Caliendo, Marco; Hujer, Reinhard

Working Paper

The microeconometric estimation of treatment effects: an overview

IZA Discussion Papers, No. 1653

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Caliendo, Marco; Hujer, Reinhard (2005) : The microeconometric estimation of treatment effects: an overview, IZA Discussion Papers, No. 1653, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
http://hdl.handle.net/10419/33651

DISCUSSION PAPER SERIES

IZA DP No. 1653

The Microeconometric Estimation of Treatment Effects - An Overview

Marco Caliendo
Reinhard Hujer

July 2005

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

The Microeconometric Estimation of Treatment Effects - An Overview

Marco Caliendo
DIW Berlin and IZA Bonn

Reinhard Hujer
University of Frankfurt,
ZEW Mannheim and IZA Bonn

Discussion Paper No. 1653
July 2005

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of the institute. Research disseminated by IZA may include views on policy, but the institute itself takes no institutional policy positions.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 1653
July 2005

ABSTRACT

The Microeconometric Estimation of Treatment Effects - An Overview[*]

The need to evaluate the performance of active labour market policies is not questioned any longer. Even though OECD countries spend significant shares of national resources on these measures, unemployment rates remain high or even increase. We focus on microeconometric evaluation which has to solve the fundamental evaluation problem and overcome the possible occurrence of selection bias. When using non-experimental data, different evaluation approaches can be thought of. The aim of this paper is to review the most relevant estimators, discuss their identifying assumptions and their (dis-)advantages. Thereby we will present estimators based on some form of exogeneity (selection on observables) as well as estimators where selection might also occur on unobservable characteristics. Since the possible occurrence of effect heterogeneity has become a major topic in evaluation research in recent years, we will also assess the ability of each estimator to deal with it. Additionally, we will also discuss some recent extensions of the static evaluation framework to allow for dynamic treatment evaluation.

JEL Classification: C40, H43, J68

Keywords: evaluation, effect heterogeneity, matching, dynamic treatments

Corresponding author:

Marco Caliendo
Department of Public Economics
DIW Berlin
Königin-Luise-Str. 5
14195 Berlin
Germany
Email: mcaliendo@diw.de

[*] The authors thank Stephan L. Thomsen and Christopher Zeiss for valuable comments. The usual disclaimer applies.

1 Introduction

The need to evaluate the performance of active labour market policies (ALMP) is not questioned any longer. Even though OECD countries spend significant shares of national resources on these measures, unemployment rates remain high or even increase. The ideal evaluation process can be looked at as a series of three steps (Fay, 1996): First, the impacts of the programme on the individual should be estimated (MICROECONOMETRIC EVALUATION). Second, it should be examined if the impacts are large enough to yield net social gains (MACROECONOMIC EVALUATION). Third, it should be answered if this is the best outcome that could have been achieved for the money spent (COST-BENEFIT ANALYSIS). In this paper we focus on the first step. The main question in microeconometric evaluation is if the outcome for an individual is affected by the participation in an ALMP programme or not. We would like to know the difference between the value of the participant's outcome in the actual situation and the value of the outcome if he had not participated in the programme. The fundamental evaluation problem arises because we can never observe both states (participation and non-participation) for the same individual at the same time, i.e. one of the states is counterfactual. Therefore finding an adequate control group is necessary to make a comparison possible. This is not an easy task because participants in programmes usually differ in more aspects than just participation from non-participants. Taking simply the difference between their outcomes after treatment will not reveal the true treatment impact, i.e. will lead to a selection bias.

Microeconometric evaluation has been hampered for a long time, since the available data was not rich enough to allow drawing policy-relevant conclusions. However, in recent years there has been made a lot of progress in enabling researchers to use administrative data for evaluation purposes.[1] Depending on the data at hand, different evaluation strategies can be thought of. Whenever feasible, experimental evaluation will provide the most compelling evidence in most cases. The basic idea of this approach is to assign individuals randomly to the participants' and the control group. Both groups then differ only with respect to participation and the differences in the outcomes can be taken as treatment effects. However, in most European countries experimental data are not available and therefore the evaluator must choose among non-experimental evaluation estimators.[2]

A lot of methodological progress has been made to develop and justify non-experimental evaluation estimators which are based on econometric and statistical methods to solve the fundamental evaluation problem (see e.g. Heckman and Robb (1985b), Heckman and Hotz (1989) or Heckman, LaLonde, and Smith (1999)). In non-experimental studies, the data are not derived in a process that is completely under the control of the researcher. Instead one has to rely on information how individuals actually performed after the intervention. That is, we observe the outcome with

[1] See e.g. Caliendo, Hujer, and Thomsen (2005) for Germany, Sianesi (2004) for Sweden, Gerfin and Lechner (2002) for Switzerland or van den Berg, van der Klaauw, and van Ours (2004) for the Netherlands.

[2] Hence, we will not discuss experimental estimators in this paper. The interested reader is referred to Burtless (1995) and Heckman and Smith (1995).

treatment for participants and the outcome without treatment for non-participants. The objective of observational studies is to use this information to restore the comparability of both groups by design. The aim of this paper is to give an overview of the most relevant evaluation approaches and provide some guidance to researchers on how to choose between them. Thereby we will also discuss the possible occurrence of effect heterogeneity, which has become a major focus in evaluation research in the last years, and the ability of each estimator to deal with it.

The different estimators can be classified with respect to two dimensions. The first dimension is the required data for their implementation, where we can distinguish between longitudinal and cross-sectional methods. The second dimension concerns the handling of selection bias, where two categories arise. The first category contains approaches that rely on the so-called unconfoundedness or selection on observables assumption, like matching, regression and regression discontinuity models. If one believes that the available data is not rich enough to justify this assumption, he has to rely on the second category of estimators which explicitly allows selection on unobservables, too. The methods of instrumental variables and selection models as well as the difference-in-differences estimator will be presented for that situation. Additionally, we will also present some recent extensions of the static evaluation framework to dynamic concepts. Here we will discuss sequential matching estimators, duration models and matching with time-varying treatment indicators.

The paper is organised as follows: First, we are going to introduce the evaluation framework in section 2, where we especially present the potential outcome approach, discuss parameters of interest, selection bias on observable and on unobservable characteristics as well as heterogeneous treatment effects. In section 3 we will present different evaluation strategies and section 4 concludes.

2 The Evaluation Framework

2.1 Potential Outcome Approach and the Fundamental Evaluation Problem

Inference about the impact of a treatment on the outcome of an individual involves speculation about how this individual would have performed in the labour market, had he not received the treatment.[3] The framework serving as a guideline for the empirical analysis of this problem is the potential outcome approach, also known as the Roy(1951)-Rubin(1974)-model.

The main pillars of this model are individuals, treatment (participating in a programme or not) and potential outcomes, that are also called responses.[4] In the basic

[3] This is clearly different from asking whether there is an empirical association between treatment and the outcome (Lechner, 2000). See Holland (1986) for an extensive discussion of concepts of causality in statistics, econometrics and other fields.

[4] It should be clear, that this framework is not restricted to the evaluation of labour market programmes. It applies for every situation where one group of units, e.g. individuals or firms or other entities, receive some form of treatment and others do not.

model there are two potential outcomes (Y^1, Y^0) for each individual, where Y^1 indicates a situation with treatment and Y^0 without. To complete the notation, we additionally denote variables that are unaffected by treatments by X. Attributes X are exogenous in the sense that their potential values for different treatment states coincide (Holland, 1986). Furthermore we define a binary assignment indicator D, indicating whether an individual actually received treatment $(D = 1)$, or not $(D = 0)$. The treatment effect for each individual i is then defined as the difference between his potential outcomes:

$$\Delta_i = Y_i^1 - Y_i^0. \tag{1}$$

The fundamental problem of evaluating this individual treatment effect arises because the observed outcome for each individual is given by:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0. \tag{2}$$

This means that for those individuals who participated in treatment we observe Y^1 and for those who did not participate we observe Y^0. Unfortunately, we can never observe Y^1 and Y^0 for the same individual simultaneously and therefore we cannot estimate (1) directly. The unobservable component in (1) is called the counterfactual outcome.

Concentration on a single individual requires that the effect of the intervention on each individual is not affected by the participation decision of any other individual, i.e. the treatment effect Δ_i for each person is independent of the treatment of other individuals. In statistical literature this is referred to as the stable unit treatment value assumption (SUTVA)[5] and guarantees that average treatment effects can be estimated independently of the size and composition of the treatment population. In particular, it excludes peer-effects as well as cross-effects and general equilibrium effects (Sianesi, 2004).

2.2 Treatment Effects and Selection Bias

Since there will never be an opportunity to estimate individual effects in (1) with confidence, we have to concentrate on population averages of gains from treatment. Two treatment effects are most frequently estimated in empirical studies. The first one is the (population) average treatment effect (ATE), which is simply the difference of the expected outcomes after participation and non-participation:

$$\Delta_{ATE} = E(\Delta) = E(Y^1) - E(Y^0). \tag{3}$$

This effect answers the question which would be the outcome if individuals in the population were randomly assigned to treatment. Heckman (1997) notes that this estimate might not be of relevance to policy makers because it includes the effect on persons for whom the programme was never intended. Therefore, the most important evaluation parameter is the so called average treatment effect on the treated

[5] See Holland (1986) for a further discussion of this concept.

(ATT), which focusses explicitly on the effects on those for whom the programme is actually intended.[6] It is given by:

$$\Delta_{ATT} = E(\Delta \mid D = 1) = E(Y^1 \mid D = 1) - E(Y^0 \mid D = 1). \tag{4}$$

In the sense that this parameter focuses directly on participants, it determines the realised gross gain from the programme and can be compared with its costs, helping to decide whether the programme is successful or not (Heckman, LaLonde, and Smith, 1999). Given equation (4), the problem of selection bias can be straightforwardly seen. Remember that the second term on the right hand side of equation (4) is unobservable as it describes the hypothetical outcome without treatment for those individuals who received treatment. If the condition $E(Y^0 \mid D = 1) = E(Y^0 \mid D = 0)$ holds, we can use the non-participants as an adequate control group. This identifying assumption is likely to hold only in randomised experiments. With non-experimental data it will usually not hold, i.e. $E(Y^0 \mid D = 1) \neq E(Y^0 \mid D = 0)$. Consequently, estimating ATT by the difference in sub-population means of participants $E(Y^1 \mid D = 1)$ and non-participants $E(Y^0 \mid D = 0)$ will therefore lead to a selection bias. Selection bias arises because participants and non-participants are selected groups that would have different outcomes, even in absence of the programme. It might be caused by observable factors, like age or skill differences, or unobservable factors like motivation.

2.3 Potential Outcome Framework and Econometric Methodology

For the further discussion it will be helpful to relate the potential outcome framework to familiar econometric notation. To do so, we follow Blundell and Costa Dias (2002) and define the following outcome equations:

$$Y_{it}^1 = g_t^1(X_i) + U_{it}^1 \quad \text{and} \quad Y_{it}^0 = g_t^0(X_i) + U_{it}^0, \tag{5}$$

where the subscripts i and t index the individual and the time period, respectively. The functions g^0 and g^1 represent the relationship between potential outcomes and the set of observable characteristics. U^0 and U^1 are error terms which have zero mean and are assumed to be uncorrelated with regressors X. For the familiar case of linear regression, the g functions specialise to $g^1(X) = X\beta_1$, and $g^0(X) = X\beta_0$.

Heckman and Robb (1985a) note that the decision to participate in treatment may be determined by a prospective trainee, by a programme administrator, or both. Whatever the specific content of the rule, it can be described in terms of an index function framework. Let IN_i be an index of benefits to the relevant decision

[6] Heckman, LaLonde, and Smith (1999) discuss further parameters, like the proportion of participants who benefit from the programme or the distribution of gains at selected base state values or the treatment effect on the untreated. For distributions of programme impacts, the interested reader is referred to Heckman, Smith, and Clements (1997).

maker from participating in the programme. It is a function of observed (Z_i) and unobserved (V_i) variables. Therefore

$$IN_i = f(Z_i) + V_i, \tag{6}$$

with enrolment in the programme D_i given by

$$\begin{aligned} D_i &= 1 \quad \text{if } IN_i > 0 \text{ and} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Under this specification and the further assumption that treatment takes place in period k, one can define the individual-specific treatment effect for any X_i as:

$$\Delta_{it}(X_i) = Y_{it}^1 - Y_{it}^0 = [g_t^1(X_i) - g_t^0(X_i)] + [U_{it}^1 - U_{it}^0] \qquad \text{with } t > k. \tag{7}$$

The ATT measured in the post-treatment period $t > k$ is then defined as:

$$\Delta_{ATT} \quad = \quad E(\Delta_{it} \mid X = X_i, D_i = 1). \tag{8}$$

The assignment process to treatment is most probably not random. Consequently, the assignment process will lead to non-zero correlation between enrolment (D_i) and the outcome's error term (U^1, U^0). This may occur because of stochastic dependence between (U^1, U^0) and V_i in (6) or because of stochastic dependence between (U^1, U^0) and Z_i. In the former case we have selection on unobservables, whereas in the latter case selection on observables is prevalent (Heckman and Robb, 1985b).

2.4 Homogeneous vs. Heterogeneous Treatment Effects

The discussion in the previous subsection allows us to consider the problem of homogeneous and heterogeneous treatment effects in a common and intuitive way. In recent years a substantial conceptual progress has been made to introduce models in which the impact of a programme is allowed to differ across individuals.[7] The impacts are homogeneous if we assume the effect to be constant across individuals, i.e.:

$$\Delta_t = \Delta_{it}(X_i) = g_t^1(X_i) - g_t^0(X_i) \qquad \text{with } t > k \tag{9}$$

for any i. This implies that g^1 and g^0 are two parallel curves that differ only in the level and, furthermore, that participation-specific error terms are not affected by the treatment status. The outcomes stated in equation (5) can therefore be re-written as (Blundell and Costa Dias, 2002):

$$Y_{it} = g_t^0(X_i) + \Delta_t D_{it} + U_i. \tag{10}$$

However, if the treatment impact varies across individuals this may come systematically through the observables' component or be part of the unobservables. In this case equation (5) changes to:

$$Y_{it} = g_t^0(X_i) + \Delta_t(X_i) D_{it} + [U_{it}^0 + D_{it}(U_{it}^1 - U_{it}^0)], \tag{11}$$

[7] See e.g. the discussion in Smith (2000).

where

$$\Delta_t(X_i) = E[\Delta_{it}(X_i)] = g_t^1(X_i) - g_t^0(X_i) \tag{12}$$

is the expected treatment effect at time t for individuals characterised by X_i (Blundell and Costa Dias, 2002). If programme impacts are heterogeneous, negative average effects must not apply for all strata of the population (Manski, 1997 and 2000). Since abandoning the 'common effect' assumption and identifying the individuals that benefit from programmes provides some scope to improve their future efficiency, we will assess for each estimation method that we will present in the following its capability to deal with heterogeneous treatment effects.

3 Non-Experimental Evaluation Methods

The discussion in subsections 2.2 and 2.3 has made clear that the problem of selection bias is a severe one and cannot be solved with more data, since the fundamental evaluation problem will not disappear. We have a distorted representation of a true population in a sample as a consequence of a sampling rule, which is the essence of the selection problem (Heckman, 2001). Hence, we have to use some identifying assumptions to draw inference about the hypothetical population based on the observed population. In the following subsections we will present several evaluation approaches. Each approach invokes different identifying assumptions to construct the required counterfactual outcome. We will start the following discussion with two estimators (matching and regression) that are based on the selection on observables assumption.[8] Following that we introduce three estimators that allow for selection on unobservables, too, namely difference-in-differences, instrumental variables and selection models. Finally, we also briefly discuss regression discontinuity models and the estimation of treatment effects in a dynamic framework.

3.1 Matching Estimator

Matching is based on the identifying assumption that conditional on some covariates X, the outcome Y is independent of D. In the notation of Dawid (1979), it assumes that:

Assumption 1 *Unconfoundedness:* $Y^0, Y^1 \amalg D \mid X,$

where $\amalg$ denotes independence. If assumption 1 is true, then $F(Y^0 \mid X, D = 1) = F(Y^0 \mid X, D = 0)$ and $F(Y^1 \mid X, D = 1) = F(Y^1 \mid X, D = 0)$. This means, that conditional on X, non-participant outcomes have the same distribution that participants would have experienced if they had not participated in the programme and

[8] See Imbens (2004) for an extensive overview of estimating average treatment effects under unconfoundedness.

vice versa (Heckman, Ichimura, and Todd, 1997). Similar to randomisation in a classical experiment, matching balances the distributions of all relevant, pre-treatment characteristics X in the treatment and comparison group.[9] Thus it achieves independence between the potential outcomes and the assignment to treatment. Hence, if the mean exists, $E(Y^0 \mid X, D = 1) = E(Y^0 \mid X, D = 0) = E(Y^0 \mid X)$, $E(Y^1 \mid X, D = 1) = E(Y^1 \mid X, D = 0) = E(Y^1 \mid X)$ and the missing counterfactual means can be constructed from the outcomes of non-participants and participants. In order for both expressions to be well defined simultaneously for all X, it is additionally assumed that

Assumption 2 *Overlap:* $0 < Pr(D = 1 \mid X) < 1$,

for all X. This implies that the support of X is equal in both groups, i.e. $S = \text{Support}(X|D = 1) = \text{Support}(X|D = 0)$.[10] Assumption 2 prevents X from being a perfect predictor in the sense that we can find for each participant a counterpart in the non-treated population and vice versa. If there are regions where the support of X does not overlap for the treated and non-treated individuals, matching has to be performed over the common support region only. The estimated effects have then to be redefined as the mean treatment effect for those individuals falling within the common support (Blundell, Dearden, and Sianesi, 2004). Rosenbaum and Rubin (1983) call assumptions 1 and 2 together 'strong ignorability'. Under 'strong ignorability' ATE in (3) and ATT in (4) can be defined for all values of X. Heckman, Ichimura, and Todd (1998) demonstrate that these conditions are overly strong and all that is needed for estimation of (3) and (4) is mean-independence. Additionally, if one is interested in estimating ATT only, the assumptions can be relaxed in a different direction, because one needs only to assume $Y^0 \amalg D \mid X$ and the weaker overlap assumption $P(D = 1 \mid X) < 1$. The mean impact of treatment on the treated can be written as:

$$\Delta_{ATT}^{MAT} = E(Y^1|X, D = 1) - E_X[E(Y^0|X, D = 0)|D = 1], \tag{13}$$

where the first term can be estimated from the treatment group and the second term from the mean outcomes of the matched comparison group. The outer expectation is taken over the distribution of X in the treated population. The method of matching can also be used to estimate ATT at some points $X = x$, where x is a particular realisation of X. Before we move on to the next estimator, let us briefly mention two things. First, it should be clear that conditioning on all relevant covariates is limited in case of a high dimensional vector X. For that case Rosenbaum and Rubin (1983) suggest the use of so-called balancing scores to overcome this dimensionality problem. One possible balancing score is the propensity score and matching procedures based on this balancing score are known as propensity score matching.[11]

[9] If we say relevant we mean all those covariates that influence the assignment to treatment as well as the potential outcomes.

[10] The support is a statistical term meaning the set of values for which a density function is non-zero, i.e. the set of values of a variable that one observes with positive probability (Hui and Smith, 2002).

[11] See Rosenbaum (2002) or Caliendo and Kopeinig (2005) for an introduction into propensity score matching estimators and some guidance for their implementation.

Second, there are several different matching algorithms suggested in the literature, e.g. kernel or nearest-neighbour matching, and the choice between them is not trivial since it involves a trade-off between bias and variance. We will not discuss these algorithms here (see Smith and Todd (2005) for an overview.)

3.2 Linear Regression Approach

Even though regression and matching both rely on the unconfoundedness assumption, there are some key differences between both approaches which are worth discussing. One key difference is that matching, due to its non-parametric nature, avoids functional form assumptions implicit in linear regression models. Basically, linear regression makes the additional assumption that simply conditioning linearly on X suffices to eliminate selection bias. We have already introduced the linear regression notation in subsection 2.3, let us reconsider the main points. For notational convenience we drop the individual subscript i and the time subscript t. Hence, the potential outcomes in a linear regression framework can be written as $Y^1 = X\beta_1 + U^1$ and $Y^0 = X\beta_0 + U^0$ and ATT under regression is given by:

$$\Delta_{ATT}^{Reg} = E(Y^1 - Y^0|X, D = 1) = X(\beta_1 - \beta_0) + E(U^1 - U^0|X, D = 1). \tag{14}$$

The identifying assumption needed to justify regression under unconfoundedness is analogue to assumption 1 and can be re-written as:

Assumption 3 *Unconfoundedness in Regression:* $U^0, U^1 \amalg D \mid X.$

In the matching framework, the goal is to set the bias $B(X) = 0$ which basically only requires that the mean of the error terms in the treatment group given a covariate cell X equals the corresponding mean in the control group, that is $B(X) = E(U^1|X, D = 1) - E(U^0|X, D = 0) = 0$. Basically, this means that it is possible to match on variables that are correlated with the error term in the outcome equation (Hui and Smith, 2002). In the regression framework, however, we need to eliminate the dependence between (U^0, U^1) and X, that is $E(U^1|X, D = 1) = E(U^0|X, D = 0) = 0$ (Heckman, Ichimura, Smith, and Todd, 1998). Of course, as Smith (2000) notes, the difference between both approaches fades with the inclusion of a sufficient number of higher-order and interaction terms in the regression. However, not only is such an inclusion not very common in practice, it is also not straightforward to choose these terms. Moreover, whereas matching estimators do rely on the common support assumption, regression estimators do not. Matching ensures that comparisons between treated and control individuals are only made over the region of common support. In contrast, standard regression approaches will produce estimates even in the absence of similar comparison units since the linear functional form assumption fills in for the missing data. The regression identifies the untreated outcome model in the region of the data where the untreated observations lie and then projects it out into the region of the data where the treated units lie, thereby implicitly estimating the counterfactual (Smith, 2004). Another key difference between regression and matching is the way both approaches handle heterogeneous treatment effects. As Lechner (2002) notes, the non-parametric matching approach leaves the individual causal effect unrestricted and allows individual effect heterogeneity in the population. This is not true for the regression approach which will not recover ATT, although, at times it might provide a close approximation as shown by Angrist (1998) and Blundell, Dearden, and Sianesi (2004).

3.3 Instrumental Variables Estimator

Let us now turn to estimators that account for selection on unobservables, too. We will start with the method of instrumental variables (IV). Its underlying identification strategy is to find a variable which determines treatment participation but does not influence the outcome equation. The instrumental variable affects the observed outcome only indirectly through the participation decision and hence causal effects can be identified through a variation in this instrumental variable. IV methods are extensively discussed in Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) among others. In terms of the discussion in subsection 2.3, IV requires the existence of at least one regressor to the decision rule, Z^*, that satisfies the following three conditions (Blundell and Costa Dias, 2000):

1. Z^* determines programme participation. For that to be true, it has to have a non-zero coefficient in the decision rule in equation 6.
2. We can find a transformation, s, such that $s(Z^*)$ is uncorrelated with the error terms (U^1, V) and (U^0, V), given the exogenous variables X.
3. Z^* is not completely determined by X.

The variable Z^* is then called the instrument. In providing variation that is correlated with the participation decision but does not affect potential outcomes from treatment directly, it can be used as a source of exogenous variation to approximate randomised trials (Blundell and Costa Dias, 2000). For a binary instrument $Z^* \in \{0, 1\}$, we can write the IV estimator as:

$$\Delta^{IV} = \frac{E(Y|X, Z^* = 1) - E(Y|X, Z^* = 0)}{P(D = 1|X, Z^* = 1) - P(D = 1|X, Z^* = 0)}. \tag{15}$$

Clearly, a major problem with this estimator is to find a good instrument. In the treatment evaluation problem it is hard to think of variables that satisfy all three above mentioned assumptions. The difficulty lies mainly in the simultaneous requirement that the variable has to predict participation but does not influence the outcome equation.[12] As pointed out by Blundell and Costa Dias (2000), a second drawback arises when considering the heterogeneous treatment framework. Recall that the error term from equation (11) in subsection 2.4 is given by $[U_{it}^0 + D_{it}(U_{it}^1 - U_{it}^0)]$. Even if Z^* is uncorrelated with U_{it}, the same cannot be true by definition for $U_{it}^0 + D_{it}(U_{it}^1 - U_{it}^0)$ since Z^* determines D_i by assumption. The violation of this assumption invalidates the application of IV methodology in a heterogeneous

[12] Bound, Jaeger, and Baker (1995) additionally point out that using 'weak instruments' may lead to inefficiency and possibly also inconsistency of the IV estimates.

framework (Blundell and Costa Dias, 2000). However, in this situation it might still be possible to provide a potentially interesting parameter of the IV estimation - called local average treatment effect (LATE) by Imbens and Angrist (1994). This estimator identifies the treatment effect for those individuals (with characteristics X) who are induced to change behaviour because of a change in the instrument.[13] It should be clear that each instrument implies its own LATE, and LATEs for two different instruments may differ substantially depending on the impacts realised by the persons each instrument induces to participate (Hui and Smith, 2002). Hence, LATE should not be confused with ATE or ATT.[14]

3.4 Selection Model

This method is also known as the Heckman selection estimator (Heckman, 1978). It is more robust than the IV method but also more demanding in the sense that it imposes more assumptions about the structure of the model. Two main assumptions are required (Blundell and Costa Dias, 2000):

1. There has to be one additional regressor in the decision rule which has a non-zero coefficient and which is independent of the error term V.
2. Additionally, the joint density of the distribution of the errors U_{it} and V_i has to be known or can be estimated.

The basic idea of this estimator is to control directly for the part of the error term in the outcome equation that is correlated with the participation dummy variable. It can be seen as a two-step-procedure. First, the part of the error term U_{it} that is correlated with D_i is estimated. Second, this term is then included in the outcome equation and the effect of the programme is estimated. By construction, the remains of the error term in the outcome equation are not correlated with the participation decision any more (Blundell and Costa Dias, 2000). To explain this estimator, we re-write the outcome equation in (10) in linear terms as $Y_i = X_i\beta_0 + \alpha D_i + U_i$, where we have dropped the time subscript for convenience. If we assume the popular special case where U_i and V_i are assumed to follow a joint normal distribution, we may write the conditional outcome expectations as:

$$E(Y_i|D_i = 1) \quad = \quad \beta + \alpha + \rho\frac{\phi(Z_i\gamma)}{\Phi(Z_i\gamma)} \tag{16}$$

and

$$E(Y_i|D_i = 0) \quad = \quad \beta - \rho\frac{\phi(Z_i\gamma)}{1 - \Phi(Z_i\gamma)}. \tag{17}$$

[13] Additionally to those assumptions already made, we further have to assume that the instrument has the same directional effect on all those whose behaviour it changes. This assumption rules out the co-existence of defiers and compliers and is known as 'monotonicity assumption' (Imbens and Angrist, 1994).

[14] For continuous instruments Heckman and Vytlacil (2001) define marginal treatment effects (MTE). This is the effect on the person just indifferent to participate at his current value of the instrument. They also show that all common treatment effects, e.g. ATE and ATT, can be written as particular integrals of such MTEs.

The new regressor includes the part of the error term that is correlated with the decision process in the outcome equation, allowing us to separate the true impact of the treatment from the selection process. Thus it is possible to identify α as outlined above, by replacing γ with $\hat{\gamma}$ and running a least-squares regression on the conditional outcome expectations (Blundell and Costa Dias, 2000). Blundell and Costa Dias (2000) also show that this approach is capable of identifying ATT if effects are assumed to be heterogeneous.[15] The Heckman selection estimator is not without critique, which rests mainly on the following point (see e.g. Puhani (2000)): If there are no exclusion restrictions, the models are identified only by assumptions about functional form and error distributions. This may lead to large standard errors and results that are very sensitive to the particular distributional assumptions invoked. This point of criticism is very closely related to the problem of finding a good instrument as described for the IV method. In fact, in a recent paper Vytlacil (2002) shows that the identifying assumptions for the selection model are equivalent to those invoked by Imbens and Angrist (1994) in the linear instrumental variables context.

3.5 Difference-in-Differences Estimator

The difference-in-differences (DID) estimator requires access to longitudinal data and can be seen as an extension to the classical before-after estimator (BAE). Whereas BAE compares outcomes of participants after they participate in the programme with their outcomes before they participate, the DID estimator eliminates common time trends by subtracting the before-after change in non-participant outcomes from the before-after change for participant outcomes.

The DID-estimator forms simple averages over the group of participants and non-participants between pre-treatment period t' and post-treatment period t, that is, changes in the outcome variable Y for treated individuals are contrasted with the corresponding changes for non-treated individuals (Heckman, Ichimura, Smith, and Todd, 1998):

$$\Delta^{DID} = [Y_t^1 - Y_{t'}^0 \mid D = 1] - [Y_t^0 - Y_{t'}^0 \mid D = 0]. \tag{18}$$

The identifying assumption of this method is:

$$E(Y_t^0 - Y_{t'}^0 \mid D = 1) = E(Y_t^0 - Y_{t'}^0 \mid D = 0). \tag{19}$$

The DID estimator is based on the assumption of time-invariant linear selection effects, so that differencing the differences between participants and non-participants eliminates the bias (Heckman, Ichimura, Smith, and Todd, 1998). To make this point clear, we can re-write the outcome for an individual i at time t as: $Y_{it} = \pi_{it} + D_{it} \cdot Y_{it}^1 + (1 - D_{it}) \cdot Y_{it}^0$, where π_{it} captures the effects of selection on unobservables. The validity of the DID estimator then relies on the assumption: $\pi_{it} = \pi_{it'}$, where it is not required that the bias vanishes completely, but that it remains constant (Heckman, Ichimura, Smith, and Todd, 1998). One problem when

[15] However, ATE is not identified.

using DID is Ashenfelter's dip, i.e. a situation where shortly before participation in an ALMP programme the employment situation of future participants deteriorates (Ashenfelter, 1978). If the 'dip' is transitory and the dip is eventually restored even in the absence of participation in the programme, the bias will not average out. To allow a more detailed discussion, Blundell and Costa Dias (2002) further decompose π_{it} in three parts: an individual-specific fixed effect, a common macroeconomic effect and a temporary individual-specific effect. Clearly, for the DID to be unbiased it is sufficient that selection into treatment is independent from the temporary individual-specific effect, since the other two effects vanish in the sequential differences. They also discuss the case where the macroeconomic effect has a differential impact across the group of participants and non-participants. This may happen when both groups differ on unobserved characteristics which make them react differently to macroeconomic shocks. To overcome this problem they propose a differential trend adjusted DID estimator (Blundell and Costa Dias, 2002). Heckman, Ichimura, Smith, and Todd (1998) combine the DID approach with the already presented matching estimator by comparing the before-after outcome of participants with those of matched non-participants. Smith and Todd (2005) show that this 'conditional DID estimator' is more robust than traditional cross-section matching estimators, as it allows selection on observables as well as time-invariant selection on unobservables.

3.6 Regression Discontinuity Model

The regression discontinuity model (RDM) can be seen as a particular type of instrumental variable identification strategy. It uses discontinuities in the selection process to identify causal effects. In this model, treatment depends on some observed variable, Z, according to a known, deterministic rule, such as $D = 1$ if $Z > \overline{Z}$ and $D = 0$ otherwise (Heckman, LaLonde, and Smith, 1999). The variable Z has direct impact on Y as well as an indirect impact on Y through D. This indirect impact is the causal effect we would like to identify. Frölich (2002) notes that this effect is identified if the direct and indirect impacts of Z on Y can be separated. A good example is given by Angrist and Lavy (1999). They estimate the effects of the size of school classes on test scores of pupils. The discontinuity stems from a rule that classes with more than 40 pupils have to be divided in two, such that in each class remain on average 20.5 pupils. The treatment in this case is the fact of being in a class which has been divided due to reaching the maximum size of 41 pupils. Clearly, in this example class size affects the test scores of the pupils directly but also indirectly through D. Other examples are given by Hahn, Todd, and Van der Klaauw (1999), who analyse the effect of antidiscrimination laws of minority workers by exploiting the fact that only firms with more than 15 employees are subject to these laws. Barnow, Cain, and Goldberger (1980) consider a hypothetical enrichment programme for disadvantaged children, where children with a family income below a cut-off level receive the programme and all others do not. Whenever such deterministic rules are in place, RDD might be a possibility to estimate causal effects.

There are several things to note about RDM (see e.g. Heckman, LaLonde, and Smith (1999)). First, it is assumed that selection is on observable characteristics only. Second, it should be clear that there is no common support for participants and non-participants making matching impossible. Hence, RDM takes over when there is selection on observables (here: the deterministic rule) but the overlapping support condition required for matching breaks down (with a certain Z you either belong to the participant or the non-participant group). Finally, the selection rule is assumed to be deterministic and known and that variation in the relevant variable Z is exogenous (Hahn, Todd, and Van der Klaauw, 2001).

3.7 Dynamic Evaluation Concepts

3.7.1 Sequential Matching Estimators

What we have discussed so far is basically a static evaluation framework where an individual can participate in one programme (or not). A recent extension of this framework for matching estimators considers the case, where individuals can participate in subsequent treatments. Lechner and Miquel (2002) discuss identifying assumptions for so-called sequential matching estimators. These estimators mimic the matching estimators described above but allow to estimate effects in a dynamic causal model. Their framework can be made clear in a three-periods-two-treatments model. We follow the discussion in Lechner (2004) and present the needed additional notation in the following. First, we introduce a time index $t \in \{0, 1, 2\}$ and extend the treatment indicator D by this time index, that is $D = (D_0, D_1, D_2)$. It is further assumed that in period 0 everybody is in the same treatment state $D_0 = 0$, whereas from the second period on D_t can take two values. Realisations of D_t are denoted by $d_t \in \{0, 1\}$. So in period 1 an individual is observed in exactly one of these two treatments $(0, 1)$, whereas in period 2 an individual participates in one of four possible treatment sequences $\{(0, 0), (1, 0), (0, 1), (1, 1)\}$. Additionally, the history of variables up to period t are denoted by a bar below a variable, e.g. $\underline{d}_2 = (d_1, d_2)$. The potential outcomes are indexed by treatments and the time period, i.e. $Y^{\underline{s}_t} = (Y_0^{\underline{d}_t}, Y_1^{\underline{d}_t}, Y_2^{\underline{d}_t})$. The observed outcomes are given by the following equation:

$$\begin{aligned} Y_t = D_1 Y_t^1 + (1 - D_1) Y_t^0 = D_1 D_2 Y_t^{1,1} + D_1(1 - D_2) Y_t^{1,0} + \\ (1 - D_1) D_2 Y_t^{0,1} + (1 - D_1)(1 - D_2) Y_t^{0,0}. \end{aligned} \tag{20}$$

As in the static model, variables that influence treatment selection and potential outcomes are called attributes and are denoted by X. An important distinction has to be made regarding the exogeneity of these variables. Whereas in the static model exogeneity is assumed, in the dynamic model the X-variables in later periods can be influenced by treatment realisations. Hence, there are potential values of these variables as well: $X^{d_t} = (X_0^{d_t}, X_1^{d_t}, X_2^{d_t})$, where e.g. $X_1^{d_1}$ may contain $Y_1^{d_1}$ or functions of it. The sequential matching framework is a powerful tool and is applicable for situations where individuals can participate more than once in a programme and where it is possible to identify treatment sequences. It allows intermediate outcomes to play a role in the participation decision for sequential participation and thus allows estimation in a dynamic context. Since Lechner (2004) presents the only application so far, practical experiences with this identifying strategy are rather limited.

3.7.2 Duration Models

Another methodology for modelling dynamically assigned treatments is the application of duration models (Abbring and van den Berg, 2003). This approach of the evaluation of an ALMP programme is done with a bivariate duration model. Normalising the point in time when an individual enters unemployment to zero, T_e denotes the duration until the individual enters a regular employment and T_p the duration until the individual enters an ALMP training programme with realisations t_u and t_p respectively. The durations T_e and T_p are assumed to vary with observable characteristics x and unobservable characteristics v_e and v_p. Abbring and van den Berg (2003) assume that the realisation t_p affects the distribution of T_e in a deterministic way from t_p onwards and consider not only the binary information if an individual participates in a programme, but also the additional information on the timing of the treatment within the unemployment spell.

For the specification of the hazard rates a mixed proportional hazard model is used. Basic feature of this model is that the duration dependence, observable and unobservable covariates enter the hazard rate multiplicatively:

$$\theta_e(t|t_p, x, v_e) = \lambda_e(t) \exp[x'\beta_e + \mu(t - t_p)I(t > t_p) + v_e] \tag{21}$$

$$\theta_p(t|x, v_p) = \lambda_p(t) \exp[x'\beta_p + v_p] \tag{22}$$

The hazard rate for the transition into regular employment at time t consists of the baseline hazard $\lambda_e(t)$ that determines the duration dependence, the systematic part $\exp(x'\beta_e)$ that determines the individual level of the hazard rate conditional on the observable characteristics and the unobserved heterogeneity term $\exp(v_e)$ that determines the level of the hazard conditional on the unobserved characteristics. The treatment effect $\exp[\mu(t-t_p)I(t > t_p)]$ with $I(t > t_p)$ as an indicator function taking the value 1 if $t > t_p$, is specified as a function of the difference $t - t_p$. In general we allow the treatment effect to vary over the time after the treatment has started. The treatment effect can be interpreted as a shift of the hazard rate by $\exp(\mu(t - t_p))$. The transition rate from unemployment into programme participation is analogously specified as a mixed proportional hazard model with the baseline hazard $\lambda_p(t)$, the systematic part $\exp(x'\beta_p)$ and the unobserved heterogeneity term $\exp(v_p)$.

As stated by Abbring and van den Berg (2003) for the identification of the treatment effect, the selectivity of the programme participation must be considered. Selectivity is present if individuals with a relatively high transition rate into employment also have a relatively high transition into programme participation. If we observe such a pattern this can result from two sources. First, a positive treatment effect rises for the group of participants the transition rate into regular employment. Second, individuals with treatment have a relatively high v_e and therefore have a higher transition rate anyway. In the second case we obviously would observe a positive correlation between v_e and v_p. Therefore the joint distribution $G(v_e, v_p)$ has to be specified. Abbring and van den Berg (2003) show that the bivariate model (21) and (22) and especially the treatment effect is identified. The identification is nonparametric, since it does not require any parametric assumptions with respect to the baseline hazard and the unobserved heterogeneity distribution (Abbring and van den Berg, 2003). Furthermore the identification does not require exclusion restrictions on x which are often hardly to justify from a theoretical point of view.

The specified model for the transition rates θ_e and θ_p rules out any anticipatory effects of vocational training programmes. An anticipatory effect is given if the realisation t_p has an effect on θ_e before t_p. However these effects can be modelled by an additional treatment state in the framework of a duration model (Abbring and van den Berg, 2003).

3.7.3 Matching with Time-Varying Treatment Indicators

An alternative concept of modelling dynamic treatment effects is presented by Fredriksson and Johansson (2004) and Sianesi (2004). They introduce a non-parametric matching estimator that takes the timing of events into account but does not rely on proportionality assumptions. An important topic in this framework is the choice of an appropriate control group. Instead of defining control individuals as those who never participate, Sianesi (2004) defines control individuals as those who did not participate until a certain time period. Fredriksson and Johansson (2004) formalise her approach and argue that the standard way of defining a control group, i.e. those individuals who never participated in a given time interval, might lead to biased results, because the CIA might be violated as the treatment indicator itself is defined conditional on future outcomes. Following Sianesi (2004), the key choice faced by the unemployed in this framework is not whether to participate at all, but whether to participate in a programme or not now. In the latter case, the individual searches longer in open unemployment. The corresponding parameter of interest in this setting is then defined as the effect of joining a programme now in contrast to waiting longer. The population of interest at time u are those still openly unemployed after u months. Treatment receipt in u is denoted by $D^{(u)} = 1$. The comparison group consists of all persons who do not join at least up to u, denoted by $D^{(u)} = 0$. The outcome of interest is defined over time t and is given by $Y_t^{(u)}$. The potential outcome if an individual joins in u is denoted by $Y_t^{1(u)}$ and if he does not join at least up to u by $Y_t^{0(u)}$. For each point of elapsed unemployment duration the parameter of interest is:

$$\begin{aligned} \Delta_u^t = E(Y_t^{1(u)} - Y_t^{0(u)} | D^{(u)} = 1) = E(Y_t^{1(u)} | D^{(u)} = 1) \\ -E(Y_t^{0(u)} | D^{(u)} = 1), \quad \text{for} \quad t = u, u+1, \ldots, T. \end{aligned} \tag{23}$$

This is the average impact at time t, for those joining a programme in their u^{th} month of unemployment compared to waiting longer in open unemployment. Sianesi (2004) notes that the treatment effects are based on a comparison of individuals who have reached the same elapsed duration of unemployment. Measurement starts at time u, the start of the programme and therefore possible locking-in effects might encounter. The second term on the right hand side is not identified and the CIA needed in that case is given by:

$$Y_t^{0(u)} \amalg D^{(u)} | X = x \quad \text{for} \quad t = u, u+1, \ldots, T, \tag{24}$$

which means that given a set of observed characteristics X, the counterfactual distribution of $Y_t^{0(u)}$ for individuals joining in u is the same as for those not joining in u and waiting longer. The estimated treatment effect is then the effect for those who participate in a programme at some time in their unemployment spell instead of waiting longer. Even though this is not a standard evaluation parameter of interest, it still shows whether a programme was effective or not. In a recent paper, Steiger (2004) compares the sensitivity of the effects of Swiss labour market policy with respect to the definition of the non-participant group and makes clear that further research in this direction is fruitful for evaluating labour market policies. Bergemann, Fitzenberger, and Speckesser (2004) combine duration modelling with the already discussed conditional DID estimator to a 'dynamic conditional DID estimator in hazard rates' and estimate the employment effects of vocational training in East Germany.

4 Summary - Which Estimator to Choose?

We have presented several different evaluation strategies in this paper. The final question to be answered is: Which strategy to choose when evaluating labour market programmes? Unfortunately, there is no 'one' answer to this question because there is no 'magic bullet' that will solve the evaluation problem in any case. As described above, different strategies invoke different identifying assumptions and also require different kinds of data for their implementation. When those assumptions hold, a given estimator will provide consistent estimates of certain parameters of interest (Smith, 2004). The literature provides a lot of guidance for making the right choice, based either on experimental datasets to benchmark the performance of alternative evaluation estimators or Monte-Carlo simulations.[16]

The different estimators can be classified with respect to two dimensions. The first dimension is the required data for their implementation. Except the DID estimator, the presented methods for the static evaluation framework require only cross-sectional information for the group of participants and non-participants. However, longitudinal information might help to justify the unconfoundedness assumption, enables the researcher to combine e.g. matching with DID estimators and allows an extension to dynamic concepts of treatment evaluation. The second dimension

[16] See LaLonde (1986), Heckman and Hotz (1989), Heckman, Ichimura, Smith, and Todd (1998) and Dehejia and Wahba (1999, 2002) among others for experimental evidence and e.g. Heckman, LaLonde, and Smith (1999) for Monte-Carlo simulations.

concerns the handling of selection bias. We have presented three estimators that are based on the unconfoundedness assumption. Clearly, the most crucial point for these estimators is that the identifying assumption is in general a very strong one and they are only as good as the used control variables X (Blundell, Dearden, and Sianesi, 2004). If the assumption holds, both matching and regression, can be used. Since regression analysis ignores the common support problem, imposes a functional form for the outcome equation, and is not as capable as matching of handling effect heterogeneity, matching might be preferred. If there is no common support at all, regression discontinuity models can be applied. For the situation where there is selection on unobservables, too, we have presented three strategies. Whereas selection models try to model the selection process completely, IV methods focus on searching a source of independent variation affecting the participation decision (but not the outcome) and DID methods erase a time-invariant selection effect by differencing outcomes of participants and non-participants before and after treatment took place. The crucial assumption for the latter approach is that the selection bias is time invariant. Finding a suitable and credible instrument and heterogeneous treatment effects are possible drawbacks for the IV method. The latter point is not a problem for selection models[17], even though this flexibility comes at a price, because a full specification of the assignment rule and stronger assumptions are required. Hence, if the common effect assumption is plausible in a given context, the IV estimator might be preferred (Smith, 2004). Finally, we have also presented some recent extensions of the static evaluation framework to analyse dynamic treatment effects. The sequential matching estimator is a powerful tool, if individuals can participate in subsequent treatments. Duration models play an important role if the timing of events, i.e. the time until the treatment and/or the transition to a job occurs, shall be taken into account. Finally, we have also discussed matching estimators with a time-varying treatment indicator where the estimated treatment effect is the effect for those individuals who participate in a programme at some time in their unemployment spell instead of waiting longer.

Let us conclude this paper with a somewhat obvious quote from Smith (2000): ‘Better data helps a lot!’. The discussion has shown that each non-experimental estimation strategy relies on identifying assumptions and has to be justified case-by-case. In an ideal world, the evaluator is already involved at early stages of the programme design and has influence on the data collected for later evaluation. In that case, one can make sure to collect those data needed to justify either the unconfoundedness assumption or to create an instrument (exclusion restriction) that allows the use of IV methods or selection models. If the evaluator is instead faced with an ongoing programme, he carefully has to assess which identification strategy works for the situation at hand, taking the design of the programme, the selection process, and the available data into account.

[17] Björklund and Moffit (1987) were the first who discussed this extension.

References

Abbring, J., and G. van den Berg (2003): "The Nonparametric Identification of Treatment Effects in Duration Models," *Econometrica*, 71(5), 1491–1517.

Angrist, J. (1998): "Estimating the Labor Market Impact of Voluntary Military Sevice using Social Security Data on Military Applicants," *Econometrica*, 66(2), 249–288.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–472.

Angrist, J. D., and D. Lavy (1999): "Using Maimodis' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114(2), 533–575.

Ashenfelter, O. (1978): "Estimating the Effects of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47–57.

Barnow, B., G. Cain, and A. Goldberger (1980): "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies Vol.5*, ed. by E. Stromsdorfer, and G. Farkas, pp. 290–317. Sage Publications.

Bergemann, A., B. Fitzenberger, and S. Speckesser (2004): "Evaluating the Dynamic Employment Effects of Training Programs in East Germany Using Conditional Difference-in-Differences," Discussion Paper No. 04-41, ZEW.

Björklund, A., and R. Moffit (1987): "Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics*, 69(1), 42–49.

Blundell, R., and M. Costa Dias (2000): "Evaluation Methods for Non-Experimental Data," *Fiscal Studies*, 21(4), 427–468.

——— (2002): "Alternative Approaches to Evaluation in Empirical Microeconomics," *Portuguese Economic Journal*, 1, 91–115.

Blundell, R., L. Dearden, and B. Sianesi (2004): "Evaluating the Impact of Education on Earnings in the UK: Models, Methods and Results from the NCDS," Working Paper No. 03/20, The Institute for Fiscal Studies.

Bound, J., D. Jaeger, and R. Baker (1995): "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Burtless, G. (1995): "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, 9, 63–84.

Caliendo, M., R. Hujer, and S. Thomsen (2005): "The Employment Effects of Job Creation Schemes in Germany - A Microeconometric Evaluation," Discussion Paper No. 1512, IZA, Bonn.

Caliendo, M., and S. Kopeinig (2005): "Some Practical Guidance for the Implementation of Propensity Score Matching," Discussion Paper No. 1588, IZA, Bonn.

DAWID, A. (1979): "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society, Series B*, 41, 1–31.

DEHEJIA, R. H., AND S. WAHBA (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94(448), 1053–1062.

——— (2002): "Propensity Score Matching Methods for Nonexperimental Causal Studies," *The Review of Economics and Statistics*, 84(1), 151–161.

FAY, R. (1996): "Enhancing the Effectiveness of Active Labor Market Policies: Evidence from Programme Evaluations in OECD Countries," Labour Market and Social Policy Occasional Papers, OECD.

FREDRIKSSON, P., AND P. JOHANSSON (2004): "Dynamic Treatment Assignment - The Consequences for Evaluations Using Observational Data," Discussion Paper No. 1062, IZA.

FRÖLICH, M. (2002): *Programme Evaluation and Treatment Choice.* Springer, Lecture Notes in Economics and Mathematical Systems.

GERFIN, M., AND M. LECHNER (2002): "A Microeconometroc Evaluation of the Active Labour Market Policy in Switzerland," *The Economic Journal*, 112, 854–893.

HAHN, J., P. TODD, AND W. VAN DER KLAAUW (1999): "Evaluating the Effect of an Antidiscrimination Law Using a Regression-Discontinuity Design," Working Paper No. 7131, NBER.

——— (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69(1), 201–209.

HECKMAN, J. (1978): "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46, 931–959.

——— (1997): "Instrumental Variables - A Study of the Implicit Behavioral Assumptions Used in Making Program Evaluations," *The Journal of Human Resources*, 32(3), 441–462.

HECKMAN, J. (2001): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy*, 109(4), 673–748.

HECKMAN, J., AND J. HOTZ (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862–874.

HECKMAN, J., H. ICHIMURA, J. SMITH, AND P. TODD (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

HECKMAN, J., H. ICHIMURA, AND P. TODD (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64, 605–654.

——— (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

Heckman, J., R. LaLonde, and J. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics Vol.III*, ed. by O. Ashenfelter, and D. Card, pp. 1865–2097. Elsevier, Amsterdam.

Heckman, J., and R. Robb (1985a): "Alternative Methods for Evaluating the Impact of Interventions - An Overview," *Journal of Econometrics*, 30, 239–267.

——— (1985b): "Alternative Models for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman, and B. Singer, pp. 156–245. Cambridge University Press.

Heckman, J., and J. Smith (1995): "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9, 85–110.

Heckman, J., and E. Vytlacil (2001): "Policy-Relevant Treatment Effects," *American Economic Review*, 91(2), 107–111.

Heckman, J. J., J. Smith, and N. Clements (1997): "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *The Review of Economic Studies*, 64(4), 487–535.

Holland, P. (1986): "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960.

Hui, S., and J. Smith (2002): "The Labor Market Impacts of Adult Education and Training in Canada," Report prepared for hrdc.

Imbens, G. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86(1), 4–29.

Imbens, G., and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

LaLonde, R. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review*, 76, 604–620.

Lechner, M. (2000): "An Evaluation of Public Sector Sponsored Continuous Vocational Training Programs in East Germany," *Journal of Human Resources*, 35, 347–375.

——— (2002): "Some practical issues in the evaluation of heterogenous labour market programmes by matching methods," *Journal of the Royal Statistical Society, A*, 165, 59–82.

——— (2004): "Sequential Matching Estimation of Dynamic Causal Models," Discussion Paper, No. 1042, IZA.

Lechner, M., and R. Miquel (2002): "Identification of Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions," Working Paper, SIAW.

Manski, C. (1997): "The Mixing Problem in Programme Evaluation," *The Review of Economic Studies*, 64, 537–553.

Manski, C. (2000): "Using Studies of Treatment Response to Inform Treatment Choice in Heterogeneous Populations," Technical Working Paper, No. 263, NBER.

PUHANI, P. A. (2000): “The Heckman Correction for Sample Selection and Its Critique,” *Journal of Economic Surveys*, 14(1), 53–68.

ROSENBAUM, P., AND D. RUBIN (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–50.

ROSENBAUM, P. R. (2002): *Observational Studies.* Springer, New York.

ROY, A. (1951): “Some Thoughts on the Distribution of Earnings,” *Oxford Economic Papers*, 3, 135–145.

RUBIN, D. (1974): “Estimating Causal Effects to Treatments in Randomised and Nonrandomised Studies,” *Journal of Educational Psychology*, 66, 688–701.

SIANESI, B. (2004): “An Evaluation of the Active Labour Market Programmes in Sweden,” *The Review of Economics and Statistics*, 86(1), 133–155.

SMITH, J. (2000): “A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies,” *Schweizerische Zeitschrift für Volkswirtschaft und Statistik*, 136(3), 1–22.

——— (2004): “Evaluating Local Development Policies: Theory and Practice,” Working Paper, University of Maryland.

SMITH, J., AND P. TODD (2005): “Does Matching Overcome LaLonde’s Critique of Nonexperimental Estimators?,” *Journal of Econometrics*, 125(1-2), 305–353.

STEIGER, H. (2004): “Is Less More? A Look at Nonparticipation in Swiss Active Labour Market Programmes,” Working Paper, University of St.Gallen.

VAN DEN BERG, G., B. VAN DER KLAAUW, AND J. VAN OURS (2004): “Punitive Sanctions and the Transition Rate from Welfare to Work,” *Journal of Labor Economics*, 22(1), 211–241.

VYTLACIL, E. (2002): “Independence, Monotonicity and Latent Index Models: An Equivalence Result,” *Econometrica*, 70, 331–341.