and energy transfer and is hence invariant with respect to the behavioral determinants of their realized values at a particular site and time. In contrast, measure adoption is a behavioral outcome for which there is no known scientific explanation. Measure savings and measure adoption can be said without exaggeration to represent opposite extremes of the possible states of scientific knowledge of an empirical phenomenon.

Equation (4) is a statement of the knowledge that, for instance, operation of a connected load of 0.8 kW for 1000 h will increase metered consumption by 800 kWh, *and* that if equipment of the same capacity and a connected load of 1.0 kW had been installed and operated during the same time interval instead of the more efficient equipment, metered consumption would have been incrementally higher by the amount of 200 kWh. That is to say, measurement and verification of the demand for useful energy and the installed measure efficiency can be utilized to quantify the measure *savings* at the individual site. Holland (1986) refers to this method, which "exploits various homogeneity or invariance assumptions," as the "scientific solution" to the fundamental problem of causal inference, to differentiate it from the statistical solution.

There is, however, no prior knowledge to sustain a parallel statement concerning measure adoption. There is no way of knowing whether the same individual who purchased the more efficient product at a program-discounted price would have purchased the same product at the baseline undiscounted price or the standard efficiency product. Unlike the energy consumption function in Eq. (4), every model of measure adoption is a formulation of a specific set of untested hypotheses regarding the data generating process that relates the latent variables (potential outcomes) to the observed data. As an illustration, consider the common structural assumption in EE applications that the consumer selects the level of product efficiency which minimizes the life-cycle cost of service, i.e., the sum of the capital and discounted operating costs of producing the same level of end-use service over the useful life of the equipment, in economic terms the "conditional demand" for energy efficiency. Given two alternative levels of energy efficiency, this criterion entails a comparison of the difference between the respective capital and operating costs:

$$MA_i = 1(IV_i \bullet MS_i > IC_i)$$

IV_i is the present value imputed by each individual to a unit of annual energy savings and IC_i is the incremental cost of the energy efficiency measure. MS_i and IC_i are, by definition, functions of observable covariates. IV_i is a hypothetical unobserved covariate that represents the imputed value, i.e., the implicit individual willingness to pay for a unit of energy savings. Under this structural assumption, measure savings and the imputed value of measure savings are the principal sources of heterogeneity in the target population. Conditioning on incremental cost and measure savings yields the conditional average:

$$\overline{MA}|MS, IC = 1 - F_{IV|MS,IC}(IC/MS)$$

$F_{IV|MS,IC}(IC/MS)$ is the cumulative proportion (conditional distribution function) of individuals for whom the unobserved imputed present value is less than the incremental measure cost per unit of measure savings. Hence, $1 - F_{IV|MS,IC}(IC/MS)$ is equal to the proportion of baseline adopters within the population stratum (MS, IC). Under this model of measure adoption, the interpretation of the "no confounding" assumption of unobserved covariate balance between target (eligible/participant) and surrogate (ineligible/

Table 1 Latent structure of program impact

Class	$MA_i(0)$	$MA_i(1)$	NA_i	$p(MA_i(0), MA_i(1))$	$\overline{NS}\|MA(0), MA(1)$
Non-adopters	0	0	0	$p(0,0)$	0
Net adopters	0	1	1	$p(0,1)$	$\overline{MS}_{(0,1)}$
Net non-adopters	1	0	-1	$p(1,0)$	$-\overline{MS}_{(1,0)}$
Adopters	1	1	0	$p(1,1)$	0
Average	$\overline{MA}(0)$	$\overline{MA}(1)$	$\overline{NA}$		$\overline{NS}$

nonparticipant) populations is that the distributions of the imputed value, conditional on measure savings and cost, in the two populations are identical, i.e.,$F^S{}_{IV|MS,IC} = F^T{}_{IV|MS,IC}$, thus establishing a sufficient condition for the validity of the identifying assumption of equal rates of baseline measure adoption within strata, here defined by measure savings and cost.

$$\begin{aligned}\text{Selection Bias}\left(\widehat{\overline{NS}}|MS,\ IC\right) &\equiv \left(\overline{EC}^S\ (0)|MS,\ IC) - \overline{EC}^T\ (0)|MS,\ IC\right)\\ &= MS\left[(\overline{MA}^T\ (0)|MS,\ IC) - (\overline{MA}^S\ (0)|MS,\ IC)\right]\end{aligned} \tag{6}$$

Equation (6) defines the *conditional* selection bias of estimated net savings as the difference between the average baseline energy consumption of the surrogate and target populations within each stratum. The superscripts indicate the surrogate and target populations. Equality of average baseline measure adoption thus eliminates selection bias. But even if one had reason to believe that all consumers base their energy efficiency preferences on a life-cycle cost comparison of the available alternatives, without prior knowledge of the principal factors governing consumer willingness to pay for energy savings, the assumption of unobserved covariate balance is not plausible. The underlying determining factors are largely unknown and hence unobservable. If we do not know the determinants of the imputed value, then there is no basis for the assumption that its distribution is the same in the target and surrogate populations, nor is it possible to test the validity of the assumption based on an analysis of the observed data. If the determinants were known and could be translated into observable data, then a stratified analysis could, in principle, be applied to adjust for selection bias; but no method of data analysis — SOO or SOU — can compensate for the deficit in our scientific knowledge of the behavioral process of consumer decision.

Confounding of measure savings and net adoption

On the basis of Eq. (5), every individual in the target population can be classified according to the four possible values of the potential measure adoption outcomes $(MA_i(0), MA_i(1))$. Table 1 summarizes the defining characteristics of each latent class in terms of potential measure adoption, net adoption, the joint distribution of the potential adoption outcomes, and the average net energy savings.

Each class is homogeneous in potential measure adoption and net adoption, which reduces the class average net savings to the product of class average measure savings and net adoption:

$$\overline{NS}|MA(0), MA(1) = \left(\overline{MS} \bullet NA\right)|MA(0), MA(1)$$

Averaging the class net savings and net adoption over the target population yields:

$$\overline{NS} = p(0,1)\overline{MS}_{(0,1)} - p(1,0)\overline{MS}_{(1,0)} \tag{7}$$

$$\overline{NA} = p(0,1) - p(1,0) \tag{8}$$

This canonical representation of population heterogeneity reveals the latent structure of program

Table 2 Joint distribution of potential measure adoption outcomes

$MA_i(PI)$	$MA_i(1) = 0$	$MA_i(1) = 1$	Average
$MA_i(0) = 0$	$p(0,0)$	$p(0,1)$	$1 - \overline{MA}(0)$
$MA_i(0) = 1$	$p(1,0)$	$p(1,1)$	$\overline{MA}(0)$
Average	$1 - \overline{MA}(1)$	$\overline{MA}(1)$	$\overline{\boldsymbol{NA}} = \overline{\boldsymbol{MA}}(1) - \overline{\boldsymbol{MA}}(0)$

impact.[5] Non-adopters and adopters do not contribute to program net impact because measure adoption is independent of exposure to program influence, whereas net adopters and net non-adopters make respectively a positive or a negative contribution to net impact. Table 2 presents the joint and marginal distributions of the potential measure adoption outcomes $(MA_i(0), MA_i(1))$ in the standard configuration of a 2×2 contingency table. The column and row averages represent the marginal distributions of measure adoption under the respective program and baseline conditions of exposure to the intervention.

The program marginal distribution $\overline{MA}(1)$ is observable. The baseline marginal distribution $\overline{MA}(0)$ is only observable in the exceptional case of selective exposure, i.e., if program eligibility is restricted to a subset of the target population. As discussed above, in the special case where selective exposure is randomized, the calculated average measure adoption in the exposed and non-exposed sub-populations is an unbiased estimate of the corresponding program and baseline population parameter (marginal mean), thus yielding an unbiased estimate of net adoption. In the typical case of unrestricted eligibility, the program population average measure adoption is, in principle, observable, but recourse must be had to a surrogate population in order to quantify baseline adoption.

As discussed previously, the net savings and net adoption impact parameters represent the intention to treat (ITT) parameter, which means that the program intervention is designed to promote measure adoption by all targeted individuals. The ITT parameter quantifies the average impact of program implementation on the entire target population. It accounts for measure adoption and energy savings of both program participants and nonparticipants, viz., spillover. As such, it produces a valid summative measure of program impact. But what exactly is the informational content of the value of the net savings parameter?

Taking as an example the idealized upstream program design considered earlier, given access to utility billing data and vendor data identifying all targeted individuals, i.e., all purchasers of central air conditioners of a given capacity, and indicating (a) whether they were offered a discounted price and (b) whether they purchased the standard or higher efficiency product, the evaluator applies the statistical solution to estimate the net savings and net adoption parameters. In this case, the evaluator can reasonably infer that the calculated average (or total) net energy savings and net adoption are attributable to the randomly assigned program price discount. However, the estimate of net savings does not provide the information required to assess the accuracy of the average energy savings attributed by the program administrator to observed market adoption by targeted individuals who receive a price discount if they purchase the more efficient product. This is because the net savings impact parameter, as revealed by Eq. (5), confounds measure savings and net measure adoption. The standard metric to assess the accuracy of claimed savings is the gross realization rate (*GRR*), equal to the ratio of *realized* gross energy savings (*GS*) to claimed energy savings. The latent structure of average gross savings over the target population is expressed by the following equation:

$$\overline{GS} \equiv \left(\overline{MS}|PI = 1\right) = p(0,1)\overline{MS}_{(0,1)} + p(1,1)\overline{MS}_{(1,1)} \tag{9}$$

It is obvious from a comparison of Eqs. (7) and (9) that estimated net savings, i.e., the average impact of program implementation on energy consumption, is not a valid metric to quantify the average impact of measure adoption on energy consumption. The two parameters are incommensurable because they are designed to address fundamentally different questions of attribution. Nevertheless, some analysts have

[5] In a seminal paper in the field of epidemiology, Greenland and Robbins (1986) introduced this formulation of the latent structure of the attributable effect of an exposure to a putative causal factor on disease risk in a target population characterized by inherent differences in risk between exposed and unexposed individuals. Their analysis of the problem of heterogeneity of response, independently of Rubin, provides unique insight into the fundamental importance of the unconfoundedness assumption to the identification of causal parameters. Pearl (2011) characterized the formulation as a "canonical partition" of the population according to equivalent response functions.

presented patently spurious interpretations of evaluation findings derived from such comparisons. In a widely cited paper, Nadel and Keating (1991, p. 24) compared the results of 42 impact evaluations, derived from statistical analysis of participant and nonparticipant billing data, to engineering estimates of program savings. The authors offered the following justification for this analysis:

> While it may seem unfair to judge engineering estimates of gross savings by comparing them to impact evaluations of net savings, given the importance of net savings for determining program impacts and program cost-effectiveness, we chose to make the comparison, in an effort to encourage program planners and implementers to devote increased attention to estimating the net impacts of programs.

The comparison revealed significant discrepancies between engineering and billing analysis estimates of program savings, which the authors attribute to flaws in the engineering estimates. The premise of this comparison is that billing analysis estimates provide a valid benchmark to assess the accuracy of engineering estimates of measure savings. But the ratio of net to claimed savings confounds the program gross realization rate and the program energy savings net-to-gross ratio (*ESNTGR*), the ratio of net savings to (realized) gross savings. For example, a ratio of net to claimed savings of 0.6 cannot differentiate between the following parameter values: (a) a gross realization rate of 1.0 and a net-to-gross ratio of 0.6, (b) a gross realization rate of 0.6, and a net-to-gross ratio of 1.0, (c) a gross realization rate of 2.0 and a net-to-gross ratio of 0.3. The three possibilities are observationally equivalent. If the net-to-gross ratio does not exceed 1.0, then the calculated net to claimed savings ratio is the lower bound of possible values of the gross realization rate and the intrinsic negative bias is equal to $GRR \bullet (ESNTGR\text{-}1)$.

The idealized randomized upstream design, with one additional assumption, could enable the identification of the average measure savings within the latent class of net adopters, $\overline{MS}_{(0,1)}$, which is a component of both net and gross savings. This parameter is the previously defined local average treatment effect (LATE), to distinguish it from the ITT parameter, i.e., net savings averaged over the entire population. An estimate of the LATE can be formed by calculating the ratio of the unbiased estimates of net savings and net adoption, provided that the evaluator is willing to make the assumption that net adoption is non-negative for all targeted individuals. Given this monotonicity assumption, the population proportion of net non-adopters, $p(1,0)$, which appears in Eqs. (7) and (8), is equal to zero and the ratio of the two parameters is equal to: $p(0,1)\ \overline{MS}_{(0,1)}/p(0,1)$.

The empirical decomposition of the average net savings parameter into its average measure savings and net adoption components is critical to the evaluation of measure and program cost-effectiveness. Net energy savings is the principal determinant of the incremental resource benefits of program implementation. However, quantification of the incremental capital cost of measure adoption requires knowledge of the population rate of net adoption in order to account for the net impact of program implementation on the life-cycle cost of end-use service, which is the sum of the cost of energy consumption and the cost of capital investment. The net benefit of measure adoption, i.e., the reduction in the life-cycle cost of service attributable to program influence (achieved by net adopters), is the fundamental driver of program cost-effectiveness. The magnitude of the average net benefit of measure adoption, which is the difference between the avoided cost of the average energy savings of net adopters (LATE) and the incremental cost of the installed measure, must be positive and the magnitude of net adoption must be sufficient to produce aggregate benefits that exceed program implementation (administration, marketing, etc.) costs.

Moreover, the separation of net measure savings and net adoption enables the evaluation of program logic and the effectiveness of key design elements in the realization of program objectives. Separate estimates of net adoption and measure savings provide ongoing feedback to program administrators which can inform decisions to improve program performance. The magnitude of net adoption is a direct indicator of the effectiveness of program incentives and information to induce efficiency improvements that would not have occurred in the absence of the program, whereas the average energy savings of induced adopters allows for an assessment of the life-cycle benefits realized by that latent class of the target population. Departures from expectations in the two impact parameters have different implications

for adaptive adjustment of program design or implementation.

If program incentive budgets, energy savings goals, and projected net benefits are based on a projection of net measure adoption that is substantially higher than the estimated actual level, the program administrator may increase the amount of the product discount or introduce a marketing or stocking incentive for equipment vendors. Or if the claimed average measure savings per net adopter is substantially higher than the estimated amount, then reevaluation of measure and program cost-effectiveness may be in order. Modification of measure or participant eligibility requirements could be investigated in order to more effectively target those market transactions which present the greatest opportunity to realize the cost-effective potential of increased end-use efficiency.

While the separate estimation of net adoption and average measure savings of net adopters affords some insight into the factors underlying the realization of program energy savings and the corollary implications for program improvement, the formative value of these parameters is severely limited by the absence of site-specific measurement and verification data that can be used to identify and quantify the discrepancies between the assumptions employed in measure savings projections and the corresponding observed conditions during the post-adoption reporting period. While, under ideal conditions of program design and data collection, statistical estimates based on longitudinal or cross-sectional comparisons of whole-building metered energy consumption may provide a valid summative measure of program impact, program administrators and regulators are left in the dark to speculate regarding the reasons for the empirical results.

M&V studies routinely conducted to estimate gross energy and demand savings are able to separately quantify measure savings at a single site via direct measurement of the determinants of end-use consumption. These on-site assessments can provide a quantitative breakdown of the site and population (gross) energy savings realization rates into separate factors to explain the sources of differences between reported and evaluated savings, such as installed capacity and efficiency and hours of operation. The sample data can also be used to characterize the population distributions of measure savings and determinants of savings, information that is critical to a clear understanding of the corresponding distribution of net benefits within the target population. The policy relevance of these findings is obvious given the program and portfolio objectives to maximize all cost-effective opportunities and to achieve a broad distribution of benefits to diverse customer sectors.

Structural models of selection bias

In cases where it is feasible to condition on the known determinants of measure savings, then one source of covariate imbalance between the participant and surrogate nonparticipant subpopulations can be eliminated. The *conditional* selection bias of the quasi-experimental stratified estimator parallels Eq. (6):

$$\begin{aligned}\text{Selection Bias}\left(\widehat{\overline{NS}}|MS,\ PP=1\right) &\equiv \left(\overline{EC}(0)|MS,\ PP=0\right) - \overline{EC}\,(0)|MS,\ PP\ =\ 1\Big) \\ &= MS\left[(\overline{MA}\,(0)|MS,\ PP=1) - (\overline{MA}\,(0)|MS,\ PP=0)\right]\end{aligned} \tag{10}$$

As before, the structural assumption of equal rates of baseline measure adoption within strata, in addition to monotonicity and no nonparticipant spillover, is sufficient to identify the participant net savings impact parameter. This assumption, the SOO condition formulated by Eq. (3) of conditional independence of baseline adoption and program participation, is the basis for most matching and regression methods employed to quantify program net savings. The problem is that the assumption is not consistent with the basic logic of program design and the market research by which it is informed.

Energy efficiency programs are designed to address specific barriers to measure adoption. The design process is accordingly guided by two questions: (1) What factors account for different measure adoption decisions among consumers who are homogeneous with respect to potential measure savings? (2) What specific program design elements can effectively target individuals for whom cost-effective measure adoption

is inhibited by one or more of these factors? Market barriers may take many forms, but can be generally subsumed under two categories: informational and financial barriers. Informational barriers include lack of awareness of the commercial availability of alternative products and services to upgrade end-use efficiency, lack of knowledge of the potential energy savings achievable by a specific upgrade and the potential reduction of annual operating expenses, uncertainty regarding equivalence of performance of alternative technologies, such as capacity, reliability, and quality of end-use service, and lack of awareness of non-energy costs and benefits of measure adoption. Financial barriers include limited access to capital, transaction costs of financing, and inability to capitalize building efficiency improvements.

A fundamental design objective is to minimize participation by baseline adopters, equivalently to maximize the program participant measure adoption net-to-gross ratio (*PPMANTGR*). The challenge is to offer program benefits that selectively motivate non-baseline adopters to participate. If, for example, program planners have conducted market research that identifies capital market imperfections that limit baseline measure adoption by a significant segment of the target population, then a program which facilitates access to capital could influence consumers within this segment to invest in a more efficient alternative, whereas consumers without capital constraints would have little incentive to seek alternative financing via program participation. In this case, the quasi-experimental estimator would be negatively biased, understating the true net impact of the program.

Product subsidies can also be designed to selectively exclude eligibility of applications in which the simple payback of measure installation is below a prescribed threshold determined by survey research to assess the target population distribution of consumer energy efficiency investment criteria. From the inception of retrofit/early replacement programs, this strategy has been employed to limit the number of participant "free riders" who receive an incentive payment for measures that consumers would adopt without a program subsidy. Implementation of this design element of course depends upon the availability of site-specific data required to project annual measure savings. Such programs which incorporate initial site assessments can also disqualify measure installations that replace existing equipment of the same or similar efficiency as the program measure or less-efficient equipment that would have been replaced with a non-qualifying product that operates at the current market standard of efficiency.

Program interventions that are primarily focused on the provision of information target consumers and trade allies who may be unaware of the availability of more efficient alternatives and uninformed about the potential benefits of measure adoption. This category includes a diverse range of program designs, including product labeling, facility operator and contractor training, home energy reports, residential energy management systems, alternative pricing, retro-commissioning, and audit programs that perform on-site assessments of energy savings potential and non-energy benefits (e.g., avoided equipment replacement costs, other resource costs, waste disposal costs, increased health, and safety and productivity benefits). The intended effect of the provision of information about the benefits of measure adoption is to enable each consumer to accurately value the impact of improved efficiency on the cost of end-use service, assuring that the willingness to pay for the improvement is not distorted by a lack of awareness of available alternatives or misconceptions about potential savings or equivalence of service. Like financing programs, informational programs provide little incentive for consumers to participate who have procured the relevant measure-specific and site-specific information, or have installed data collection and analysis systems that generate such information.

However, some programs are designed to selectively attract participants who have higher-than-average potential energy savings in order to maximize gross energy savings and measure adoption. Many states have established performance incentives to motivate program administrators to achieve or exceed pre-approved energy savings targets utilizing resources that are limited by approved program-specific budgets. Performance metrics consist of program, sector, or portfolio energy savings targets and may incorporate additional components tied to program cost-effectiveness. Such metrics can have the unintended consequence of rewarding program design and implementation which selectively encourages participation by high-use consumers that generate higher than average energy savings per participant. Selective marketing by customer account representatives and "standard offer" programs which pay a fixed

Table 3 Participant baseline measure adoption, net-to-gross ratio, and selection bias

Self-selection	$\overline{MA}(0)\|PP=1$	PPMANTGR	Selection bias
Random selection (MA(0)⫫PP)	$=\overline{MA}(0)\|PP=0$	$=1-\overline{MA}(0)\|PP=0$	$=0$
Incentive-compatible selection	$<\overline{MA}(0)\|PP=0$	$>1-\overline{MA}(0)\|PP=0$	$-\overline{MA}(0)\|PP=0<0$
Adverse selection	$>\overline{MA}(0)\|PP=0$	$<1-\overline{MA}(0)\|PP=0$	$0<1-\overline{MA}(0)\|PP=0$

incentive amount per unit of energy savings are notable examples of program design and implementation practices that are likely to result in a high rate of participation by baseline adopters (free riders) and a corresponding low *PPMANTGR*.

Prescriptive rebates and subsidized measure installation incentives reduce the incremental cost of measure adoption by a fixed amount that does not vary within the eligible population. Baseline non-adopters who are aware of the more efficient alternative may be influenced to adopt by the availability of an incentive that substantially reduces the difference in installed cost. Baseline adopters, on the other hand, are confronted with a fundamentally different tradeoff. Because they have determined that measure adoption is worth the additional (unsubsidized) cost, their baseline is the life-cycle cost of the more efficient alternative. Their decision criterion is not a comparison of the benefit and subsidized cost of measure adoption, but rather a comparison of the benefit and cost of program participation. The incremental benefit is the amount of the incentive; the incremental cost is the transaction cost of program participation, which is not observable.

These program design examples are presented to articulate the problem of knowledge in terms of the unobserved determinants of baseline measure adoption and program participation. While, in practice, most programs incorporate some combination of informational and financial inducements, the cited alternative design elements indicate the scope of possible interventions that may be employed to influence diverse segments of a heterogeneous target population consisting of individuals who are confronted with some unknown combination of barriers to measure adoption and program participation. The examples reveal that the magnitude and direction of selection bias, as defined by Eq. (10), depend critically on the nature of the program intervention and the specific factors that account for variation in adoption and participation decisions among the individuals who comprise the target population.

Programs which favor self-selection by nonbaseline adopters may be characterized as incentive-compatible because the intervention induces eligible individuals to implicitly reveal their unobservable willingness to pay the unsubsidized incremental cost of the efficiency upgrade via their observed participation outcomes. Random selection implies equality of average participant and nonparticipant baseline adoption and, consequently, zero selection bias; as such it is a boundary point between incentive compatibility and adverse selection, which characterizes a program that selectively attracts baseline adopters to participate. The magnitude and direction of the selection bias reflect the nature of the dependence between baseline measure adoption and participation. Negative bias indicates negative dependence and positive bias indicates positive dependence. The negative and positive limits of dependence correspond respectively to the cases in which participant baseline adoption rates equal zero and one, or equivalently the participant measure adoption net-to gross ratios (PPMANTGR) equal one and zero. Table 3 presents the bounds on these parameters for the three self-selection scenarios.

Under the SOO assumption of random self-selection, baseline measure adoption and program participation are independent (MA(0)⫫PP) and there is no selection bias. Under incentive compatibility, participant baseline adoption is less than nonparticipant baseline adoption and selection bias is negative; under adverse selection, participant baseline adoption is greater than nonparticipant baseline adoption and selection bias is positive. The lower and upper bounds of self-selection bias are respectively equal to $(-\overline{MA}(0)|PP=0)$ and $(1-\overline{MA}(0)|PP=0)$.

The previously mentioned method of control functions is designed to account for unobservable differences between participant and nonparticipant baseline behavior (Heckman & Navarro-Lozano, 2004). The

technical approach is to formulate separate models of the individual program participation decision process and the potential outcome of interest which, in energy efficiency program impact evaluations, is typically energy consumption or measure adoption. The participation model is formulated as a random utility model, according to which an individual participates if the utility of participation exceeds the utility of nonparticipation:

$$\begin{aligned} PP_i &= 1(V_{PP,i} + U_{PP,i} > V_{NP,i} + U_{NP,i}) \\ &= 1(V_{PP,i} - V_{NP,i} > U_{NP,i} - U_{PP,i}) \end{aligned}$$

In this formulation, utility is the sum of observable and unobservable components. The observable component, also known as "representative utility," is a specified function of measured covariates and the unobservable component is a random variable with a specified distribution. The probability of participation (propensity score) is determined by the difference between the observable components and the distribution function (F) of the difference between the unobservable components:

$$\Pr(PP_i) = F_{U_{NP,i}-U_{PP,i}}(V_{PP,i} - V_{NP,i}) \tag{11}$$

SOU models of energy consumption incorporate a correction factor for unobservable differences between participant and nonparticipant baseline energy consumption which is derived from the random utility model probability of participation. The correction factor, termed the control function by Heckman and Robb (1985), is the expectation of the unobserved residual component of the regression of energy consumption on measured covariates, conditional upon the participation decision. Violette and Ozog (1989) provide a review of the application of this approach, first developed by Heckman (1979), to impact evaluation of energy efficiency programs, utilizing an adaptation by Dubin and McFadden (1984), who characterize it as a "discrete/continuous choice" model.

Train (1988, 1994) developed an alternative approach to correction for selection bias, in both audit and incentive programs, which combines the participation model with a discrete choice model of measure adoption, instead of a model of energy consumption. This approach has the virtue of separate quantification of measure adoption and measure savings, which avoids selection bias in the latter component of net savings, whereas billing analysis of participant and nonparticipant energy consumption confounds two sources of selection bias. As discussed by Agnew and Goldberg (2017), participant and nonparticipant comparison groups can be expected to differ with respect to observable as well as unobservable characteristics that contribute to differences in baseline energy consumption. The observable differences may be characterized as differences in the determinants of measure savings. The unobservable differences represent differences in baseline measure adoption.

Figure 1 illustrates the problem for a single stratum of imputed hours of operation. The top and bottom lines depict the energy consumption of a single device of a given capacity as a function of two alternative levels of rated efficiency and annual equivalent full-load hours of operation. The slope of each line is equal to the corresponding connected load. The difference between the values of the MA = 0 and MA = 1 lines at the same hours of operation is equal to the amount of measure savings. The other two lines represent the average energy consumption at intermediate values of average baseline measure adoption, and consequently connected load, corresponding to participant and nonparticipant subpopulations. The average value of participant net savings of individuals homogeneous in hours of operation is equal to the difference between the MA(0)|PP = 1 and MA = 1 consumption values, because it is assumed that participation is contingent upon measure adoption. The corresponding value of net adoption is equal to the ratio of net savings to measure savings at the same hours of use.

In this hypothetical example, the quasi-experimental estimator of participant net savings is the difference between the observed average energy consumption values of nonparticipants and participants within the single stratum of imputed hours of operation. Selection bias is the difference between nonparticipant and participant average baseline consumption, indicated by the horizontal solid lines.[6] All quasi-experimental methods seek to account for observable differences between the determinants of participant and nonparticipant baseline consumption in order to

[6] The maintained assumption of no program impact on nonparticipants implies that observed consumption is equal to baseline consumption.

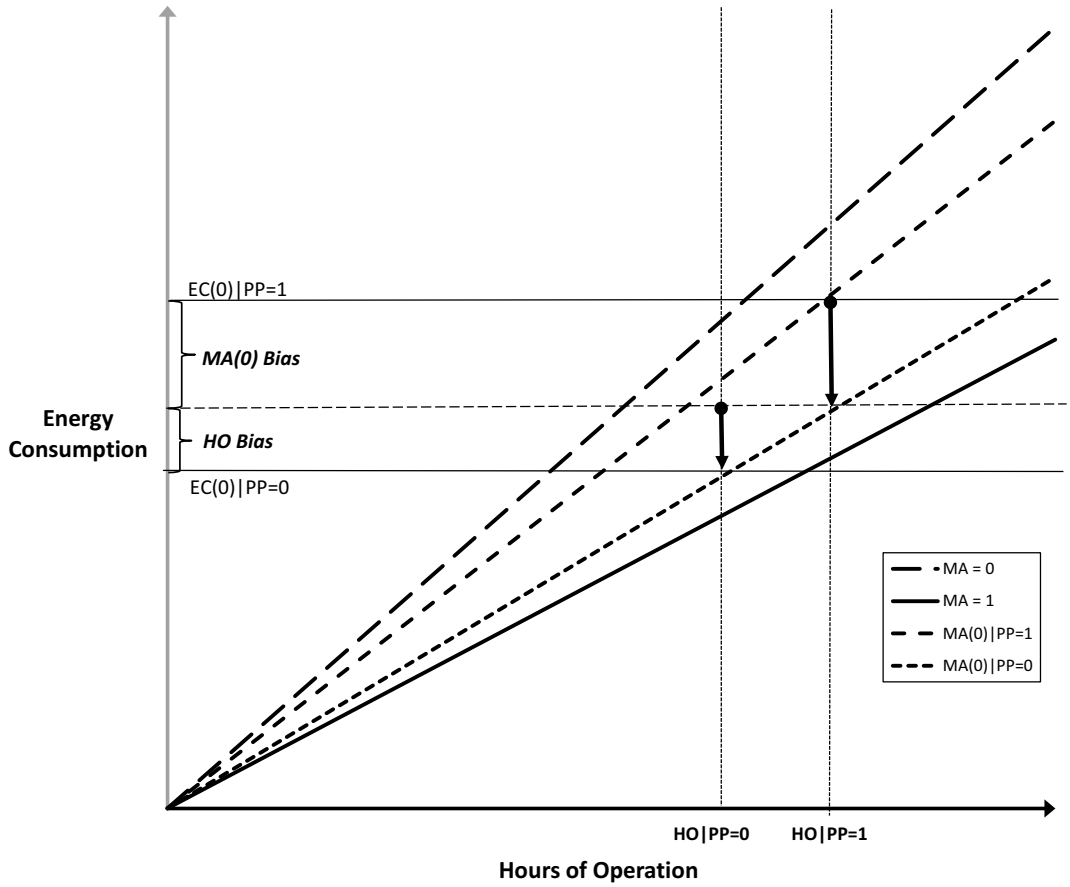

Fig. 1 Sources of self-selection bias

eliminate or minimize this source of bias. The principal *observable* determinants of the consumption of the single device are the rated efficiency, which in this example is limited to two possible values, and individual hours of operation. The analysis must therefore condition energy consumption on measured hours of operation, utilizing either a non-parametric (e.g., stratification, matching) or a parametric regression method to eliminate this source of bias.

As discussed earlier, a serious deficiency of most billing analyses is the lack of site measurement and verification of key covariates for both participants and nonparticipants. Evaluators must accordingly rely upon available data that can serve as proxy values for site-specific measurements. Proxies for device hours of operation, depending on the end use, include building operating schedule obtained via customer surveys, utility records of building type, weather data collected at the closest site, or values imputed from site measurements conducted in other populations. Figure 1 depicts a case in which the true average hours of operation of nonparticipants in one particular stratum is less than the corresponding true participant average, after conditioning on one or more proxy covariates to adjust for this source of bias. The residual difference in this illustrative example creates a negative bias in the estimate of participant baseline consumption and participant net savings.

The negative bias in hours of operation reduces the observed nonparticipant energy consumption by the amount bracketed on the vertical axis.[7] This component of bias is reflected in the difference in measure savings determined by the difference in average hours of operation. The second component of bias is

[7] The bracketed hours of operation bias is the incremental bias after accounting for baseline adoption bias.

the reduction in nonparticipant energy consumption associated with higher nonparticipant baseline adoption. For clarity of illustration, the two components of bias are negative, but both could be positive or one positive and the other negative. It must also be noted that the participant and nonparticipant baseline rates of adoption can assume any value between 0 and 1 in any stratum, so that the slopes of the intermediate lines can differ among strata.

The discrete/continuous energy consumption model must address both sources of bias in the estimation of net savings. The data problem which is the source of measure savings bias can, in principle, be solved via accurate site-specific measurement of the observable determinants, in this example hours of operation. However, the proposed correction for latent baseline adoption bias entails the inclusion of a selectivity covariate in the energy consumption regression model which is derived from the modeled probability of participation. The key assumption is that the difference between nonparticipant and participant average baseline consumption, conditional on measure savings, is equal to the product of the difference between the participant and nonparticipant average values of the selectivity covariate and its regression coefficient. The covariate, commonly known as the inverse Mills ratio, is the expected value of the unknown component of the utility of participation conditional on the observed participation decision.

The unknown component of utility is the difference between the hypothetical random utility of the participation decision made by each targeted individual and the modeled representative utility, which is a specified function of observable covariates. The individual values of the selectivity covariate included in the consumption model are determined by the assumptions about the set of relevant covariates, the functional form of the relationship between the observed covariates and representative utility, and the distribution of the unknown random components of utility, as formulated in Eq. (11). Misspecification of the representative utility function undermines the validity of the assumptions concerning this distribution. Omission of relevant covariates or misspecification of the form of the utility function will violate the common assumptions that the modeled errors are identically distributed as normal or logistic random variables with zero mean and common variance. These distributional assumptions, moreover, are inconsistent with heterogeneity of representative utility implied by variation in the relative importance that a consumer attaches to the observable variables specified in the evaluator's model of representative utility.

As discussed above, due to heterogeneity of market barriers to measure adoption, different latent classes of targeted individuals can be expected to respond differently to program design elements, depending on the individual baseline threshold of adoption and the incremental utility of participation. In particular, it was noted that baseline adopters and non-adopters can be expected to employ different decision criteria and further that different segments of baseline adopters and non-adopters may be selectively attracted by different program design elements, such that some baseline adopters may be more likely to participate than some non-adopters and vice versa. In light of these considerations, the scope of uncertainty concerning selection bias may be characterized in terms of an unknown mixture of some number of latent classes of targeted individuals with selection probabilities that can vary over the entire range of incentive-compatibility and adverse selection presented in Table 3. The selectivity covariate is thus derived from a participation model that is subject to the same types of specification error that it is designed to correct in the consumption model. The "solution" to the problem of knowledge of the determinants of baseline adoption has simply been transposed via a different set of assumptions regarding the unobserved components of the utility of participation.

Train (1994, p. 440) rejects the application of discrete/continuous choice methods to the problem of estimation of net energy savings because this approach does not separately model the measure adoption decision as well as the program participation decision:

> Estimation of net savings necessarily requires identifying causation. The central question is: to what extent did participants install measures because of the program? Yet regressions of consumption against program participation dummies are not causal. Program participation does not cause consumption to change. It is the implementation of measures that causes consumption to change. Program participation is expected to increase the chance that a customer implements conservation measures, and then

Table 4 Joint distribution of potential measure adoption/participation outcomes (random selection)

$MA_i(PI)$	$[MA_i(1)\|PP=0]=0$	$[MA_i(1)\|PP=0]=1$	$[MA_i(1)\|PP=1]=1$	Average
$MA_i(0)=0$	0.60	0	0.15	0.75
$MA_i(0)=1$	0	0.20	0.05	0.25
Average	0.60	0.20	0.20	$\overline{NA}=.15$
$\overline{MA}(1)\|PP$	0.75 (=0.60/0.80)	0.25 (=0.20/0.80)	1.00 (=0.20/0.20)	

this installation changes the consumption of the customer. A causal model would represent these two steps of causation explicitly.

In contrast to the discrete/continuous modeling approach, discrete choice methods explicitly model the measure adoption decision as well as the participation decision. Train et al. (1994) formulated a nested logit model of the three options represented in Table 4: non-adoption, nonparticipant adoption, and participant adoption, applicable to programs which condition participation on measure adoption. The nested logit probability of each adoption/participation decision is derived from a random utility model composed of the sum of the observable components of the utilities of participation and measure adoption and an unobservable random component. Each choice probability can be expressed as the product of two binary logit probabilities. The conditional probability of participation contingent upon measure adoption takes the same binary logit form adopted by Dubin and McFadden (1984) and subsequently applied by energy efficiency program evaluators to estimate net energy savings using the discrete/continuous approach (Violette & Ozog, 1989). The marginal probability of measure adoption is also binary logit, which is a function of the individual representative utility of adoption, regardless of which participation option is selected, and the expected incremental utility afforded by the availability of the option to participate.

The energy consumption (discrete/continuous choice) and measure adoption (discrete choice) approaches differ in important respects. One fundamental difference is the separation of the two questions of attribution. The discrete choice methods model the individual probability of measure adoption under program and baseline conditions. Net savings is estimated by calculating the average of the product of the individual probability of net adoption and individual measure savings estimates derived from onsite surveys of sampled participants and nonparticipants (Train et al., 1994), whereas discrete/continuous choice billing analysis is designed to provide a comprehensive solution to selection bias which corrects for participant-nonparticipant differences in both measure savings and baseline measure adoption, with the consequent confounding of measure savings and measure adoption bias. Discrete choice analysis generates separate estimates of the two components of net savings in Eq. (5), providing an unbiased estimate of the average gross measure savings impact parameter and eliminating one source of bias in the estimate of net savings. However, a serious practical limitation of this method is the cost of site measurement of measure savings and verification of measure adoption for nonparticipant as well as participant samples, which, as noted above, is a primary motivation for the use of the discrete/continuous billing analysis. The solution to the data problem may be prohibited by regulatory requirements to maintain evaluation budgets that do not exceed a specified percent of total program expenditures.

In addition to avoiding measure savings bias, the expansion of the observed source data to include a characterization of the available efficiency alternatives, verification of measure adoption, estimation of end-use demand and potential measure savings, and other site-specific observations enables the evaluator to explicitly model the measure adoption decision by participants and nonparticipants. The decomposition of the nested logit probabilities into separate binary logit models provides a transparent formulation of the distinct structural assumptions underlying the adoption and participation decision processes. The representative utility of measure adoption is specified as a function of covariates that do not vary with participation, e.g., measure savings, whereas the representative utility of participation depends on covariates that do vary with participation, e.g., program awareness and amount of program incentive to adopt a measure. The individual probability of baseline adoption is "simulated" by recalculation of the marginal probability of adoption with the option to participate removed from the model, by setting the expected incremental utility

Table 5 Participant baseline measure adoption, net-to-gross ratio, and selection bias

Self-selection	$\overline{MA}(0)\|PP=1$	PPMANTGR	Selection bias
Random selection (MA(0)⫫PP)	0.25	0.75	0
Incentive-compatible selection	[0, 0.25)	(0.75, 1.0]	[−0.25, 0)
Adverse selection	(0.25, 1.0]	[0, 0.75)	(0, 0.75]

of participation to zero. The individual probability of net adoption is accordingly calculated as the difference between the marginal probability of adoption with and without this term in the equation.

The joint distribution of the potential outcomes under the program and baseline conditions presented in Table 2 can be expanded to disaggregate program measure adoption into participant and nonparticipant subpopulations. Under the monotonicity and exclusion restriction assumptions, this joint distribution reduces to the joint distribution of participation and baseline adoption. Table 4 presents a numerical example under these assumptions and the SOO assumption of conditional independence of participation and baseline adoption, in which case the participant and nonparticipant rates of baseline adoption are equal and baseline adopters and non-adopters are equally likely to participate. Table 5 presents the bounds on program participant baseline measure adoption, net-to-gross ratio, and the selection bias under the self-selection assumptions presented in Table 3.

It is clear from Tables 4 and 5 that every possible value of average participant baseline measure adoption, from 0 to 1.0, is consistent with the observed measure adoption data for non-adoption, nonparticipant adoption, and participant adoption (presented in the row labeled "Average"). While nonparticipant baseline adoption is directly observable under the maintained assumptions of monotonicity and nonparticipant spillover, alternative values of participant baseline adoption are observationally equivalent. The source of this indeterminacy is the absence of baseline measure adoption data to estimate the population value (presented in the column labeled "Average"). The statistical solution is not available because there are no observations of the marginal distribution of baseline adoption in the target population. Structural assumptions must therefore be invoked to determine a solution.

In the three-option nested logit model, the joint distribution of the unobserved random components of utility is assumed to follow a specific parametric form that determines the nature of the dependence between baseline measure adoption and program participation. This specification of the joint distribution limits the assumed relationship to one of nonnegative dependence. In other words, the model specification *rules out* the possibility of incentive compatibility, thus incorporating an intrinsic negative bias in the estimation of net adoption. While more flexible models of discrete choice are available, there remains the basic problem of knowledge. In the quasi-experimental setting, every SOO or SOU model represents an alternative set of untestable assumptions about the structure of the adoption and participation decision processes. Both approaches are fundamentally flawed because there is no scientific foundation for the assumptions that are required to determine a unique solution to the problem of attribution.

Quantification of bias

As noted in the "Introduction" section, program administrators and regulators, lacking a consensus methodological standard for valid attribution, must assess the credibility of the reported findings generated by diverse approaches to data collection and analysis. Given that most programs are not designed to randomize exposure to program influence, the presence of exposure or selection bias cannot be ruled out. The presumption of attribution bias accordingly raises the practical question of the magnitude of the effect on estimates of program impact.

Unfortunately, industry evaluation protocols governing statistical precision and physical measurement error cannot be applied to the quantification of these sources of bias. In certain cases, the direction of bias may be posited via plausible assumptions: in studies which employ targeted nonparticipants as a baseline surrogate, the assumption of positive spillover results in a negatively biased estimate of net impact. Furthermore, the magnitude of the negative bias is twice the level of spillover, because positive spillover leads to

an overstatement of participant baseline adoption and an understatement of nonparticipant net adoption. For example, even a modest level of 10% spillover will result in a negative bias of 20% of net impact. Yet even in applications where the assumption of no exposure bias may be plausible, the preceding discussion of the sources of selection bias reveals the *unrestricted* scope of the uncertainty of impact estimates derived from evaluation data which do not include observations of the baseline measure adoption or energy consumption of targeted individuals: the observed data are consistent with every net-to-gross value between zero and one. In this common situation, the only recourse left to the evaluator is the explicit or implicit adoption of untestable structural assumptions regarding the determinants of baseline behavior.

Discussion

In this examination of the problem of attribution, I have concentrated on the formal underpinnings of valid causal inference when using experimental and quasi-experimental methods to analyze program impact. Definition of policy-relevant impact parameters in terms of potential outcomes, following Neyman and Rubin, reveals the intrinsic limitations on valid attribution when all targeted individuals are subject to program influence or selective exposure is not randomized within the target population. This categorical distinction between quasi-experimental and experimental program design elements signifies the presence or absence of exposure and selection bias because, with quasi-experimental designs, there are no observable data that can adequately represent the potential outcomes under the baseline condition of no intervention.

In EM&V practice, the limited scope of application of randomized control in program design is problematic given the importance of quantifying the benefits of programs funded by energy consumers to public policy decisions governing energy resource planning and environmental protection. In default of the program design conditions that enable a statistical solution to the problem, evaluators must rely upon structural assumptions to justify a causal interpretation of estimates of net energy savings, net adoption, and net-to-gross ratios. These assumptions are necessary to fill the empirical void created by the lack of valid baseline data. The posited underlying behavioral processes that determine measure adoption and program participation may be intuitively plausible or consistent with a theory of consumer preference, but they are devoid of empirical content, i.e., not testable from the available data. The model assumptions specifying the set of observed determinants, the distribution of the unobserved determinants, and their functional relationship to potential outcomes represent one out of many alternative hypotheses that can be devised to explain the potential outcomes under the mutually exclusive conditions of program implementation.

The observational equivalence of alternative sets of plausible assumptions undermines the credibility of any one of the modeled hypothetical alternatives as a valid basis for attribution. As illustrated by numerical example, different analyses of the same observed choice data can lead to conflicting, indeed diametrically opposed conclusions about program net impact, depending on the structural assumptions selected by the evaluator. The deficit in knowledge of the relevant behavioral determinants provides full scope for latent heterogeneity of consumer response to program marketing, information, and financial incentives which will vary according to the differential effects of program design elements on diverse segments of the target population.

Vine et al., (2014, p. 628), in a plea for the adoption of the experimental paradigm in EE program evaluation, note the contrast between engineering calculations to quantify measure savings, which are "relatively uncontroversial," with problems of attribution in the assessment of program impact "that have resisted resolution and create a persistent climate of uncertainty about the effectiveness of energy efficiency programs." The authors argue for the use of experiments to resolve the uncertainties intrinsic to estimates of net impact, including spillover effects, of programs that have not typically employed randomized selection for eligibility. They further propose the explicit incorporation of a research dimension into program evaluation which emphasizes the formative, as well as the summative value of randomized control, in that program administrators "can learn what is working and what is not and can thereby develop more innovative and effective programs."

When the object of inquiry is consumer preference, a controlled experiment represents an attempt

to acquire a rudimentary understanding of the factors that influence observed market behavior. For example, randomized eligibility for a high, low, or no incentive or a 2×2 factorial design to analyze the effects of two different types of intervention, e.g., informational and/or financial, could provide some limited insight into the influence of program design elements on measure adoption, as well as produce a valid estimate of net impact. As in any field of scientific inquiry, progress toward knowledge of behavioral determinants would be incremental and contingent upon replication of findings in other populations. Such efforts, however limited in scope, would make better use of ratepayer funds than the continued reliance on methods which can only perpetuate the "climate of uncertainty" sustained by uncritical acceptance of "heroic and unwarranted assumptions" incorporated within models of the science (Rubin, 2005). Indeed it was the explicit intention to remove the burden of untestable assumptions concerning potential outcomes that motivated Rubin's insistence on a posited assignment mechanism as the *sine qua non* of causal inference from observed data.

Randomization is a confession of ignorance. In Fisher's (1971, p. 44) words, "Randomisation properly carried out … relieves the experimenter from the anxiety of considering and estimating the magnitude of the innumerable causes by which his data may be disturbed." There are no intermediate solutions to the problem of attribution. The *hypothesis* of randomization cannot be construed as an approximation to *knowledge* of the physical act of random selection for eligibility, and analysis of the data as if it were generated by such a process cannot produce a credible estimate of net impact. By the same token, models based on assumptions about the observable and unobservable determinants of measure adoption or program participation, however plausible, cannot provide a scientific solution in the absence of prior knowledge of the relevant causal factors, knowledge which can only be gained by an experimental approach. However, alternative randomized designs which selectively control exposure to interventions other than program implementation, notably RED, cannot identify the policy-relevant population parameters that define program impact, and therefore, the findings of such studies cannot be relied upon to support a valid inference of attribution.

Estimation of net energy savings via comparison of the metered consumption of participants and nonparticipants, or some other baseline surrogate, confounds two questions of attribution. As distinct populations, participants and nonparticipants have different rates of baseline measure adoption and different potential measure savings. The primary conclusion of this analysis is that, when randomized exposure is not feasible, there is no solution to the problem of attribution of measure adoption — and hence energy savings — to program influence. Even when sufficient site measurement and verification data are collected to condition on the known determinants of measure savings, Eq. (10) reveals that the net savings estimate within each stratum will be biased by the unknown difference between participant and nonparticipant rates of baseline measure adoption. However, in practice, the bias may be compounded by conditioning on covariates that cannot adequately control for differences between participant and nonparticipant site-specific end-use determinants of measure savings, e.g., equipment capacity and annual hours of operation.

An important implication of this confounding of measure savings and net adoption is that comparisons of participant and nonparticipant metered consumption cannot serve as a valid standard to evaluate the accuracy of projections of measure savings, and hence the potential benefits of measure adoption, which are critical determinants of measure and program cost-effectiveness. Such comparisons confound net adoption and gross realization rates and, even in applications in which the assumption of negligible baseline adoption may be credible, there remains the bias generated by the failure to control for systematic variation between participants and nonparticipants in the demand for end-use service.

The statistical and scientific solutions can provide valid answers to the respective questions of attribution that they are designed to address. Randomized exposure yields observations of potential outcomes that are representative of the target population under the baseline as well as the program conditions of intervention. Measurement and verification of end-use demand and installed efficiency at program participant sites provide the data required to estimate individual energy consumption under the alternative conditions of installed and baseline efficiency, which determine measure savings defined as the reduction in energy required to supply the same level of service.

This analysis has focused on methods that are based on a comparison of the energy consumption or measure

adoption of program participants and a baseline surrogate population: ROI, RED, and quasi-experimental analyses of opt-in or upstream programs that compare participant outcomes to outcomes of eligible nonparticipants or ineligible customers located in other service territories. However, the most common approach to net savings estimation in the USA is to adjust site-specific estimates of measure savings by an estimated net-to-gross ratio derived from survey methods which solicit responses from participants, nonparticipants, and trade allies to hypothetical questions regarding market behavior under the counterfactual condition of no program implementation.

A key advantage of the survey approach is that separate estimation of site measure savings and net-to-gross factors avoids the problem of confounding and therefore can produce valid estimates of individual measure and gross program energy savings. Furthermore, as discussed above, site M&V data can be used to identify the sources of low or high gross realization rates, by comparing the tracking assumptions used to calculate projected measure savings to observed values of equipment-rated capacity, thermostat settings, building occupancy, and other determinants of end-use service demand. Analysis of discrepancies in these assumptions also enables quantification of their separate contributions to the value of the realization rate. However, like other nonexperimental methods, the use of interview data to derive estimates of net-to-gross adjustments depends on the untestable assumption that program participant responses to hypothetical questions can serve as a valid substitute for observations of baseline market behavior; nevertheless, the bias is confined to the quantification of net impact.

In conclusion, these findings refute the presumption of validity accorded to methods of attribution that do not effectively randomize the exposure to the influence of the evaluated program as designed. This presumption has been sustained by an understandable reluctance of evaluators to raise fundamental questions about evaluation methods that have been in common use and generally approved by regulators in jurisdictions that require estimation of net savings to assess program performance. Nevertheless, it is incumbent upon evaluators to advise program administrators and regulators against unwarranted interpretations of impact evaluation findings which may misinform program implementation and funding decisions that are not aligned with public policy objectives. Evaluators and regulators are further encouraged to work collaboratively with other interested parties toward a reorientation of evaluation protocols and reallocation of EM&V resources to make better use of ratepayer EM&V funding. Evaluators could propose an overall shift in emphasis from production of summative net impact metrics to expanded site measurement and verification that can provide a more comprehensive understanding of potential measure savings opportunities in diverse segments of customer populations. This highly disaggregated information could be complemented by focused experimental studies of alternative program design elements to provide a strong empirical basis for enhanced program design and the development of portfolio plans which advance the long-term policy objectives of energy efficiency programs.

Data availability No empirical research was conducted to support the analysis and conclusions presented in this article.

Declarations The author declares no competing interests.

References

Abadie, A., & Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics, 10*, 465–503.

Agnew, K. & Goldberg, M. (2017). Chapter 8: Whole-building retrofit with consumption data analysis evaluation protocol, the uniform methods project: Methods for determining energy- efficiency savings for specific measures. National Renewable Energy Laboratory. NREL/SR-7A40–68564. http://www.nrel.gov/docs/fy17osti/68564.pdf

Angrist, J., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*, 444–472.

Copas, J., & Li, H. (1997). Inference for non-random samples. *J. r. Statist. Soc. B, 59*(1), 55–95.

Dubin, J., & McFadden, D. (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica, 52*(2), 345–362.

Fisher, R. (1971). *The design of experiments*. The University of Adelaide.

Freedman, D. (2006). Statistical models for causation: What inferential leverage do they provide? *Evaluation Review, 30*(6), 691–713.

Greenland, S., & Robins, J. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology, 15*(3), 413–419.

Greenland, S., & Robins, J. (2009). Identifiability, exchangeability, and epidemiological confounding revisited. *Epidemiologic Perspectives & Innovations, 6*(3), 1–9.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153–161.

Heckman, J. (2005). The scientific model of causality. *Sociological Methodology, 35*, 1–97.

Heckman, J. (2010a). The assumptions underlying evaluation estimators. *Brazilian Review of Econometrics, 30*(2), 369–449.

Heckman, J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature, 48*(2), 356–398.

Heckman, J., & Navarro-Lozano, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics, 86*(1), 30–57.

Heckman, J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions. *Journal of Econometrics, 30*, 239–267.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

Imbens, G. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua. *Journal of Economic Literature, 48*, 399–423.

Imbens, G., & Rubin, D. (2015). *Causal inference for statistics, social and biomedical sciences*. Cambridge University Press.

Kushler, M., Nowak, S., Witte, P. (2014). Examining the net savings issue: A national survey of state policies and practices in the evaluation of ratepayer-funded energy efficiency programs. American Council for an Energy-Efficient Economy (ACEEE) Report Number U1401. https://www.aceee.org/sites/default/files/publications/researchreports/u1401.pdf

Nadel, S. & Keating, K. (1991). Engineering estimates vs. impact evaluation results: How do they compare and why? *Proceedings of the 1991 International Energy Program Evaluation Conference, Chicago.*

Neyman, J. (1923). On the application of probability theory to agricultural experiments essay on principles. Section 9. *Roczniki Nauk Rolniczych Tom X translated in Statistical Science, 5*(4), 465–480.

Pearl, J. (1996). Causation, action and counterfactuals. *Proceedings of TARK V I, 1996*, 51–73.

Pearl, J. (2011). Principal Stratification – A goal or a tool? *The International Journal of Biostatistics, 7*(1), 1–13.

Provencher, B., Vittetoe-Glinsmann, B., Dougherty, A., Randazzo, K., Moffitt, P., Prahl, R. (2013). "Some insights on matching methods in estimating energy savings for an opt-in, behavioral-based energy efficiency program." *Proceedings of the 2013 International Energy Program Evaluation Conference, Chicago*.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal Effects. *Biometrika, 70*(1), 41–55.

Rubin, D. B. (1990). Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science, 5*(4), 472–480.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100*(469), 322–331.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics, 2*(3), 808–840.

Sobel, M. (2005). Discussion: The scientific model of causality. *Sociological Methodology, 35*, 99–133.

Sobel, M. (2009). Chapter 1: The causal inference in randomized and non-randomized studies: The definition, identification, and estimation of causal parameters. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 3–22). SAGE Publications.

Sociology, 25, 659–706.

Stewart, J. & Todd, A. (2017). Chapter 17: Residential behavior protocol, the uniform methods project: Methods for determining energy-efficiency savings for specific measures. National Renewable Energy Laboratory. NREL/SR-7A40–68573. http://www.nrel.gov/docs/fy17osti/68573.pdf

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21.

Train, K. (1994). Estimation of net savings from energy-conservation programs. *Energy, 19*(4), 423–441.

Train, K., Buller, S., Mast, B., Parikh, K., Paquette, E. (1994). "Estimation of net savings for rebate programs: A three-option nested logit approach." Proceedings of the 1994 ACEEE Summer Study on Energy Efficiency in Buildings. Washington, DC: *American Council for an Energy Efficient Economy* (pp 7–239 to 7–248).

Train, K. (1988)."Correcting for self-selection bias in the estimation of audit program impacts." Proceedings of the 1988 ACEEE Summer Study on Energy Efficiency in Buildings. Washington, DC: *American Council for an Energy Efficient Economy* (pp 9–182 to 9–194).

Vine, E., Hall, N., Keating, K. M., Kushler, M., & Prahl, R. (2012). Emerging issues in the evaluation of energy-efficiency programs: The US experience. *Energy Efficiency, 5*, 5–17.

Vine, E., Sullivan, M., Lutzenhiser, L., Blumstein, C., & Miller, B. (2014). Experimentation and the evaluation of energy efficiency programs. *Energy Efficiency, 7*, 627–640.

Violette, D. & Ozog, M. (1989). Correction for self-selection bias: Theory and application. *Proceedings of the 1989 Energy Program Evaluation Conference, Chicago*.

Violette, D. M. & Rathbun, P. (2017). Chapter 21: Estimating net savings – Common practices: Methods for determining energy-efficiency savings for specific measures. National Renewable Energy Laboratory. NREL/SR-7A40–68578. http://www.nrel.gov/docs/fy17osti/68578.pdf

Winship, C. & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25*, 659–706.

Yamamoto, T. (2012). Understanding the past: Statistical analysis of causal attribution. *American Journal of Political Science, 56*(1), 237–256.