DIW Berlin

German Institute
for Economic Research

Discussion Papers

485

Marco Caliendo*
Sabine Kopeinig**

Some Practical Guidance for the Implementation of Propensity Score Matching

Berlin, April 2005

* DIW Berlin, IZA, Bonn
** University of Cologne

IMPRESSUM

DIW Berlin
Deutsches Institut für Wirtschaftsforschung
Königin-Luise-Str. 5
14195 Berlin
Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
www.diw.de

ISSN 1433-0210 (Druck) 1619-4535 (elektronisch)

Some Practical Guidance for the Implementation of Propensity Score Matching*

Marco Caliendo† Sabine Kopeinig‡

DIW, Berlin
IZA, Bonn

University
of Cologne

Working Paper
This draft: April 26, 2005

Abstract

Propensity Score Matching (PSM) has become a popular approach to estimate causal treatment effects. It is widely applied when evaluating labour market policies, but empirical examples can be found in very diverse fields of study. Once the researcher has decided to use PSM, he is confronted with a lot of questions regarding its implementation. To begin with, a first decision has to be made concerning the estimation of the propensity score. Following that one has to decide which matching algorithm to choose and determine the region of common support. Subsequently, the matching quality has to be assessed and treatment effects and their standard errors have to be estimated. Furthermore, questions like 'what to do if there is choice-based sampling?' or 'when to measure effects?' can be important in empirical studies. Finally, one might also want to test the sensitivity of estimated treatment effects with respect to unobserved heterogeneity or failure of the common support condition. Each implementation step involves a lot of decisions and different approaches can be thought of. The aim of this paper is to discuss these implementation issues and give some guidance to researchers who want to use PSM for evaluation purposes.

Keywords: Propensity Score Matching, Implementation, Evaluation, Sensitivity
JEL Classification: C40, H43

*The authors thank Sascha O. Becker for valuable comments. All remaining errors are our own.

†Marco Caliendo is Senior Research Associate at the German Institute for Economic Research (DIW Berlin) and Research Affiliate of the IZA, Bonn, e-mail: mcaliendo@diw.de. Corresponding author: Marco Caliendo, DIW Berlin, Dep. of Public Economics, Königin-Luise-Str. 5, 14195 Berlin, phone: +49-30-89789-154, fax: +49-30-89789-9154.

‡Sabine Kopeinig is Research Assistant at the Department of Marketing and Market Research, University of Cologne, e-mail: kopeinig@wiso.uni-koeln.de.

1 Introduction

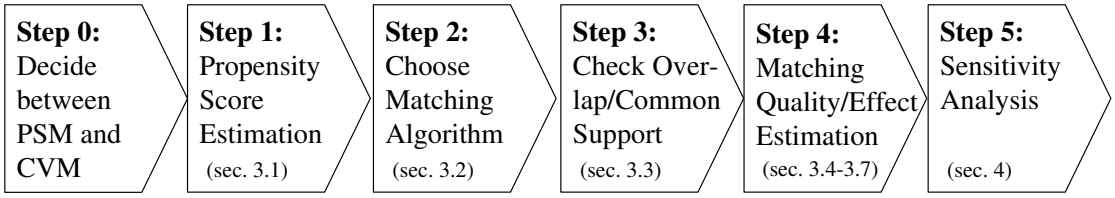
Matching has become a popular approach to estimate causal treatment effects. It is widely applied when evaluating labour market policies (see e.g. Dehejia and Wahba (1999) or Heckman, Ichimura, and Todd (1997)), but empirical examples can be found in very diverse fields of study. It applies for all situations where one has a treatment, a group of treated individuals and a group of untreated individuals. The nature of treatment may be very diverse. For example, Perkins, Tu, Underhill, Zhou, and Murray (2000) discuss the usage of matching in pharmacoepidemiologic research. Hitt and Frei (2002) analyse the effect of online banking on the profitability of customers. Davies and Kim (2003) compare the effect on the percentage bid-ask spread of Canadian firms being interlisted on an US-Exchange, whereas Brand and Halaby (2003) analyse the effect of elite college attendance on career outcomes. Ham, Li, and Reagan (2003) study the effect of a migration decision on the wage growth of young men and Bryson (2002) analyse the effect of union membership on wages of employees. Every microeconometric evaluation study has to overcome the fundamental evaluation problem and address the possible occurrence of selection bias. The first problem arises because we would like to know the difference between the participants' outcome with and without treatment. Clearly, we cannot observe both outcomes for the same individual at the same time. Taking the mean outcome of non-participants as an approximation is not advisable, since participants and non-participants usually differ even in the absence of treatment. This problem is known as selection bias and a good example is the case, where motivated individuals have a higher probability of entering a training programme and have also a higher probability of finding a job. The matching approach is one possible solution to the selection problem. It originated from the statistical literature and shows a close link to the experimental context.[1] Its basic idea is to find in a large group of non-participants those individuals who are similar to the participants in all relevant pre-treatment characteristics X. That being done, differences in outcomes of this well selected and thus adequate control group and of participants can be attributed to the programme.

Since conditioning on all relevant covariates is limited in case of a high dimensional vector X ('curse of dimensionality'), Rosenbaum and Rubin (1983) suggest the use of so-called balancing scores $b(X)$, i.e. functions of the relevant observed covariates X such that the conditional distribution of X given $b(X)$ is independent of assignment into treatment. One possible balancing score is the propensity score, i.e. the probability of participating in a programme given observed characteristics X. Matching procedures based on this balancing score are known as propensity score matching (PSM) and will be the focus of this paper. Once the researcher has decided to use PSM, he is confronted with a lot of questions regarding its implementation. Figure 1 summarises the necessary steps when implementing PSM.[2]

[1] See e.g. Rubin (1974), Rosenbaum and Rubin (1983, 1985a) or Lechner (1998).

[2] The decision whether to apply PSM or covariate matching (CVM) will not be discussed in this paper. With CVM distance measures like the Mahalanobis distance are used to calculate similarity of two individuals in terms of covariate values and the matching is done on these distances. The interested reader is referred to Imbens (2004) or Abadie and Imbens (2004) who develop covariate and bias-adjusted matching estimators. Zhao (2004) discusses the basic differences between PSM and covariate matching.

Figure 1: PSM - Implementation Steps



CVM: Covariate Matching, PSM: Propensity Score Matching

The aim of this paper is to discuss these issues and give some practical guidance to researchers who want to use PSM for evaluation purposes. The paper is organised as follows. In section 2 we will describe the basic evaluation framework and possible treatment effects of interest. Furthermore we show how propensity score matching solves the evaluation problem and highlight the implicit identifying assumptions. In section 3 we will focus on implementation steps of PSM estimators. To begin with, a first decision has to be made concerning the estimation of the propensity score (see subsection 3.1). One has not only to decide about the probability model to be used for estimation, but also about variables which should be included in this model. In subsection 3.2 we briefly evaluate the (dis-)advantages of different matching algorithms. Following that we discuss how to check the overlap between treatment and comparison group and how to implement the common support requirement in subsection 3.3. In subsection 3.4 we will show how to assess the matching quality. Subsequently we present the problem of choice-based sampling and discuss the question 'when to measure programme effects?' in subsections 3.5 and 3.6. Estimating standard errors for treatment effects will be briefly discussed in subsection 3.7, before we conclude this section with an overview of available software to estimate treatment effects (3.8). Section 4 will be concerned with the sensitivity of estimated treatment effects. In subsection 4.1 we describe an approach (Rosenbaum bounds) that allows the researcher to determine how strongly an unmeasured variable must influence the selection process in order to undermine the implications of PSM. In subsection 4.2 we describe an approach proposed by Lechner (2000b). He incorporates information from those individuals who failed the common support restriction, to calculate bounds of the parameter of interest, if all individuals from the sample at hand would have been included. Finally, section 5 reviews all steps and concludes.

2 Evaluation Framework and Matching Basics

Roy-Rubin Model: Inference about the impact of a treatment on the outcome of an individual involves speculation about how this individual would have performed

had he not received the treatment. The standard framework in evaluation analysis to formalise this problem is the potential outcome approach or Roy-Rubin-model (Roy (1951), Rubin (1974)). The main pillars of this model are individuals, treatment and potential outcomes. In the case of a binary treatment the treatment indicator D_i equals one if individual i receives treatment and zero otherwise. The potential outcomes are then defined as $Y_i(D_i)$ for each individual i, where $i = 1, ..., N$ and N denotes the total population. The treatment effect for an individual i can be written as:

$$\tau_i = Y_i(1) - Y_i(0). \tag{1}$$

The fundamental evaluation problem arises because only one of the potential outcomes is observed for each individual i. The unobserved outcome is called counterfactual outcome. Hence, estimating the individual treatment effect τ_i is not possible and one has to concentrate on (population) average treatment effects.[3]

Parameter of Interest: The parameter that received the most attention in evaluation literature is the 'average treatment effect on the treated' (ATT), which is defined as:

$$\tau_{ATT} = E(\tau|D = 1) = E[Y(1)|D = 1] - E[Y(0)|D = 1]. \tag{2}$$

As the counterfactual mean for those being treated - $E[Y(0)|D = 1]$ - is not observed, one has to choose a proper substitute for it in order to estimate ATT. Using the mean outcome of untreated individuals $E[Y(0)|D = 0]$ is in non-experimental studies usually not a good idea, because it is most likely that components which determine the treatment decision also determine the outcome variable of interest. Thus, the outcomes of individuals from treatment and comparison group would differ even in the absence of treatment leading to a 'self-selection bias'. For ATT it can be noted as:

$$E[Y(1)|D = 1] - E[Y(0)|D = 0] = \tau_{ATT} + E[Y(0)|D = 1] - E[Y(0)|D = 0]. \tag{3}$$

The difference between the left hand side of equation (3) and τ_{ATT} is the so-called 'self-selection bias'. The true parameter τ_{ATT} is only identified, if:

$$E[Y(0)|D = 1] - E[Y(0)|D = 0] = 0. \tag{4}$$

In social experiments where assignment to treatment is random this is ensured and the treatment effect is identified.[4] In non-experimental studies one has to invoke some identifying assumptions to solve the section problem stated in equation (3). Another parameter of interest is the 'average treatment effect' (ATE), which is defined as:

$$\tau_{ATE} = E[Y(1) - Y(0)]. \tag{5}$$

The additional challenge when estimating ATE is that both counterfactual outcomes $E[Y(1)|D = 0]$ and $E[Y(0)|D = 1]$ have to be constructed.

[3]Estimation of average treatment effects requires that the treatment effect for each individual i is independent of treatment participation of other individuals ('stable unit-treatment value assumption').

[4]See Smith (2000) for a discussion about advantages and disadvantages of social experiments.

Conditional Independence Assumption: One possible identification strategy is to assume, that given a set of observable covariates X which are not affected by treatment, potential outcomes are independent of treatment assignment:

$$\text{(Unconfoundedness)} \quad Y(0), Y(1) \amalg D|X, \quad \forall \quad X. \tag{6}$$

This implies, that selection is solely based on observable characteristics and that all variables that influence treatment assignment and potential outcomes simultaneously are observed by the researcher. Clearly, this is a strong assumption and has to be justified by the data quality at hand. For the rest of the paper we will assume that this condition holds.[5] It should also be clear, that conditioning on all relevant covariates is limited in case of a high dimensional vector X. For instance if X contains s covariates which are all dichotomous, the number of possible matches will be 2^s. To deal with this dimensionality problem, Rosenbaum and Rubin (1983) suggest to use so-called balancing scores. They show that if potential outcomes are independent of treatment conditional on covariates X, they are also independent of treatment conditional on a balancing score $b(X)$. The propensity score $P(D = 1|X) = P(X)$, i.e. the probability for an individual to participate in a treatment given his observed covariates X, is one possible balancing score. The conditional independence assumption (CIA) based on the propensity score (PS) can be written as:

$$\text{(Unconfoundedness given the PS)} \quad Y(0), Y(1) \amalg D|P(X), \quad \forall \quad X. \tag{7}$$

Common Support: A further requirement besides independence is the common support or overlap condition. It rules out the phenomenon of perfect predictability of D given X:

$$\text{(Overlap)} \quad 0 < P(D = 1|X) < 1 \tag{8}$$

It ensures that persons with the same X values have a positive probability of being both participants and non-participants (Heckman, LaLonde, and Smith, 1999).

Estimation Strategy: Given that CIA holds and assuming additional that there is overlap between both groups (called 'strong ignorability' by Rosenbaum and Rubin (1983)), the PSM estimator for ATT can be written in general as[6]:

$$\tau_{ATT}^{PSM} = E_{P(X)|D=1}\{E[Y(1)|D = 1, P(X)] - E[Y(0)|D = 0, P(X)]\}. \tag{9}$$

To put it in words, the PSM estimator is simply the mean difference in outcomes over the common support, appropriately weighted by the propensity score distribution of participants. Based on this brief outline of the matching estimator in the general evaluation framework, we are now going to discuss the implementation of PSM in detail.

[5] See Blundell and Costa Dias (2002) or Caliendo and Hujer (2005) for evaluation strategies when selection is also based on unobservable characteristics.

[6] For the identification of ATT it is sufficient to assume that $Y(0) \amalg D|P(X)$ and $P(D = 1|X) < 1$.

3 Implementation of Propensity Score Matching

3.1 Estimating the Propensity Score

When estimating the propensity score, two choices have to be made. The first one concerns the model to be used for the estimation, and the second one the variables to be included in this model. We will start with the model choice before we discuss which variables to include in the model.

Model Choice: Little advice is available regarding which functional form to use (see e.g. the discussion in Smith (1997)). In principle any discrete choice model can be used. Preference for logit or probit models (compared to linear probability models) derives from the well-known shortcomings of the linear probability model, especially the unlikeliness of the functional form when the response variable is highly skewed and predictions that are outside the $[0, 1]$ bounds of probabilities. However, when the purpose of a model is classification rather than estimation of structural coefficients, it is less clear that these criticisms apply (Smith, 1997). For the binary treatment case, where we estimate the probability of participation vs. non-participation, logit and probit models usually yield similar results. Hence, the choice is not too critical, even though the logit distribution has more density mass in the bounds. However, when leaving the binary treatment case, the choice of the model becomes more important. The multiple treatment case (as discussed in Imbens (2000) and Lechner (2001)) constitutes of more than two alternatives, e.g. when an individual is faced with the choice to participate in job-creation schemes, vocational training or wage subsidy programmes or do not participate at all. For that case it is well known that the multinomial logit is based on stronger assumptions than the multinomial probit model, making the latter one the preferable option.[7] However, since the multinomial probit is computational more burdensome, a practical alternative is to estimate a series of binomial models like suggested by Lechner (2001). Bryson, Dorsett, and Purdon (2002) note that there are two shortcomings regarding this approach. First, as the number of options increases, the number of models to be estimated increases disproportionately (for L options we need $0.5(L(L-1))$ models). Second, in each model only two options at a time are considered and consequently the choice is conditional on being in one of the two selected groups. On the other hand, Lechner (2001) compares the performance of the multinomial probit approach and the series estimation and finds little difference in their relative performance. He suggests that the latter approach may be more robust since a mis-specification in one of the series will not compromise all others as would be the case in the multinomial probit model.

Variable Choice: More advice is available regarding the inclusion (or exclusion) of covariates in the propensity score model. The matching strategy builds on the

[7] Especially the 'independence from irrelevant alternatives' assumption (IIA) is critical. It basically states that the odds ratio between two alternatives are independent of other alternatives. This assumption is convenient for estimation but not appealing from an economic or behavioural point of view (for details see e.g. Greene (2003)).

CIA, requiring that the outcome variable(s) must be independent of treatment conditional on the propensity score. Hence, implementing matching requires choosing a set of variables X that credibly satisfy this condition. Heckman, Ichimura, and Todd (1997) show that omitting important variables can seriously increase bias in resulting estimates. Only variables that influence simultaneously the participation decision and the outcome variable should be included. Hence, economic theory, a sound knowledge of previous research and also information about the institutional settings should guide the researcher in building up the model (see e.g. Smith and Todd (2005) or Sianesi (2004)). It should also be clear that only variables that are unaffected by participation (or the anticipation of it) should be included in the model. To ensure this, variables should either be fixed over time or measured before participation. In the latter case, it must be guaranteed that the variable has not been influenced by the anticipation of participation. Heckman, LaLonde, and Smith (1999) also point out, that the data for participants and non-participants should stem from the same sources (e.g. the same questionnaire). The better and more informative the data are, the easier it is to credibly justify the CIA and the matching procedure. However, it should also be clear that 'too good' data is not helpful either. If $P(X) = 0$ or $P(X) = 1$ for some values of X, then we cannot use matching conditional on those X values to estimate a treatment effect, because persons with such characteristics either always or never receive treatment. Hence, the common support condition as stated in equation (8) fails and matches cannot be performed. Some randomness is needed that guarantees that persons with identical characteristics can be observed in both states (Heckman, Ichimura, and Todd, 1998).

In cases of uncertainty of the proper specification, sometimes the question may arise if it is better to include too many rather than too few variables. Bryson, Dorsett, and Purdon (2002) note that there are two reasons why over-parameterised models should be avoided. First, it may be the case that including extraneous variables in the participation model exacerbate the support problem. Second, although the inclusion of non-significant variables will not bias the estimates or make them inconsistent, it can increase their variance. The results from Augurzky and Schmidt (2000) point in the same direction. They run a simulation study to investigate propensity score matching when selection into treatment is remarkably strong, and treated and untreated individuals differ considerably in their observable characteristics. In their setup, explanatory variables in the selection equation are partitioned into two sets. The first set includes variables that strongly influence the participation and the outcome equation, whereas the second set does not (or only weakly) influence the outcome equation. Including the full set of covariates in small samples might cause problems in terms of higher variance, since either some treated have to be discarded from the analysis or control units have to be used more than once. They show that matching on an inconsistent estimate of the propensity score (i.e. the one without the second set of covariates) produces better estimation results of the average treatment effect.

On the other hand, Rubin and Thomas (1996) recommend against 'trimming' models in the name of parsimony. They argue that a variable should only be excluded from analysis if there is consensus that the variable is either unrelated to the outcome or not a proper covariate. If there are doubts about these two points, they explicitly advise to include the relevant variables in the propensity score estimation.

By these criteria, there are both reasons for and against including all of the reasonable covariates available. Basically, the points made so far imply that the choice of variables should be based on economic theory and previous empirical findings. But clearly, there are also some formal (statistical) tests which can be used. Heckman, Ichimura, Smith, and Todd (1998) and Heckman and Smith (1999) discuss two strategies for the selection of variables to be used in estimating the propensity score.

Hit or Miss Method: The first one is the 'hit or miss' method or prediction rate metric, where variables are chosen to maximise the within-sample correct prediction rates. This method classifies an observation as '1' if the estimated propensity score is larger than the sample proportion of persons taking treatment, i.e. $\hat{P}(X) > \overline{P}$. If $\hat{P}(X) \leq \overline{P}$ observations are classified as '0'. This method maximises the overall classification rate for the sample assuming that the costs for the misclassification are equal for the two groups (Heckman, Ichimura, and Todd, 1997).[8] But clearly, it has to be kept in mind that the main purpose of the propensity score estimation is not to predict selection into treatment as good as possible but to balance all covariates (Augurzky and Schmidt, 2000).

Statistical Significance: The second approach relies on statistical significance and is very common in textbook econometrics. To do so, one starts with a parsimonious specification of the model, e.g. a constant, the age and some regional information, and then 'tests up' by iteratively adding variables to the specification. A new variable is kept if it is statistically significant at conventional levels. If combined with the 'hit or miss' method, variables are kept if they are statistically significant and increase the prediction rates by a substantial amount (Heckman, Ichimura, Smith, and Todd, 1998).

Leave-one-out Cross-Validation: Leave-one-out cross-validation can also be used to choose the set of variables to be included in the propensity score. Black and Smith (2003) implement their model selection procedure by starting with a 'minimal' model containing only two variables. They subsequently add blocks of additional variables and compare the resulting mean squared errors. As a note of caution they stress, that this amounts to choosing the propensity score model based on goodness-of-fit considerations, rather than based on theory and evidence about the set of variables related to the participation decision and the outcomes (Black and Smith, 2003). They also point out an interesting trade-off in finite samples between the plausibility of the CIA and the variance of the estimates. When using the full specification, bias arises from selecting a wide bandwidth in response to the weakness of the common support. In contrast to that, when matching on the minimal specification, common support is not a problem but the plausibility of the CIA is. This trade-off also affects the estimated standard errors, which are smaller for the minimal specification where the common support condition poses no problem. Finally, checking the matching quality can also help to determine which variables

[8] See e.g. Breiman, Friedman, Olsen, and Stone (1984) for theory and Heckman, Ichimura, Smith, and Todd (1998) or Smith and Todd (2005) for applications.

should be included in the model. We will discuss this point later on in subsection 3.4.

Overweighting some Variables: Let us assume for the moment that we have found a satisfactory specification of the model. It may sometimes be felt that some variables play a specifically important role in determining participation and outcome (Bryson, Dorsett, and Purdon, 2002). As an example, one can think of the influence of gender and region in determining the wage of individuals. Let us take as given for the moment that men earn more than women and the wage level is higher in region A compared to region B. If we add dummy variables for gender and region in the propensity score estimation, it is still possible that women in region B are matched with men in region A, since the gender and region dummies are only a sub-set of all available variables. There are basically two ways to put greater emphasis on specific variables. One can either find variables in the comparison group who are identical with respect to these variables, or carry out matching on sub-populations. The study from Lechner (2002) is a good example for the first approach. He evaluates the effects of active labour market policies in Switzerland and uses the propensity score as a 'partial' balancing score which is complemented by an exact matching on sex, duration of unemployment and native language. Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd (1998) use the second strategy and implement matching separately for four demographic groups. That implies that the complete matching procedure (estimating the propensity score, checking the common support, etc.) has to be implemented separately for each group. This is analogous to insisting on a perfect match e.g. in terms of gender and region and then carrying out propensity score matching. This procedure is especially recommendable if one expects the effects to be heterogeneous between certain groups.

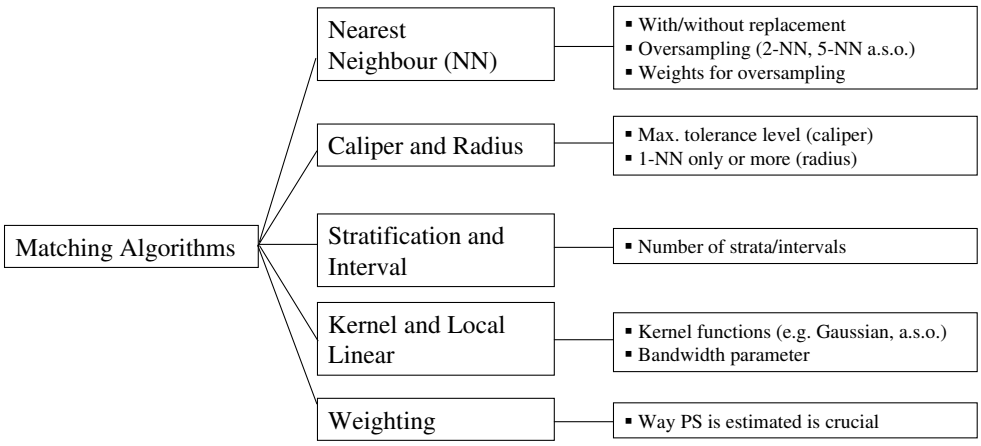
Alternatives to the Propensity Score: Finally, it should also be noted that it is possible to match on a measure other than the propensity score, namely the underlying index of the score estimation. The advantage of this is that the index differentiates more between observations in the extremes of the distribution of the propensity score (Lechner, 2000a). This is useful if there is some concentration of observations in the tails of the distribution. Additionally, in some recent papers the propensity score is estimated by duration models. This is of particular interest if the 'timing of events' plays a crucial role (see e.g. Brodaty, Crepon, and Fougere (2001) or Sianesi (2004)).

3.2 Choosing a Matching Algorithm

The PSM estimator in its general form was stated in equation (9). All matching estimators contrast the outcome of a treated individual with outcomes of comparison group members. PSM estimators differ not only in the way the neighbourhood for each treated individual is defined and the common support problem is handled, but also with respect to the weights assigned to these neighbours. Figure 2 depicts different PSM estimators and the inherent choices to be made when they are used.

We will not discuss the technical details of each estimator here at depth but rather present the general ideas and the involved trade-offs with each algorithm.[9]

Figure 2: Different Matching Algorithms



NN: Nearest Neighbour, PS: Propensity Score

Nearest Neighbour Matching: The most straightforward matching estimator is nearest neighbor (NN) matching. The individual from the comparison group is chosen as a matching partner for a treated individual that is closest in terms of propensity score. Several variants of NN matching are proposed, e.g. NN matching 'with replacement' and 'without replacement'. In the former case, an untreated individual can be used more than once as a match, whereas in the latter case it is considered only once. Matching with replacement involves a trade-off between bias and variance. If we allow replacement, the average quality of matching will increase and the bias will decrease. This is of particular interest with data where the propensity score distribution is very different in the treatment and the control group. For example, if we have a lot of treated individuals with high propensity scores but only few comparison individuals with high propensity scores, we get bad matches as some of the high-score participants will get matched to low-score non-participants. This can be overcome by allowing replacement, which in turn reduces the number of distinct non-participants used to construct the counterfactual outcome and thereby increases the variance of the estimator (Smith and Todd, 2005). A problem which is related to NN matching without replacement is that estimates depend on the order in which observations get matched. Hence, when using this approach it should be ensured that ordering is randomly done.

It is also suggested to use more than one nearest neighbour ('oversampling'). This form of matching involves a trade-off between variance and bias, too. It trades reduced variance, resulting from using more information to construct the counterfactual for each participant, with increased bias that results from on average poorer

[9]See Smith and Todd (2005) or Imbens (2004) for more technical details.

matches (see e.g. Smith (1997)). When using oversampling, one has to decide how many matching partners should be chosen for each treated individual and which weight (e.g. uniform or triangular weight) should be assigned to them.

Caliper and Radius Matching: NN matching faces the risk of bad matches, if the closest neighbour is far away. This can be avoided by imposing a tolerance level on the maximum propensity score distance (caliper). Imposing a caliper works in the same direction as allowing for replacement. Bad matches are avoided and hence the matching quality rises. However, if fewer matches can be performed, the variance of the estimates increases. Applying caliper matching means that those individual from the comparison group is chosen as a matching partner for a treated individual that lies within the caliper ('propensity range') and is closest in terms of propensity score. As Smith and Todd (2005) note, a possible drawback of caliper matching is that it is difficult to know a priori what choice for the tolerance level is reasonable.

Dehejia and Wahba (2002) suggest a variant of caliper matching which is called radius matching. The basic idea of this variant is to use not only the nearest neighbour within each caliper but all of the comparison members within the caliper. A benefit of this approach is that it uses only as many comparison units as are available within the caliper and therefore allows for usage of extra (fewer) units when good matches are (not) available. Hence, it shares the attractive feature of oversampling mentioned above, but avoids the risk of bad matches.

Stratification and Interval Matching: The idea of stratification matching is to partition the common support of the propensity score into a set of intervals (strata) and to calculate the impact within each interval by taking the mean difference in outcomes between treated and control observations. This method is also known as interval matching, blocking and subclassification (Rosenbaum and Rubin, 1983). Clearly, one question to be answered is how many strata should be used in empirical analysis. Cochrane and Chambers (1965) shows that five subclasses are often enough to remove 95% of the bias associated with one single covariate. Since, as Imbens (2004) notes, all bias under unconfoundedness is associated with the propensity score, this suggests that under normality the use of five strata removes most of the bias associated with all covariates. One way to justify the choice of the number of strata is to check the balance of the propensity score (or the covariates) within each stratum (see e.g. Aakvik (2001)). Most of the algorithms can be described in the following way: First, check if within a stratum the propensity score is balanced. If not, strata are too large and need to be split. If, conditional on the propensity score being balanced, the covariates are unbalanced, the specification of the propensity score is not adequate and has to be re-specified, e.g. through the addition of higher-order terms or interactions (Dehejia and Wahba, 1999).

Kernel and Local Linear Matching: The matching algorithms discussed so far have in common that only a few observations from the comparison group are used to construct the counterfactual outcome of a treated individual. Kernel matching (KM) and local linear matching (LLM) are non-parametric matching estimators that use weighted averages of all individuals in the control group to construct the

counterfactual outcome. Thus, one major advantage of these approaches is the lower variance which is achieved because more information is used. A drawback of these methods is that possibly observations are used that are bad matches. Hence, the proper imposition of the common support condition is of major importance for KM and LLM. Heckman, Ichimura, and Todd (1998) derive the asymptotic distribution of these estimators and Heckman, Ichimura, and Todd (1997) present an application. As Smith and Todd (2005) note, kernel matching can be seen as a weighted regression of the counterfactual outcome on an intercept with weights given by the kernel weights. Weights depend on the distance between each individual from the control group and the participant observation for which the counterfactual is estimated. It is worth noting that if weights from a symmetric, nonnegative, unimodal kernel are used, then the average places higher weight on persons close in terms of propensity score of a treated individual and lower weight on more distant observations. The estimated intercept provides an estimate of the counterfactual mean. The difference between KM and LLM is that the latter includes in addition to the intercept a linear term in the propensity score of a treated individual. This is an advantage whenever comparison group observations are distributed asymmetrically around the treated observation, e.g. at boundary points, or when there are gaps in the propensity score distribution. When applying KM one has to choose the kernel function and the bandwidth parameter. The first point appears to be relatively unimportant in practice (DiNardo and Tobias, 2001). What is seen as more important (see e.g. Silverman (1986) or Pagan and Ullah (1999)) is the choice of the bandwidth parameter with the following trade-off arising: High bandwidth-values yield a smoother estimated density function, therefore leading to a better fit and a decreasing variance between the estimated and the true underlying density function. On the other hand, underlying features may be smoothed away by a large bandwidth leading to a biased estimate. The bandwidth choice is therefore a compromise between a small variance and an unbiased estimate of the true density function.

Weighting on Propensity Score: Imbens (2004) notes that propensity scores can also be used as weights to obtain a balanced sample of treated and untreated individuals. If the propensity score is known, the estimator can directly by implemented as the difference between a weighted average of the outcomes for the treated and untreated individuals. Unless in experimental settings, the propensity score has to be estimated. As Zhao (2004) note, the way propensity scores are estimated is crucial when implementing weighting estimators. Hirano and Imbens (2002) suggest a straightforward way to implement this weighting on propensity score estimator by combining it with regression adjustment.

Trade-offs in Terms of Bias and Efficiency: Having presented the different possibilities, the question remains on how one should select a specific matching algorithm. Clearly, asymptotically all PSM estimators should yield the same results, because with growing sample size they all become closer to comparing only exact matches (Smith, 2000). However, in small samples the choice of the matching algorithm can be important (Heckman, Ichimura, and Todd, 1997), where usually a trade-off between bias and variance arises (see Table 1). So what advice can be given to researchers facing the problem of choosing a matching estimator? It should be

clear that there is no 'winner' for all situations and that the choice of the estimator crucially depends on the situation at hand. The performance of different matching estimators varies case-by-case and depends largely on the data structure at hand (Zhao, 2000). To give an example, if there are only a few control observations, it makes no sense to match without replacement. On the other hand, if there are a lot of comparable untreated individuals it might be worth using more than one nearest neighbour (either by oversampling or kernel matching) to gain more precision in estimates. Pragmatically, it seems sensible to try a number of approaches. Should they give similar results, the choice may be unimportant. Should results differ, further investigation may be needed in order to reveal more about the source of the disparity (Bryson, Dorsett, and Purdon, 2002).

Table 1: Trade-Offs in Terms of Bias and Efficiency

Decision	Bias	Variance
Nearest neighbour matching:		
multiple neighbours / single neighbour	(+)/(-)	(-)/(+)
with caliper / without caliper	(-)/(+)	(+)/(-)
Use of control individuals:		
with replacement / without replacement	(-)/(+)	(+)/(-)
Choosing method:		
NN-matching / Radius-matching	(-)/(+)	(+)/(-)
KM or LLM / NN-methods	(+)/(-)	(-)/(+)
Bandwidth choice with KM:		
small / large	(-)/(+)	(+)/(-)

KM: Kernel Matching, LLM: Local Linear Matching
NN: Nearest Neighbour
Increase: (+), Decrease: (-)

3.3 Overlap and Common Support

Our discussion in section 2 has shown that ATT and ATE are only defined in the region of common support. Hence, an important step is to check the overlap and the region of common support between treatment and comparison group. Several ways are suggested in the literature, where the most straightforward one is a visual analysis of the density distribution of the propensity score in both groups. Lechner (2000b) argues that given that the support problem can be spotted by inspecting the propensity score distribution, there is no need to implement a complicated formal estimator. However, some formal guidelines might help the researcher to determine the region of common support more precisely. We will present two methods, where the first one is essentially based on comparing the minima and maxima of the propensity score in both groups and the second one is based on estimating the density distribution in both groups. Implementing the common support condition ensures that any combination of characteristics observed in the treatment group can also be observed among the control group (Bryson, Dorsett, and Purdon, 2002). For ATT it is sufficient to ensure the existence of potential matches in the control group, whereas for ATE it is additionally required that the combinations of characteristics in the comparison group may also be observed in the treatment group (Bryson,

Dorsett, and Purdon, 2002).

Minima and Maxima comparison: The basic criterion of this approach is to delete all observations whose propensity score is smaller than the minimum and larger than the maximum in the opposite group. To give an example let us assume for a moment that the propensity score lies within the interval $[0.07, 0.94]$ in the treatment group and within $[0.04, 0.89]$ in the control group. Hence, with the 'minima and maxima criterion', the common support is given by $[0.07, 0.89]$. Observations which lie outside this region are discarded from analysis. Clearly a two-sided test is only necessary if the parameter of interest is ATE; for ATT it is sufficient to ensure that for each participant a close non-participant can be found. It should also be clear that the common support condition is in some ways more important for the implementation of kernel matching than it is for the implementation of nearest-neighbour matching. That is, because with kernel matching all untreated observations are used to estimate the missing counterfactual outcome, whereas with NN-matching only the closest neighbour is used. Hence, NN-matching (with the additional imposition of a maximum allowed caliper) handles the common support problem pretty well. There are some problems associated with the 'minima and maxima comparison', e.g. if there are observations at the bounds which are discarded even though they are very close to the bounds. Another problem arises if there are areas within the common support interval where there is only limited overlap between both groups, e.g. if in the region $[0.51, 0.55]$ only treated observations can be found. Additionally problems arise, if the density in the tails of the distribution are very thin, for example when there is a substantial distance from the smallest maximum to the second smallest element. Therefore, Lechner (2002) suggests to check the sensitivity of the results when the minima and maxima are replaced by the 10th smallest and 10th largest observation.

Trimming to Determine the Common Support A different way to overcome these possible problems is suggested by Smith and Todd (2005). They use a trimming procedure to determine the common support region and define the region of common support as those values of P that have positive density within both the $D = 1$ and $D = 0$ distributions, that is:

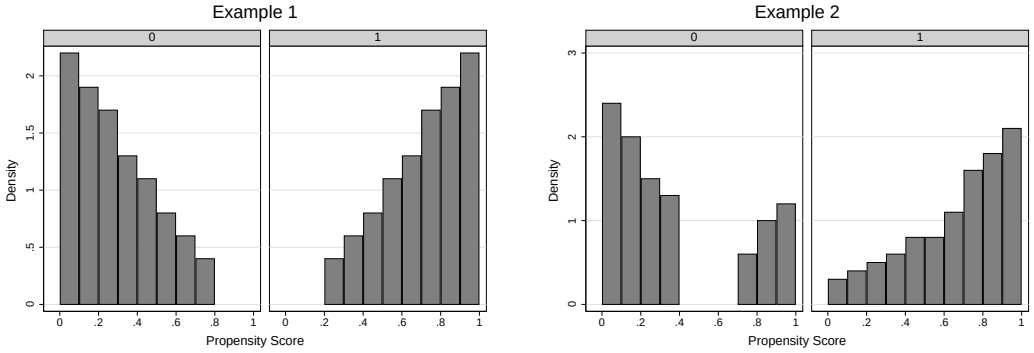
$$\hat{S}_P = \{P : \hat{f}(P|D = 1) > 0 \quad \text{and} \quad \hat{f}(P|D = 0) > 0\}, \tag{10}$$

where $\hat{f}(P|D = 1) > 0$ and $\hat{f}(P|D = 0) > 0$ are non-parametric density estimators. Any P points for which the estimated density is exactly zero are excluded. Additionally - to ensure that the densities are strictly positive - they require that the densities exceed zero by a threshold amount q. So not only the P points for which the estimated density is exactly zero, but also an additional q percent of the remaining P points for which the estimated density is positive but very low are excluded:

$$\hat{S}_{Pq} = \{Pq : \hat{f}(P|D = 1) > q \quad \text{and} \quad \hat{f}(P|D = 0) > q\}.^{10} \tag{11}$$

[10]For details on how to estimate the cut-off trimming level see Smith and Todd (2005). Galdo (2004) notes that the determination of the smoothing parameter is critical here. If the distribution is skewed to the right for participants and skewed to the left for non-participants, assuming a normal distribution may be very misleading.

Figure 3: The Common Support Problem



The left side in each example refers to non-participants (D=0), the right side to participants (D=1).
Source: Hypothetical Example

Figure 3 gives a hypothetical example and clarifies the differences between both approaches. In the first example the propensity score distribution is highly skewed to the left (right) for participants (non-participants). Even though this is an extreme example, researchers are confronted with similar distributions in practice, too. With the 'minima and maxima comparison' we would exclude any observations lying outside the region of common support given by $[0.2, 0.8]$. Depending on the chosen trimming level q, we would maybe also exclude control observations in the interval $[0.7, 0.8]$ and treated observations in the interval $[0.2, 0.3]$ with the trimming approach since the densities are relatively low there. However, no large differences between both approaches would emerge. In the second example we do not find any control individuals in the region $[0.4, 0.7]$. The 'minima and maxima comparison' fails in that situation, since minima and maxima in both groups are equal at 0.01 and 0.99. Hence, no observations would be excluded based on this criterion making the estimation of treatment effects in the region $[0.4, 0.7]$ questionable. The trimming method on the other hand would explicitly exclude treated observations in that propensity score range and would therefore deliver more reliable results. Hence, the choice of the method depends on the data situation at hand and before making any decisions, a visual analysis is recommended.

Failure of the Common Support: Once one has defined the region of common support, individuals that fall outside this region have to be disregarded and for these individuals the treatment effect cannot be estimated. Bryson, Dorsett, and Purdon (2002) note that when the proportion of lost individuals is small, this poses few problems. However, if the number is too large, there may be concerns whether the estimated effect on the remaining individuals can be viewed as representative. It may be instructive to inspect the characteristics of discarded individuals since those can provide important clues when interpreting the estimated treatment effects. Lechner (2000b) notes that both ignoring the support problem and estimating treatment effects only within the common support (subgroup effects) may be misleading. He develops an approach that can be used to derive bounds for the true treatment effect and we describe this approach in detail in subsection 4.2.

3.4 Assessing the Matching Quality

Since we do not condition on all covariates but on the propensity score, it has to be checked if the matching procedure is able to balance the distribution of the relevant variables in both the control and treatment group. Several procedures to do so will be discussed in this subsection. These procedures can also, as already mentioned, help in determining which interactions and higher order terms to include for a given set of covariates X. The basic idea of all approaches is to compare the situation before and after matching and check if there remain any differences after conditioning on the propensity score. If there are differences, matching on the score was not (completely) successful and remedial measures have to be done, e.g. by including interaction-terms in the estimation of the propensity score. A helpful theorem in this context is suggested by Rosenbaum and Rubin (1983) and states that:

$$X \amalg D | P(D = 1|X). \tag{12}$$

This means that after conditioning on $P(D = 1|X)$, additional conditioning on X should not provide new information about the treatment decision. Hence, if after conditioning on the propensity score there is still dependence on X, this suggests either mis-specification in the model used to estimate $P(D = 1|X)$ or a failure of the CIA (Smith and Todd, 2005).[11]

Standardised Bias: One suitable indicator to assess the distance in marginal distributions of the X-variables is the standardised bias (SB) suggested by Rosenbaum and Rubin (1985). For each covariate X it is defined as the difference of sample means in the treated and matched control subsamples as a percentage of the square root of the average of sample variances in both groups. The standardised bias before matching is given by:

$$SB_{before} = 100 \cdot \frac{(\overline{X}_1 - \overline{X}_0)}{\sqrt{0.5 \cdot (V_1(X) + V_0(X))}}. \tag{13}$$

The standardised bias after matching is given by:

$$SB_{after} = 100 \cdot \frac{(\overline{X}_{1M} - \overline{X}_{0M})}{\sqrt{0.5 \cdot (V_{1M}(X) + V_{0M}(X))}}, \tag{14}$$

where X_1 (V_1) is the mean (variance) in the treatment group before matching and X_0 (V_0) the analogue for the control group. X_{1M} (V_{1M}) and $X_{0M}(V_{0M})$ are the corresponding values for the matched samples. This is a common approach used in many evaluation studies, e.g. by Lechner (1999), Sianesi (2004) and Caliendo, Hujer, and Thomsen (2005). One possible problem with the standardised bias approach is that we do not have a clear indication for the success of the matching procedure, even though in most empirical studies a bias reduction below 3% or 5% is seen as sufficient.

[11] Smith and Todd (2005) note that this theorem holds for any X, including those that do not satisfy the CIA required to justify matching. As such, the theorem is not informative about which set of variables to include in X.

t-Test: A similar approach uses a two-sample t-test to check if there are significant differences in covariate means for both groups (Rosenbaum and Rubin, 1985). Before matching differences are expected, but after matching the covariates should be balanced in both groups and hence no significant differences should be found. The t-test might be preferred if the evaluator is concerned with the statistical significance of the results. The shortcoming here is that the bias reduction before and after matching is not clearly visible.

Joint significance and Pseudo-R^2: Additionally, Sianesi (2004) suggests to re-estimate the propensity score on the matched sample, that is only on participants and matched non-participants and compare the pseudo-R^2's before and after matching. The pseudo-R^2 indicates how well the regressors X explain the participation probability. After matching there should be no systematic differences in the distribution of covariates between both groups and therefore, the pseudo-R^2 should be fairly low. Furthermore, one can also perform an F-test on the joint significance of all regressors. The test should not be rejected before, and should be rejected after matching.

Stratification Test: Finally, Dehejia and Wahba (1999, 2002) divide observations into strata based on the estimated propensity score, such that no statistically significant difference between the mean of the estimated propensity score in both treatment and control group remain. Then they use t-tests within each strata to test if the distribution of X-variables is the same between both groups (for the first and second moments). If there are remaining differences, they add higher-order and interaction terms in the propensity score specification, until such differences no longer emerge.

This makes clear that an assessment of matching quality can also be used to determine the propensity score specification. If the quality indicators are not satisfactory, one reason might be mis-specification of the propensity score model and hence it may be worth to take a step back, include e.g. interaction or higher-order terms in the score estimation and test the quality once again. If after re-specification the quality indicators are still not satisfactory, it may indicate a failure of the CIA (Smith and Todd, 2005) and alternative evaluation approaches should be considered.

3.5 Choice-Based Sampling

An additional problem arising in evaluation studies is that samples used are often choice-based (Smith and Todd, 2005). This is a situation where programme participants are oversampled relative to their frequency in the population of eligible persons. We discuss this point briefly and suggest one correction mechanism introduced by Heckman and Smith (1995). First of all, note that under choice-based sampling weights are required to consistently estimate the probability of programme participation. Heckman and Smith (1995) show that with weights unknown, matching methods can still be applied, because the odds ratio estimated using the incorrect weights (those that ignore the fact of choice-based samples) is a scalar multiple of

the true odds ratio, which is itself a monotonic transformation of propensity scores. Hence, matching can be done on the (mis-weighted) estimate of the odds ratio (or of the log odds ratio). Clearly, with single nearest-neighbour matching it does not matter whether matching is performed on the odds ratio or the estimated propensity score (with wrong weights), since ranking of the observations is identical and therefore the same neighbours will be selected. However, for methods that take account of the absolute distance between observations, e.g. kernel matching, it does matter.

3.6 When to Compare and Locking-in Effects

An important decision which has to be made in the empirical analysis is when to measure the effects. The major goal is to ensure that participants and non-participants are compared in the same economic environment and the same individual lifecycle position. For example, when evaluating labour market policies one possible problem which has to be taken into account is the occurrence of locking-in effects. The literature is dominated by two approaches, either comparing the individuals from the begin of the programme or after the end of the programme. To give an example let us assume that a programme starts in January and ends in June. The latter of the two alternatives implies that the outcome of participants who re-enter the labour market in July is compared with matched non-participants in July. There are two shortcomings to this approach. First, if the exits of participants are spread over a longer time period, it might be the case that very different economic situations are compared. Second, a further problem which arises with this approach is that it entails an endogeneity problem (Gerfin and Lechner (2002)), since the abortion of the programme may be caused by several factors which are usually not observed by the researcher.[12]

The above mentioned second approach is predominant in the recent evaluation literature (see e.g. Sianesi (2004) or Gerfin and Lechner (2002)) and measures the effects from the begin of the programme. One major argument to do so concerns the policy relevance. In the above example the policy-maker is faced with the decision to put an individual in January in a programme or not. He will be interested in the effect of his decision on the outcome of the participating individual in contrast with the situation if the individual would not have participated. Therefore comparing both outcomes from begin of the programme is a reasonable approach. What should be kept in mind, however, is the possible occurrence of locking-in effects for the group of participants. Since they are involved in the programme, they do not have the same time to search for a new job as non-participants. Following van Ours (2004), the net effect of a programme consists of two opposite effects. First, the increased employment probability through the programme and second, the reduced search intensity. Since both effects cannot be disentangled, we only observe the net effect and have to take this into account when interpreting the results. As to the fall in the search intensity, we should expect an initial negative effect from any kind of participation in a programme. However, a successful programme should

[12]It may be the case for example that a participant receives a job offer, refuses to participate because he thinks the programme is not enhancing his employment prospects or because lack of motivation. As long as the reasons for abortion are not identified, an endogeneity problem arises.

overcompensate for this initial fall. So, if we are able to observe the outcome of the individuals for a reasonable time after begin/end of the programme, the occurrence of locking-in effects poses fewer problems but nevertheless has to be taken into account in the interpretation.

3.7 Estimation of Standard Errors

Testing the statistical significance of treatment effects and computing their standard errors is not a straightforward thing to do. The problem is that the estimated variance of the treatment effect should also include the variance due to the estimation of the propensity score, the imputation of the common support, and possibly also the order in which treated individuals are matched.[13] These estimation steps add variation beyond the normal sampling variation (see the discussion in Heckman, Ichimura, and Todd (1998)). For example, in the case of NN matching with one nearest neighbour, treating the matched observations as given will understate the standard errors (Smith, 2000).

Bootstrapping: One way to deal with this problem is to use bootstrapping as suggested e.g. by Lechner (2002). This method is a popular way to estimate standard errors in case analytical estimates are biased or unavailable.[14] Even though Imbens (2004) notes that there is little formal evidence to justify bootstrapping, it is widely applied, see e.g. Black and Smith (2003) or Sianesi (2004). Each bootstrap draw includes the re-estimation of the results, including the first steps of the estimation (propensity score, common support, etc.). Repeating the bootstrapping N times leads to N bootstrap samples and in our case N estimated average treatment effects. The distribution of these means approximate the sampling distribution (and thus the standard error) of the population mean. Clearly, one practical problem arises because bootstrapping is very time-consuming and might therefore not be feasible in some cases.

Variance Approximation by Lechner: An alternative is suggested by Lechner (2001). For the estimated ATT obtained via NN-matching the following formula applies:

$$Var(\hat{\tau}_{ATT}) = \frac{1}{N_1} Var(Y(1) \mid D = 1) + \frac{(\Sigma_{j \in I_0}(w_j)^2)}{(N_1)^2} \cdot Var(Y(0) \mid D = 0), \quad (15)$$

where N_1 is the number of matched treated individuals. w_j is the number of times individual j from the control group has been used, i.e. this takes into account that matching is performed with replacement. If no unit is matched more than once, the formula coincides with the 'usual' variance formula. By using this formula to estimate the variance of the treatment effect at time t, we assume independent observations and fixed weights. Furthermore we assume homoscedasticity of the variances of the outcome variables within treatment and control group and that the outcome variances do not depend on the estimated propensity score. This approach

[13] This matters only when matching is done without replacement as discussed in subsection 3.2.
[14] See Brownstone and Valletta (2001) for a discussion of bootstrapping methods.

can be justified by results from Lechner (2002) who finds little differences between bootstrapped variances and the variances calculated according to equation (15).

3.8 Available Software to Implement Matching

The bulk of software tools to implement matching and estimate treatment effects is growing and allows researchers to choose the appropriate tool for their purposes. The most commonly used platform for these tools is Stata and we will present the three most distributed tools here. Becker and Ichino (2002) provide a programme for PSM estimators (*pscore, attnd, attnw, attr, atts, attk*) which includes estimation routines for nearest neighbour, kernel, radius, and stratification matching. To obtain standard errors the user can choose between bootstrapping and the variance approximation proposed by Lechner (2001). Additionally the authors offer balancing tests (blocking, stratification) as discussed in subsection 3.4.

Leuven and Sianesi (2003) provide the programme *psmatch2* for implementing different kinds of matching estimators including covariate and propensity score matching. It includes nearest neighbour and caliper matching (with and without replacement), kernel matching, radius matching, local linear matching and Mahalanobis metric (covariate) matching. Furthermore, this programme includes routines for common support graphing (*psgraph*) and covariate imbalance testing (*pstest*). Standard errors are obtained using bootstrapping methods.

Finally, Abadie, Drukker, Leber Herr, and Imbens (2004) offer the programme *nnmatch* for implementing covariate matching, where the user can choose between several different distance metrics.

4 Sensitivity Analysis

4.1 Unobserved Heterogeneity - Rosenbaum Bounds

We have outlined in section 2 that the estimation of treatment effects with matching estimators is based on the CIA, that is selection on observable characteristics. However, if there are unobserved variables which affect assignment into treatment and the outcome variable simultaneously, a 'hidden bias' might arise. It should be clear that matching estimators are not robust against this 'hidden bias'. Since it is not possible to estimate the magnitude of selection bias with non-experimental data, we address this problem with the bounding approach proposed by Rosenbaum (2002). The basic question to be answered is, if inference about treatment effects may be altered by unobserved factors. In other words, we want to determine how strongly an unmeasured variable must influence the selection process in order to undermine the implications of matching analysis. Recent applications of this approach can be found in Aakvik (2001), DiPrete and Gangl (2004) or Caliendo, Hujer, and Thomsen (2005). We outline this approach briefly, an extensive discussion can be found in Rosenbaum (2002).

Let us assume that the participation probability is given by $P(x_i) = P(D_i = 1 \mid x_i) = F(\beta x_i + \gamma u_i)$, where x_i are the observed characteristics for individual i, u_i is the unobserved variable and γ is the effect of u_i on the participation decision. Clearly, if the study is free of hidden bias, γ will be zero and the participation probability will solely be determined by x_i. However, if there is hidden bias, two individuals with the same observed covariates x have differing chances of receiving treatment. Let us assume we have a matched pair of individuals i and j and further assume that F is the logistics distribution. The odds that individuals receive treatment are then given by $\frac{P(x_i)}{(1-P(x_i))}$ and $\frac{P(x_j)}{(1-P(x_j))}$, and the odds ratio is given by:

$$\frac{\frac{P(x_i)}{1-P(x_i)}}{\frac{P(x_j)}{1-P(x_j)}} = \frac{P(x_i)(1-P(x_j))}{P(x_j)(1-P(x_i))} = \frac{\exp(\beta x_j + \gamma u_j)}{\exp(\beta x_i + \gamma u_i)} = exp[\gamma(u_i - u_j)]. \tag{16}$$

If both units have identical observed covariates - as implied by the matching procedure - the x-vector is cancelled out. But still, both individuals differ in their odds of receiving treatment by a factor that involves the parameter γ and the difference in their unobserved covariates u. So, if there are either no differences in unobserved variables ($u_i = u_j$) or if unobserved variables have no influence on the probability of participating ($\gamma = 0$), the odds ratio is one, implying the absence of hidden or unobserved selection bias. It is now the task of sensitivity analysis to evaluate how inference about the programme effect is altered by changing the values of γ and $(u_i - u_j)$. We follow Aakvik (2001) and assume for the sake of simplicity that the unobserved covariate is a dummy variable with $u_i \in \{0, 1\}$. A good example is the case where motivation plays a role for the participation decision and the outcome variable, and a person is either motivated ($u = 1$) or not ($u = 0$). Rosenbaum (2002) shows that (16) implies the following bounds on the odds-ratio that either of the two matched individuals will receive treatment:

$$\frac{1}{e^\gamma} \leq \frac{P(x_i)(1-P(x_j))}{P(x_j)(1-P(x_i))} \leq e^\gamma. \tag{17}$$

Both matched individuals have the same probability of participating only if $e^\gamma = 1$. If $e^\gamma = 2$, then individuals who appear to be similar (in terms of x) could differ in their odds of receiving the treatment by as much as a factor of 2. In this sense, e^γ is a measure of the degree of departure from a study that is free of hidden bias (Rosenbaum, 2002).

Aakvik (2001) suggests to use the non-parametric Mantel and Haenszel (MH, 1959) test-statistic, which compares the successful number of persons in the treatment group against the same expected number given the treatment effect is zero. He notes that the MH test can be used to test for no treatment effect both within different strata of the sample and as a weighted average between strata. Under the null-hypothesis the distribution of the outcomes Y is hypergeometric. We notate N_{1s} and N_{0s} as the numbers of treated and untreated individuals in stratum s, where $N_s = N_{0s} + N_{1s}$. Y_{1s} is the number of successful participants, Y_{0s} is the number of successful non-participants, and Y_s is the number of total successes in stratum s. The test-statistic $Q_{MH} = (Y_{1s} - E(Y_{1s})/Var(Y_{1s}))$ follows the chi-square

distribution with one degree of freedom and is given by:

$$Q_{MH} = \frac{U^2}{Var(U)} = \frac{[\sum_{s=1}^{S}(Y_{1s} - \frac{N_{1s}Y_s}{N_s}]^2}{\sum_{s=1}^{S}\frac{N_{1s}N_{0s}Y_s(N_s - Y_s)}{N_s^2(N_s-1)}}. \tag{18}$$

To use such a test-statistic, we first have to make treatment and control group as equal as possible since this test is based on random sampling. Since this is done by our matching procedure, we can proceed to discuss the possible influences of $e^\gamma > 1$. For fixed $e^\gamma > 1$ and $u \in \{0, 1\}$, Rosenbaum (2002) shows that the test-statistic Q_{MH} can be bounded by two known distributions. As noted already, if $e^\gamma = 1$ the bounds are equal to the 'base' scenario of no hidden bias. With increasing e^γ, the bounds move apart reflecting uncertainty about the test-statistics in the presence of unobserved selection bias. Two scenarios can be thought of. Let Q_{MH}^+ be the test-statistic given that we have overestimated the treatment effect and Q_{MH}^- the case where we have underestimated the treatment effect. The two bounds are then given by:

$$Q_{MH}^{+(-)} = \frac{[\sum_{s=1}^{S}(Y_{1s} - \widetilde{E}_s^{+(-)}]^2}{\sum_{s=1}^{S} Var(\widetilde{E}_s^{+(-)})}, \tag{19}$$

where $\widetilde{E_s}$ and $Var(\widetilde{E_s})$ are the large sample approximations to the expectation and variance of the number of successful participants when u is binary and for given γ.

4.2 Failure of Common Support - Lechner Bounds

In subsection 3.3 we have presented possible approaches to implement the common support restriction. Those individuals that fall outside the region of common support have to be disregarded. But, deleting such observations yields an estimate that is only consistent for the subpopulation within the common support. However, information from those outside the common support could be useful and informative especially if treatment effects are heterogeneous.

Lechner (2000b) describes an approach to check the robustness of estimated treatment effects due to failure of common support. He incorporates information from those individuals who failed the common support restriction, to calculate non-parametric bounds of the parameter of interest, if all individuals from the sample at hand would have been included. To introduce his approach some additional notation is needed. Define the population of interest with Ω which is some subset from the space defined by treatment status ($D = 1$ or $D = 0$) and a set of covariates X. Ω^{ATT} is defined by $\{(D = 1) \times X\}$ and W^{ATT} is a binary variable which equals one if an observation belongs to Ω^{ATT}. Identification of the effect is desired for $\tau_{ATT}(\Omega^{ATT})$. Due to missing common support the effect can only be estimated for $\tau_{ATT}(\Omega^{ATT*})$. This is the effect ignoring individuals from the treatment group without a comparable match. Observations within common support are denoted by the binary variable W^{ATT*} equal one. The subset for whom such effect is not identified is $\widetilde{\Omega}^{ATT}$.

Let $Pr(W^{ATT*} = 1|W^{ATT} = 1)$ denote the share of participants within common support relative to the total number of participants and λ_0^1 be the mean of

Y(1) for individuals from the treatment group outside common support. Assume that the share of participants within common support relative to the total number of participants as well as ATT for those within the common support, and λ_0^1 are identified. Additionally, assume that the potential outcome Y(0) is bounded: $Pr(\underline{Y} \leq Y(0) \leq \overline{Y}|W^{ATT*} = 0|W^{ATT} = 1) = 1$.[15] Given these assumptions, the bounds for ATT $\tau_{ATT}(\Omega^{ATT}) \in [\underline{\tau}_{ATT}(\Omega^{ATT}), \overline{\tau}_{ATT}(\Omega^{ATT})]$ can be written as:

$$\begin{aligned} \underline{\tau}_{ATT}(\Omega^{ATT}) &= \tau_{ATT}(\Omega^{ATT*})Pr(W^{ATT*} = 1|W^{ATT} = 1) \\ &+ (\lambda_0^1 - \overline{Y})[1 - Pr(W^{ATT*} = 1|W^{ATT} = 1)] \end{aligned} \tag{20}$$

$$\begin{aligned} \overline{\tau}_{ATT}(\Omega^{ATT}) &= \tau_{ATT}(\Omega^{ATT*})Pr(W^{ATT*} = 1|W^{ATT} = 1) \\ &+ (\lambda_0^1 - \underline{Y})[1 - Pr(W^{ATT*} = 1|W^{ATT} = 1)] \end{aligned} \tag{21}$$

Lechner (2000b) states that either ignoring the common support problem or estimating ATT only for the subpopulation within the common support can both be misleading. He recommends to routinely compute bounds analysis in order to assess the sensitivity of estimated treatment effects with respect to the common support problem and its impact on the inference drawn from subgroup estimates.

5 Conclusion

The aim of this paper was to give some guidance for the implementation of propensity score matching. Basically five implementation steps have to be considered when using PSM (as depicted in Figure 1). The discussion has made clear that a researcher faces a lot of decisions during implementation and that it is not always an easy task to give recommendations for a certain approach. Table 2 summarises the main findings of this paper and also highlights sections where information for each implementation step can be found.

The first step of implementation is the estimation of the propensity score. We have shown, that the choice of the underlying model is relatively unproblematic in the binary case whereas for the multiple treatment case one should either use a multinomial probit model or a series of binary probits (logits). After having decided about which model to be used, the next question concerns the variables to be included in the model. We have argued that the decision should be based on economic theory and previous empirical findings, and we have also presented several statistical strategies which may help to determine the choice. If it is felt that some variables play a specifically important role in determining participation and outcomes, one can use an 'overweighting' strategy, for example by carrying out matching on sub-populations.

The second implementation step is the choice among different matching algorithms. We have argued that there is no algorithm which dominates in all data situations. The performance of different matching algorithms varies case-by-case

[15] For example, if the outcome variable of interest is a dummy variable, $Y(0)$ is bounded in $[0, 1]$.

Table 2: Implementation of Propensity Score Matching

Step	Decisions, Questions and Solutions	Chapter
1. Estimation of Propensity Score		
Model Choice	◇ Unproblematic in the binary treatment case (logit or probit)	3.1
	◇ In the multiple treatment case multinomial probit or series of binomial models should be preferred	3.1
Variable Choice	◇ Variables should not be influenced by participation (or anticipation) and must satisfy CIA	3.1
→ Economic Issues	Choose variables by economic theory and previous empirical evidence	3.1
→ Statistical Issues	'Hit or miss'-method, stepwise augmentation, leave-one-out cross validation	3.1
→ Key Variables	'Overweighting' by matching on sub-populations or insisting on perfect match	3.1
2. Choice Among Alternative Matching Algorithms		
Matching Algorithms	◇ The choice (e.g. NN matching with or without replacement, caliper or kernel matching) depends on the sample size, the available number of treated/control observations and the distribution of the estimated PS → Trade-offs between bias and efficiency!	3.2
3. Check Overlap and Common Support		
Common Support	◇ Treatment effects can be estimated only over the CS region!	3.3
→ Tests	Visual analysis of propensity score distributions	3.3
→ Implementation	'Minima and maxima comparison' or 'trimming' method Alternative: Caliper matching	3.3
4.1 Assessing the Matching Quality		
Balancing Property	◇ Is the matching procedure able to balance the distribution of relevant covariates?	3.4
	◇ If matching was not successful go back to step 1 and include higher-order terms, interaction variables or different covariates	↩ Step 1
	◇ After that, if matching is still not successful → Reconsider identifying assumption and consider alternative estimators	
→ Tests	Standardised bias, t-test, stratification test, joint significance and Pseudo-R^2	3.4
4.2 Calculation of Treatment Effects		
Choice-Based Sample	◇ Sample is choice-based? Match on the odds-ratio instead on the propensity score	3.5
When to Compare	◇ Compare from begin of the programme to avoid endogeneity problems!	3.6
	→ Pay attention to the possible occurrence of locking-in effects!	3.6
Standard Errors	◇ Calculate standard errors by bootstrapping or variance approximation	3.7
5. Sensitivity Analysis		
Hidden Bias	◇ Test the sensitivity of estimated treatment effects with respect to unobserved covariates → Calculate Rosenbaum-bounds. If results are very sensitive reconsider identifying assumption and consider alternative estimators	4.1
Common Support	◇ Test the sensitivity of estimated treatment effects with respect to the common support problem	4.2
	→ Calculate Lechner-bounds. If results are very sensitive reconsider variable choice	↩ Step 1

CS: Common Support, NN: Nearest Neighbour, PS: Propensity Score, CIA: Conditional Independence Assumption

and depends largely on the data sample. If results among different algorithms differ, further investigations may be needed to reveal the source of disparity.

The discussion has also emphasised that treatment effects can only be estimated

in the region of common support. To identify this region we recommend to start with a visual analysis of the propensity score distributions in the treatment and comparison group. Based on that, different strategies can be applied to implement the common support condition, e.g. by 'minima and maxima comparison' or 'trimming', where the latter approach has some advantages when observations are close to the 'minima and maxima' bounds and if the density in the tails of the distribution are very thin.

Since we do not condition on all covariates but on the propensity score we have to check in step 4 if the matching procedure is able to balance the distribution of these covariates in the treatment and comparison group. We have presented several procedures to do so, including standardised bias, t-tests, stratification, joint significance and pseudo-R^2. If the quality indicators are not satisfactory, one should go back to step 1 of the implementation procedure and include higher-order or interaction terms of the existing covariates or choose different covariates (if available). If, after that, the matching quality is still not acceptable, one has to reconsider the validity of the identifying assumption and possibly consider alternatives.

However, if the matching quality is satisfactory one can move on to estimate the treatment effects. The estimation of standard errors should either be done by bootstrapping methods or by applying the variance approximation proposed in Lechner (2001). Another important decision is when to measure the effects. We argue that it is preferable to measure the effects from the beginning of the programme. Clearly, what has to be kept in mind for the interpretation is the possible occurrence of locking-in-effects.

Finally, a last step of matching analysis is to test the sensitivity of results with respect to 'hidden bias'. We have presented an approach (Rosenbaum bounds) that allows a researcher to determine how strongly an unmeasured variable must influence the selection process in order to undermine implications of matching analysis. If the results are sensitive and if the researcher has doubts about the CIA he should reconsider to use alternative identifying assumptions. Furthermore, we have presented an approach (Lechner bounds) that allows the researcher to assess how sensitive treatment effects are with respect to the common support problem.

To conclude, we have discussed several issues surrounding the implementation of PSM. We hope to give some guidance for researchers who believe that their data is strong enough to credibly justify CIA and who want to use PSM.

References

Aakvik, A. (2001): "Bounding a Matching Estimator: The Case of a Norwegian Training Program," *Oxford Bulletin of Economics and Statistics*, 63(1), 115–143.

Abadie, A., D. Drukker, J. Leber Herr, and G. W. Imbens (2004): "Implementing Matching Estimators for Average Treatment Effects in STATA," *The Stata Journal*, 4(3), 290–311.

Abadie, A., and G. Imbens (2004): "Large Sample Properties of Matching Estimators for Average Treatment Effects (previous version: Simple and Bias-Corrected Matching Estimators for Average Treatment Effects)," Working Paper, Harvard University.

Augurzky, B., and C. Schmidt (2000): "The Propensity Score: A Means to An End," Working Paper, University of Heidelberg.

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

Black, D., and J. Smith (2003): "How Robust is the Evidence on the Effects of the College Quality? Evidence from Matching," Working Paper, Syracuse University, University of Maryland, NBER, IZA.

Blundell, R., and M. Costa Dias (2002): "Alternative Approaches to Evaluation in Empirical Microeconomics," *Portuguese Economic Journal*, 1, 91–115.

Brand, J., and C. Halaby (2003): "Regression and Matching Estimates of the Effects of Elite College Attendance on Career Outcomes," Working Paper, University of Wisconsin, Madison.

Breiman, L., J. Friedman, R. Olsen, and C. Stone (1984): *Classification and Regression Trees*. Wadsworth International Group, Belmont.

Brodaty, T., B. Crepon, and D. Fougere (2001): "Using Matching Estimators to Evaluate Alternative Youth Employment Programs: Evidence from France, 1986-1988," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 85–123. Physica-Verlag.

Brownstone, D., and R. Valletta (2001): "The Bootstrap and Multiple Imputations: Harnessing Increased Computing Power for Improved Statistical Tests," *Journal of Economic Perspectives*, 15(4), 129–141.

Bryson, A. (2002): "The Union Membership Wage Premium: An Analysis Using Propensity Score Matching," Discussion Paper No. 530, Centre for Economic Performance, London.

Bryson, A., R. Dorsett, and S. Purdon (2002): "The Use of Propensity Score Matching in the Evaluation of Labour Market Policies," Working Paper No. 4, Department for Work and Pensions.

Caliendo, M., and R. Hujer (2005): "The Microeconometric Estimation of Treatment Effects - An Overview," Working Paper, J.W.Goethe University of Frankfurt.

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.

Cochrane, W., and S. Chambers (1965): "The Planning of Observational Studies of Human Populations," *Journal of the Royal Statistical Society, Series A*, 128, 234–266.

Davies, R., and S. Kim (2003): "Matching and the Estimated Impact of Interlisting," Discussion Paper in Finance No. 2001-11, ISMA Centre, Reading.

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.

DiNardo, J., and J. Tobias (2001): "Nonparametric Density and Regression Estimation," *Journal of Economic Perspectives*, 15(4), 11–28.

DiPrete, T., and M. Gangl (2004): "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments," Working Paper, WZB.

Galdo, J. (2004): "Evaluating the Performance of Non-Experimental Estimators: Evidence from a Randomized UI Program," Working Paper, Centre for Policy Research, Toronto.

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

Greene, W. H. (2003): *Econometric Analysis*. New York University, New York.

Ham, J., X. Li, and P. Reagan (2003): "Propensity Score Matching, a Distance-Based Measure of Migration, and the Wage Growth of Young Men," Working Paper, Department of Economics and CHRR Ohio State University, Columbus.

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

——— (1999): "The Pre-Program Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies," Working Paper No. 6983, National Bureau of Economic Research.

Hirano, K., and G. Imbens (2002): "Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catherization," *Health Services & Outcomes Research Methodology*, 2, 259–278.

Hitt, L., and F. Frei (2002): "Do Better Customers Utilize Electronic Distribution Channels? The Case of PC Banking," *Management Science*, 48, No. 6, 732–748.

Imbens, G. (2000): "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87(3), 706–710.

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

Lechner, M. (1998): "Mikrokonometrische Evaluationsstudien: Anmerkungen zu Theorie und Praxis," in *Qualifikation, Weiterbildung und Arbeitsmarkterfolg. ZEW-Wirtschaftsanalysen Band 31*, ed. by F. Pfeiffer, and W. Pohlmeier. Nomos-Verlag.

——— (1999): "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification," *Journal of Business Economic Statistics*, 17, 74–90.

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

——— (2000b): "A Note on the Common Support Problem in Applied Evaluation Studies," Discussion Paper, SIAW.

——— (2001): "Identification and estimation of causal effects of multiple treatments under the conditional independence assumption," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 1–18. Physica-Verlag, Heidelberg.

——— (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.

Leuven, E., and B. Sianesi (2003): "PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing," Software, http://ideas.repec.org/c/boc/bocode/s432001.html.

Mantel, N., and W. Haenszel (1959): "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease," *Journal of the National Cancer Institute*, 22, 719–748.

Pagan, A., and A. Ullah (1999): *Nonparametric Econometrics*. Cambridge University Press, Cambridge.

Perkins, S. M., W. Tu, M. G. Underhill, X. Zhou, and M. D. Murray (2000): "The Use of Propensity Scores in Pharmacoepidemiologic Research," *Pharmacoepidemiology and Drug Safety*, 9, 93–101.

Rosenbaum, P., and D. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–50.

——— (1985): "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *The American Statistican*, 39, 33–38.

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.

Rubin, D. B., and N. Thomas (1996): "Matching Using Estimated Propensity Scores: Relating Theory to Practice," *Biometrics*, 52, 249–264.

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

Silverman, B. (1986): *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Smith, H. (1997): "Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies," *Sociological Methodology*, 27, 325–353.

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

Smith, J., and P. Todd (2005): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics*, 125(1-2), 305–353.

van Ours, J. (2004): "The Locking-in Effect of Subsidized Jobs," *Journal of Comparative Economics*, 32(1), 37–52.

Zhao, Z. (2000): "Data Issues of Using Matching Methods to Estimate Treatment Effects: An Illustration with NSW Data Set," Working Paper, China Centre for Economic Research.

——— (2004): "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence," *The Review of Economics and Statistics*, 86(1), 91–107.