Evaluating the Effectiveness of an Institutional Training Program in Slovenia: A Comparison of Methods

Laura Juznik Rotar *

Abstract:

This paper aims to estimate the effect of an institutional training program on participants' chances of finding a job, using a rich dataset which comes from the official records of the Employment Service of Slovenia and taking into account the potential bias due to the existence of unobserved confounding factors. To deal with these selection biases, three methods are implemented in a comparative perspective: (1) instrumental variable (IV) regression; (2) Heckman's two-stage approach and (3) propensity score matching. This paper underlines important divergences between the results of parametric and non-parametric estimators. Some of the results, however, show the impact of the institutional training program on participants' chances of finding a job, especially in the short run. In the long run, however, the results are not so obvious.

Keywords: active employment policies, evaluation, institutional training program, IV and Heckman models, propensity score matching, Slovenia.

JEL: C50, J38, J68 **DOI:** 10.2478/v10033-012-0004-8

1. Introduction

Active employment policies are essentially public interventions in the labour market. For the measures of active employment policies in Slovenia in the year 2008 there was a budget of approximately 98,6 mio EUR (current prices). Measures of active employment policies are financed from the state budget, with some of these resources being European resources. Most heavily financed are training programs and programs dealing with social inclusion. Based on scientific methodology there are few studies on the evaluation of the effectiveness of employment programs in Slovenia. This leads to the overestimation of results and the inadequate distribution of resources.

The aim of the empirical analysis is to determine the effectiveness of an institutional training program for future employment probability. We are interested in the future employment probability of the young unemployed who participated in an institutional training program compared with the young unemployed who did not participate in the employment program. Because it is not possible to identify the individual causal effect for inclusion in the employment program, it is necessary to introduce certain assumptions.

In the empirical analysis, which is based on a rich database, we implement three ways to deal with selection/endogeneity biases (instrumental variables, Heckman two stages, and propensity score matching). Matching means pairing individuals from different groups, where program participants are similar in terms of their observable characteristics. In this case the estimates of the employment program are unbiased. The main assumption on which the matching method is based is a conditional independence assumption. Considering this assumption, the participation in a program and outcome are conditionally independent

*** Laura Juznik Rotar**
Higher Education Centre Novo mesto, School of Business and Management, Slovenia.
E-mail: laura_juznik@yahoo.com

according to the set of observable variables. When facing a highly dimensioned vector of observable variables, the propensity score can ease the calculations greatly. Moreover, results from the matching analysis are compared to results obtained from instrumental variables and the Heckman procedure.

The rest of this paper is organised as follows: (2) Estimation of an institutional training effect: a special case of the evaluation problem, (3) Data set and estimation strategy, (4) Results and discussion, (5) Conclusion.

2. Estimation of an Institutional Training Effect: A Special Case of the Evaluation Problem

We are interested in measuring the effect of an institutional training program (our treatment variable) on participants' chances of finding a job. This problem can be seen as a specific case of the more general evaluation problem dealing with causality (Angrist and Krueger, 1999; Barnow, Cain and Goldberger, 1980; Briggs, 2004; Caliendo and Hujer, 2006; Cameron and Trivedi, 2005; Dehejia and Wahba, 1999; Heckman, 1998). The main hindrance for modelling the causality arises from the basic problem of causal inference, which is to say that for an individual, we cannot simultaneously observe (1) the outcome when the individual receives the treatment and (2) the outcome when the individual does not receive the treatment, and as a result, we cannot observe the outcome for such an individual at the same time in the event of the treatment and in the event of the absence of treatment. In short, each causal inference includes a comparison of the actual outcome with the counterfactual outcome. We cannot say anything about the causal effect if we do not have a record of the counterfactual status. The problem of treatment effect assessment may actually be defined as a problem of missing data (Baltagi, 1995; Hujer and Caliendo, 2000; Ichino, 2006).

We observe the outcomes of individuals participating in an institutional training program and the outcomes of those not participating in the institutional training program. To know the »true« effect of the institutional training program on a particular individual, we must compare the observed outcome with the outcome that would have resulted had that individual not participated in the institutional training program. However, only one outcome is actually observed. What would have resulted had the individual not been treated -the counterfactual- cannot be observed (Ackum, 1991; Ashenfelter, 1978; Barron, Berger and Black, 1997; Fraker and Maynard, 1987). And this is precisely what gives rise to the evaluation problem. Yet, information on non-participants can be used to derive the counterfactual for participants.

Before stating how this idea can be implemented, it is important to specify the parameters of interest when estimating the treatment effect. Three types of estimates are mentioned in the literature (Heckman et al., 1996; Imbens and Angrist, 1994). In this paper, we will focus on the impact that the institutional training program has on individuals who were actually treated – the average effect of treatment on the treated (hereafter referred to as ATT). However, one could also be interested in the effect of the institutional training program on a random individual – the average treatment effect (ATE). These two effects are identical if we assume homogeneous responses to treatment among individuals; should the responses be allowed to vary across individuals, ATT and ATE would differ. The third parameter of interest is known as the local average treatment effect or LATE (Imbens and Angrist, 1994); it measures how a treatment affects people at the margin of participation, that is, it gives the mean effect of a program on those people whose participation changes as a result of the program.

Of these three parameters (ATT, ATE and LATE), ATT constitutes an obvious start: it easily makes sense for policy makers, who may consider it the most relevant. The first question policy makers want to see addressed is, of course, whether a program has any impact. Very often, they also want to know whether the expansion of a given program is worth considering (for instance, increasing the number of individuals participating in an institutional training program). While ATT may provide answers to these questions, other measures (ATE, for instance) are needed to go further. For instance, if only individuals with the largest expected gains participate in an institutional training program, ATE will be smaller than ATT. A generalisation of the program may thus produce a lower effect than the one measured by ATT (Lee, 2005; Vandenberghe and Robin, 2004; Verbeek, 2004) The empirical analysis outlined in this paper, however, is mostly exploratory. It will therefore focus on ATT only, but will propose different ways of measuring it.

2.1. Ordinary Least Squares Model

Since the outcome of participating in an institutional training program is defined by the probability of being employed, the ordinary least squares model to estimate

the effect of a treatment on probability of being employed (EMP_i) can be written as:

$$\mathrm{EMP_i} = \beta \mathrm{X_i} + \delta \mathrm{ITP_i} + \varepsilon_i \quad (1)$$

where ITP_i is a dummy variable indicating whether or not the *i*th individual participated in an institutional training program. In this basic »benchmark« case, the dummy has a constant coefficient, which gives the ATT.

If the independent variables X_i perfectly control for the other determinants of participation (usually the individual's background and other characteristics), then estimating equation (1) with OLS yields unbiased estimates of ATT. In this case ATT and ATE are equivalent since a homogeneous and constant response to the treatment is assumed. Implicit in this approach is the assumption that (having controlled for X_i), the treatment is independent of the process-determining outcomes (ITP_i and ε_i are uncorrelated) (Amemiya, 1985). The rest of this paper focuses on the sensitivity of OLS results to the relaxation of two assumptions: first, the absence of any selection bias beyond what is observed by the statistician and second, the linearity of the institutional training program effect across individuals.

2.2. Cross-Section Estimators Dealing with Selection on Unobserved Variables

Since the early 1980s, the literature has repeatedly emphasized that the OLS approach to the treatment effect is likely to be biased by the imperfect measurement or omission of some variables (Heckman, 1979; Heckman, 1990; Heckman and Robb, 1986; Vandenberghe and Robin, 2004). For example, more able or motivated individuals – dimensions that remain unobserved by the statistician – could select themselves into the institutional training program. On the other hand, administrative procedures to define eligibility to participate in an institutional training program can be such to select such individuals (e.g., through a selection procedure via consultation with the representative of the employment office). Technically, the OLS measure of ATT – the parameter associated with the ITP_i dummy in equation (1) – could be confounded with the effect of the unobserved (selection) variables. Means of controlling for this selection bias (i.e., for the endogeneity of ITP_i) consists of implementing the Instrumental Variable (IV) and the Heckman Selection estimators.

2.2.1. Instrumental Variables Two-Stage Least Square

The IV method consists of estimating a two-stage regression model. The second stage equation (eq. (3)) uses the linear prediction $ITPHAT_i$, obtained by regressing ITP_i against all other exogenous variables plus one D_i (eq. (2)). This variable, known as the »instrument«, introduces an element of randomness into the assignment, which approximates the effect of an experiment (Lee, 2005, Wooldridge, 2010)

$$\mathrm{ITP_i} = \gamma \mathrm{X_i} + \theta \mathrm{D_i} + \mu_i \quad (2)$$

$$\mathrm{EMP_i} = \beta \mathrm{X_i} + \delta \mathrm{ITPHAT_i} + \varepsilon_i \quad (3)$$

Provided D_i exists, the estimation of equations (2) and (3) gives an estimate of ATT. The main drawback to the IV approach, however, is that it will often be difficult to find a suitable instrument. To be valid as an instrument candidate, D_i should influence the probability to be treated, without being itself determined by any confounding factors affecting outcome (i.e., without being correlated with the error term ε_i (Wooldridge, 2010). Since this last condition can never be tested, the choice of a valid instrument largely depends on intuition and economic reasoning.

2.2.2. Heckman Two-Step Procedure

The Heckman Selection estimator is the other extensively used method to control for selection on unobserved variables. It relies on the assumption that a specific distribution of the unobservable characteristics jointly influences participation and outcome. By explicitly modelling the participation decision (estimating the first step equation similar to equation (2), generally using a Probit specification), it is possible to derive a variable that can be used to control for the potential correlation between the residual of the outcome equation and that of the selection equation. By including this new variable alongside the observable variables (X_i) and the institutional training program dummy in the second step (outcome) equation, Heckman can generate unbiased estimates of ATT. However, as with the IV approach, credible implementation requires the selection equation to contain an instrument and the identification of a suitable instrument is often an obstacle to proper implementation (Heckman, 1998; Vandenberghe and Robin, 2004).

2.3. Non-Parametric Estimators: Propensity Score Matching

A major drawback of the IV and Heckman methods (as well as OLS) is that they impose a linear form on the outcome equation. The institutional training program effect is assumed to be uniform across the distribution of covariates and adequately captured by the (constant) coefficient of a dummy variable. But economic theory provides no justification for such a linear restriction. We therefore complement our analysis with the non-parametric matching approach (Dehejia and Wahba, 2002; Heckman, Ichimura and Todd, 1997; Larsson, 2003; Rosenbaum and Rubin, 1983; Rosenbaum and Rubin, 1984; Rubin, 1977).

The underlying principle consists of matching treatment with comparison units (i.e., individuals participating in the institutional training program vs. those not participating in the institutional training program) that are similar in terms of their observable characteristics. This approach has an intuitive appeal, but rests on a very strong assumption: that any selection of unobserved variables is trivial, in the sense that the latter do not affect outcomes in the absence of treatment (Vandenberghe and Robin, 2004). This identifying assumption for matching, which is also the identifying assumption for OLS regression, is known as the Conditional Independence Assumption (CIA).

Under the CIA, estimators relying on matching techniques can yield unbiased estimates of ATT. They allow the counterfactual outcome for the treatment group to be inferred and therefore for any differences between the treated and non-treated to be attributed to the treatment. To make this approach credible, a very rich dataset is needed as the evaluator should be confident that all variables affecting both participation and outcome are observed. Matching individuals directly on their vector of covariates would be computationally demanding, especially when the number of covariates to control is large. The number of »cells« into which the data has to be divided would then augment exponentially. Rosenbaum and Rubin (1984) suggest a way to overcome this problem. They demonstrate that matching can be done on a single-index variable, the propensity score, defined as $p(X_i) \sim \Pr(ITP_i = 1 | X_i)$, which considerably reduces the dimensionality problem, as conditioning is done on a scalar rather than a vector basis (Vandenberghe and Robin, 2004).

The propensity score, however, must verify the balancing property. This means that individuals with the same propensity score must have the same distribution of observed covariates. The function used to compute the propensity score should be such that individuals with a similar propensity score and participate in an institutional training program, display, on average, similar values of X_i.

When doing propensity score matching, it is possible that for a particular individual in the treatment group no match can be found (i.e., no one in the non-tratment group has a propensity score that is »similar« to that particular individual). This is known as the common support problem. One way of addressing it is to drop treatment observations whose propensity score is higher than the maximum or less than the minimum of the within the common support. Enforcement of the common support can result in the loss of a sizeable proportion of the treated population. For these discarded individuals, the program effect cannot be estimated (Becker and Ichino, 2002; Vandenberghe and Robin, 2004).

Finally, even within the common support, the probability of observing two individuals with exactly the same value of $p(ITP_i = 1 | X_i)$ is in principle zero, since this index is a continuous variable. Various methods have been proposed to overcome this difficulty (e.g., nearest neighbour matching, kernel matching, radius matching). According to the sensitivity analysis implemented (using the Nannicini program code), we choose to present the results with the nearest neighbour matching approach (Nannicini, 2007). The nearest neighbour matching approach consists of an algorithm that matches each individual participating in an institutional training program with an individual not participating in an institutional training program displaying the nearest propensity score. The resulting match is as good as it is possible to achieve, in that the bias across the treatment and comparison groups is minimised. However, this method disregards potentially useful observations. Over reliance on a reduced number of individuals (the nearest neighbours) can result in ATT with large standard errors (Vandenberghe and Robin, 2004).

3. Data Set and Estimation Strategy

3.1. Data and variables

The data used in this empirical study is a random sample of approximately 3000 unemployed persons collected from an unemployment register kept by the

Variable	Non-participants	Participants in the institutional training program
Gender (%)		
Male	53,4	34,0
Female	46,6	66,0
Age (average)	24,4	25,3
Region (%)		
Pomurska	5,3	15,6
Podravska	16,9	29,8
Koroska	5,1	2,1
Savinjska	12,7	13,2
Zasavska	2,2	4,4
Spodnjeposavska	2,9	5,2
JV Slovenija	4,8	5,1
Osrednjeslovenska	27,6	8,0
Gorenjska	9,8	7,3
Notranjsko-kraska	2,6	2,7
Goriska	5,8	3,9
Obalno-kraska	4,3	2,7
Education (%)		
Unfinished/finished elementary school	15,8	7,3
Lower vocational school (2 years)	3,1	4,1
Lower vocational school (3 years)	0,7	0,6
Vocational school (4 years)	18,7	8,3
High school	41,4	42,2
University (2 year degree)	1,7	5,4
University (4 year degree)	18,5	32,1
Master's degree	0,1	0
PhD	0	0
Duration of unemployment before the program's start in months (average)	4,3	18,2
Number of observations	1346	1456

Note: "Program start" for the group of non-participants is a hypothetical date randomly assigned by a procedure suggested by Lechner (see Lechner, 1999). "Duration of unemployment before the program's start" for the group of non-participants is based on this hypothetical date.

Table 1: Descriptive statistics

Employment Service of Slovenia. The unemployment register includes records of all individuals who have been registered with the Employment Service as unemployed persons and are actively searching for a job. The advantages of this source of data are the availability and accuracy of data, that the data can be shown at the lowest possible level (with regard to the protection of personal data), whereas the disadvantage of such a database is that the data do not allow for international comparisons. The target group in our empirical analysis represents young unemployed persons aged from 20 to 29. For each person used in this study we have data on registration dates, data on labour market status: unemployed persons not included in employment programs and unemployed persons included in the institutional training program, and individual characteristics. Descriptive statistics used in this empirical analysis are presented in Table 1.

Table 1 presents descriptive statistics of some selected variables for the group of non-participants and the group of participants in the institutional training program. The data are for the years 2002 and 2003. Among the group of non-participants and the group of participants in the institutional training program there are differences in program characteristics as well as in individual characteristics. The duration of unemployment before the program start is shorter for the group of non-participants compared with the group of participants in the institutional training program. Secondly, the group of participants in the institutional training program consists of persons who registered quite early with the Employment Service, and thus also started earlier in the program. There are also differences in gender, age, education and region.

Variable	Marginal effect	Std. err.	z	P
Gender	0,136	0,0216	6,33	0,000
Age	0,179	0,0724	2,47	0,013
Age²	-0,004	0,0015	-2,53	0,012
Education	0,055	0,0062	8,87	0,000
Unemployment before the program's start	0,036	0,0015	20,14	0,000

Table 2: Sensitivity of probability of being employed to the variable region (probit estimates)

Algorithm	ATT – basic result	ATT – simulated result	Outcome effect	Selection effect
Measure of success: probability of being employed one year after the program's start				
ATTND	0,408	0,412	1,161	0,325
ATTK	0,452	0,452	1,186	0,334
ATTR	0,301	0,309	1,124	0,333
Measure of success: probability of being employed two years after the program's start				
ATTND	0,160	0,196	1,127	0,334
ATTK	0,215	0,214	1,117	0,327
ATTR	0,077	0,083	1,102	0,328

Note: *ATTND indicates nearest neighbour matching algorithm; ATTK indicates kernel matching algorithm; ATTR indicates radius matching algorithm.*
Table 3: Sensitivity analysis: average effect of treatment on the treated (ATT)

3.2. Estimation Strategy

We focus on the impact of the institutional training program on participants' chances of finding a job (measured as the probability of being employed one/two years after the program start and expressed as the differential being equal to participant mean minus the non-participant mean). Using the independent variables presented above, we run a traditional OLS model to get a first estimate of ATT. The next step is to implement the IV and Heckman models in order to control for the potential endogeneity of the treatment. As stated in section 2, both models crucially depend on the presence of a proper instrument in the first equation (choice equation). We have opted for a dummy variable »region«, equal to 1 (osrednjeslovenska, gorenjska, notranjsko-kraska, goriska and obalno-kraska region) and to 0 otherwise.

This variable fulfills the first condition of an instrumental variable candidate (Wooldridge, 2002) to be correlated with the endogenous variable or choice variable *ITP*, ceteris paribus. As can be seen in Table 2, the (marginal) effect of region on participants' chances of finding a job is strongly significant and evident.

As stated in section 2, the second condition for a variable to be an instrumental candidate (non-correlation with the residuals of the outcome equation) cannot be tested, which makes the choice of an instrument largely dependent on sensible arguments. We believe that there are plausible circumstances that would make »region« a valid instrument. If participation in an institutional training program is more frequent in regions that are less developed and have more rural areas, the risk of overestimating the effectiveness of an institutional training program is serious. Since the values of coefficients in Table 2 are not very large and the sign of coefficients is not the same across variables, we could consider that this somehow reduces the risk of overestimating the effectiveness of the institutional training program. But it could still be the case that the relative prevalence of the institutional training program according to region somehow reflects demand-side factors, in which case the endogeneity problem would remain.

The last step is to implement the propensity score matching approach presented in section 2. This is done (using program code written by Becker and Ichino, 2002) by using a Logit model to compute a propensity score, and the nearest neighbour as a matching algorithm, under the condition that common support is satisfied. The matching algorithm uses the same set of covariates (X_i) as in all previous estimations. As stated in section 2, due to the verification of common support less than 2% of individuals are discarded from the sample. Therefore, we estimate that our sample is still representative enough.

As we mentioned above, we use the nearest neighbour as a matching algorithm due to the results of sensitivity analysis. We follow the procedure suggested by Nannicini (2007). Identification of the ATT relies crucially on the validity of the CIA. Since the data are completely uninformative about the distribution of the outcome in the case of no treatment for treated

OLS		IV		Heckman				Nearest neighbour matching	
ATT	t	ATT	t	ATT	P(rho=0)	rho	P	ATT	t
0,250	13,95	0,167	3,24	0,981	0,000	0,488	0,000	0,396	5,99

Table 4: Probability of being employed one year after the program's start

OLS		IV		Heckman				Nearest neighbour matching	
ATT	t	ATT	t	ATT	P(rho=0)	rho	P	ATT	t
0,025	1,33	-0,168	-3,16	0,164	0,009	0,389	0,000	0,169	2,60

Table 5: Probability of being employed two years after the program's start

Variable	Experimental group	Control group	% of std. err.
Gender	1,66	1,61	9,5
Age	24,32	24,48	-5,9
Region	4,55	4,62	-1,9
Education	5,18	5,17	0,6
Unemployment before the program's start	16,72	14,91	11,0

Table 6: Balancing of covariates: mean standardised bias/difference

individuals, the CIA is untestable. Credibility of the CIA can be supported/rejected by theoretical reasoning and additional evidence. The Nannicini procedure for sensitivity analysis is a way to estimate whether average tratment effects are robust with possible deviations from the CIA. According to this procedure the results are presented in Table 3.

Since the basic and simulated results are for all three of the abovementioned algorithms very close to one another, we can conclude that matching algorithms are robust with possible deviations from the CIA. The outcome effect (which reports the magnitude and the sign of the simulated confounder in the case of no treatment) is positive and relatively small, whereas the selection effect (which reports the magnitude and the sign of the simulated confounder in the case of treatment) is negative and relatively small. This is not the problem, as the outcome effect and selection effect somehow balance, so that the basic and simulated results are very close to each other. According to the evidence presented, we chose the nearest neighbour as the matching algorithm to present our results.

4. Results and discussion

In Tables 4-5, four types of results of interest are detailed: (1) ATT as captured by the ITP dummy (δ) in an OLS regression model without control for selection biases; (2) ATT estimated via IV two-stage least squares; (3) ATT obtained with the Heckman two-stage estimates; (4) ATT from the nearest neighbour propensity score matching.

It is worth noting that all of the methods investigated here lead to estimates of ATT that diverge from the OLS results. When we measure the probability of being employed two years after the program's start, the Heckman two-stage method and nearest neighbour propensity score matching generate results that are relatively similar. The real differences emerge with IV and Heckman estimates and quite logically for cases where selection biases (as detected by the correlation between error terms in the Heckman model – rho in Tables 4-5) are significant. The correction can be positive (rho>0 with Heckman), suggesting that OLS exaggerates the effectiveness of the institutional training program. It can also be significantly negative (rho<0 with Heckman), suggesting that OLS underestimates the effectiveness of the institutional training program.

The results, however, show the impact of the institutional training program on participants' chances of finding a job, especially in the short run. With a propensity score matching method the probability of being employed one year after the program start is approximately 40%, In the long run, however, the results are not so obvious.

Table 6 provides some diagnostics of the performance of the match. Mean standardised bias/difference (MSB) provides a test of the balancing of covariates between the experimental and control groups (Rosenbaum, 2010; Rosenbaum and Rubin, 1983). The smaller the value of the MSB, the greater the similarity between the

experimental and control groups. There is no clear theoretical foundation for the limit to which these differences are still acceptable or the matching procedure still adequate. Oakes and Kaufman (2006) report that differences exceeding 10% of the standard error are unacceptable. On the other hand, Caliendo and Hujer (2006) report that acceptable differences are between 3% and 5% of the standard error. Only for the variable unemployment before the program start does the MSB exceed the abovementioned limits. In general, however, the results show that the balancing of covariates is satisfied (Table 6).

5. Conclusion

The objective of this paper was to estimate the effectiveness of an institutional training program on participants' chances of finding a job. The target group in our empirical analysis was young unemployed persons aged from 20 to 29. The problem of measuring the effectiveness of active employment programs was formulated as a specific case of the evaluation problem dealing with causality. The major problem represents constructing the proper counterfactual. The methods used were essentially twofold: IV and Heckman, on the one hand, in an attempt to control for potential selection on unobserved variables (ability, motivation), and propensity score matching (using the nearest neighbour algorithm) on the other to depart from the linearity restriction imposed by the OLS estimator.

From a methodological perspective, this paper underlines the main obstacle to the implementation of the IV and Heckman approaches, namely the difficulty of finding a valid instrument. The propensity score matching method helps overcome this obstacle, but at the cost of a risky assumption that the differences between the treated and control groups are fully embedded in the observed variables.

With regard to the results of our empirical study, we found that the effectiveness of the institutional training program on participants' chances of finding a job is different depending on the method used. However, the results show the impact of the institutional training program on participants' chances of finding a job, especially in the short run. With the propensity score matching method the probability of being employed one year after the program's start is approximately 40%, In the long run, however, the results are not so obvious.

The results presented in this paper can be of great help to the Slovenian government in deciding how to spend financial resources effectively. Moreover, this study also contributes to a lack of literature on the topic of evaluation of the active employment policies of Slovenia.

References

Ackum, S. (1991), "Youth Unemployment, Labour Market Programs and Subsequent Earnings." *Scandinavian Journal of Economics*, 93, (4), pp. 531-543.

Amemiya, T. (1985), *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Angrist, J., Krueger, A. B. (1999), "Empirical Strategies in Labor Economics," in Ashenfelter, O., Card, D., ed., *Handbook of Labor Economics*. Amsterdam: North Holland, pp. 1277-1366.

Ashenfelter, O. (1978), "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics*, 6, (1), pp. 47-57.

Baltagi, B. (1995), *Econometric Analysis of Panel Data*. New York: Wiley.

Barnow, B., Cain, G., Goldberger, A. (1980), "Issues in the Analysis of Selectivity Bias," in Stromsdorfer, E., Farkas, G., ed., *Evaluation Studies Vol. 5*. Beverly Hills, CA: Sage Publications, pp. 1324-1355.

Barron, J., Berger, M., Black, D. (1997), "How Well Do We Measure Training?" *Journal of Labour Economics*, 15, (3), pp. 507-528.

Becker, S. O., Ichino, A. (2002), "Estimation of Average Treatment Effects Based on Propensity Scores." *The Stata Journal*, 2, (4), pp. 358-377.

Briggs, D. C. (2004), "Causal Inference and the Heckman Model." *Journal of Education and Behavioral Statistics*, 29, pp. 397-420.

Caliendo, M., Hujer, R. (2006), "The Microeconometric Estimation of Treatment Effects. An Overview. " *Allgemeines Statistisches Archiv*, 90, pp. 197-212.

Cameron, A. C., Trivedi, P. K. (2005), *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Dehejia, R., Wahba, S. (1999), "Causal Effects in non-experimental Studies: Re-evaluating the Evaluation of Training Programs." *Journal of the American Statistical Associatio*n, 94, pp. 1053-1062.

Dehejia, R., Wahba, S. (2002), "Propensity Score Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics*, 84, pp. 151-161.

Fraker, T., Maynard, R. (1987), "The Adequacy of Comparison Group Designs for Evaluations of Employment-related Programs." *Journal of Human Resources*, 22, (2), pp. 194-227.

Heckman, J. (1979), "Sample Selection Bias as a Specifications Error." *Econometrica*, 47, (1), pp. 153-161.

Heckman, J. (1990), "Varieties of Selection Bias." *American Economic Review*, 80, (2), pp. 313-318.

Heckman, J. (1998), "The Economic Evaluation of Social Programs," in Heckman, J., Leamer, E., ed., *Handbook of Econometrics Vol. 5.* Amsterdam: Elsevier, pp. 1673-1694.

Heckman, J., Robb, R. (1986), "Alternative Identifying Assumptions in Econometric Models of Selection Bias," in Rhodes, G., ed., *Advances in Econometrics Vol. 5.* Greenwich, CT: JAI Press, pp. 1534-1576.

Heckman, J. et al. (1996), "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching as a Program Evaluation Method." *Proceedings of the National Academy of Sciences USA*, 93, (23), pp. 13416-13420.

Heckman, J., Ichimura, H., Todd, P. (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies*, 64, (4), pp. 605-654.

Hujer, R., Caliendo, M. (2000), "Evaluation of Active Labour Market Policy: Methodological Concepts and Empirical Estimates. IZA Discussion Paper no. 236.

Ichino, A. (2006), *The problem of Causality in Microeconometrics*. Bologna: EUI.

Imbens, G., Angrist, J. (1994), "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62, (4), pp. 467-476.

Larsson, L. (2003), "Evaluation of Swedish youth Labour Market Programmes." *Journal of Human Resources*, 38, (4), pp. 891-927.

Lechner, M. (1999), "Earnings and Employment Effects of Continuous Off-the-job Training in East Germany After Unification." *Journal of Business and Economic Statistics*, 17, pp. 74-90.

Lee, M. J. (2005), *Microeconometrics for Policy, Program and Treatment Effects*. Oxford: Oxford University Press.

Nannicini, T. (2007), "A Simulation-Based Sensitivity Analysis for Matching Estimators." *The Stata Journal*, 7, (3), pp. 334-350.

Oakes, M.J., Kaufman, J.S. (2006), *Methods in Social Epidemiology*. New York: Wiley.

Rosenbaum, P. (2010), *Observational Studies*. New York: Springer.

Rosenbaum, P., Rubin, D. (1983), "The Central Role of Propensity Score in Observational Studies for Causal Effects." *Biometrik*a, 70, (1), pp. 41-55.

Rosenbaum, P., Rubin, D. (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association*, 79, pp. 516-524.

Rubin, D. (1977), "Assignment to Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics*, 2, pp. 1-26.

Vandenberghe, V., Robin, S. (2004), "Evaluating the Effectiveness of Private Education Across Countries: A Comparison of Methods." *Labour Economics,* 11, pp. 487-506.

Verbeek, M. (2004), *A Guide to Modern Econometrics*. Chichester: John Wiley & Sons.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross-Section and Panel Data*. *2nd ed.* Cambridge, MA: MIT Press.