Schmidt, Christoph M.

Working Paper

Knowing What Works The Case for Rigorous Program Evaluation

IZA Discussion Papers, No. 77

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Schmidt, Christoph M. (1999) : Knowing What Works The Case for Rigorous Program Evaluation, IZA Discussion Papers, No. 77, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/20911

DISCUSSION PAPER SERIES

IZA DP No. 77

Knowing What Works The Case for Rigorous Program Evaluation

Christoph M. Schmidt

December 1999

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Knowing What Works
The Case for Rigorous Program Evaluation

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

Discussion Paper No. 77
December 1999

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. 77
December 1999

ABSTRACT

Knowing What Works
The Case for Rigorous Program Evaluation *

Since interventions by the public sector generally commit substantial societal resources, the evaluation of effects and costs of policy interventions is imperative. This paper outlines why program evaluation should follow well-respected scientific standards and why it should be performed by independent researchers. Moreover, it outlines the three fundamental elements of evaluation research, the choice of the appropriate outcome measure, the assessment of the direct and indirect cost associated with the intervention, and the attribution of effects to underlying causes. The paper proceeds to outline in intuitive terms that the construction of a credible counterfactual situation is at the heart of the formal statistical evaluation problem. It introduces several approaches, based on both experiments and on non-experimental data, that have been proposed in the literature to solve the evaluation problem, and illustrates them numerically.

JEL Classification: H43, C40, C90

Keywords: Experiments, observational studies, counterfactual

Christoph M. Schmidt
University of Heidelberg
Alfred Weber Institute
Grabengasse 14
69117 Heidelberg
Germany
Fax: +49 6221 543640
Email: Christoph.Schmidt@urz.uni-heidelberg.de

* I am grateful to Boris Augurzky, David Card, Jochen Kluve, Ralph Würthwein, and Klaus F. Zimmermann for their comments and to the Center for Labor Economics, UC Berkeley for its hospitality.

1 The Context of Evaluation

Interventions by the public sector generally commit substantial societal resources that are then unavailable for alternative programs, as well as for private expenditures. The evaluation of effects and costs of an intervention is therefore imperative, with the principal objectives of ranking alternative candidate interventions and gauging their cost-effectiveness. This is especially true in times of moderate economic growth, when there are relatively few uncommited resources. Tight budgets make us painfully aware that policy makers as well as administrators are more and more considered accountable for what happens with the taxpayers' money. For instance, in their recent article for the *Handbook of Labor Economics*, Heckman et al. (1999) write

> "An emphasis on objective publicly accessible evaluations is a distinctive feature of the modern welfare state, especially in an era of limited funds and public demands for accountability."

Far from being a simple matter of accounting, the evaluation of policy interventions faces serious theoretical and methodological problems, though[2]. Recent developments in the field of evaluation research can be of help in identifying these problems and offer some guidance in their solution. The goal of this paper is to outline in intuitive terms what constitutes the *evaluation problem*. Furthermore, the principal solution possibilities offered in the literature are discussed informally.

In the second section, the general problem of program evaluation will be placed into perspective: why does one need *scientific* evaluation, what are the standards according to which evaluation efforts should be judged, and who should perform the analysis? Furthermore, the fundamental elements of evaluation research, (i) the choice of the appropriate outcome measure, (ii) the assessment of the direct and indirect cost associated with the intervention, and (iii) the attribution of effects to underlying causes are briefly discussed. The third section 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

[2] Henceforth the terms *program*, *treatment* and *policy intervention* will be used synonymously.

essential task for any evaluation analysis is the construction of a credible counterfactual – a precise statement of what would have happened in the absence of the policy intervention. In technical terms this is known as the *identification problem*.

The fourth section proceeds to introduce several approaches proposed in the literature to solve the evaluation problem, both experimentally based and observational, and places them into perspective by outlining explicitly the underlying identification assumptions that have to be made to justify their application. The fifth section concludes.

2 Fundamental Aspects of Program Evaluation

2.1 The Requirement for Scientific Evaluation

As a first step, it is important for policy makers to realize why any serious evaluation effort has to follow a set of standards well-accepted in the scientific community – the strict reliance on evidence, a careful statement of data sources, consideration of sources of possible errors in inference, and the standard of publicness –, and cannot be done in-house as an addendum to the usual accounting procedures[3]. As a matter of principle, one could question, and many experienced practitioners will certainly do so, why one should have to scrutinize at all the honest planning, meticulous administration and careful delivery of policy measures, if one can rely on a well-trained and well-intentioned staff and on institutions that have withstood the test of time. Undeniably, though, convincing arguments for this requirement are not in short supply. One major reason is that even the best can be in error. The inception and design of the policy measure might rest on a false premise about the causes of the problem at hand. Moreover, not all practitioners and administrators engaged in implementing policy measures are equally knowledgeable or competent.

Second, one can hardly expect all these participants in the design and delivery of policy measures to be completely impartial as to the importance of their tasks and to the efficacy of the activities they are spending their professional lifes with. Given that the effects of policy measures are often quantitatively small, even slight and inadvertent tendencies to

[3] A lucid introduction into research in the social sciences is given by Katzer et al. (1998).

emphasize positive aspects might invalidate their conclusions. By contrast, scientists are more likely to be impartial to the policy intervention under study. Finally, attributing an effect to an underlying cause with considerable confidence is a task that is far more complex than is generally appreciated: in all instances, it requires the construction of a plausible counterfactual situation – identical to what is observed, apart from the absence of the intervention – against which the actual situation has to be compared. Thus, at best the effect can only be estimated with confidence, but never measured with certainty. Indeed, the tradition of evaluating policy interventions scientifically on the basis of publicly accessible data and with publicized accounts of research methods and evidence, is underdeveloped in most economies. Thus, in countries with little tradition of independent scientific evaluation, previous conclusions about the efficacy and efficiency of public policy measures are often left unsupported by any empirical evidence.

To provide a simple example, consider a situation in which a decision maker has to decide which of two different active labor market programs will receive the sum earmarked in this year's budget for such policies, one being a training program and one being a program of wage subsidies for make-shift work. One of the fundamental principles of any economic problem applies here as well: the budget can only afford to support one of those programs, not both of them (and, for the sake of the argument, there are indivisibilities preventing any compromise solution). How can the decision maker receive the information necessary to make the best decision, which we will assume is implementation of the program that brings more unemployed workers back into stable employment?

One way to approach the question is to resort to plausibility considerations, more or less supported by explicit behavioral, not necessarily economic, theory. The troubling fact is, though, that there are plausible arguments supporting both policy interventions. Training programs intend to enhance the human capital endowments of trainees and, since low human capital seems to be a major source of unemployment risk, more human capital will improve the labor market situation of the workers undergoing the program. This would speak in favor of the first intervention. On the other hand, unemployed workers might foremost need the first step of getting back again into the labor market, that is, finding a new employer. Then they will receive the opportunity to display their favorable characteristics and are

likely to hold on to employment. Moreover, since wage subsidies may apply to more workers than could be supported in a training program, this would argue for undertaking the second intervention. Unfortunately, one might still find many professional economists who would be satisfied with a single plausible and consistent theoretical model; those will be of little help as policy advisors.

A second approach would be to rely on the advice of experienced practitioners who were involved in the implementation of comparable policy interventions in the past. Even leaving questions of impartiality aside – one can hardly expect the advice that both measures should better not be undertaken and the money spent differently altogether – one has to realize that these experiences were usually not obtained in a controlled situation, but rather reflect the idiosyncratic histories of the practitioners as well as the circumstances under which they made their observations of reality. Moreover, instead of objectively evaluating effective output, it is more likely than not that those experts would equate previous success – whose accurate measurement had not been their priority – with efforts spent, both in terms of money and man-hours. As a consequence, the decision of which program to fund can hardly be decided with confidence using either of the first two approaches.

If the policy maker were to take up solving the decision problem on the basis on her own analysis, she will inevitably discover that no strategy exists for ascertaining the true efficacy of either program. For instance, one might survey participants of former training programs and ask them how, in their preception, program participation had helped their labor market situation. Former participants are a very select group, though, as their voluntary participation was undoubtfully a reflection of their desire to improve their situations. They might have excelled in the labor market irrespective of the program. Instead, one might ask currently unemployed what type of program they would prefer if given the choice, and whether their answer is guided by the idea of re-gaining stable employment or by alternative motives. Unfortunately, as the consequences of mis-representation of true intentions are nil, it is unlikely that too much confidence can be placed on the answers to this question. Also, it is unlikely that unemployed workers are any better informed about the true costs and benefits of the alternative interventions than the policy maker herself. Since there does not seem to be a straightforward way to solve these problems, and in fact all relevant approaches

one could think of are prone to some error, the policy maker certainly would like to know which of all possible methods would likely generate the smallest errors.

Finally, as a first step in pursuing one's own analysis, one could consult the growing body of research on the evaluation of policy interventions. Unfortunately, a search of the published literature would very likely yield evidence and arguments for and against the two possible interventions considered here, thus raising the question into which of the results one should place more confidence. Moreover, it would become obvious that even scholarly research is unable to eliminate all error. Finally, none of the available studies would address exactly the same decision problem – technically identical interventions, covering the same type of workers, operating in the same environment – as in the case under consideration, thus raising the additional question of their generalizability to the current context. Again, without further guidance, it would not be clear to the policy maker which decision to adopt.

In essence, instead of a clear-cut answer to the decision problem, the best that a policy maker can hope for is a summary of the available evidence according to the well-respected standards of scientific research. One of the cornerstones of scientific research is the idea that only the weight of the evidence is able to answer any research question. If the same qualitative result – not the same numerical estimates, but a coherent set of results close to each other – were to be obtained in a variety of analyses, say training programs generally lead to more favorable outcomes, then with some confidence, albeit not with certainty, one could conclude that training programs should be the first choice. Ideally, these analyses would have been conducted by independent researchers, with slight variation in the particular choices regarding the methods of their analysis, the sources of their data, and the precise details of the various types of programs.

If there is no such conclusive evidence, however, the policy maker still has to make a decision. She will then usually not be able to wait until the weight of the evidence is sufficient to reach a confident decision. Therefore, she not only has to take a guess, but also has to admit to the state of her uncertainty to the public. That one cannot ascertain ultimate truth is neither a failure of science nor of the decision maker, it is an undeniable fact of life. Certainly, it would be much worse to take a firm opinion of somebody partial to the process as the correct answer, simply because it happens to be available.

On the other hand, since the ultimate truth is unavailable, it is important to realize that all results are derived with more or less confidence or reliability, depending on the particular research approach chosen. Unfortunately, it is not enough to simply compare the standard measures of sampling variability that are reported in any competent scientific study (see section 3.2 below). If that were the case, one could simply concentrate on those studies reporting a high degree of statistical reliability. To derive an explicit assessment of remaining uncertainty, any study must invoke more or less restrictive assumptions – so-called *identification assumptions* – which are *assumed* to be true for the purposes of the analysis, and whose validity is not reflected in the usual measures of sampling variability. Indeed, more restrictive assumptions will generally lead to smaller sampling errors. But this raises the question whether the identification assumptions were correct to begin with. Thus some basic understanding of the nature of specific identification assumptions and their applicability is imperative for an assessment of the available evidence.

Finally, since replication of evidence is a cornerstone of scientific research, well-respected scientific standards require that researchers make public their data and methods. Basically, any other researcher who were to decide that she wants to replicate a study should be able to do so on the basis of what is given in the report and of access to the data material – with the development of the *Internet* technically a minor request. It is this kind of discipline that creates the basis for modern scientific research. In essence, this requires also that the data pertaining to the policy maker's fictional problem should be publicly accessible for research purposes. As it turns out, advances in the extraction of anonymous information from individual records allow this research ideal to be pursued with ease.

In sum, evaluation of the impact of policy intervention should be undertaken by independent researchers, on the basis of publicly accessible data, with an emphasis on the publication of research methods, in particular of the identification assumptions underlying the derivation of a set of results, and on statements regarding the extent of any remaining uncertainty. Most importantly, only the weight of the evidence will allow any reliable conclusion on the possible impact of policy interventions. Since following a wrong route with confidence but without justifiable cause cannot be a serious option, the only reasonable choice for policy makers is to embrace the idea of scientific evaluation.

2.2 Major Elements of Program Evaluation

Any evaluation effort requires that relevant outcomes, measured in terms which are suitably defined, be compared in situations that differ in their relevant aspects only in the fact that one is with and the other without the intervention. Then any impact attributed to the intervention should be compared further to the costs involved. The first issue discussed here is the suitable choice of outcome measure; this discussion comprises several alternative single outcomes, multiple outcome variables, and the theoretical and practical aspects of integrative outcome measures. Second, since a complete evaluation of interventions necessitates a comparison of both effects and costs of the program, various cost components arising from interventions will be discussed. Program costs generally comprise both direct and indirect (opportunity) costs. Similar to estimating the impact of interventions on outcomes, estimation of costs could rely on experimental and non-experimental studies. Special emphasis will be given in the discussion here to the measurement of indirect cost components.

The third issue taken up in this section is the evaluation of the impact of interventions. This is the aspect of the *evaluation problem* that enjoys most of the attention by scientists, mostly because it seems to pose the largest intellectual challenge. This aspect will also be the topic of extensive discussion in sections 3 and 4. Wherever possible, one would like to base the evaluation on a suitable experiment, for reasons that will become transparent in these sections. However, since the experimental evaluation of policy interventions is frequently precluded by political, ethical or cost considerations and is often hampered by conceptual problems, the potential and the limitations of both experimental and non-experimental evaluation strategies will be explored.

2.2.1 What Outcomes?

Yet, the first question to clarify in any evaluation study is what should be considered as a success. Often a natural outcome measure suggests itself. In the example of interventions tailored to unemployed workers, this outcome measure would arguably be a measure that captures whether the intervention was bringing the unemployed back into (stable) employment. But the same qualitative outcome (improvement in the employment situation, say)

may plausibly be measured in several ways, for instance by hours per week in the new job, by a simple distinction between employment and unemployment, or even by a self-reported scale indicating satisfaction.

Moreover, it would be very demanding to expect that any intervention affected only a single outcome measure. By contrast, it is very likely that a single intervention affects several outcomes. In the example, this could be employment and wages in the new job. It might well be that one intervention raises employment substantially, albeit at low wages, whereas another intervention brings fewer workers into better jobs. By the same token, it is important to consider possible side effects on outcomes other than the primary ones that attract immediate policy attention. For instance, participation in a wage subsidy program may bring the unemployed participants back into employment successfully, but at the expense of accepting poor working conditions. The problem of directly comparing various qualitatively different outcomes is even more obvious, when several measures are competing at the same time for a share of the budget. One may be a program targeting the employment prospects of elderly workers, another may be attempting to raise the wages of employed women, say.

Thus, although it may be obvious which outcome measure to use for an intervention targeted at a narrow population with a specific problem, in many other cases the choice is more complicated. In addition, outcomes may not be comparable across interventions. Finally, perhaps the most severe difficulty is that, although a comparison of benefits and costs of an intervention has to rely on some understanding of the value of the program impact, program outcomes are normally not translated into money terms. It is more common to express results in terms of measurable outcomes, such as employment rates. However, as cost-effectiveness analysis measures effects in a wide range of outcomes, such as reductions of unemployment incidence or increase in wages, this analysis is not suitable for comparing interventions with different kinds of effects. Therefore one might want to carry out so-called cost-utility analyses, using integrated effect variables such as, to borrow an example from health economics, "quality-adjusted life years". In the context of health interventions, such integrative measures incorporate premature mortality and disability due to disease into a single measure. Naturally, integrative concepts will spark considerable debate as to their explicit value judgments with regard to the weighting of different types of outcomes as well

as the *time preferences* (determining the trade-off between changes in outcomes occurring at different points in the future) or even *group preferences* (determining the trade-off between changes benefitting different societal groups). The choice of such parameter values is not trivial since it will strongly influence the rank order of policy interventions.

One might want to base the parameter choice on evidence from survey data. However, a further problem for the translation from survey information into numerical estimates of the key parameters is the qualitative and self-reported nature of the underlying data, in contrast to the easily measurable quantitative data one would like to process. In particular, in their self-assessment individuals might imply different meanings with the same statement (e.g. being impaired in some societal function to some degree) depending on their societal and cultural context. Comparable problems have plagued the literature on contingent valuation which tries to estimate the values of non-market goods from survey data. Thus, particular emphasis should be given to the impact of measurement errors on the conclusions – the potential trade-off between a theoretically satisfying outcome variable prone to mis-measurement and theoretically less attractive proxies should always be kept in mind.

2.2.2 Measuring Program Costs

For the economic evaluation of policy interventions, valid estimates of the ensuing costs are as important as estimates of their impact. The second key aspect of the evaluation problem is therefore the estimation of the costs of interventions. In contrast to the evaluation of program impact or *efficacy*, so far relatively little methodological work has been done on how to assess the *efficiency* of policy interventions, that is their impact per dollar spent. Although in principle the estimation of resource use can be carried out simultaneously with the estimation of effectiveness, and therefore roughly the same tools can be applied, there are some issues specific to evaluation of efficiency that demand special emphasis. For example, for reasons of data availability and because of measurement problems the difficulty of extrapolating results to other situations is even more profound in efficiency analysis. Most importantly, in most economic evaluations of policy interventions so far only the direct outlays for the program have been determined, whereas the costs for the program participants and their families have been neglected.

The full costs of program participation include three components: the time cost of participation (the opportunity cost of wages foregone by the unemployed person, including the time costs of seeking treatment); the time cost of administrators and those involved in the delivery of the program (the opportunity costs of individuals' time spent on treating the unemployed or accompanying the unemployed to the place of treatment); and the financial costs of treatment (the expenditures incurred by the household in seeking treatment, including out-of-pocket expenditures for treatment, fees, transport, and the costs of subsistence at a distant treatment site). The second of these cost components – including costs that arise from administrative overhead – are often neglected. However, it is quite naive to think that the efforts of existing administrative agencies are costless. Instead, in a comparison of costs and effects of different policy interventions, it is the total consumption of societal resources that should be incorporated.

While most of the emphasis in evaluation research is on the self-selective nature of program participation regarding program impact (workers with higher expected impact being more likely to participate, see sections 3 and 4), one could even explore the consequences of heterogeneous participation cost on the endogenous selection of treatment. For instance, highly educated individuals may benefit more from a given intervention, thus tending to distort the sample of treated individuals to display a disproportionately high average education level. On the other hand, they also face higher opportunity cost of participation and may be underrepresented instead. It is not clear *a priori* what direction the resulting biases will ultimately take. The conditions leading to positively or negatively biased assessment of cost via some sort of self-selection and the potential remedies offered by a variation in empirical strategies could be an avenue for further research.

2.2.3 Evaluation Strategies

Since labor market success and, in particular, the successful return to employment from a spell of unemployment is influenced by numerous factors such as the regional structure of labor demand or individual search effort, a major scientific challenge is the attribution of labor market success to specific policy interventions. In the evaluation of interventions, the randomized controlled trial (RCT) is generally considered as the *gold standard*. In the

natural sciences, this is especially true for interventions that can be implemented in a tightly controlled environment, such as medical interventions at the hospital level. In this context, the interventions are randomized to different people on an individual basis. The impact of an intervention can then be evaluated by comparing average outcomes of those provided the intervention (the so-called "treatment group") versus those provided some alternative intervention or no specific treatment (the "control group").

Policy interventions affecting the labor market are certainly different. Often they do not lend themselves to controlled implementation, and even more often they are implemented before a controlled experiment can be designed and executed. One does not even have to emphasize the constraints on the controlled implementation of policy interventions in the context of the labor market, though, to make a case for the general relevance of non-experimental or *observational* approaches to evaluation (see section 4.2 below). Even in the health sector there are many interventions which are targeted at the community, rather than at the individual, for instance water supply and sanitation programs, or health education to reduce heart disease. One major drawback of applying the randomized design to the evaluation of community based programs is the required sample size for finding statistical significance (see for instance DONNER ET AL. (1981)). The effectiveness of randomization – its inherent ability to generate an overall balance pertaining to all aspects of relevance – decreases with a low number of communities included, because of the increasing chance that intervention and control groups differ with respect to important characteristics that bias the estimates. Thus, observational approaches form a serious alternative to experimental analysis also in this context.

The next two sections discuss this third element in more detail. First, the *evaluation problem* is stated more formally in terms of modern evaluation research. Then, this framework is used to contrast experimental and observational approaches to evaluation.

3 The Evaluation Problem: A Formal Statement

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 both

experimental and non-experimental evaluation strategies, following origins for instance to be found in RUBIN (1974). The current evaluation literature in economics has emphasized the potential of social experiments and the limits of non-experimental approaches (see for instance LALONDE (1995) or HECKMAN ET AL. (1999)), but before these approaches will be compared in section 4, the fundamental *evaluation problem* will be stated explicitly. For the purposes of this section, we abstract from all considerations of cost and also presume that it is clear what are the relevant outcomes to investigate.

3.1 The Construction of Counterfactuals

To address the problem in a sufficiently abstract way so as to appreciate the arguments made by modern evaluation research, some degree of formalism will be unavoidable. In particular, it is fruitful to describe each individual in the realm of the program under scrutiny by several key characteristics. For this purpose, denote the state associated with receiving the intervention by "1", and the state associated with not receiving the intervention by "0". Receiving the intervention is indicated by the individual indicator variable D_i. That is, if individual i receives training under this program, then $D_i = 1$. What we would like to compare is what would happen to individual i on the labor market, if i received the treatment ($D_i = 1$), say a training program, as well as if i did not ($D_i = 0$).

Specifically, the labor market outcomes in post-treatment period t are denoted by Y_{ti}, if individual i did not receive treatment, and by $Y_{ti} + \Delta_i$, if individual i received treatment. To exemplify these concepts in terms of a training program, these outcomes are defined here as indicator variables (with the possible realizations $0 =$ "*not employed*", $1 =$ "*employed*"). That is, if individual i were employed in t after receiving training ($D_i = 1$), then $Y_{ti} + \Delta_i = 1$, if not, then $Y_{ti} + \Delta_i = 0$. To summarize, for any individual, the pair of individual outcomes $(Y_{ti} + \Delta_i, Y_{ti})$ could be $(0, 0)$, $(0, 1)$, $(1, 0)$,or $(1, 1)$, respectively.

This setup directly allows the formulation of the *causal impact* of the intervention on the labor market status of individual i as Δ_i. If there are systematic differences between $(Y_{ti} + \Delta_i, Y_{ti})$, that is, a systematic pattern to the individual Δ_i's, then these are said to

be caused by the intervention[4]. Naturally, this concentration on a single individual requires that the effect of the intervention on each individual i not be affected by the participation decision of any other individual. We might refer to this as the assumption of independence of $i's$ treatment impact from the treatment status of the rest of the population. In the statistics literature (RUBIN (1986)) it is referred to as the *stable unit treatment value assumption* or SUTVA. Its validity facilitates a manageable formal setup; nevertheless, in practical applications it is frequently questionable whether it holds.

Unfortunately, and this is the core of the evaluation problem, we can never observe Y_{ti} and $Y_{ti} + \Delta_i$ simultaneously for a given individual – a worker can either participate in the training or not. Instead, only one of these two outcome variables can actually be observed for each individual i. That is, the outcome Y_{ti} is the *counterfactual outcome* for those individuals who do participate in the program ($D_i = 1$), whereas $Y_{ti} + \Delta_i$ is the *counterfactual outcome* for non-participants ($D_i = 0$). Furthermore, we cannot even make any concrete statements regarding which combinations of Y_{ti} and $Y_{ti} + \Delta_i$ are more or less prevalent across the population of workers. That is, we cannot observe the joint frequency distribution of these individual outcomes in any sample and, thus, cannot estimate the probability distribution of the Δ_i in the population.

The only frequency distributions that we can estimate, are the frequency distributions of $Y_{ti} + \Delta_i$ for participants ($D_i = 1$), and that of Y_{ti} for non-participants ($D_i = 0$), respectively. In the training example, if individual i underwent the training program, i's observed outcome in the labor market in post-treatment period t, $Y_{ti} + \Delta_i$, could take the value of 1, if i is employed in t ($Y_{ti} + \Delta_i = 1, D_i = 1$), or 0, if i is not employed in t ($Y_{ti} + \Delta_i = 0, D_i = 1$). Similarly, if individual i did not participate in the training program, i's observed outcome in the labor market in post-treatment period t, Y_{ti}, could also take the value of 1, if i is employed in t ($Y_{ti} = 1, D_i = 0$), or 0, if i is not employed in t ($Y_{ti} = 0, D_i = 0$). It is program participation, that is the value of D_i, that decides which of the two entries will be observed.

[4]For notational convenience, the discussion here is confined to a single pre-treatment period $t\prime$ (to be introduced below) and a single post-treatment period t. Thus, the causal impact of the intervention is written without time-subscript as Δ_i. In a general setting with several post-treatment periods, however, one might very well consider treatment effects that vary across post-treatment periods as well as across individuals.

To give further structure to the discussion, presume that the underlying frequency distributions of the outcomes $Y_{ti} + \Delta_i$ and Y_{ti} across the population are characterized by a set of individual characteristics X_i and by pre-intervention (period $t\prime$) outcomes $Y_{t'i}$. That is, for each and every possible configuration of the characteristics (X_i, $Y_{t'i}$), the respective conditional frequency distributions of Y_{ti} and $Y_{ti} + \Delta_i$ (which we don't know, but whose central aspects we want to estimate) describe the frequency with which every possible realization arises in the sub-population defined by X and $Y_{t'}$. Knowledge of these conditioning variables will allow to correct for "selection on observables", in a way formalized below. Suppose, for purposes of illustration, that X_i would capture education and could take on three values ($k = 0, 1, 2$, with $0 =$ "*low*", $1 =$ "*medium*", $2 =$ "*high*"). What is important is the so-called *exogeneity* of these conditioning characteristics: the participation in the program must not alter the value of (X_i, $Y_{t'i}$) for any individual i. That is, it is crucial not to condition on variables that themselves are outcomes of the treatment.

Now let us return to the real world where counterfactual outcomes are not observed. Instead, the available data comprise, in addition to observed outcomes Y_{ti} or $Y_{ti} + \Delta_i$, and characteristics X_i and $Y_{t'i}$, the indicator of treatment D_i. In general, we would be very hesitant to impose that Δ_i is equal for all workers, not even for those workers sharing the same values of X_i and $Y_{t'i}$. Some workers might be better off as a result of treatment, some worse. There will thus be no opportunity to ever estimate individual gains 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 of the frequency distributions of $Y_{ti} + \Delta_i$ and Y_{ti} can be estimated for participants and non-participants, respectively (henceforth, population averages are denoted by the mathematical expectations operator $E(.)$).

Interest in program evaluation is therefore on specific *evaluation parameters*, that is values that summarize the individual gains from treatment appropriately. The most prominent evaluation parameter is the so-called *mean effect of treatment on the treated*,

$$\begin{aligned} M_{X=k} &= E(\Delta \mid X = k, D = 1) = E\left((Y_t + \Delta) - Y_t \mid X = k, D = 1\right) \\ &= E\left(Y_t + \Delta \mid X = k, D = 1\right) - E\left(Y_t \mid X = k, D = 1\right), \end{aligned} \tag{1}$$

conditional on the specific realization of the exogenous variables, where in our example $k = 0, 1$, or 2. In this equation, individual subscripts are dropped to reflect the focus on population averages. The mean effect of treatment on the treated appropriately summarizes the individual gains in the population of those individuals who do receive the treatment, without restricting their heterogeneity.

Yet, it is not enough to define the population parameter of interest. As a final step, one has to link population averages and their estimates in a sample of limited size. Whenever a sample is used to estimate a population average, the answer given by the estimate will unlikely be exactly the true population parameter itself. Instead, the estimate can only give an approximation to the true parameter, since it has been derived on the basis of only a subset of all members of the population. A successful estimation strategy requires that, as the sample taken from the population becomes larger and larger, the approximation become more and more exact. In the limit, the approximation should be indistinguishable from the true parameter. In all samples of limited size the distinction between *bias* (a systematic deviation of the estimate from the true value that would consistently arise in independent replications of the process of data collection and estimation, if that process could indeed be repeated arbitrarily often) and *noise* (unsystematic deviations of the estimate from the true value that would wash out in such repetitions) is important (see also section 3.2). These are the conceptual ideas behind the formulations offered here.

Specifically, a population parameter is *identified* from observable data, if it could be estimated correctly with infinite precision by collecting abundantly many observations from the underlying population. If the sample size could be made abundantly large, statistical inference would simply be based on relative frequencies, since the relative frequency distribution would converge (that is, come closer and closer until complete resemblance) to the probability distribution in the population. One of the two population averages featured in equation (1) is identified from observable data, while the other is not: in principle, one could estimate $E(Y_t + \Delta_i \mid X = k, D = 1)$ with infinite precision from the available data on program participants, but one could not even hypothetically estimate the population average $E(Y_t \mid X = k, D = 1)$, since no sample size would alleviate the fact that Y_{ti} is not observed for participants.

This clarifies the nature of the fundamental problem facing program evaluation. This *evaluation problem* is the problem of finding an appropriate identification assumption that allows replacing this counterfactual population average $E(Y_t \mid X = k, D = 1)$ in (1) with an entity that is identified from observable data. It 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 more or by refined measurement. It can only be resolved by finding a plausible comparison group.

So far, the evaluation parameter has been formulated for sub-groups of individuals defined by their particular realization of the exogenous variable X. Ultimate interest of program evaluation might rather be in the provision of an estimate for the average treatment effects over all *relevant* values of X given $D = 1$, such as for instance (with $\sum_{k\epsilon\{0,1\}}$ denoting summation over workers with low and medium education, respectively)

$$\begin{aligned} M &= \frac{\sum_{k\in\{0,1\}} E(\Delta \mid X = k, D = 1) \Pr(X = k \mid D = 1)}{\sum_{k\in\{0,1\}} \Pr(X = k \mid D = 1)} \\ &= \frac{\sum_{k\in\{0,1\}} E(Y_t + \Delta \mid X = k, D = 1) \Pr(X = k \mid D = 1)}{\sum_{k\in\{0,1\}} \Pr(X = k \mid D = 1)} \\ &\quad - \frac{\sum_{k\in\{0,1\}} E(Y_t \mid X = k, D = 1) \Pr(X = k \mid D = 1)}{\sum_{k\in\{0,1\}} \Pr(X = k \mid D = 1)}. \end{aligned} \tag{2}$$

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 education groups in the population of program participants, $\Pr(X = k \mid D = 1)$. Note that the third education category, $X = 2$, was not included: in the example, high education workers do not fall within the realm of the program.

In principle, three conceptually distinct and non-exclusive errors may plague any attempt of program evaluation. First, one might not find comparable individuals who did not participate. For instance, it would be impossible to assess the impact of an intervention affecting low-skilled workers ($X = 0$), if every low-skilled worker participated in the program. In that case, the corresponding evaluation parameter $M_{X=0}$ is undefined. Second, while there might be comparable workers among participants and non-participants for every configuration of observable characteristics, their relative shares might be disproportionate. For instance, if

more low-skilled workers are among the treatment participants, but one were to take simple averages over low-skilled and medium-skilled workers, then the average of the participants' outcome would be relatively disfavorable. In expression (2) this problem is solved by the appropriate weighting with $\Pr(X = k \mid D = 1)$.

Third, there might be *selection bias*. Even when one compares comparable individuals for all relevant configurations of observable characteristics *and* weighs the corresponding means appropriately, there might be unobservable factors that invalidate the comparison. In formal terms, we would have

$$E(Y_t \mid X = k, D = 1) \neq E(Y_t \mid X = k, D = 0)$$

for at least one of the relevant k. For instance, more motivated workers might perform better in terms of employment, but might also be more likely to participate in a training program. Then, in the absence of treatment, the population average of the counterfactual outcomes Y_{ti} would have been higher among the participants ($D_i = 1$) than the average observable outcome is among the non-participants ($D_i = 0$).

Throughout this essay, the emphasis is on the mean effect of treatment on the treated. However, there are alternative evaluation parameters one could be interested in, for instance the *mean effects of treatment on individuals randomly drawn from the population*,

$$\widetilde{M}_{X=k} = E(\Delta \mid X = k) = E\left((Y_t + \Delta) - Y_t \mid X = k\right), \tag{3}$$

conditional on exogenous variables, and

$$\widetilde{M} = \frac{\sum_{k \in \{0,1\}} E(\Delta \mid X = k) \Pr(X = k)}{\sum_{k \in \{0,1\}} \Pr(X = k)}. \tag{4}$$

Note that the major difference to the first evaluation parameter M lies in the omission of conditioning on $D = 1$. While this is a justifiable evaluation parameter, for most contexts it might be difficult to imagine that one is seriously interested in the average value of Δ_i for a population of individuals a substantial fraction of which never participates in the treatment. For instance, the effect of a training program designed to improve basic literacy skills on

unemployed university graduates is arguably completely irrelevant. Hence, concentration in much of the evaluation literature is on evaluation parameter M as in expression (2).

3.2 Sampling Distributions

In this section, the discussion will return briefly to the distinction between bias and noise. It was stated above that whenever a sample is used to estimate a population average, the answer given by the estimate will unlikely be exactly the population parameter itself. Instead, the estimate can only give an approximation to the true parameter, since it has been derived on the basis of only a subset of all members of the population. Moreover, although a successful estimation strategy requires that, as the sample taken from the population becomes larger and larger, the approximation become more and more exact, it does not mean that one will ever receive the correct answer in any given estimation attempt.

Instead, what one would have with such a strategy is the confidence that, if one were to perform many repetitions of the sequence *drawing a random sample - estimating the population parameter - storing the estimated parameter value*, then the central tendency of the resulting frequency distribution would be on the correct value. That is, in following this estimation strategy one would be correct on average, but irrespective of the sample size in each of these replications, there would be some dispersion around the true population parameter. Generally, this dispersion would decrease with growing sample size, but never vanish completely. This remaining uncertainty or *noise* about the true value (or better said, some estimate of it) should be reported in any decent empirical study. Only then will the recipient of the results be able to assess, whether large confidence should be placed in the conclusions of the study or not. Typically, researchers report *standard errors* or *confidene intervals* to this effect.

Yet, the remaining uncertainty that will be reported will always reflect the researcher's conviction that all systematic deviations between the answer given by the estimation strategy and the true population parameter have been successfully eliminated by invoking the correct identification assumption. Stricter identification conditions typically lead to lower assessments of remaining uncertainty. To take an extreme example, a researcher could decide to estimate the program impact always by the number 0.3, irrespective of the context

and the data material at hand. Asked to provide an assessment of the variability of this estimate as various samples are taken from the population, one could truthfully argue that this remaining uncertainty is nil (the estimate never changes at all), although it would be completely absurd to proceed in such a way. Thus, small noise is not the only important aspect of an empirical study.

Instead, any evaluation effort that wants to be taken seriously should aim at a convincing strategy that eliminates all systematic tendencies to deviate from the correct population parameter or *bias*. For this reason, the identification assumptions underlying any empirical evaluation analysis take center stage in this essay. Section 4 is devoted to the choice of identification assumption, issues of sample size – as important as they are in practice – are de-emphasized in the remainder of the discussion.

4 Empirical Approaches to the Evaluation Problem

All empirical approaches that will be discussed in this section follow a common principle of analogy. In order to formulate an estimate of population parameters, one searches for the corresponding concept in the sample at hand. If the population parameter of interest is in fact identified from observable data, there will be noise around the estimate in each and every practical application, but this noise would vanish in the hypothetical case of an abundantly large sample. It is therefore not a conceptual hurdle for finding the correct population average. Instead, 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; this discussion will always be in terms of the corresponding population averages.

Actual estimation in the sample is then performed by taking the appropriate averages. The estimator will therefore always be given in terms of observable entries in the sample. In what follows N_1 is the number of individuals in the sample of participants, with indices $i \in I_1$. The sample of nonparticipants consists of N_0 individuals, with indices $j \in I_0$. Subsets of these samples are denoted in a straightforward fashion. For instance, the number of medium-skilled participants is $N_{1,X=1}$, the set of indices of all non-participants with characteristics

$X_i = 1$ and $Y_{t'i} = 0$ is $I_{0,X=1,Y_{t'}=0}$. Accordingly, the corresponding number of observations is $N_{0,X=1,Y_{t'}=0}$.

4.1 Experimental Studies

4.1.1 Randomization

Under the fundamental requirement that an experiment completely replicate the intervention that will be implemented in the field, experimental studies generally provide for a convincing approach to the *evaluation problem*. The key concept of any experiment is the *randomized assignment* of individuals into treatment and control groups. For workers who voluntarily would be participants in the program ($D_i = 1$) the random mechanism decides whether they are in fact allowed to enter the program or whether they are excluded from the program instead. This assignment mechanism is a process that is completely beyond the workers' 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 level with considerable control by the researcher about the delivery and about individual compliance with the program. 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.

Let R_i be an indicator of randomization status, that is, the status given to the individual in the assignment procedure ($1 =$ "*in*" and $0 =$ "*out*"), and concentrate all attention on the population of would-be participants ($D_i = 1$). Then, the identification assumption made here is

$$E(Y_t \mid D = 1, R = 1) = E(Y_t \mid D = 1, R = 0). \tag{5}$$

To reiterate, all observations are on individuals which applied for treatment but then were assigned to treatment or control groups by a random mechanism, all individuals in the control group have been randomized out, but would have chosen to be in had there been no random mechanism. Thus, one can infer the average treatment effect from the difference of

the average outcomes of these randomly selected individuals (with $\widehat{\beta}$ denoting the estimated value of any parameter β),

$$\widehat{M}_{X=k}^{experiment} = \frac{1}{N_{1,X=k}} \sum_{i \in I_{1,X=k,R=1}} (Y_{ti} + \Delta_i) - \frac{1}{N_{0,X=k}} \sum_{j \in I_{0,X=k,R=0}} Y_{tj}. \tag{6}$$

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.

4.1.2 Limitations of Randomized Trials

It is important to keep in mind, however, that a randomized controlled trial might not be a feasible approach at all, for political, ethical, logistic, or financial reasons, or a randomized trial might be contaminated by influences beyond the control of the researcher designing the study. This holds *a fortiori* for community-based interventions. Since these play a prominent role in the real world, emphasis in this sub-section will be on the limitations of group-randomization. For instance, the applied literature on the evaluation of community-based interventions documents serious ethical objections against group-randomization: In the evaluation of treatments that have a high probability of being effective, it may be considered unethical to carry out an evaluation study involving a control group or area from which the effective intervention is withheld.

Randomization of communities might also face strong political objections - communities are not simply large-sized individuals, their decisions are rather the consequence of the complex aggregation of their members. Thus, it might be significantly more difficult to generate the widespread pre-intervention support for the randomized study across a sizeable number of communities that would be the prerequisite for a group-randomized design and that is so easily ensured with individual patients in a clinical setting. Furthermore, the political influence on the assignment process can also take more subtle forms: communities that suffer more from a particular problem or simply more wealthy communities will lobby for better access to promising interventions. In addition, there might be strong indirect

influence of political pressure through the influence of the status quo on the assignment choice. Administrators might feel tempted to assign those communities to the treatment group which display favorable characteristics such as average education, the participation in a previous trial or a well-developed infrastructure. This is not merely a problem of sample size and, thus, cannot be solved by involving more communities in the trial.

Moreover, whenever the program is to be delivered at a large scale, it may be impossible for logistic reasons to generate a setting in which neighboring communities can actually be assigned to treatment and control groups by a random process. For instance, consider the case of mass-media campaigns that have to be delivered at a regional level. Again, this is not simply a problem of sample size. Finally, one might simply not be able to acquire the appropriate sample size: in many situations the cost of assuring randomization of sufficiently many units at the community level might be prohibitive. If for any of these reasons randomization cannot be attempted seriously to begin with, appealing to any superior theoretical properties of group-randomized trials is not of any value for actual applications.

Even where one can engage into group-randomization, the community context can work against the construction and preservation of randomized treatment and control groups. That is, what is set up as a group-randomized study might be contaminated by processes beyond the control of the researcher designing the analysis. Such problems arise at various stages of treatment assignment and intervention delivery, since compliance with the assignment and the program are difficult to monitor. At every stage of the process, communities might explicitly decide to drop out from treatment or control groups altogether or they might reduce or increase their effort in supporting the delivery in a less conspicuous fashion. In terms of the formal setup, as a consequence of both these fundamental problems of randomization or subsequent contamination the assumption underlying impact estimation in group-randomized settings, equation (5), is no longer justified.

4.1.3 Small Sample Bias

Randomized assignment of individuals into or out of treatment is a very powerful approach to inference: on average, a simple difference in mean outcomes across treatment and control groups yields an unbiased estimate of the true evaluation parameter, even without particular

attention to observable confounders. In most practical applications, though, there are issues of so-called *small sample bias*. This usually concerns the possible presence of observable factors X_i, which are not balanced completely by the process of randomization, due to sample size limitations. An imbalance of observable confounders in the sample used for analysis would not provide for the correct weighting of conditional means in the treatment and control groups.

Failure to account for this would lead to a poor approximation of estimates to the true value in small samples, but – due to the very nature of randomized assignment –, if sample sizes were to grow beyond any limit, randomization would serve to eliminate this bias completely. Thus, it is often suggested that analysts should adjust their data for the presence of observable confounders, exactly as in expression (6) which conditions on the observable confounding variable X. The advantage of doing so will be a smaller non-systematic deviation of the estimated program impact from its true value, a smaller noise. By contrast, evaluation strategies that do not rely on randomized assignment might not be able to afford the luxury of controlling for observable confounding factors just to achieve a reduction in noise. Instead, in non-experimental settings simple averages might suffer from an imbalance of observable confounders even in large samples (see section 4.2).

Even in an experiment involving a large number of observations on both treated individuals and controls, there will be some remaining noise around the true parameter value. Given that a researcher takes all appropriate measures to eliminate the small-sample bias discussed here, this problem only concerns any remaining imbalance of unobservable confounders. While clearly being relevant in any sample of limited size, in a randomized study the bias emerging from this source would also diminish as sample size was growing to be large.

4.2 Observational Studies

By contrast to experimental analyses, in non-experimental or *observational* studies (a seminal source is ROSENBAUM (1995)) the data are not derived in a process that is completely under the control of the researcher. Instead, administrators might have offered the program to individuals for whom they held favorable expectations regarding the program's impact. Or

participants decided to apply for program participation because they (correctly) anticipated the intervention as particularly beneficial for them, while non-participants shied away from the program for symmetric reasons.

What is collected instead of the desired experimental data, is an account of how individual workers actually performed after the intervention. For participants this means observation of $Y_{ti} + \Delta_i$, for non-participants observation of Y_{ti}. The objective of any observational study is to use this information in an appropriate way such as to replace the comparability of treatment and control groups *by design* – the cornerstone of experimental analyses – by a plausible alternative identification condition.

In experiments, random assignment of treatment ensured a balancing between treatment and control groups of all aspects relevant to the process, observable and unobservable. The desire in any observational study is to use the observable information (on X_i and on $Y_{t\prime i}$) such that in sub-populations defined by these observables, for instance low-skilled workers who were employed in $t\prime$, any remaining differences between participants and non-participants can be attributed to chance. Then, using a random sample from this sub-population, the impact of the program can be estimated by forming the difference between means of actual outcomes for participants and non-participants. One of the considerations in choosing an appropriate identification strategy is sample size. Neither would one place high confidence in averages taken only over a handful of individuals, nor would one be able to derive any result for a configuration of characteristics X_i such that every worker in this sub-population decided to participate.

The following sub-sections introduce four observational approaches, characterizing the identification assumptions necessary to justify their application and possible reasons for their failure. To facilitate the concentration on the idea of identification, the issue of precision or sampling variability is not discussed at length. Nevertheless, this is an issue of considerable relevance for all practical applications – a reported impact estimate that is not accompanied by an indication of the sampling variability around it is absolutely worthless. In section 4.3, the four identification approaches outlined in detail below will be illustrated by a small numerical example.

4.2.1 Comparison of Exact Matches

The principal idea of *exact matching* is to assign to one or more of the individuals i 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 individual characteristics. That is, the exact match procedure specifies the most general possible model of post-intervention outcomes in terms of the observable data (pre-intervention histories and education, say). The central identification assumption is that for individuals that are characterized by any specific configuration of observable characteristics, the participation decision is independent of any unobservable determinant of the post-intervention outcome,

$$E(Y_t \mid X, Y_{\mathbf{t}'}, D = 1) = E(Y_t \mid X, Y_{\mathbf{t}'}, D = \mathbf{0}). \tag{7}$$

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. Then, the desired estimate of the program impact $M_{X=k}$ is given by a weighted average over these sample means,

$$\begin{aligned}\widehat{M}^{exact-match}_{X=k} = \; & \tfrac{1}{N_{1,X=k}} \textstyle\sum_{Y_{t'}} N_{1,X=k,Y_{t'}} \Big(\tfrac{1}{N_{1,X=k,Y_{t'}}} \textstyle\sum_{i \in I_{1,X=k,Y_{t'}}} (Y_{ti} + \Delta_i) \\ & - \tfrac{1}{N_{0,X=k,Y_{t'}}} \textstyle\sum_{j \in I_{0,X=k,Y_{t'}}} Y_{tj} \Big),\end{aligned} \tag{8}$$

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

Matching estimators thereby approximate the virtues of randomization by balancing the observables. First, the population of workers who participate in the program is partitioned into strata defined by all relevant combinations or *cells* of individual characteristics X_i and pre-intervention outcomes $Y_{t\prime i}$. Only those cells will enter the calculation of program effects, for which one will also be able to find non-participants. Thus, matching ensures that the averages over treated and untreated individuals are taken over the same cells (in techni-

cal terms the same *support* of the frequency distribution of the observable characteristics). Second, in forming the weighted average that estimates $M_{X=k}$, the relative shares of each of these cells among the participants are taken as weights, ensuring that the mean effects of treatment on the treated is identified. Thus, whatever the relative shares of these cells are in the population of non-participants, their outcomes are reweighted such as to generate balance with the population of participants.

Validity of this evaluation approach requires that the decision to participate and the outcome should not be influenced jointly by factors beyond previous labor market outcomes $Y_{t'i}$ and current characteristics X_i. If, for instance, a favorable mix of unobservable factors were particularly common among the population of participants, then, via the identification assumption (7) the low population average of the Y_{ti} among non-participants in at least one (X, $Y_{t\prime}$)-cell would lead to too low an estimate of the average of Y_{ti} among the participants in this cell and, thus, an overstated estimate $\widehat{M}_{X=k}^{exact-match}$.

It would not be detrimental for this estimator if workers' employment success followed a cyclical pattern: in period t participants' counterfactual outcomes under no treatment would then typically be higher than in $t\prime$, but one would only treat those non-participants as comparable whose outcome in $t\prime$ was equally unsatisfactory. Intertemporal changes in economic conditions that affected, on average, the change between $t\prime$ and t equally for participants and non-participants, as long as they shared a common set of characteristics (X, $Y_{t\prime}$), would not be consequential either. What would be consequential were changes that affect participants differently from untreated even within these narrowly defined cells.

4.2.2 Difference-in-differences Estimation

Forming exact matches is computationally demanding and, depending on the application, might lead to the problem of relatively small samples in each relevant population cell. One might therefore entertain an alternative approach that retains the ideas of using non-participants as a control group and of accounting for changes in the macroeconomic environment. The *difference-in-differences* approach continues to focus on a comparison of the changes between $t\prime$ and t. The major difference to the exact matching procedure is that one does no longer condition on the exact pre-intervention outcome in defining comparable

strata of the population. Instead, it is postulated that population average of the change in the no-program outcome participants between $t\prime$ and t is equal to that experienced by non-participants,

$$E(Y_t - Y_{t'} \mid X, D = 1) = E(Y_t - Y_{t'} \mid X, D = \mathbf{0}). \tag{9}$$

The corresponding estimator then implements a sample analogue of this idea, by comparing the sample averages of the changes in outcomes for random samples of participants and non-participants (with individual characteristics X_i),

$$\widehat{M}_{X=k}^{diff.\text{-}in\text{-}diff.} = \frac{1}{N_{1,X=k}} \sum_{i \in I_{1,X=k}} ((Y_{ti} + \Delta_i) - Y_{t'i}) - \frac{1}{N_{0,X=k}} \sum_{j \in I_{0,X=k}} (Y_{tj} - Y_{t'j}). \tag{10}$$

As the exact matching approach, this strategy is vulnerable to the presence of unobservable factors affecting the participation decision. Estimator $\widehat{M}_{X=k}^{exact-match}$ was able, though, to account for all such unobservable factors that were associated with the labor market outcome in the pre-treatment period. It is not unlikely that workers who are unemployed in $t\prime$ are more likely to participate in the program and that, at the same time, the economic upswing benefits more those being unemployed in $t\prime$. As a consequence, the difference-in-differences estimator $\widehat{M}_{X=k}^{diff-in-diff}$ will attribute the relatively large change in observed outcomes for participants exclusively to the program.

4.2.3 Before-After Comparisons

Perhaps the most common evaluation strategy for attempting the construction of a plausible counterfactual is a comparison of treated individuals with themselves at a time $t\prime$ preceding the intervention, a *before-after* comparison. In this approach, what is defined as the comparable non-participants are the participants themselves before the program was implemented. Based on longitudinal data, information on the effects of the program is then extracted exclusively from changes between $t\prime$ and post-treatment period t. To estimate treatment impact, one forms the average over all pairs of before-after observations for the individuals

in the training sample. The underlying identification assumption is that, taken over the population of all treated individuals, the population average of actual outcomes in period $t\prime$ (by definition, these are the counterfactual outcomes in $t\prime$ in the absence of treatment) is equal to the population average of what these individuals would have experienced had they not participated in the program in period t, that is, equal to counterfactual outcomes in t in the absence of treatment. Formally, this is

$$E(Y_t \mid X, D = 1) = E(Y_{t\prime} \mid X, D = 1). \tag{11}$$

That is, for each individual pair of observations before and after treatment, there might be differences in the counterfactual outcomes for the no-treatment state, even substantial ones, but on average these are cancelling out. In effect, one can take a mean over the corresponding individual pairs of pre-/post-intervention observations in a random sample of the population of the treated and estimate the impact of the intervention consistently as

$$\widehat{M}_{X=k}^{before-after} = \frac{1}{N_{1,X=k}} \sum_{i \in I_{1,X=k}} ((Y_{ti} + \Delta_i) - Y_{t'i}) \tag{12}$$

Validity of this identification assumption requires that the outcomes before treatment not be influenced by an anticipated intervention. For instance, if members of the treatment sample know in $t\prime$ that they will be trained, they might not try to be very successful in t'. A low $Y_{t'i}$ would be the consequence, and, via the identification assumption (11) the low population average of the $Y_{t'i}$ would lead to too low an estimate of the average of Y_{ti} and, thus, an overstated estimate $\widehat{M}_{X=k}^{before-after}$. Similarly, if workers' employment success follows a cyclical pattern, and bad times (that is a spell of unemployment) are typically temporary episodes followed by improvements, this approach is problematic. In period t participants' counterfactual outcomes under no treatment would then typically be higher than in $t\prime$, the so-called *Ashenfelter's dip* (Ashenfelter (1978)), and, thus, the natural tendency to return to a long-run average (so-called *mean reversion*) would lead estimator (12) to overstate the program impact. By contrast, permanent dips in labor market performance are not detrimental to the approach, because they will affect $Y_{t'i}$ and Y_{ti} in an identical way.

In addition, this approach requires considerable stability of the economic environment: if period t' was a bust period, say, and period t that of an economic upswing, then counterfactual outcomes in the absence of treatment, Y_{ti}, would exceed the outcomes in the pre-treatment period $t\prime$ because of the business cycle – yet, estimator (12) would attribute this improvement to the intervention.

4.2.4 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 macroeconomic conditions shift substantially over time. In that case, the population average of the observed outcome of non-participants could serve as the entity to replace the population average of the unobservable Y_{ti} for participants. The formal statement of this identification condition would read

$$E(Y_t \mid X, D = 1) = E(Y_t \mid X, D = \mathbf{0}). \tag{13}$$

That is, although the populations of participants and non-participants might be quite different in size and within each of those populations the (counterfactual) outcomes for the no-treatment state might differ widely, on average these are cancelling out. In effect, one can take the means over the corresponding observations in random samples of participants and non-participants, and estimate the impact of the intervention as

$$\widehat{M}_{X=k}^{cross\text{-}section} = \frac{1}{N_{1,X=k}} \sum_{i \in I_{1,X=k}} (Y_{ti} + \Delta_i) - \frac{1}{N_{0,X=k}} \sum_{j \in I_{0,X=k}} Y_{tj}. \tag{14}$$

For this identification assumption to be valid, selection into treatment has to be statistically independent of its effects given X, the case of *exogenous selection*. That is, no unobservable factor (such as "motivation" which would increase the desire to participate in a training program but also increase the (counterfactual) no-treatment outcomes) should lead individual workers to participate. Otherwise equality (13) would not hold any longer. This property

exact matches

$$\widehat{M}^{exact-match}_{X=k} = \frac{1}{N_{1,X=k}} \sum_{Y_{t'}} N_{1,X=k,Y_{t'}} \left(\frac{1}{N_{1,X=k,Y_{t'}}} \sum_{i \in I_{1,X=k,Y_{t'}}} (Y_{ti} + \Delta_i) - \frac{1}{N_{0,X=k,Y_{t'}}} \sum_{j \in I_{0,X=k,Y_{t'}}} Y_{tj} \right)$$

identifying assumption

$$E(Y_t \mid X = k, Y_{\mathbf{t'}}, D = 1) = E(Y_t \mid X = k, Y_{\mathbf{t'}}, D = \mathbf{0})$$

difference-in-differences

$$\widehat{M}^{diff.\text{-}in\text{-}diff.}_{X=k} = \frac{1}{N_{1,X=k}} \sum_{i \in I_{1,X=k}} ((Y_{ti} + \Delta_i) - Y_{t'i}) - \frac{1}{N_{0,X=k}} \sum_{j \in I_{0,X=k}} (Y_{tj} - Y_{t'j})$$

identifying assumption

$$E(Y_t - Y_{\mathbf{t'}} \mid X = k, D = 1) = E(Y_t - Y_{\mathbf{t'}} \mid X = k, D = \mathbf{0})$$

before-after

$$\widehat{M}^{before-after}_{X=k} = \frac{1}{N_{1,X=k}} \sum_{i \in I_{1,X=k}} ((Y_{ti} + \Delta_i) - Y_{t'i})$$

identifying assumption

$$E(Y_t \mid X = k, D = 1) = E(Y_{\mathbf{t'}} \mid X = k, D = 1)$$

cross-section

$$\widehat{M}^{cross\text{-}section}_{X=k} = \frac{1}{N_{1,X=k}} \sum_{i \in I_{1,X=k}} (Y_{ti} + \Delta_i) - \frac{1}{N_{0,X=k}} \sum_{j \in I_{0,X=k}} Y_{tj}$$

identifying assumption

$$E(Y_t \mid X = k, D = 1) = E(Y_t \mid X = k, D = \mathbf{0})$$

Table 1: A Summary of Selected Observational Approaches.

was ensured in a controlled randomized trial by randomizing some individuals 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. When working with non-experimental data, however, individuals who received treatment and those who did not might have been selected into these two groups in a systematic fashion. The underlying selection process might, among other aspects, reflect individual gains from treatment. Consequently, a cross-sectional approach might be a very poor evaluation strategy. On the other hand, Ashenfelter's dip will not pose a relevant problem for this approach, since the temporary dip in labor market performance in period t' will not play a role in the construction of the counterfactual $E(Y_t \mid X, D = 1)$.

The various evaluation approaches discussed in sections 4.2.1 to 4.2.4 are summarized in **Table 1** that displays the estimators together with the appropriate identification assumptions. Note that the estimators always implement the sample analogue to the appropriate

population averages implied by the identification assumption. None of the identification assumptions (7) to (13) is superior to the others. Choosing the appropriate strategy has to depend on outside knowledge of the processes of program implementation and participation. Furthermore, given sample size, some of the approaches, in particular *exact matching*, but also *difference-in-differences* are more demanding, that is involve fewer individuals in the formation of the relevant averages than others.

4.2.5 Parametric Approaches and Bounding

All four approaches introduced above do rest on identification assumptions, sets of minimal assumptions to generate an estimate of the impact of the policy interventions. Conventional econometric research usually went further, resting on *a priori* information about various aspects of the process, either in terms of functional forms, in forms of restrictions on the impact of treatment or in terms of information on the determinants of the choice of treatment regime. In particular, parametric modelling, implying an explicit assumption of the functional form of the frequency distribution of unobservable factors across the population of participants and on an assumption on the homogeneity of program effects, has been so prevalent that many outside observers think of regression adjustments or standard nonlinear discrete choice models as the *econometric approach*.

The variety of parametric approaches is purposefully de-emphasized in this essay to clarify that, in principle, there is no partiality across disciplines as to how stringent the identification assumptions have to be in applied work. To the contrary, the body of research on program evaluation has grown together in recent decades, leading to the common probabilistic approach at the issue that is discussed in section 3 and 4. Furthermore, since all scientific results are derived with some remaining uncertainty, it is clear that stronger identification assumptions – if correct – will lead on average to more precise results. That does certainly not imply that a parametric constant-effects approach (a linear regression model, say, which imposes $\Delta_i = \Delta$ for all i) has to be preferred to a non-parametric heterogeneous-effects approach (such as an application of expression (8), which allows for heterogeneous Δ_i), merely because the remaining uncertainty surrounding the estimated impact is smaller. After all, the strict identification assumption of the constant-effects analysis might be wrong, biasing

the results systematically. On the other hand, in applied work there are many situations in which small sample sizes or the sheer complexity of the calculations require additional identification assumptions to be invoked, for instance in the condensation of the probability of program participation into the so-called *propensity score*.

Yet another approach to the evaluation problem (see for instance MANSKI (1995)) is to avoid making strict identification assumptions as those embodied in equations (7) to (13) altogether. Instead, one might find a milder set of restrictions on the process, restrictions that can be justified without any knowledge of the behavioral side of program participation or of implementation issues, which facilitate to bound the genuine impact of the program from above and below. In the example of workers' employment success, the absolute difference $\mid (Y_{ti} + \Delta_i) - Y_{ti} \mid$ cannot exceed unity, for instance. Starting from this idea, one might find refinements that facilitate narrowing the corresponding interval $[-1, 1]$ further. These would not be identfication assumptions in the sense introduced above, since even in an abundantly large sample they would not allow pinpointing the genuine effect exactly. One question in program evaluation would then be whether the bounded intervall would only comprise positive entries. Then, one could conclude with confidence that the program at least had some positive impact on the labor market outcomes of participants. Generally, it is difficult to tell whether one will be able to find restrictions mild enough to be palatable, yet strong enough to provide a sufficiently narrow set of bounds. For the purposes of this essay, this approach will not be discussed at further length.

4.3 A Numerical Example

To illustrate the different estimators introduced in section 4.2, this section provides a small numerical example. Continue to suppose that the population of workers comprises three kinds of workers, low-skilled ($X_i = 0$), medium-skilled ($X_i = 1$), and high-skilled ($X_i = 2$), and consider a program primarily being designed to improve workers' labor market skills. High-skilled workers are considered not to be in the realm of the program; the program is also not taken up by many medium-skilled workers. The evaluator's interest lies therefore in the mean impact of the program on treated low-skilled workers, $M_{X=0}$, the mean impact of the program on treated medium-skilled workers, $M_{X=1}$, and the mean impact of the program on

low-skilled and medium-skilled workers as a group, $M_{X\in\{0,1\}}$. The two relevant skill groups will be taken one at a time.

The first two columns of **Table** 2 display the potential labor market outcomes under no treatment for periods $t\prime$ (prior to the program) and t (after the program) for those low-skilled workers who do not participate in the program (thus, these are also their actual or observed outcomes). Some 60% of these workers are employed in both periods; in addition, there is some improvement in the labor market situation over time, since one of those workers being unemployed in $t\prime$ is employed in t (worker #1). Thus, if the program has any impact, the labor market situation of the participants must improve by at least as much, on average, as it did for the non-participants.

Columns three to five of the table pertain to those unskilled workers who received the intervention between $t\prime$ and t. The third and fourth column display the potential labor market outcomes under no treatment in these two periods. Whereas the outcome in $t\prime$, $Y_{t'i}$, is observed, the counterfactual outcome under no treatment in t, Y_{ti}, is not. Instead, one observes $Y_{ti}+\Delta_i$ which is displayed in the last column of the table. The genuine impact of the program becomes apparent when one compares the fifth and the fourth column and forms $(Y_{ti}+\Delta_i)-Y_{ti}$ (something a researcher will not be able to do, since the fourth column will remain unobserved). The genuine impact of the program is 0.2, that is, of the 10 workers who receive treatment, two actually experience an improvement in their employment situation, as compared to the unobservable counterfactual (workers #13 and #19).

A researcher who were to form *exact matches* to calculate $\widehat{M}_{X=0}^{exact-match}$ would invoke the identification assumption that for low-skilled workers who participate in the program **and** who happen to experience a particular pre-program outcome $Y_{t'}$, we can replace the population average of their unobservable counterfactual outcome under no treatment in t with the population average of the observable actual outcome of comparable individuals in the group of non-participants (see equation (7)). Estimation is then simply the sample analogue of this idea. That is, one would impose here that, on average, workers 5 to 10 and 17 to 20 would be comparable, and that the same holds for workers 1 to 4 and 11 to 16. For all low-skilled workers taken together, that is those with $Y_{t'}=1$ and those with $Y_{t'}=0$, the

	$D_i = 0$			$D_i = 1$		
Individual	$Y_{t'i}$	Y_{ti}	Individual	$Y_{t'i}$	Y_{ti}	$Y_{ti} + \Delta_i$
1	0	1	11	0	0	0
2	0	0	12	0	0	0
3	0	0	13	0	0	1
4	0	0	14	0	1	1
5	1	1	15	0	1	1
6	1	1	16	0	1	1
7	1	1	17	1	1	1
8	1	1	18	1	1	1
9	1	1	19	1	0	1
10	1	1	20	1	1	1

Table 2: Employment & training of low-skilled workers.

evaluation parameter of interest is then a weighted (with the relative shares of treated workers experiencing $Y_{t'} = 1$ and $Y_{t'} = 0$) average of the impacts in both subgroups, $\widehat{M}^{exact-match}_{X=0,Y_{t'}=1}$ and $\widehat{M}^{exact-match}_{X=0,Y_{t'}=0}$. The result in this example would be an estimate of $\widehat{M}^{exact-match}_{X=0} = 0.25$. The estimate would not quite capture the genuine impact correctly, because the average counterfactual outcome under no treatment for participants (the fourth column of the table) would be different for workers 17 to 20 in comparison to workers 5 to 10, and for workers 11 to 16 in comparison to workers 1 to 4.

An even stronger identification assumption would have been the basis for a calculation of the *difference-in-differences* estimator $\widehat{M}^{diff.\text{-}in\text{-}diff.}_{X=0}$. It would assert that for low-skilled workers who participate in the program, the population average of the difference between their unobservable counterfactual outcome under no treatment in t and their observable actual outcome in $t\prime$ would equal the population average of the observable actual outcomes in t and $t\prime$ of individuals in the group of non-participants (see equation (9)). Again, estimation is then simply the sample analogue of this idea. That is, one would impose here that, on average, workers 11 to 20 over time experience an increase in potential no-treatment outcomes that equals that for workers 1 to 10. The evaluation parameter of interest is then a simple average of the differences in average growth in observable outcomes across groups. The result in the example would be an estimate of $\widehat{M}^{diff.\text{-}in\text{-}diff.}_{X=0} = 0.30$. The estimate would fail to approximate the genuine impact closely, because the improved labor market situation

in t would benefit participants far more than non-participants. To the extent that those workers out of employment in $t\prime$ benefitted more as a group than those already holding a job at that time, estimator $\widehat{M}_{X=0}^{exact-match}$ above yields a better approximation to $M_{X=0}$ than $\widehat{M}_{X=0}^{diff.\text{-}in\text{-}diff.}$. It fails to yield the exact impact, though, because even among those without a job in $t\prime$, participants would have performed better in the absence of treatment.

Yet another evaluation approach would lie in the calculation of the *before-after comparison* $\widehat{M}_{X=0}^{before-after}$. The relevant identification assumption that would have to hold here concentrates on the population of participants. For those low-skilled workers who participate in the program, the population average of their unobservable counterfactual outcome under no treatment in t would be taken to be identical to the population average of their observable actual outcome in pre-treatment period $t\prime$ (see equation (11)). The estimator would be implemented by comparing the sample averages in actual outcomes in t and $t\prime$ for workers 11 to 20, with the result being an estimate of $\widehat{M}_{X=0}^{before-after} = 0.40$. Clearly, this estimate completely attributes the improvement in the overall employment situation between the two periods to the program. Since in this example the economic upswing was constructed to be substantial, the corresponding discrepancy between $\widehat{M}_{X=0}^{before-after}$ and $M_{X=0}$ is large.

Finally, the researcher might disregard the pre-program information on participants and non-participants altogether, and implement the *cross-section* estimator $\widehat{M}_{X=0}^{cross\text{-}section}$. To be a valid evaluation strategy, this would require that in period t the population average of the unobservable no-participation outcome of participants would equal the observable outcome of non-participants (see equation (13)). This estimator would be implemented by comparing the sample averages in actual outcomes in t for workers 11 to 20 and 1 to 10, respectively. The result, $\widehat{M}_{X=0}^{cross\text{-}section} = 0.10$ would closely approximate the genuine impact $M_{X=0}$, since as of t, the labor market situation of participants in the absence of training and that of non-participants is indeed constructed to be very similar.

A similar example is constructed for medium-skilled workers in **Table 3**, with two important differences. First, the population of medium-skilled workers is taken to be smaller, and second, a disproportionately small share of medium-skilled workers participates in training. The example is constructed so that the genuine impact $M_{X=1}$ of the program is 0. Both the estimator imposing exact matching, $\widehat{M}_{X=1}^{exact-match}$, and the cross-section estimator

	$D_i = 0$			$D_i = 1$		
Individual	$Y_{t'i}$	Y_{ti}	Individual	$Y_{t'i}$	Y_{ti}	$Y_{ti} + \Delta_i$
21	0	1	27	0	1	1
22	1	1	28	1	1	1
23	1	1	29	1	1	1
24	1	1	30	1	1	1
25	1	1				
26	1	1				

Table 3: Employment & training of medium-skilled workers.

$\widehat{M}_{X=1}^{cross\text{-}section}$ lead to the correct conclusion of no treatment effect, because in the sample the underyling identification assumptions are met exactly. The difference-in-differences estimator is slightly off, $\widehat{M}_{X=1}^{diff.\text{-}in\text{-}diff.} = 0.08$, simply because among the non-participants relatively more workers did not improve over time, thus understating the average difference between Y_{ti} and $Y_{t'i}$ for the participants in the calculations. The before-after comparison, however, again performs relatively poorly, $\widehat{M}_{X=1}^{before-after} = 0.25$, again because the example was constructed to display a substantial improvement in the employment situation between $t\prime$ and t even in the absence of the program.

This variety of estimates of the program impact on low- and medium-skilled workers is summarized in the first two columns of **Table 4**. The third column reports the weighted averages (4 out of 14 program participants are medium-skilled) of these estimates, that is $\widehat{M}_{X\in\{0,1\}}^{exact-match}$ to $\widehat{M}_{X\in\{0,1\}}^{cross\text{-}section}$. This example was designed as an illustration not only of the various approaches to evaluation, but also of the fact that identification assumptions rarely do hold exactly in sample. By contrast, the best a researcher can hope for is a close approximation to the true population value. It is therefore important to select a strategy

	low-skilled	medium-skilled	low & medium
genuine impact	0.20	0.00	0.14
estimates			
exact matches	0.25	0.00	0.18
diff.-in-differences	0.30	0.08	0.24
before-after	0.40	0.25	0.36
cross-section	0.10	0.00	0.07

Table 4: Estimated program impact.

that will be right on average – there will always remaining uncertainty around the estimate. In a sample as small as in the example, this remaining uncertainty will be enormous. Thus, while the emphasis on the choice of identification assumption is warranted, one should not forget that sufficient sample sizes are necessary as well.

5 Conclusions

Modern evaluation research teaches several important lessons to policy makers, administrators and researchers alike. First, whenever a policy intervention is undertaken, a serious evaluation effort is required, since the issue is too complex to be solved by introspection or by a casual glance at the economic outcomes of program participants. Second, to be a valid guidance for policy decisions, evaluation has to follow well-respected standards of scientific research. This include that a research question can only be decided upon the weight of the evidence and that any reported results should make both the identification assumptions and the degree of remaining uncertainty transparent. Moreover, data material, methods, and evidence should be made public to allow replication of the results by other researchers. Third, program evaluation entails more than the attribution of consequences to underlying causes, although this is intellectually a very challenging problem; but before this problem can be tackled satisfactorily, it has to be clarified which outcome measures to focus upon and which costs are involved in participating in the program.

Finally, 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, not simply of generating larger quantities of unsatisfactory data or of devoting more manpower to analyzing the data. Since the question is always what the program contributed, on average, to the outcomes of participants over and above the hypothetical outcomes they had experienced, if they had not participated, it is the construction of that *counterfactual* that is at issue. Several approaches have been discussed in this essay, experimental and observational, but none is preferable under every circumstance. As a consequence of its convincing approach to the identification problem, whenever possible one should consider

conducting an experimental study. Experimental approaches go a long way in solving the evaluation problem, but experimentation is often not possible. Furthermore, to infer from experimental results on real-world implementations of the same intervention is often difficult.

Performed appropriately, observational approaches are powerful competitors to experimental studies. They 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 factors just as the ideal experiment would. In this statement, the term *truly comparable* is operational – this is exactly the point where untestable identification assumptions enter the evaluation process. In particular, contrary to the beliefs of many observers, observational approaches are not confined to simple linear regression on observable characteristics. In fact, they generally allow for treatment effects that are heterogenous across the population of participants. Although good econometricians have always known this, the statistical framework presented in this essay has helped to clarify these issues. Since so much rests on the identification assumption, though, perhaps the most important lesson emerging from this new literature is summarized by Heckman et al. (1999) in writing

> "The best solution to the evaluation problem lies in improving the quality of the data on which evaluations are conducted and not in the development of formal econometric methods to circumvent inadequate data."

Thus, while we can expect the formal methods of evaluation research to be refined further, if the objective is an improved body of evidence on the effects of policy interventions, policy makers and administrators have to work more closely together with researchers already at the stage of designing the interventions. Experimental evidence should be collected wherever possible. A final, long-term objective of future research should be the feedback into the interdisciplinary exchange. While the current evaluation literature in economics emphasizes the necessity for social experimentation, in the natural sciences the limitations of experimental approaches in a community setting should be a focus of the analysis. This effort might therefore lead to important qualifications of arguments for and against experimental and observational evaluation strategies and to their synthesis into a unified approach across disciplines.

References

Ashenfelter, Orley (1978) Estimating the Effect of Training Programs on Earnings, *Review of Economics and Statisitics* **60**, 47-57.

Donner, Allan, N. Birkett, and C. Buck (1981) Randomization By Cluster: Sample Size Requirements and Analysis, *American Journal of Epidemiology* **114**, 906-914.

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

Katzer, Jeffrey, Kenneth H. Cook, and Wayne W. Crouch (1998): *Evaluating Information. A Guide for Users of Social Science Research*, 4th edition, Boston et al.: McGraw-Hill.

LaLonde, Robert J. (1995) The Promise of Public Sector-Sponsored Training Programs, *Journal of Economic Perspectives* **9**, 149-168.

Manski, Charles F. (1995) *Identification Problems in the Social Sciences*, Cambridge, Mass. et al.: Harvard University Press.

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.

Rubin, Donald B. (1986) Which Ifs Have Causal Answers?, *Journal of the American Statistical Association* **81**, 961-962.