DISCUSSION PAPER SERIES

IZA DP No. 397

Evaluating Environmental Programs:
The Perspective of Modern Evaluation Research

Manuel Frondel
Christoph M. Schmidt

November 2001

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Evaluating Environmental Programs:
The Perspective of Modern Evaluation Research

Manuel Frondel
Centre for European Economic Research, Mannheim

Christoph M. Schmidt,
University of Heidelberg, CEPR, London and IZA, Bonn

Discussion Paper No. 397
November 2001

IZA

P.O. Box 7240
D-53072 Bonn
Germany

Tel.: +49-228-3894-0
Fax: +49-228-3894-210
Email: iza@iza.org

IZA Discussion Paper No. 397
November 2001

ABSTRACT

Evaluating Environmental Programs: The Perspective of Modern Evaluation Research[*]

Large-scale environmental programs generally commit substantial societal resources, making the evaluation of their actual effects on the relevant outcomes imperative. As the example of the subsidization of energy-saving appliances illustrates, much of the applied environmental economics literature has yet to confront the problem of proper attribution of effects to underlying causes on a convincing methodological basis. This paper argues that recent results in the econometrics and statistics literature on program evaluation could be utilized to advance considerably in this context. In particular, the construction of a credible counterfactual situation is at the heart of the formal statistical evaluation problem. Even when controlled experiments are not a viable option, appropriate approaches might succeed where traditional empirical strategies fail to uncover the effects of environmental interventions.

JEL Classification: H43, C40, C90

Keywords: Environmental policy, energy-conservation programs, experiments, observational studies, counterfactual

Christoph M. Schmidt
Department of Economics
University of Heidelberg
Grabengasse 14
69117 Heidelberg
Germany
Tel.: +49-6221-542955
Fax: +49-6221-543640
Email schmidt@uni-hd.de

[*] We are grateful to Herbert Buscher and Andreas Ziegler for a careful reading of the manuscript.

Non-technical Summary. Large-scale environmental programs generally commit substantial societal resources, making the evaluation of their actual effects on the relevant outcomes imperative. As the example of the subsidization of energy-saving appliances illustrates, much of the applied environmental economics literature has yet to confront the problem of proper attribution of effects to underlying causes on a convincing methodological basis. This paper argues that recent results in the econometrics and statistics literature on program evaluation could be utilized to advance considerably in this context. In particular, the construction of a credible counterfactual situation is at the heart of the formal statistical evaluation problem. To illustrate our arguments with a practical example, we consider energy-conservation programs of the type typically implemented and publicly subsidized by electricity utilities throughout the US. Participants in the program are offered new and more efficient appliances such as energy-saving light bulbs or refrigerators at a subsidized price. The new technology, it is hoped, will replace old-fashioned energy-inefficient technology which is available in the majority of households at the time the program commences. Naturally, before funding a program on a large scale, decision makers would like to know whether or not such an environmental intervention is successful in terms of, first, ecological efficacy and, second, economic efficiency. Several approaches to the first issue, suggested by modern evaluation research, are surveyed in this essay, experimental and observational. In an ideal experiment with a randomized assignment of households into and out of treatment the simple difference in mean outcomes (electricity consumptions) across treatment and control groups would yield an unbiased estimate of the true evaluation parameter, even without particular attention to observable characteristics. Hence, whenever possible, one should consider conducting an experimental study, where the randomized control group solves the problem of identifying the counterfactual, and should collect experimental evidence: A randomized experiment is the most convincing approach to the identification and self-selection problem. Performed appropriately, observational approaches are powerful competitors to experimental studies, not only because experimentation is sometimes not possible. In any case, however, environmental regulators and utilities have to work more closely together with researchers already at the stage of designing the interventions.

1 Introduction

A naive recipient of any of the publicized results on the efficacy of environmental interventions[1] would certainly expect these assertions to rest on scientific analysis. Yet, evaluating policy interventions scientifically on the basis of publicly accessible data with appropriately chosen research methods such as randomized experiments or non-experimental difference-in-differences approaches is unfortunately far from being the standard in the debate on environmental policy. Current empirical research in environmental economics concentrates on a few selected issues, such as the prediction of emission levels - either on the basis of computable general equilibrium models or on the basis of reduced form regressions or the estimation of aggregate production functions, with energy as one of the inputs.

Thus, it seems that further development of empirical strategies in environmental economics is set to follow the course outlined by other applied fields - rising emphasis of micro data over aggregate time series data, and increasing use of structural econometric models applied to general-purpose individual-level data. These changes in emphasis on empirical research occur at a time, though, in which precisely this menu of traditional econometric techniques faces mounting scepticism regarding its credibility. Answers to this sceptic view have included a shifting emphasis away from structuralist to quasi-experimental approaches, the incidental collection of data material appropriate for the research questions, and a careful analysis of measurement issues (prominent proponents of this view in the labor economics literature are, for instance, Angrist and Krueger 1999).

This paper attempts to embed empirical research in environmental economics into this current discussion, with particular focus on the evaluation of environmental conservation programs. The ultimate economic evaluation of an environmental intervention requires the assessment of both ecological efficacy and the cost – direct as well as indirect – of the intervention. The issue of economic efficiency (benefits per $ spent) of policy

[1] Energy taxes on fuels like oil or gas, often called CO_2-taxes, are such interventions as well as subsidies for installing solar-energy and photovoltaic applications.

interventions, though, cannot be addressed unless questions of efficacy are solved appropriately: While the determination of cost often poses its own, complicated problems, they are easily gauged in many circumstances. Yet, finding a confident value of the ecological effect, is nearly always a serious intellectual challenge. We thus focus on program efficacy and ignore issues of program cost and economic efficiency for the purposes of this paper. In discussing the potential and limits of recent advances in evaluation strategies, this contribution emphasizes the necessity of constructing a credible counterfactual situation without the intervention, a theme that has successfully been discussed in recent contributions to other fields of economics (see Heckman *et al.* 1999 or Schmidt 1999).

Contrary to the spirit in much of the published accounts on the results of environmental interventions, e. g. Lovins (1985,1988), but in line with some critics, e. g. Wirl (2000), we argue in Section 2 of this paper that the decisive shortcoming in the current assessment of environmental policies is the flagrant lack of *scientific* evaluation. Section 3 provides a formal account of the *evaluation problem* as it is stated by modern evaluation research, supplemented by intuitive explanation. In that section, it will be clarified that the essential task for any evaluation analysis is the construction of a credible counterfactual situation – a precise statement of what economic agents would have done in the absence of the policy intervention. In deviation from the previous literature, Section 4 proceeds to introduce several approaches proposed in the literature to solve this *identification problem*, both non-experimental approaches and controlled experiments. These approaches are placed into perspective by outlining explicitly the respective underlying identification assumptions that have to be made to justify their application. The fifth section concludes.

2 Scientific Evaluation of Environmental Programs

To illustrate our arguments with a practical example, consider energy-conservation programs of the type typically implemented and publicly subsidized by electricity utilities throughout the US. Participants in the program are offered new and more efficient appliances such as energy-saving light bulbs or refrigerators at a subsidized price, sometimes accompanied by free technological advice. The new technology, it is hoped, will replace old-fashioned energy-inefficient technology which is available in the majority of households at the time the program commences. Naturally, before funding a program on a large scale, decision makers would like to know whether or not such an environmental intervention is successful in terms of, first, ecological efficacy and, second, economic efficiency.

At the present time, the typical way to approach these questions is to resort to plausibility considerations, more or less supported by explicit economic theory. Disconcertingly, there are plausible arguments for both supporting and rejecting the implementation of this example's policy intervention. Lovins' (1985) theoretical argument that "a kilowatt-hour saved is just like a kilowatt-hour generated" seems to confirm the efficacy of energy-conservation programs, while Wirl (1997) casts doubt on this equivalence. He argues that a higher technological efficiency increases the demand for energy services. This so-called *rebound effect* reduces the amount of energy saved and, hence, the efficacy of the program. In principle, it is not excluded that more efficient appliances even may increase the total energy consumption (*energy-saving paradox*, see Wirl (1989)).

The rebound effect of energy efficiency improvements was first mentioned by Khazzoom (1980). Its importance for energy-policy interventions has been hotly debated ever since: While Lovins (1988) maintains that the rebound effect is so insignificant that it can safely be neglected, Khazzoom (1987,1989) argues that it might be so large as to nearly defeat the purpose of energy efficiency improvements. Hence, the magnitude of the rebound effect is the key to the absolute effectiveness of technological efficiency improvements and relative to energy price or tax policies in reducing energy use. This debate is but one example showing that the evaluation of environmental programs such as energy-conservation programs is far from being an easy matter: "Most serious analysts recognize that it is quite difficult to measure accurately the energy savings resulting from utility conservation efforts" (Joskow and Marron 1992:62).

An alternative approach which suggests itself would be to rely on the advice of experienced practitioners, physicists or engineers for example, who were involved in the development of the energy-saving technique in the past: "[E]nergy savings are frequently based on engineering estimates rather than adjusted by ex post measurement of actual changes in consumer behavior." (Joskow and Marron 1992:50). Behavioral changes of consumers, as reflected by the rebound effect, remain unassessed and, thus, are not considered. Therefore, estimates of engineers rather represent the maximum program effect with respect to ecological efficacy: "It has been common for ex post evaluations to reveal that [...] engineering estimates were significantly overstated. [...] In a survey of 42 program evaluations Nadel and Keating (1991) found that ex post impact evaluations estimated lower savings than ex ante engineering projections about 75 % of the time"(Joskow and Marron 1992:63).

Estimating energy savings is thus clearly an empirical issue. It might be addressed by the following econometric model, describing household i's energy demand Y_{ti} at time t for given observable household characteristics $\mathbf{X}_{ti}$ like family size etc.:

$$Y_{ti} = \boldsymbol{\beta}^T \mathbf{X}_{ti} - d_i \cdot p_t + \varepsilon_{it}, \tag{1}$$

where

$$\varepsilon_{ti} = \nu_i + \eta_{ti} \qquad \text{and} \quad E(\nu_i) = 0, E(\eta_{ti}) = 0, E(\nu_i \eta_{ti}) = 0 \quad \text{for all } i, t. \tag{2}$$

d_i in model (1) reflects household-specific price effects on electricity demand, while, as a part of the error term ε_{ti}, ν_i captures unobservable household characteristics such as environmental consciousness. A low environmental consciousness, for example, would imply a high ν_i. Energy-efficiency improvements achieved by participation in conservation programs offering subsidized appliances result in a reduced demand for electricity, but also in declining prices for each unit of energy services. This price effect may reduce the theoretical maximum amount of electricity savings (rebound effect).

Unfortunately, participation in the program is not the choice of all customers to whom the offer of program participation is made. By contrast, it seems likely that the reluctance for taking up such an offer is higher for consumers with a low environmental consciousness (high ν_i) and a low price-response of electricity demand (low d_i). That is, ν_i and d_i might be correlated negatively: Low ν_i and high d_i should characterize the type of households which is more likely to participate in the program. In other words, there is self-selection. In Section 4 below, possible experimental or non-experimental approaches are discussed regarding their potential to overcome the problems due to unobservable characteristics ν_i and of self-selection.

Clearly, the modern environmental regulator can rely on a large in-house staff to address the evaluation problem directly. Yet, more often than not the complexity of this task and recent advances in its treatment are not appreciated – emphasis is on technological expertise rather than on econometric or behavioral analysis. Moreover, even these caveats set aside, no strategy exists for ascertaining the true efficacy of the program *directly*. For instance, one might simply survey participants of a pilot program and ask them how, in their perception, program participation has reduced their energy consumption. The answer to this question might be less informative than desired, though: Even if answers are completely truthful, voluntary participation in the program might be a reflection of, say, energy consciousness. These respondents might have invested in better technology irrespective of the program. In other words, the respondents would not be representative of the population for which a large-scale implementation of the environmental intervention is at issue, and this problem carries over to their responses.

As a consequence, the decision of whether or not it is worthwhile to fund the program on a large scale can hardly be decided with confidence without turning to scientific program evaluation. That is, researchers, ideally being independent, should – on the basis of their own, openly discussed research strategy (see Section 4 below) – analyze publicly available data and present their data, methods, and results openly for debate. It is the weight of the evidence derived under these well-accepted standards which should be the basis for any conclusions, not infatuation with technical advances or political convictions. It is the overarching theme of this contribution that there is no real alternative to this ideal, since attributing an effect to an underlying cause with considerable confidence is a task that is far more complex than is generally appreciated. But this is the nature of the problem, not the deficiency of the tools addressing it.

At best, the efficacy of a program can only be estimated with confidence, but never measured with certainty: In all instances, it requires the construction of a plausible counterfactual situation – identical to what is observed in the absence of the intervention – against which the actual situation has to be compared. That is, counterfactuals are principally unobservable and have to be replaced by invoking specific identification assumptions and constructing plausible counterfactual situations. This is the *evaluation problem* that will be stated more formally in the following section.

3 The Evaluation Problem

In recent years the evaluation literature in statistics and econometrics has developed a unified formal framework that facilitates the exploration of the potential and the limits of evaluation strategies, following origins for instance to be found in Rubin (1974). To illustrate the formal ideas, throughout the following sections, we retain the example of an energy-conservation program, the subsidized provision of energy-efficient appliances, which is implemented by electricity utilities. The relevant units of observation are individual households which may or may not participate in the program, that is, may or may not purchase subsidized energy-efficient appliances. Participation is voluntary and depends upon the efficiency of households' energy-consumption technology. The energy-conservation program might be part of the provider's demand-side management which aims at avoiding the construction of new expensive power plants. Instead, households are offered more efficient electricity-consumption technology at a subsidized price, energy-saving light bulbs or refrigerators for example. Obviously, the relevant outcome to investigate is the households' energy consumption in the form of electricity.

In this section, we address the problem in a sufficiently abstract way so as to appreciate the arguments made by modern evaluation research. The formal setup describes each household in the realm of the program under scrutiny by several key characteristics. For this purpose, denote the state associated with *receiving* the intervention (that is, again, receiving the offer to purchase a subsidized appliance) by "1", and the state associated with *not receiving* the intervention by "0". Receiving the intervention is indicated by the indicator variable D_i. That is, if household i receives subsidized access to the new technology under the energy-conservation program, then $D_i = 1$. What we would like to compare is what would happen to household i's energy consumption if i participated in an energy-conservation program, that is, if it had received the treatment $(D_i = 1)$ as well as if i did not $(D_i = 0)$.

Specifically, the energy consumption in post-treatment period t is denoted by Y_{ti} if household i did not receive treatment, and by $Y_{ti} - \Delta_i$ if it received treatment. This setup directly allows the formulation of the *causal impact* of the energy-conservation program on household i as Δ_i, the unit-level treatment effect. The discussion here is deliberately confined to a single pre-intervention period indicated by t' and a single post-intervention period t. Thus, the causal impact of the intervention is written without time-subscript as Δ_i. In a general setting one might very well consider several post-treatment periods. Naturally, this concentration on a single household requires that the effect of the program on each household i is not affected by the participation decision of any other household. In the statistics literature, it is referred to as the *stable unit treatment value assumption.* Its validity facilitates a manageable formal setup. Nevertheless, in many practical applications it might be questionable whether it holds.

Unfortunately, and this is the core of the evaluation problem, we can never observe both potential energy consumption outcomes Y_{ti} and $Y_{ti} - \Delta_i$ simultaneously for a given household, since it can either participate in the program or not. Instead, merely one of these two outcome variables can actually be observed for each household. That is, for those households who do participate in the program, the outcome $Y_{ti} - \Delta_i$ is to be observed, whereas the outcome Y_{ti} is the *counterfactual outcome* which, inherently, is not observable. On the other hand, in the population of non-participants merely Y_{ti} can be observed, while the counterfactual outcomes $Y_{ti} - \Delta_i$ are basically not available. It is program participation, that is, the value of D_i that decides which of either outcomes, Y_{ti} or $Y_{ti} - \Delta_i$ will be observed.

From the perspective of the analyst, the available data comprise, in addition to observed outcomes Y_{ti} or $Y_{ti} - \Delta_i$, a set of (time-invariant) household characteristics X_i and pre-intervention outcomes $Y_{t'i}$, and the indicator of treatment D_i. Suppose, for instance, that X_i captures the number or persons living in household i, thus taking on a discrete number of values: $k = 1, 2, 3, 4, ...$. In general, the higher is the number of household members the larger is a household's energy consumption. Since for uncovering the program effect it is crucial not to condition on variables that themselves are outcomes of the treatment, we impose the *exogeneity* of these conditioning characteristics: the participation in the program must not alter the value of (X_i, $Y_{t'i}$) for any household i. Note that, in principle, the Δ_i vary across households, even across those sharing the same number of household members X_i and the same pre-intervention consumption $Y_{t'i}$. Some households will display a lower energy consumption as a result of treatment, others might consume the same or even a higher amount of energy (energy-saving paradox).

Definitely, due to the inherent problems of observability there will never be an opportunity to estimate individual program impacts upon household consumption with confidence. Yet, one might still hope to be able to assess the population average of gains from treatment, since we know that the population averages $E(.)$ of the frequency distributions of $Y_{ti} - \Delta_i$ and Y_{ti} can be estimated for participants and non-participants, respectively. Interest in program evaluation is therefore on specific *evaluation parameters* measuring appropriately average individual behavioral changes resulting from treatment.

The most prominent evaluation parameter is the so-called *mean effect of treatment on the treated*,

$$M_{X=k} := E(Y_t - \Delta | X = k, D = 1) - E(Y_t | X = k, D = 1), \tag{3}$$

conditional on the specific realization of the exogenous variables, where in our example $k = 1, 2, 3, 4, ...$ denotes the number of household members. In definition (3) and henceforth, individual subscripts are dropped to reflect the focus on population averages. The mean effect of treatment on the treated appropriately summarizes the individual behavioral changes in the population of those households with $X = k$ who do receive the treatment, without restricting a priori the impact of program participation to be the same across households. If the environmental intervention indeed exerts an effect, $M_{X=k}$ should be a negative number.

Ultimate interest of program evaluation, specifically in our example, might be in the average treatment effect over all *relevant* values of X given $D = 1$. On the basis of definition (3), this average effect over all types of households is given by

$$M := \sum_k M_{X=k} \frac{\Pr(X = k \mid D = 1)}{\sum_k \Pr(X = k \mid D = 1)}. \tag{4}$$

This is nothing else but a weighted average of the conditional (on X) program effects, with the weights being the relative frequencies of the different household groups in the population of program participants.

There are alternative evaluation parameters one could be interested in, for instance the *mean effects of treatment on individuals randomly drawn from the population* of households,

$$\widetilde{M}_{X=k} := E(Y_t - \Delta | X = k) - E(Y_t | X = k), \tag{5}$$

also conditional on exogenous variables. Note that the major difference to the evaluation parameter M lies in the omission of conditioning on $D = 1$. Thus this would be the appropriate evaluation parameter in a mandatory program. Parameter $\tilde{M}$ will not be discussed further in this paper, since in our specific example participation in the environmental program is voluntary.

Like any other population parameter, the population average $E(Y_t - \Delta | X = k, D = 1)$ is unknown in general and thus has to be estimated from a sample of limited size, that is, the households involved in the pilot project. Yet, whenever a sample is used to estimate a population average, it is unlikely that this estimate will exactly mimic the true population parameter itself. Rather, the estimate can only give an approximation to the true parameter, since it has been derived on the basis of only a subset of the population. An estimation strategy is successful if the approximation tends to become more and more exact while the sample taken from the population becomes larger and larger. In the limit, the approximation should be indistinguishable from the true parameter. In that case when a population parameter could be estimated correctly with infinite precision by collecting abundantly many observations from the underlying population, the parameter is said to be *identified* from observable data.

While one could, at least in principle, estimate $E(Y_t - \Delta | X = k, D = 1)$ with infinite precision from the available data on program participants, one could not even hypothetically estimate the population average $E(Y_t | X = k, D = 1)$: No sample size would alleviate the fact that, principally, Y_t cannot be observed for participants – they "produce" only data for the variable $Y_t - \Delta$. That is, merely the population average $E(Y_t - \Delta | X = k, D = 1)$ but not $E(Y_t | X = k, D = 1)$ is identified from observable data.

This clarifies the nature of the fundamental *evaluation problem*. It is the problem of *finding* an appropriate *identification assumption* that allows replacing the counterfactual population average $E(Y_t | X = k, D = 1)$ in definition (3) with an entity that is identified from observable data. $E(Y_t | X = k, D = 1)$ is a counterfactual because it indicates what would have happened to participants, on average, if they had not participated in the program. It is a problem that cannot be solved by collecting more of the same data or by refined measurement of the observed variables. It can only be resolved by finding a plausible comparison group.

In *before-after comparisons*, for instance, the comparison group is the group of participants themselves, yet before the program was implemented. $E(Y_t | X = k, D = 1)$, the principally undetectable population average is replaced by $E(Y_{t'} | X = k, D = 1)$ in before-after comparisons. That is,

$$E(Y_t | X = k, D = 1) = E(Y_{t'} | X = k, D = 1) \tag{6}$$

is the identification assumption absolutely necessary for this approach. Yet, the identification assumption might be wrong. There might be unobservable factors affecting the behavior of all households such as the number of rainy days which differ from t' to t. This would invalidate the before-after comparison, since then equation (6) is not fulfilled for at least one of the relevant k.

The subsequent section documents the precise identification assumptions required by the various principal empirical strategies suggested in the literature, and discusses potential reasons for their failure. Following one of the fundamental statistical principles that sample means are natural estimates of population averages, this section formulates the problem of estimating the desired evaluation parameter as a problem of analogy. Actual estimation in the sample of households collected in the pilot project is formulated as the appropriate adaption of the identification assumption in the sample data.

If the population parameter of interest is in fact identified from observable data, there will nevertheless be noise around the estimate in each and every practical application. This noise would vanish in the hypothetical case of an abundantly large sample, though. Noise is therefore not a conceptual hurdle for finding the correct population average. Rather, what is decisive on a conceptual level, is the choice of identification strategy. Therefore, in what follows, each evaluation approach will be characterized by the identification assumption that justifies its application. Since there does not seem to be a straightforward way to solve the evaluation problem, a variety of relevant approaches is introduced. All approaches one could think of are prone to some error, though. Certainly, and first of all, one would therefore have to develop an informed judgement which of all possible methods would likely generate the smallest errors.

4 Empirical Evaluation Approaches

In the course of this section, it becomes apparent that under many circumstances, for the evaluation of environmental interventions, the *randomized controlled trial* is the most desirable empirical strategy. While experimental analyses are rather the rule than the exception in the natural sciences, energy policy interventions designed to mitigate global climate change, for example, often do not lend themselves to controlled implementation. Even more often environmental programs are implemented before a controlled experiment can be designed and executed. In these cases researchers have to utilize non-experimental or *observational* approaches to evaluation – a seminal source is Rosenbaum (1995). We first discuss the menu of observational approaches before we assess the potential and the limits of experimental analyses of environmental interventions.

In what follows N_1 is the number of households in the sample of participating households, with indices $i \in I_1$. The sample of non-participants consists of N_0 households, with indices $j \in I_0$. Subsets of these samples are denoted in a straightforward fashion. For instance, the number of participating households with three members is $N_{1,X=3}$, the set of indices of all non-participant households with four members is $I_{0,X=4}$. Accordingly, the corresponding number of observations is $N_{0,X=4}$.

4.1 Observational Studies

What is collected in our specific example is an account of the actual energy use of individual households after the implementation of the pilot project. For participants this means observation of $Y_{ti} - \Delta_i$, for non-participants observation of Y_{ti}. In order to estimate the program effect, one would then like to calculate

$$\widehat{M}_{X=k} := \frac{1}{N_{1,X=k}} \sum_{i \in I_1, X_i = k} (Y_{ti} - \Delta_i) - \frac{1}{N_{1,X=k}} \sum_{i \in I_1, X_i = k} Y_{ti} \ , \tag{7}$$

which would be an estimator of the mean effect of treatment on the treated defined in (3). Yet, the fundamental evaluation problem is that Y_{ti} is not observable for participants and $\frac{1}{N_{1,X=k}} \sum_{i \in I_{1,X_i=k}} Y_{ti}$ is thus unknown. To solve this problem, this inherently undetectable mean has to be replaced by an observable average. The kind of replacement and its respective underlying identification assumption are uniquely characterizing any evaluation approach. The following subsections introduce four observational evaluation approaches, the respective identification assumptions necessary to justify their application, and possible reasons for failures of these approaches with particular respect to environmental program evaluation.

4.1.1 Before-After Comparisons

Perhaps the most common evaluation strategy for attempting the construction of a plausible counterfactual is a *before-after comparison*, a comparison of treated households with themselves at the pre-intervention period t'. Necessarily based on longitudinal data, information on the effects of the program is then extracted exclusively from changes between the pre-intervention period t' and post-treatment period t: Frequently, for simplicity, just the difference between pre- and post-participation energy use in kWh/year are calculated on the basis of electricity bills (see U. S. 1993:107). That is, in practice, in the second sum of fundamental definition (7), the inherently unobservable Y_{ti} is replaced by $Y_{t'i}$, household i's energy consumption before participating in the energy-conservation program. Thus, in order to estimate the impact of treatment on households of size k, one forms the following average over all pairs of before-after observations for the sample of participating households:

$$\widehat{M}_{X=k}^{b-a} := \frac{1}{N_{1,X=k}} \sum_{i \in I_1, X_i=k} (Y_{ti} - \Delta_i) - \frac{1}{N_{1,X=k}} \sum_{i \in I_1, X_i=k} Y_{t'i}. \tag{8}$$

This is the before-after estimator of the mean effect of treatment on the treated.[2] Note that persistent unobservable household characteristics, captured in (1) by error term ν_i, do not harm this estimator, since ν_i is just differenced out in (8) .

The identification assumption underlying estimator (8), is, in formal terms,

$$E(Y_t|X = k, D = 1) = E(Y_{t'}|X = k, D = 1). \tag{9}$$

That means, taken over the population of all treated households, the population average of actual outcomes in period t' is equal to the population average of what these households would have experienced had they not participated in the program in period t, that is, equal to the average of counterfactual outcomes in t in the absence of treatment. Assumption (9) is a necessary, though not a sufficient, precondition for $\frac{1}{N_{1,X=k}} \sum_{i \in I_1 X_i=k} Y_{t'i}$ to equal $\frac{1}{N_{1,X=k}} \sum_{i \in I_1 X_i=k} Y_{ti}$: If assumption (9) does not hold, both means will not even be equal in the limiting case of an infinitely large sample. At the individual level, $Y_{t'i}$ will not exactly equal Y_{ti} for any participating household i.

In our example, by identification assumption (9) it is supposed that the average energy consumption of the population of participating households would not have changed over time if the energy-conservation program had not been implemented. That is, identification assumption (9) requires no less in our example than stability of the average

[2] Adapting this formula to the case where household sizes vary between the periods t' and t is completely straightforward. This applies also to the estimators introduced in Sections 4.1.3 and 4.1.4.

consumption behavior. This stability might be disturbed, for example, by changes in the environment: If period t' was a a rainy period, say, and period t contained many sunny days, then, speaking in terms of an individual household i, the counterfactual consumption Y_{ti} in the absence of participation might be lower than the consumption in the pre-treatment period t' because of the environmental conditions – yet, estimator (8) would attribute the reduction of average consumption to the intervention and, hence, would overestimate the effect of the program.

In addition to environmental stability, identification assumption (9) of the before-after comparison approach necessitates the further assumption that the outcomes before treatment are not influenced by anticipation of the intervention. If households of the treatment sample know already in t' that they may, for instance, participate in a program in which a subsidized energy-saving technology will be provided, these households will hardly invest into new technology in t' at their own cost – a strategic behavior characterized by Wirl (2000:96) as moral hazard. A high average $Y_{t'i}$ would be the consequence among participating households, and this would lead to an overstated estimate $\widehat{M}^{b-a}_{X=k}$. If those households had not perceived the benefits of the program when deciding on their consumption pattern in period t', they might have invested into better technology and, as a consequence, would probably not have participated in the program. The issue of *program anticipation* will be discussed repeatedly in this section due to its fundamental potential to affect empirical evaluation strategies based on *longitudinal data*.

4.1.2 Cross-Section Estimators

In before-after comparisons, program participants serve as their own controls. Following this approach might be precluded, either because no longitudinal information on participants is available or because environmental or macroeconomic conditions fluctuate substantially over time, thus contaminating identification assumption (9). In that case, in effect, the mean of the observed outcome of non-participants could serve as the entity to replace the mean of the unobservable Y_{ti} for participants in the basic estimator (7). The intervention impact is then estimated by the cross-section estimator

$$\widehat{M}_{X=k}^{c-s} := \frac{1}{N_{1,X=k}} \sum_{i \in I_1, X_i = k} (Y_{ti} - \Delta_i) - \frac{1}{N_{0,X=k}} \sum_{j \in I_0, X_i = k} Y_{tj}. \tag{10}$$

The formal statement of the identification condition underlying estimator (10) would read

$$E(Y_t | X = k, D = 1) = E(Y_t | X = k, D = 0). \tag{11}$$

That is, although the populations of participants and non-participants might be quite different in size and within each of those populations the outcomes for the no-treatment state might differ widely, on average, these are cancelling out. For identification assumption (11) to be valid, selection into treatment has to be statistically independent of the no-treatment outcome Y_t given X. That is, no unobservable factor such as "environmental consciousness" should motivate individual households to participate. This factor would in our example increase the desire to participate in an energy-conservation program but would also decrease the (counterfactual) no-treatment energy-consumption if the program had not been implemented and energy-conscious households had substantively invested into better technology. In this instance, equation (11), the condition for invoking the idea of *exogenous selection*, cannot be assumed to hold. The cross-section estimator would then overestimate $M_{X=k}$, that is, there would be *selection bias.* According to Wirl (2000:95), this "adverse selection" of participants, who do not represent the average of the consumers, results in little conservation.

4.1.3 Difference-in-Differences Estimation

An alternative approach that retains the idea of using non-participants as a control group, yet takes a longitudinal perspective is the *difference-in-differences* approach. As opposed to the before-after comparison approach, apparently, it has never been employed in the practical evaluation of conservation programs despite the advantage that it accounts for changes for instance in the environmental and macroeconomic conditions between t' and t. That is, this evaluation strategy does not require intertemporal stability which, by identification assumption (9), is explicitly assumed in the before-after comparison approach.

The difference-in-differences estimator compares the sample averages of the changes in outcomes for random samples of participating and non-participating households (with individual characteristics X_i),

$$\widehat{M}^{d-d}_{X=k} := \frac{1}{N_{1,X=k}} \sum_{i \in I_1, X_i = k} ((Y_{ti} - \Delta_i) - Y_{t'i}) - \frac{1}{N_{0,X=k}} \sum_{j \in I_0, X_i = k} (Y_{tj} - Y_{t'j}). \tag{12}$$

The corresponding identification assumption underlying this estimator postulates a population analogue of this idea: The population average of the change in the no-program outcome of participating households between t' and t is equal to that experienced by non-participating households,

$$E(Y_t - Y_{t'}|X = k, D = 1) = E(Y_t - Y_{t'}|X = k, D = 0). \tag{13}$$

That is, while the cross-sectional perspective accounts for the general environment, the longitudinal perspective ensures that persistent household heterogeneity is differenced out.

What might be problematic, though, is the relatively restrictive form this estimator imposes upon the effects of unobserved heterogeneity on energy-consumption change. On average, the milder climate of period t might induce energy-conscious households (likely to participate in the program) to save some energy even in the absence of the intervention, but there might be little room for changes. On the other hand, energy-wasting households might reduce their generous consumption quite a bit in period t. The difference-in-differences estimator will underestimate the impact of the program by assuming that the counterfactual change of participating households can be approximated by those of non-participants. This potential shortcoming is addressed in the exact matching approach discussed next.

4.1.4 Accounting for the History of Energy Use

The principal idea of accounting for the history of energy use is to assign to one or more of the households with equal specific characteristics (X, $Y_{t\prime}$) in the intervention sample as matching partners one or more individuals from the non-experimental control sample who are similar in terms of their observed characteristics. That is, the exact matching procedure specifies the most general possible model of post-intervention outcomes in terms of the observable data (pre-intervention energy use histories and number of household members, say). For any population cell (X, $Y_{t\prime}$) for which at least one match could be found, we estimate the impact of the intervention within this cell by a comparison of sample averages. The desired estimate of the program impact $M_{X=k}$ is thus given by a weighted average over these sample means,

$$\widehat{M}_{X=k}^{match} := \frac{1}{N_{1,X=k}} \sum_{Y_{t'}} N_{1,X=k,Y_{t'}} \left[\frac{1}{N_{1,X=k,Y_{t'}}} \sum_{i \in I_1, X_i=k, Y_{t'}} (Y_{ti} - \Delta_i) - \frac{1}{N_{0,X=k,Y_{t'}}} \sum_{j \in I_0, X_i=k, Y_{t'}} Y_{tj} \right] . \tag{14}$$

$N_{1,X=k,Y_{t'}}$ is the number of individuals with characteristics $X = k$ and pre-treatment outcome $Y_{t'}$ who receive the intervention ($N_{1,X=k} := \sum_{Y_{t'}} N_{1,X=k,Y_{t'}}$) and $N_{0,X=k,Y_{t'}}$ is the corresponding number of control observations.

The central identification assumption underlying estimator (14) is: For households that are characterized by any specific configuration of observable characteristics, in particular energy consumption in pre-program period t' the participation decision is independent of any unobservable determinant of the post-intervention no-treatment outcome. This assumption implies

$$E(Y_t|X = k, Y_{t'} = y_{t'}, D = 1) = E(Y_t|X = k, Y_{t'} = y_{t'}, D = 0). \tag{15}$$

Similar to the difference-in-differences approach, intertemporal changes in environmental and economic conditions that affect, on average, the change between t' and t equally for participants and non-participants, as long as they shared a common set of characteristics ($X = k$, $Y_{t'} = y_{t'}$), would not be consequential. Exact matching is less restrictive than the difference-in-differences approach, though, since the average change is allowed to differ across various values of $Y_{t'}$.

The major difference of the exact matching procedure to the difference-in-differences approach is that it conditions on the *exact* pre-intervention outcome by defining comparable strata of the population. The exact matching estimator is thus able to account for all those persistent unobservable factors that were associated with the pre-treatment outcome, and for their impact on the change of energy consumption over time. However, this strategy is not able to deal with anticipation effects satisfactorily: Households who would invest into better energy-consumption technology in period t' at their own cost but anticipate program participation thus tend to have a relatively high energy-consumption $Y_{t'}$, yet for entirely different reasons than non-participating households with the same $Y_{t'}$. Those participating households would have a relatively low no-treatment energy consumption in period t when they realize their original plan to install new energy-efficient consumption technology. Intertemporal changes in no-treatment energy consumption could thus be different for participating and nonparticipating households, even for those displaying the same $Y_{t'}$. This invalidates identification assumption (15). The impact of the environmental program would then be overestimated by the exact matching estimator.

In sum, the objective of any observational study is to use ex-post information on $Y_t - \Delta$ for participants, on Y_t for non-participants, respectively, and on $Y_{t'}$, if available, in an appropriate way such as to ensure the comparability of treatment and comparison groups by a plausible alternative identification condition. Choosing the appropriate strategy has to depend on outside knowledge of the processes of program implementation and participation, and on data requirements. Furthermore, some of the approaches, in particular exact matching, require relatively large samples for their implementation. Thus, problems of identification will be accompanied by problems of reliable statistical inference. Specifically, exact matching might be precluded due to the problem of relatively small samples in each relevant population cell.

4.2 Experimental Studies

The key concept of any experiment is the *randomized assignment* of households into treatment and control groups. For households who voluntarily would be participants in the subsidized-appliance program ($D_i = 1$), a *random mechanism* decides whether they are in fact allowed to enter the program (indicated by $R_i = 1$) or whether they are excluded from the program instead ($R_i = 0$). This assignment mechanism is a process that is completely beyond the households' control and that also does not discriminate as to who will receive treatment. Thus, interventions are particularly good candidates for experimental evaluation if treatment is delivered on the individual household level with considerable control by the researcher about the delivery and about individual compliance with the program.

Under the fundamental requirement that an experiment completely replicates the energy conservation-promoting intervention that will be implemented in the field, experimental studies generally provide for a convincing approach to the *evaluation problem.* In effect, if sample sizes are sufficiently large, randomization will generate a complete balancing of all relevant observable and unobservable characteristics across treatment and control groups, thus facilitating comparability between experimental treatment and control groups. All individuals in the control group have been applicants to the program but were randomized out. One can then infer the average treatment effect from the difference of the average outcomes of these randomly selected households,

$$\widehat{M}^{trial}_{X=k} := \frac{1}{N_{1,X=k}} \sum_{i \in I_1, X_i = k, R_i = 1} (Y_{ti} - \Delta_i) - \frac{1}{N_{0,X=k}} \sum_{j \in I_0, X_i = k, R_i = 0} Y_{tj}. \qquad (16)$$

As long as the randomization is uncompromised and samples are not outrageously small, there is no need for any sophisticated statistical analysis. Generations of natural scientists have been raised in their training with the conviction that if an experiment needs any statistics, one simply ought to have done a better experiment.

The identification assumption underlying an experimental estimator is

$$E(Y_t|X = k, D = 1, R = 1) = E(Y_t|X = k, D = 1, R = 0). \qquad (17)$$

This property is ensured in a controlled randomized trial by randomizing some households out of the potential treatment group into a control group and by preserving the composition of treatment and control groups by close monitoring as the experiment proceeds. Yet, it might fail, just as the identification assumptions of the observational evaluation estimators might fail. In particular, the process of randomization might disrupt the usual course of affairs (Heckman 1996), making participants particularly energy-conscious. This would lead to an overstatement of the impact of the energy-conservation program.

On balance, however, the advantages of the experimental evaluation of environmental programs are undeniable and dominate the comparative assessment of the relative benefits of experimental and observational approaches: Since the randomization is performed on volunteering households no issue of residual unobserved heterogeneity can taint the conclusions of experimental estimates. In other words, self-selection or, as Wirl (2000) put it, adverse selection cannot contaminate the results of a randomized trial.

5 Conclusions

Whenever a policy intervention is undertaken to alter aspects of household behavior relevant to the environment, a serious evaluation effort is required. The issue of program impact is too complex to be solved by introspection or by a casual glance at the consumption choices of program participants. Moreover, since following a wrong route with confidence but without justifiable reason cannot be a serious option, the only sensible choice for environmental regulators is to embrace the idea of scientific evaluation. The body of literature on program evaluation that has evolved in econometrics, statistics and other scientific disciplines offers a framework for guiding program evaluation. In particular, the fundamental *evaluation problem* is revealed to be a problem of observability, in technical terms, of identification, not simply of generating larger quantities of unsatisfactory data or of devoting more manpower to analyzing the data. Since the question is always how the program altered, on average, the consumption patterns of participating households over and above the hypothetical choices they had made if they had not participated, it is the construction or identification of that *counterfactual* that is at issue (*identification problem*).

Several approaches have been discussed in this essay, experimental and observational. In an ideal experiment with a randomized assignment of households into and out of treatment the simple difference in mean outcomes across treatment and control groups would yield an unbiased estimate of the true evaluation parameter, even without particular attention to observable characteristics. It is merely the finite sample size that provides for some noise, but – due to the very nature of randomized assignment – if sample sizes were to grow beyond any limit, randomization would serve to eliminate this noise completely. Hence, whenever possible, one should consider conducting an experimental study, where the randomized control group solves the problem of identifying the counterfactual, and should collect experimental evidence: A randomized experiment is the most convincing approach to the identification and self-selection problem.

Performed appropriately, observational approaches are powerful competitors to experimental studies, not only because experimentation is sometimes not possible. Observational studies rest on the idea that a suitable comparison of participants with non-participants who are *truly comparable* can lead to a balancing of all relevant observable factors, just as the ideal experiment would. The term "truly comparable" is operational – this is exactly the point where untestable identification assumptions enter the evaluation process. Thus, environmental regulators and utilities have to work more closely together with researchers already at the stage of designing the interventions.

References

Angrist, Josua D., Alan B. Krueger (1999): Empirical Strategies in Labor Economics, Chapter in Ashenfelter, Orley and David Card (eds.): *Handbook of Labor Economics*, Vol. III, Amsterdam et al.: North-Holland.

Heckman, James J. (1996) Randomization as an Instrumental Variable, *Review of Economics and Statistics* **77**, 336-341.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith (1999): The Economics and Econometrics of Active Labor Market Programs, Chapter in: Ashenfelter, Orley and David Card (eds.): *Handbook of Labor Economics*, Vol. III, Amsterdam et al.: North-Holland.

Joskow, Paul L. and Donald B. Marron (1992): What Does a Negawatt Really Cost? Evidence from Utility Conservation Programs. *The Energy Journal* **13(4)**, 41-75.

Khazzoom, Daniel J. (1989) Energy Savings from More Efficient Appliances: A Rejoinder. *The Energy Journal* **10**, 157-165.

Khazzoom, Daniel J. (1987) Energy Saving Resulting from the Adoption of More Efficient Appliances. *The Energy Journal* **8(4)**, 85-89.

Khazzoom, Daniel J. (1980) Economic Implications of Mandated Efficiency in Standards for Household Appliances. *The Energy Journal* **1(1)**, 21-40.

Lovins, Amory B. (1988) Energy Saving Resulting from the Adoption of More Efficient Appliances: Another View. *The Energy Journal* **9(2)**, 155-162.

Lovins, Amory B. (1985) Saving Gigabucks with Negawatts. *Public Utilities Fortnightly* **115**, 19-26.

Nadel, Steven and Kenneth M. Keating (1991) *Engineering Estimates versus Impact Evaluation Results: How Do They Compare and Why?* in: Energy Program Evaluation: Uses, Methods, and Results, Proceedings of the 1991 International Energy Program Evaluation Conference.

Rosenbaum, Paul R. (1995) *Observational Studies*, New York: Springer Series in Statistics.

Rubin, Donald B. (1974) Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, *Journal of Educational Psychology* **66**, 688-701.

Schmidt, Christoph. M. (1999) *Knowing what works – The Case for Rigorous Program Evaluation.* IZA-Discussion Paper No. 77, Institute for the Study of Labor, Bonn.

U.S., Congress (1993) *Energy Efficiency: Challenges and Opportunities for Electric Utilities.* Office of Technology Assessment, OTA-E-561, Washington, DC: U: S. Government Printing Office.

Wirl, Franz (2000) Lessons from Utility Conservation Programs *The Energy Journal* **21(1)**, 87-108.

Wirl, Franz (1989) Analytics of Demand-Side Conservation Programs. *Energy Systems & Policy*, **13**, 285-300.

Wirl, Franz (1997) *The Economics of Conservation Programs.* Kluwer Academic Publishers, London.

IZA Discussion Papers

No.	Author(s)	Title	Area	Date
383	D. Blau E. Tekin	The Determinants and Consequences of Child Care Subsidies for Single Mothers	3	11/01
384	D. Acemoglu J.-S. Pischke	Minimum Wages and On-the-Job Training	1	11/01
385	A. Ichino R. T. Riphahn	The Effect of Employment Protection on Worker Effort: A Comparison of Absenteeism During and After Probation	1	11/01
386	J. Wagner C. Schnabel A. Kölling	Threshold Values in German Labor Law and Job Dynamics in Small Firms: The Case of the Disability Law	3	11/01
387	C. Grund D. Sliwka	The Impact of Wage Increases on Job Satisfaction – Empirical Evidence and Theoretical Implications	1	11/01
388	L. Farrell M. A. Shields	Child Expenditure: The Role of Working Mothers, Lone Parents, Sibling Composition and Household Provision	3	11/01
389	T. Beissinger H. Egger	Dynamic Wage Bargaining if Benefits are Tied to Individual Wages	3	11/01
390	T. Beissinger	The Impact of Labor Market Reforms on Capital Flows, Wages and Unemployment	2	11/01
391	J. T. Addison P. Teixeira	Employment Adjustment in Portugal: Evidence from Aggregate and Firm Data	1	11/01
392	P. Tsakloglou F. Papadopoulos	Identifying Population Groups at High Risk of Social Exclusion: Evidence from the ECHP	3	11/01
393	S. M. Fuess, Jr.	Union Bargaining Power: A View from Japan	2	11/01
394	H. Gersbach A. Schniewind	Awareness of General Equilibrium Effects and Unemployment	2	11/01
395	P. Manzini C. Ponsatí	Stakeholders, Bargaining and Strikes	6	11/01
396	M. A. Shields S. Wheatley Price	Exploring the Economic and Social Determinants of Psychological and Psychosocial Health	5	11/01
397	M. Frondel C. M. Schmidt	Evaluating Environmental Programs: The Perspective of Modern Evaluation Research	6	11/01

An updated list of IZA Discussion Papers is available on the center‘s homepage *www.iza.org.*