Universität St.Gallen

Practical Procedures to Deal with Common Support Problems in Matching Estimation

Michael Lechner, Anthony Strittmatter

April 2014 Discussion Paper no. 2014-10

School of Economics and Political Science,
Department of Economics

University of St. Gallen

Editor: Martina Flockerzi
University of St.Gallen
School of Economics and Political Science
Department of Economics
Bodanstrasse 8
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35
Email seps@unisg.ch

Publisher: School of Economics and Political Science
Department of Economics
University of St.Gallen
Bodanstrasse 8
CH-9000 St. Gallen
Phone +41 71 224 23 25

Electronic Publication: Fax +41 71 224 31 35
http://www.seps.unisg.ch

Practical Procedures to Deal with Common Support Problems in Matching Estimation[1]

Michael Lechner[2], Anthony Strittmatter[3]

Author's address:

Prof. Dr. Michael Lechner
SEW-HSG
Varnbüelstrasse 14
CH-9000 St. Gallen
Phone +41 71 224 23 20
Fax +41 71 220 23 02
Email Michael.lechner@unisg.ch
Website www.michael-lechner.eu

Anthony Strittmatter, Ph.D.
SEW-HSG
Varnbüelstrasse 14
CH-9000 St. Gallen
Phone +41 71 224 23 05
Fax +41 71 220 23 02
Email anthony.strittmatter@unisg.ch

[1] This project is part of the project "Regional Allocation Intensities, Effectiveness and Reform Effects of Training Vouchers in Active Labor Market Policies", IAB project number 1155. This is a joint project of the Institute for Employment Research (IAB) and the University of Freiburg. We gratefully acknowledge financial and material support by the IAB. We thank Lorenzo Camponovo, Bernd Fitzenberger, and Andreas Steinmayr for helpful comments on a previous draft of the paper. The usual disclaimer applies.

[2] Michael Lechner is also affiliated with CEPR and PSI, London, CESIfo, Munich, IAB, Nuremberg, and IZA, Bonn.

[3] Anthony Strittmatter is also affiliated with the Albert-Ludwigs-University Freiburg.

Abstract

This paper assesses the performance of common estimators adjusting for differences in covariates, like matching and regression, when faced with so-called common support problems. It also shows how different procedures suggested in the literature to tackle common support problems affect the properties of such estimators. Based on an Empirical Monte Carlo simulation design, a lack of common support is found to increase the root mean squared error (RMSE) of all investigated parametric and semiparametric estimators. Dropping observa¬tions that are off support usually improves their performance, although the amount of improvement depends on the particular method used.

Keywords

Empirical Monte Carlo Study, matching estimation, regression, common support, outlier, small sample performance.

JEL Classification

C21, J68.

1 Introduction

It is a common task in applied econometrics to compare moments or distributions of random variables of two subsamples that are free of differences due to some observed variables. One example that received a lot of attention in the recent applied literature is the evaluation of active labour market programmes (ALMP) based on very informative and large administrative data (see the meta study by Card, Kluve, and Weber, 2010, for a comprehensive summary of this literature). Usually, the main goal in this literature, which is very much inspired by the treatment effect framework (e.g. Rubin, 1974), is to compare expected future reemployment and earnings ('outcomes') of participating in such programmes ('treatment') compared to not doing so. In many cases, identification is based on a selection-on-observable identification strategy exploiting the rich data available (e.g., see Imbens, 2004, for an overview).

Essentially, all estimation procedures used in this context are based on predicting average outcomes in one treatment state (e.g. for non-participants called 'non-treated' from now on) based on the distribution of exogenous variables in the other treatment state (e.g., the programme participants called 'treated' from now on). Apparently, in this example, if the values of the characteristics observed for the treated are not observable among the non-treated, the mean outcome for non-treated that would have such characteristics cannot be estimated without strong assumptions. These assumptions have to allow some extrapolation for such values of characteristics. This is the so-called *no common support problem*. A related (finite sample) problem appears when there are only few observations in some relevant part of the covariate space in the particular sample at hand. This is called the *thin common support problem*. Such areas of no or thin support may increase biases and variances of estimators (e.g., Kahn and Tamer, 2010, Crump, Hotz, Imbens, and Mitnik, 2009). Surprisingly, and

different to other aspects of identification and estimation in treatment effect models, these issues received only little attention in the literature.

In this paper, we investigate the impact of these problems on commonly used estimators, and analyse procedures proposed to deal with it in applied studies. Similar to Huber, Lechner, and Wunsch (2013), we do this in the context of a large-scale active labour market policy evaluation and use Empirical Monte Carlo methods. One of the contributions of this paper is to show that common support problems as well as the 'remedies' chosen matter for the results obtained (see also Busso, DiNardo, and McCrary, 2014a, Dehejia and Wahba, 1999, Smith and Todd, 2005). When there are support problems, different options have been used in applied studies. The most convenient one is to ignore the problem and take as given the way the specific estimators (as implemented in the software used) deals with it. If support problems are explicitly dealt with, then one option is to change the population for which the effects are estimated to the one for which the distributions of characteristics overlap. This matters if effects are heterogeneous. Alternatively, one may use a parametric model to predict mean outcomes in no-support regions. A third alternative is to give up point-identification altogether and confine oneself to a set-identified parameter (e.g., Lechner, 2008).

In this paper, we stick to the more popular point-identified case and analyse several practical adjustment procedures proposed in the literature. For example, Rosenbaum and Rubin (1983) suggest dropping treated observations for which there are no non-treated observations with the same covariate values. However, this procedure may lead to drastic reductions in sample size if the covariate space is large or if there are (almost) continuous covariates. Thus, most procedures used in practice, and analysed in this paper, focus directly on the propensity score, i.e. the conditional probability of treatment given the values of the covariates (e.g. Crump, Hotz, Imbens, and Mitnik, 2009, Dehejia and Wahba, 1999, Grzybowski et al., 2003,

Heckman, Ichimura, Smith, and Todd, 1998, Heckman, Ichimura, and Todd, 1998, and Smith and Todd, 2005, and Vincent et al., 2002).

There are also suggestions to deal simultaneously with problems of common support as well as with (too) important non-treated observations. These issues are related because in regions of thin support the predictions are based on only few observations. This potentially leads to finite-sample bias and increased variance. Therefore, Imbens (2004) and Huber, Lechner, and Wunsch (2013) develop ways to deal with this issue, which appear to be effective in the simulation study conducted by the latter authors. Finally, Crump, Hotz, Imbens, and Mitnik (2009) suggest explicitly focussing the estimation on a subsample of the data with 'strong common support' in order to maximise the precision of the estimators.

Recently, Busso, DiNardo, and McCrary (2014a) investigate some of those procedures. They find that some of the approaches (i.e. those by Crump, Hotz, Imbens, and Mitnik, 2009, Dehejia and Wahba, 1999, Ho, Imai, King, and Stuart, 2007, and Smith and Todd, 2005, to be explained in detail below) have the potential to reduce the bias and partly increase the efficiency of the estimators. However, their simulation study is based on artificial distributional assumptions, which are unlikely to be observed in reality.

To address the issue of using a realistic design in a simulation study, recently, Lechner and Wunsch (2013), and Huber, Lechner, and Wunsch (2013) advocated using what they call an Empirical Monte Carlo Study (EMCS). The main idea is to use a large data set, which is similar to the data typically used in relevant applied work. In this large data set, which is considered as the 'population' in the simulation exercise, a propensity score is estimated. Then, random samples are drawn from the subpopulation of the non-treated. For this group, the effect of treatment is known to be zero. Next, the previously estimated propensity score is used to assign a (pseudo-) treatment status to these non-treated. Such procedure reflects the

same selectivity observed in the population while insuring that the true effect is known (and zero). Importantly, it does not require a priori specifying the joint distribution of outcomes, confounders, and treatment (or conditional moments of it).[2]

Our results suggest that dropping observations off-support improves the performance of many estimators, mainly by increasing their precision. For matching estimators, this improvement can exceed 20 standard deviations in some specification. Even for parametric estimators the potential performance improvements are non-trivial and worth pursuing. The procedures of Dehejia and Wahba (1999), Grzybowski et al. (2003), and Vincent et al. (2002) appear to improve the performance of different estimators in (almost) all specifications. We suggest dropping treated observations with propensity score values above a specific threshold. Our findings suggest specifying the cut-off value at the maximum or the 99%-quantile of the propensity score in the non-treated subpopulation. This procedure might be combined with an adjustment among the non-treated. Non-treated observations with a high importance could be dropped in the first place. Afterwards, treated observations with propensity score values above a threshold are dropped, with the threshold value being specified in the remaining non-treated sample (two-step procedure suggested by Huber, Lechner, and Wunsch, 2013).

The remainder of this study is organized as follows. In the next section, we introduce the econometrics framework. In Section 3, we discuss detection of support problems and possible identifying assumptions. The underlying data and empirical set-up is presented in Section 4. In Section 5, we discuss the simulation design. The results are presented in Section 6 and

[2] There are alternatives in the literature that share the goal of making simulation studies more relevant but do not share this feature. For example, Abadie and Imbens (2011) and Busso, DiNardo and McCrary (2014b) propose to apply more structural empirical simulation designs. The dependence structures between the control, treatment and outcome variables are estimated with real data. Afterwards, the treatment and outcome variables are simulated using the distribution of control variables from real data and the coefficients estimated in the first step. This approach has the advantage that the size of the treatment effect can be restricted, but it requires assumptions about the distribution of the 'error terms'.

conclusions are drawn in Section 7. Further, we provide some details of the simulation procedures and large tables with supplementary results in the Online Appendices A-F.

2 The econometric model

2.1 Parameter of interest

Consider a setup with a binary treatment D, $d \in \{0,1\}$ (e.g. participation in a program) and an outcome variable Y (e.g. post-programme employment or earnings).[3] X is a K-dimensional vector of covariates with support $\mathcal{X} \subset \mathbb{R}^K$. The goal of the empirical analysis is to obtain mean comparisons of the outcome variable in the subsamples defined by D that are free of any differences due to the covariates X. If the covariates space is sufficiently rich, such a comparison will uncover parameters that have a causal interpretation like the average treatment effect on the treated (ATET, for details see Imbens, 2004, or Rubin, 1974). More specifically, the focus is on the following estimand:

$$\gamma = E[Y|D = 1] - E[E(Y|X = x, D = 0)|D = 1].$$

In analogy to this definition, one can define similar parameters for other subpopulations as well. Here, for simplicity, the focus is only on γ, because (i) this parameter is of interest in most evaluation studies, and (ii) because support problems appearing for the other parameters can be dealt with in a symmetric way.

Assume that an i.i.d. sample of size N is available containing measurements of y_i, d_i, and x_i. Thus, under usual regularity conditions γ can be non-parametrically estimated. However, when the dimension of X is high, the so-called curse of dimensionality makes reliable non-

[3] We use the convention that capital letters denote random variables, while small letter denote particular values. If small letters are subscripted with i, such values are observed in a random sample.

parametric estimation difficult to impossible. In this case, the balancing score property introduced by Rosenbaum and Rubin (1983) obtains practical relevance. It implies the following equality, which also holds in all subpopulations defined by *X*:

$$E[E(Y|X = x, D = 0)|D = 1] = E[E(Y|p(X) = p(x), D = 0)|D = 1].$$

p(x) = P(D=1|X=x) denotes the *propensity score*. In most applications, the propensity score is approximated by a parametric model so that the resulting non-parametric estimation problem becomes essentially one-dimensional.

2.2 The common support assumption

Implicitly, the estimand defined above requires that for every unit with *d=1* there should be a unit with the same (or a similar) value of *p(x)* among the group of units with *d=0*. Let $f(p(x)|D = d)$ be the density of the propensity score $p(x)$ conditional on the treatment status d. The density $f(p(x)|D = 1)$ can be consistently estimated in the treated subsample under some regularity conditions. The same is true for $f(p(x)|D = 0)$ in the non-treated subsample. There are 'support problems' (in the population) when $f(p(x)|D = 0) = 0$ and $f(p(x)|D = 1) > 0$. The set $\mathcal{W}_d = \{p(x) \in (0,1): f(p(x)|D = d) > 0\}$ represents the support of $p(x)$ for *d* being zero or one.

Common support assumption (CS)

$\mathcal{W}_1 \subseteq \mathcal{W}_0$.

This assumption is automatically satisfied in the population when $p(x) < 1$. However, even if this assumption holds in the population, the actual sample available may still be plagued by a lack of overlap.

2.3 Estimation

We assess the impact of support problems and their remedies on three different classes of estimators. In the following, we give a brief overview of the applied estimators. The details of the implementation can be found in Section 5.5.

The first class consists of parametric regressions. We use ordinary least squares (OLS) regressions for continuous and discrete non-binary outcome variables. For binary outcome variables, we specify parametric probit models. These estimators 'work' also without common support. However, since parametric models are usually seen as an approximation to the true model, one may suspect that its out-of-support predictions might be particularly unreliable.

The second class of estimators considered are based on inverse probability weighting (IPW) using the propensity score. The estimated propensity score is obtained from a parametric probit model. Therefore, the common support assumption is, again, not required for this estimator either. However, as before, outside the common support, the results crucially depend on the correct specification of the parametric model for the propensity score.

The third class of estimators considered are propensity score matching estimators. In particular, we investigate the modified and bias adjusted propensity score radius-matching estimator suggested by Lechner, Miquel, and Wunsch (2011).

3 Identification and estimation in the presence of support problems

3.1 Rules to detect support problems

To discuss support issues more formally, define a binary support variable $\Omega \in \{0,1\}$, which is equal to one for observations in the region of common support and zero otherwise. Note that

since the focus is on the parameter γ, common support issue only concern treated units without comparable non-treated units (and not vice versa). Thus, Ω is set to one for non-treated observations (if not explicitly indicated differently).

Several rules to detect support problems have been suggested in the literature. For example, Heckman, Ichimura, Smith, and Todd (1998), Heckman, Ichimura, and Todd (1998), and Smith and Todd (2005) drop observations with low densities of the propensity score. Let $f\big(p(x)\big)$ denote the marginal density of the propensity score and specify the τ% quantile of the sample distribution of the marginal propensity score density as follows:

$$q_\tau = F^{-1}_{f(p(x))}(\tau) = inf\left(q_\tau : F_{f(p(x))}(q_\tau) \geq \tau\right).$$

Treated observations with a density below the τ%-quantile are dropped, i.e. $f\big(p(x)\big) < q_\tau$.[4]

This restriction requires the marginal sample density to be above a specific threshold, irrespective of the value of the propensity score. This restriction allows the detection of 'holes in the support' as well as support problems at the boundary. This is an advantage in comparison to other restrictions, which detect support problems in boundary regions only. However, the choice of cut-off values τ will be ad-hoc. Furthermore, it is challenging to estimate the marginal density $f\big(p(x)\big)$ non-parametrically, especially in regions of thin support that are by definition characterised by few observations. In the simulations, we estimate $f\big(p(x)\big)$ using Gaussian kernels with Silverman's rule of thumb determining the bandwidth. Consequently, the following rule is used to determine whether observation 'i' is on the support:

[4] This procedure could also be applied using the conditional densities. Smith and Todd (2005) estimate average treatment effects for the entire population. They estimate $f(p(x)|D=1)$ and $f(p(x)|D=0)$ and rank these densities. Afterwards, they drop observations with the 2% lowest densities. Busso, DiNardo and McCrary (2014a) rank only $f(p(x)|D=1)$ to specify cut-off rules for Average Treatment Effects on the Treated (ATET).

Rule 1: $\Omega(x_i, d_i) = (1 - d_i) + d_i \, 1\{f(p(x_i)) \geq q_\tau\}$,

with $1\{\cdot\}$ being the indicator function, which is one if its argument is true.

As alternative procedure, Grzybowski et al. (2003) and Vincent et al. (2002) use a symmetric interval $[p(x) - u, p(x) + u]$ to determine common support. If at least one observation is observed within this internal, then they conclude to have (fuzzy) common support at *x*. The relevant values of *x* for which this is checked are the realised values of *X* among the treated. They suggest dropping treated observations for which non-treated units are not observed within this interval. As before, the cut-off value *u* is chosen by some ad-hoc rule. As before, this procedure allows detecting holes in the support. Below, the rule

Rule 2: $\Omega(x_i, d_i) = (1 - d_i) + d_i \max_j \left\{(1 - d_j) 1\{|p(x_i) - p(x_j)| \leq u\}\right\}$

is used with either *u = 0.01* or *u = 0.1*.

Dehejia and Wahba (1999) use upper bounds of the propensity score, $\bar{p}$, to detect support problems. They discard treated observations with propensity scores larger than $\bar{p}$. This restriction is easy to use as it does only depend on the propensity score and it requires no additional estimation steps. Clearly, it is a special case of the previous case. Dehejia and Wahba (1999) suggest using the $k\%$-highest propensity score level of the non-treated subpopulation to specify $\bar{p}$. Thus, it is not possible to detect holes in the support. The support rule applied below is:

Rule 3: $\Omega(x_i, d_i) = (1 - d_i) + d_i \, 1\{p(x_i) < \bar{p}\}$.

In the simulations below, this rule is applied with three different values for the cut-off: (i) $\bar{p}$ being the highest, (ii) the 99%- or (iii) the 95%-highest value of the propensity score in the non-treated subpopulation. Note that the specification with $\bar{p}$ being the highest value of the

propensity score in the non-treated subpopulation corresponds to the suggestions of Ho, Imai, King, and Stuart (2007) in the one-dimensional case.

The suggestions of Crump, Hotz, Imbens, and Mitnik (2009) are similar for the treated. In addition, they suggest discarding non-treated observations with propensity scores above the threshold as well. Below, two different versions of their proposal are implemented. The first one follows exactly their rule of thumb: $\Omega(x_i) = 1\{p(x_i) \geq 0.1, p(x_i) \leq 0.9\}$. Since we are only interested in γ, low propensity score levels are essentially of no concern. Therefore, we also use $\Omega(x_i) = 1\{p(x_i) \leq 0.9\}$, following Busso, DiNardo, and McCrary (2014a).

3.2 Classification of support problems

In the next step, we classify support problems using conditional expectations of potential outcomes. The potential outcome of the treated under treatment can be split into,

$$Y_{1|S} = E[Y|D = 1, \Omega = 1], \text{and}$$

$$Y_{1|N} = E[Y|D = 1, \Omega = 0],$$

where $Y_{1|S}$ equals the expected outcome of the treated population which is on the support and $Y_{1|N}$ corresponds to the expected outcome of the treated population which is off the support. The expected outcome of *all* treated is thus given by,

$$E[Y|D = 1] = Y_{1|S} \cdot p(\Omega) + Y_{1|N} \cdot \big(1 - p(\Omega)\big),$$

with $p(\Omega) = P(\Omega = 1|\mathrm{D} = 1)$.

Noting that Ω is a function of X and D only, the expected outcome of the non-treated evaluated at the characteristics of the treated can be split in a similar way:

$$Y_{0|S} = E[E(Y|X = x, D = 0)|D = 1, \Omega = 1],$$

$$Y_{0|N} = E[E(Y|X = x, D = 0)|D = 1, \Omega = 0],$$

$$E[E(Y|X = x, D = 0)|D = 1] = Y_{0|S} \cdot p(\Omega) + Y_{0|N} \cdot (1 - p(\Omega)).$$

It will be useful to use the same notation for the effects of interest:

$$\gamma_S = E[Y|D = 1, \Omega = 1] - E[E(Y|X = x, D = 0)|D = 1, \Omega = 1] = Y_{1|S} - Y_{0|S},$$

$$\gamma_N = E[Y|D = 1, \Omega = 0] - E[E(Y|X = x, D = 0)|D = 1, \Omega = 0] = Y_{1|N} - Y_{0|N},$$

$$\gamma = \gamma_S \cdot p(\Omega) + \gamma_N \cdot (1 - p(\Omega)).$$

If there are support problems, i.e. $p(\Omega) < 1$, identification of γ requires identification of $Y_{1|S}$, $Y_{1|N}$, $Y_{0|S}$, and $Y_{0|N}$ (otherwise identification of $Y_{1|S}$ and $Y_{0|S}$ is sufficient). In the following, the first three parameters are assumed to be identified (i.e. can be consistently estimated), thus γ_S is identified. $Y_{0|N}$ is unidentified because there no observations that are sample counterparts (due to a violation of Assumption CS). Therefore, if there is effect heterogeneity, i.e. $\gamma_S \neq \gamma_N$, additional assumptions are needed to identify $Y_{0|N}$ and thus γ.

3.3 Assumptions that identify the effects off support

Now consider assumptions that are sufficient to identify γ in the presence of support problems. These assumptions can be classified into two categories. Either they concern the expected outcome of the (unobserved) non-treated with characteristics off support ($Y_{0|N}$) or they relate to the effect in the no-support region, γ_N.

3.3.1 *Assumptions about expected outcomes of unobserved non-treated*

In this section three different identifying assumptions about $Y_{0|N}$ are introduced. First, assume that the expected outcomes on and off support are equal.

Assumption A (equal averages): $Y_{0|N} = Y_{0|S}$.

Obviously, Assumption A is very strong because this assumption stipulates that two populations with potentially very different values of the covariates (usually selected as jointly influencing *D* and *Y*) have the expected non-treatment outcomes.

Second, assume that the level of $Y_{0|N}$ is equal to the level of the expected outcome evaluated in the close neighbour to the cut-off value defined in terms of *Rule 3*:

$$Y_{0|p(x)\uparrow\bar{p}} = \lim_{a\to\bar{p}} E[E(Y|p(X)=a, D=0)|D=1, p(X)=a] = \lim_{a\to\bar{p}} E(Y|p(X)=a, D=0)$$

Assumption B (evaluate $Y_{0|N}$ at truncation level): $Y_{0|N} = Y_{0|p(x)\uparrow\bar{p}}$.

Assumption B reflects the implicit assumptions of matching estimators when support problems are ignored. Estimating $Y_{0|p(x)\uparrow\bar{p}}$ is similar to the one-sided regression problems around the cut-off that have been well studied in the context of the regression discontinuity design (see for example Imbens and Lemieux, 2008). Therefore, below $Y_{0|p(x)\uparrow\bar{p}}$ is estimated using a local linear kernel regression and the bandwidth selector suggested by Imbens and Kalyanaraman (2012).

Finally, assume a parametric model to predict $Y_{0|N}$ based on the dependence structure between *Y* and *X* within the common support.

Assumption C (model of expectation of outcome): $Y_{0|N} = \int g(x, \hat{\alpha}) \cdot f_X(x|D=1, \Omega=0)dx.$

Below $g(x, \hat{\alpha})$ are the fitted values of a (flexible) parametric regression of *Y* on *X* in the non-treated subpopulation.

3.3.2 *Assumptions about treatment effects off support*

The second set of assumptions is made with respect to the average treatment effect on the treated off support, γ_N. In analogy to Assumption A, assume that the average effect on the treated outside the support is equal to the average effect on the treated inside the support.

Assumption D (effect homogeneity): $\gamma_S = \gamma_N$.

This assumption is the workhorse of applied research. It drops observations outside the common support and uses the results for γ_S to extrapolate to the population parameter γ.

Second, in analogy to Assumption B, $(Y_{1|N} - Y_{0|N})$ may be restricted to be equal to the conditional effect at the truncation level $\bar{p}$ (Rule 3). Define,

$$\gamma_{p(x)\uparrow\bar{p}} = \lim_{a\to\bar{p}}\{E[Y|p(x)=a, D=1] - E[E(Y|p(x)=a, D=0)|D=1, p(x)=a]\}$$

$$= \lim_{a\to\bar{p}}\{E[Y|p(x)=a, D=1] - E(Y|p(x)=a, D=0)\},$$

to be the average effect of the treated with support conditional on the propensity score level $p(x)$ being as close as possible to $\bar{p}$.

Assumption E (local effect homogeneity)*:* $\gamma_N = \gamma_{p(x)\uparrow\bar{p}}$

The conditional treatment effect $\gamma_{p(x)\uparrow\bar{p}}$ is estimated equivalently to the conditional potential outcome $Y_{0|p(x)\uparrow\bar{p}}$ in the last subsection. We use a local linear kernel together with the bandwidth selection approach of Imbens and Kalyanaraman (2012).

Third, in analogy to Assumption C, identification might be based on a functional form assumption on the dependence of the effect on the covariates.

Assumption F (model of conditional effects): $\gamma_N = \int g(x, \hat{\gamma}_1) \cdot f_X(x|D=1, \Omega=0)dx$,

with $g(x, \hat{\gamma}_1)$ being the expected conditional effects. Below $g(x, \hat{\gamma}_1)$ is estimated with a (flexible) parametric regression of *Y* on *X* and a full set of interaction terms between *D* and *X*.

3.4 Balance of propensity score distributions

Even if the support conditions are satisfied in the population, i.e. $p(\Omega)=1$, an imbalance in the sample propensity score distributions could still lead to regions with only few (or none)

treated or non-treated observations. Imbens (2004) suggests generating measures for the importance of each observation. A high importance indicates a high probability for a large imbalance in the conditional propensity score distributions with respect to the treatment status (for a given sample size). Huber, Lechner, and Wunsch (2013) suggest dropping non-treated observations with a high importance in the first place. They use $\omega_i = \frac{(1-d_i)p(x_i)}{1-p(x_i)} / \sum_{i=1}^{N} \frac{(1-d_i)p(x_i)}{1-p(x_i)}$ as importance measure (see details of the implementation in Section 5.5). It is one advantage of this procedure that it becomes asymptotically irrelevant because the individual weights, ω_i, shrink with sample size. As a second step, they suggest applying procedures to detect support violations as described in Section 3.1. Simulation results from Huber, Lechner, and Wunsch (2013) suggest that this two-step approach works well in the context of the program evaluation literature. Therefore, in the simulations, both, one- and two-step procedures are used. For the two-step procedures, below non-treated observations with $\omega_i > 0.04$ are dropped (this value worked well in Huber, Lechner, and Wunsch, 2013; below, experimenting with other cut-off values did not change the results by much).

4 Empirical bases for the simulations: German administrative data

4.1 Data base

The German Federal Employment Agency generated the data for this study from their internal administrative social security records. It contains information on individuals receiving a training voucher in 2003 or 2004 and those who did not. Training vouchers certify the eligibility for public sponsored further training (see Doerr et al., 2014, for more details). Unemployed awarded with a voucher may redeem it at certified training providers.

The data contains extensive daily information on employment subject to social security contributions, receipt of transfer payments during unemployment, job search, and participation in different active labour market programs as well as rich individual information. It is a combination of two populations: A 3% random sample of those individuals in Germany who experience at least one switch from employment to non-employment in 2003 and are not awarded a voucher are merged with *all* voucher recipients. We account for the treatment-based sampling scheme by using sampling weights when necessary. All individuals in the sample became unemployed in 2003 after continuous employment of at least three months. They are eligible for unemployment benefits, between 20 and 65 years old, German citizens, and hold at least a lower secondary schooling degree. This type of data has been frequently used to evaluate German active labour market policies (e.g. Biewen, Fitzenberger, Osikominu, and Paul, 2014, Lechner, Miquel, and Wunsch, 2011, Rinne, Uhlendorff, and Zao, 2013). This type of data is comparable to many administrative data sets in Europe (see Lechner and Wunsch, 2013). The next section describes how it is used to simulate realistic data generating processes (DGPs).

4.2 Treatment definition

The treatment consists of being awarded with a voucher for further training during the first twelve months of unemployment. Following Lechner and Wunsch (2013) we simulate hypothetical program starts (i.e. receipts of the voucher) in the non-treated group by drawing start dates from the distribution observed among the treated. We consider the first voucher awarded during the first unemployment spell in 2003. Training vouchers indicate the particular type of course, for which they may be redeemed. We only consider vouchers awarded to obtain skills for manufacturing or service workers (in the following: *vouchers for manufacturing and service workers, VMSW*) and vouchers to obtain skills for technicians (in the follow-

ing: *vouchers for technicians, VTEC*). They are interesting because their selection processes show considerable heterogeneity: The *award of VTEC* is far more selective than the *award of VMSW*. Concerning the sample sizes, there are about 50,000 non-treated observations in both samples (note that they differ slightly across comparisons because of the way start dates are generated) and around 35,000 *VMSW* and 2,600 *VTEC* recipients.

4.3 Control variables and common support

The choice of control variables follows Lechner and Wunsch (2013). We consider all variables identified as important confounders in their study, i.e. baseline personal characteristics, timing of program start, regional dummies, benefit and unemployment insurance claims, pre-program outcomes, and the short-term labour market history. Covariates on individual characteristics refer to the time of inflow into unemployment. The full list of variables used for the propensity scores (including interaction terms) is given in Online Appendix A (Tables A.1 and A.2). Table 1 shows those control variables with a standardized mean difference between treated and controls larger than 30.

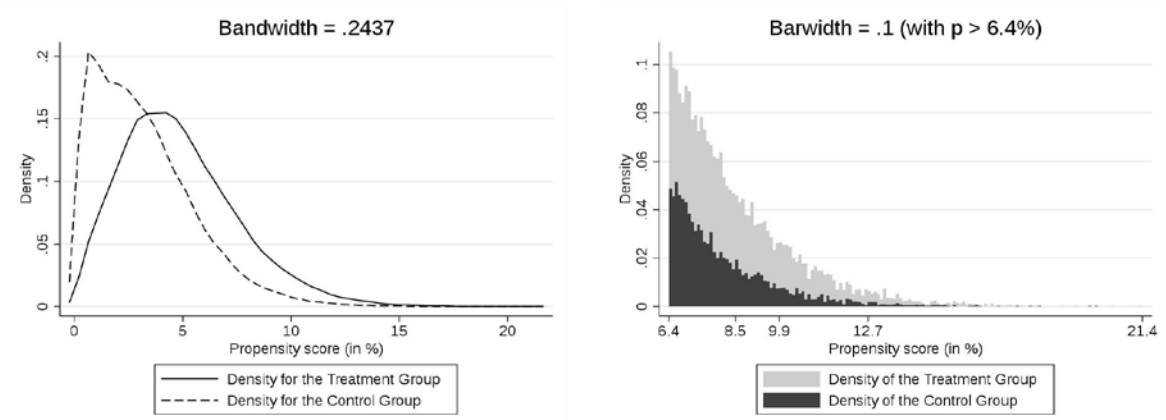
Table 1 shows that individuals with high educational or occupational degrees who work as technicians, professionals, or managers are very likely to obtain a *VTEC* (but not necessarily a *VMSW*). For simulation and modelling purposes, it is useful to relate the support problems to particular variables. Thus, three interactions between occupation and educational/vocational degree are included in the propensity score probit model: (i) being a technicians or having an associate profession interacted with higher secondary schooling degree (S_1); (ii) being a professional or manager interacted with higher secondary schooling degree (S_2); and (iii) being a professional or manager interacted with university or college degree (S_3).

Table 1: Selected statistics for the treatment 'award of VMSW' and 'award of VTEC'

	VMSW			*VTEC*		
	Treated Mean	Non-treated Mean	Stand. Diff. in %	Treated Mean	Non-treated Mean	Stand. Diff. in %
Variables	(1)	(2)	(3)	(4)	(5)	(6)
			Control Variables			
University entry school degree	.23	.19	7	.51	.19	51
Vocational degree	.83	.84	3	.98	.84	34
Craft, machine operators, related	.32	.38	10	.16	.38	36
Service workers and clerks	.41	.36	7	.08	.36	50
Technicians, associate profes-sions, professionals, managers	.27	.26	2	.76	.26	82
Half-months employm. last 24 m.	44.0	40.3	32	44.91	40.28	41
# of employm. spells last 24 m.	1.2	1.4	21	1.14	1.36	31
Half-months since last UE last 24 months	43.9	40.7	22	44.84	40.77	30
Any program in last 24 months	.07	.25	35	.07	.25	36
Amount of UI benefit	25.2	22.3	16	32.95	22.34	55
Earnings 4 years before (defl.)	52.78	47.75	9	72.35	47.72	42
Cumulated duration employed 4 years before (half-months)	79.49	72.25	23	82.19	72.27	32
Cum. earnings 4 years before (deflated, per month, in 1000)	1.83	1.57	20	2.45	1.57	62
Cum. UI benefits 4 years before (deflated, per month, in 1000)	.04	.09	31	.04	.09	30
Technicians & associate professions x higher secondary school degree S_{i1}	.21	.20	2	.65	.20	73
Professionals and managers x Higher second. school degree S_{i2}	.10	.10	1	.36	.10	45
Professionals & managers x university or college degree S_{i3}	.05	.06	3	.30	.06	45
'Support indic.' $S_i = max(S_{i1},S_{i2},S_{i3})$	.21	.20	2	.66	.20	73
			Outcome Variables			
Average earnings in year 4 after voucher receipt (Earnings)	15,034	13,098	10	20,084	13,081	32
Months employed in year 4 after voucher receipt (Months Empl.)	7.26	6.53	9	7.26	6.49	10
Employment 4 years after voucher receipt (Employment)	.64	.56	11	.64	.56	12
Observations	34,838	48,148		2,629	47,670	
Weighted Observations	34,838	1,111,813		2,629	1,099,443	

Note: Columns (1) & (2) and (4) & (5) show the mean of selected covariates and outcome variables in the treated and the corresponding non-treated subpopulations. Columns (3) & (6) present the respective standardized differences. The full set of covariates is contained in Tables A.1 and A.2 in Online Appendix A.

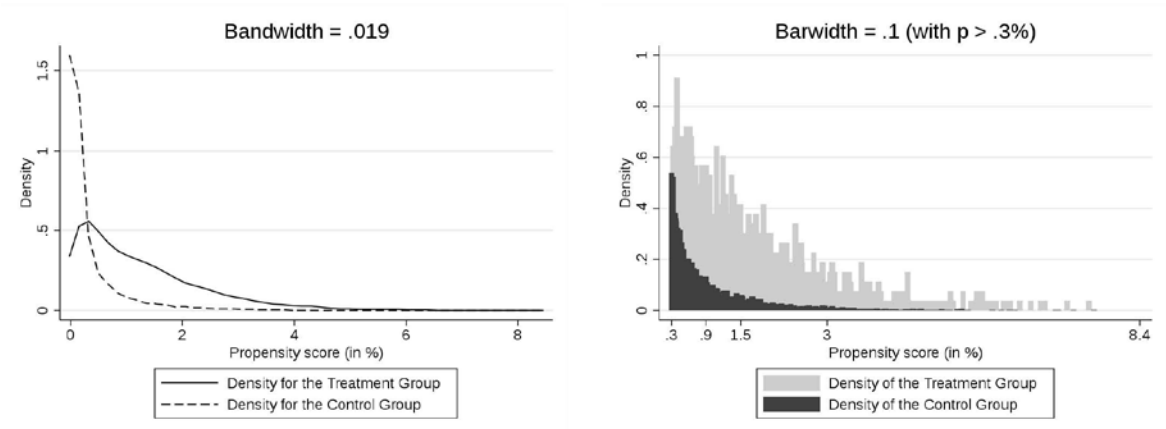
Figure 1: *Kernel and histogram density estimates of the propensity score in the full sample for the treatment 'award of VMSW'*



(a) Density (b) Histogram

Note: Kernel and histogram density estimates are reported, because support problems are difficult to detect in (smoothed) kernel estimations. Kernel estimate are based on the Gaussian kernel using Silverman's rule as bandwidth selector. In the histogram, the 1st to 4th number shown corresponds to the 75%, the 90% quantile, the 95% quantile and the 99% quantile. The last number is the maximum value of the propensity score.

Figure 2: *Kernel and histogram density estimates of the propensity score in the full sample for the treatment 'award of VTEC'*



(a) Density (b) Histogram

Note: See note below Figure 1.

The three interaction terms, which are binary, are collected in the 'support variable' *S*:

$$S_i = \max(S_{1i}, S_{2i}, S_{3i}).$$

The support variable S_i equals one for 66% of all individuals who are awarded with *VTEC*, but only for 21% of all individuals who are awarded with *VMSW* and for 20% of those not receiving a voucher at all.

In addition to sample means and standardized differences between treated and controls, column (4) of Tables A.1 and A.2 in Online Appendix A shows the average marginal effects of the estimated propensity scores. The corresponding densities among the respective treated and non-treated (shown in Figure 1 for *VMSW* and in Figure 2 for *VTEC*) together with the much higher standardized differences reiterate that selectivity is more important for *VTEC* than for *VMSW*. In fact, even in this large sample, for large values of the propensity score the common support in the *VTEC* case is at least thin, if not inexistent.

The bottom panel of Table 1 shows descriptive statistics for the three outcome variables considered. These are the earnings in the fourth year after being awarded with a voucher (in the following: earnings), months employed during the fourth year after being awarded with a voucher (in the following: months employed), and being employed four years after being awarded with a voucher (in the following: employment). This selection is made to consider outcome variables that semi-continuous, discrete, and binary, respectively. Columns (5), (6), and (7) of Tables A.1 and A.2 in Online Appendix A show the relation of these outcomes to the covariates included in the propensity score estimation.

5 Design of simulation study

5.1 Empirical Monte Carlo study

The analysis of the common support issues are based on an Empirical Monte Carlo Study (EMCS), as suggested by Huber, Lechner, and Wunsch (2013) and Lechner and Wunsch (2013). It combines the advantages of a 'classical' Monte Carlo study with the advantages of using real data. The basic idea is to use a large data set, which is similar for to data sets in the respective field of study. In order to obtain appropriate random samples for the Monte Carlo study, the first step consists in estimating the propensity score in the full data. This estimated model will reflect the 'true' selectivity. Subsequently, random samples from the non-treated subpopulation are drawn. A (pseudo) treatment is assigned according to the 'true' selection model. Since there are only non-treated individuals in the samples used in the simulations, the true effect of the assigned treatment is known to be zero. The key advantages of this procedure are that the selectivity in the samples reflects the selectivity in the larger population, and that there is no need for stipulating a model on how the outcomes depend on treatment and covariates. Adding some more structure allows varying various components of the data generating process that are deemed relevant in the particular analysis. The details of the simulation process are described in the following sections.

5.2 Baseline specifications

After being used for estimation of the 'true' propensity score, all treated observations are removed from the sample. Then, observations are randomly drawn with replacement from the non-treated subpopulation to form the Monte Carlo samples. The sample sizes of *N = 500, 2,000,* and *8,000* have been chosen to analyse differential behaviour in smaller and larger samples. Using quadrupling sampling sizes is convenient because the standard error of $\sqrt{N}$-

consistent estimators is about halved if the sample size is quadrupled. Thus, it is easy to detect deviations from $\sqrt{N}$ –convergence. Although larger sample sizes are observed in applied work as well, computing time is a limitation.

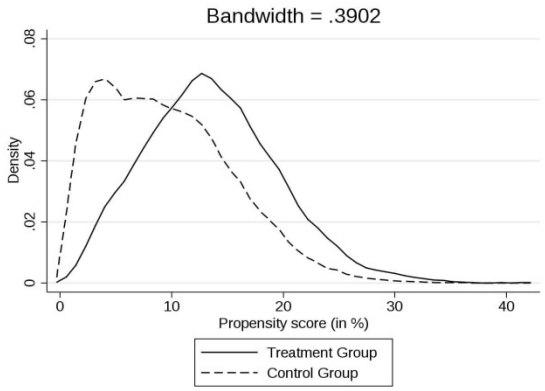
Next, a binary pseudo treatment d_i is assigned to some individuals in the simulation sample (all of which are non-treated). The assignment rule of the pseudo treatment is based on the population propensity score, with $X_i\hat{\beta}$ being the linear index of the estimated probit model in the entire sample (see column 4 in Tables A.1 and A.2 in Online Appendix A). The actual selection for the pseudo treatment d_i is determined by the following indicator function,

$$d_i = 1\{X_i\hat{\beta} + a + u_i > 0\}, \qquad u_i \sim N(0,1).$$

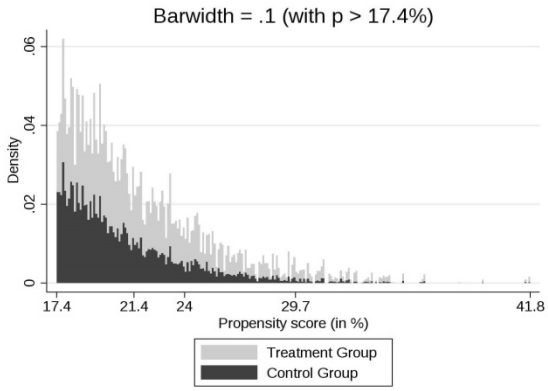
The parameter a is introduced to manipulate the share of pseudo treated. In our simulations, we investigate shares of pseudo treatment of 10%, 50%, and 90%, respectively. This share may be important for the performance of the estimators as it directly influences the thickness of the support (data generating process, DGPs, with high treatment shares are more likely to lead to support issues). Note that this simulation process implies that asymptotically there is common support (because u_i varies over the entire real line).

Figures 3 and 4 show the distribution of the propensity scores as well as histogram for their distribution above the third quartile (to focus on the region of interest for support problems when estimating γ) of the population propensity scores from the various simulated DGPs for very large samples. However, especially in the simulated specification with pseudo treatment shares of 90% even in very large samples support appears to be (very) thin in the upper part of the propensity score distribution, in particularly for *VTEC*.

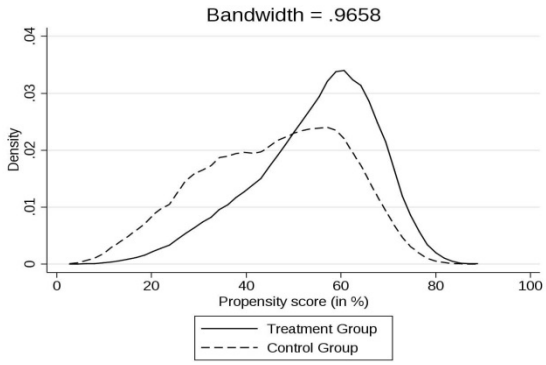
Figure 3: Kernel and histogram density estimates of the propensity score in the simulation samples without reduction of support for the treatment 'award of VMSW'



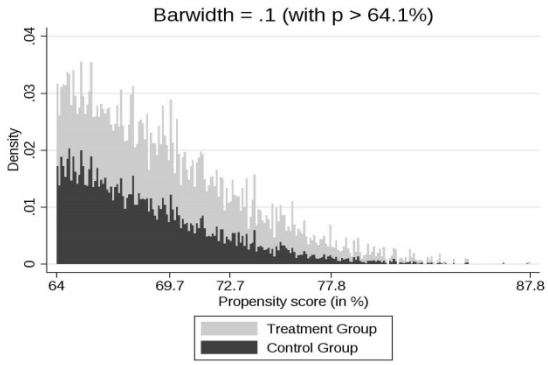
(a) Density, 10% treatment share



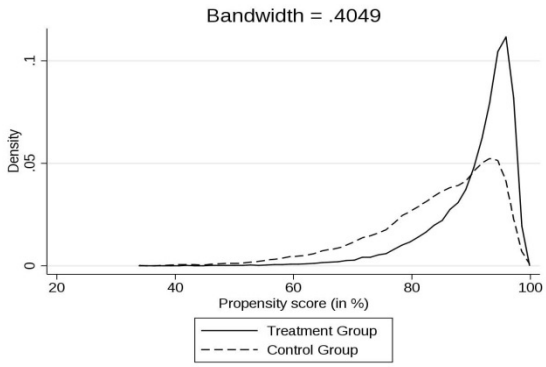
(b) Histogram, 10% treatment share



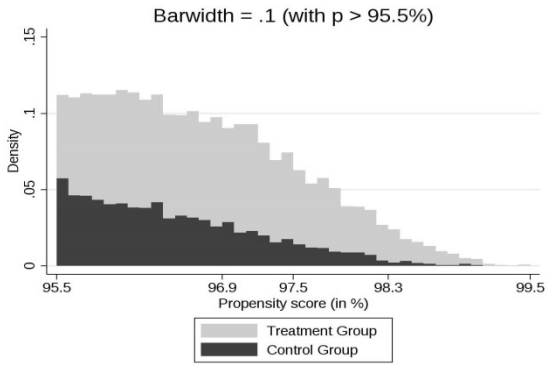
(c) Density, 50% treatment share



(d) Histogram, 50% treatment share



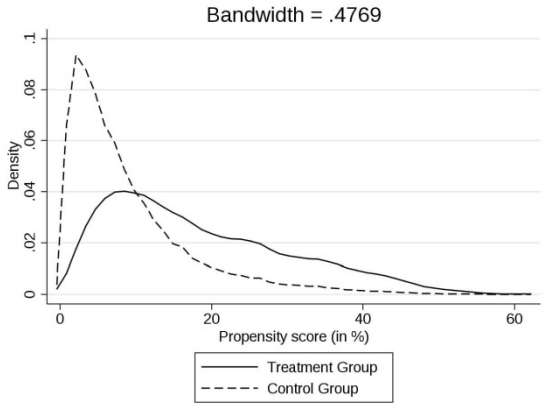
(e) Density, 90% treatment share



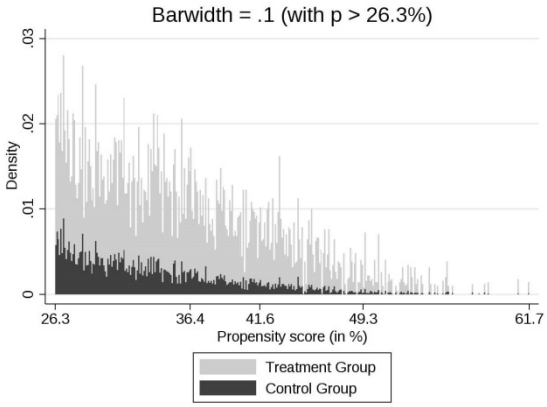
(f) Histogram, 90% treatment share

Note: All graphs are calculated using 500,000 simulated observations. We report results using the population propensity score (see description Section 5.5). See also note below Figure 1.

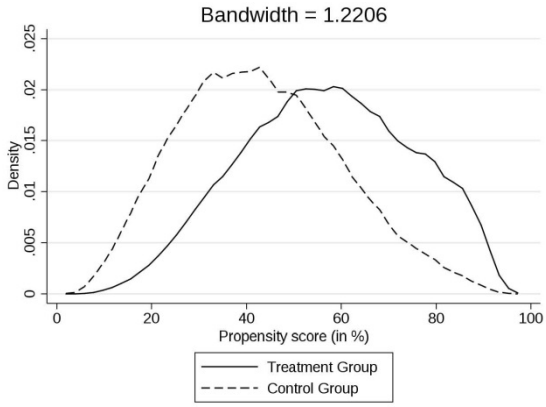
Figure 4: Kernel and histogram density estimates of the propensity score in the simulation samples without reduction of support for the treatment 'award of VTEC'



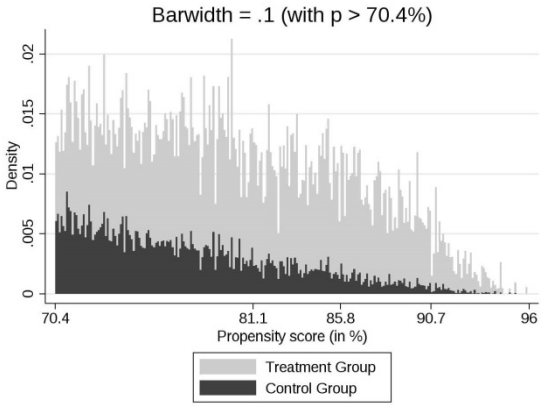
(a) Density, 10% treatment share



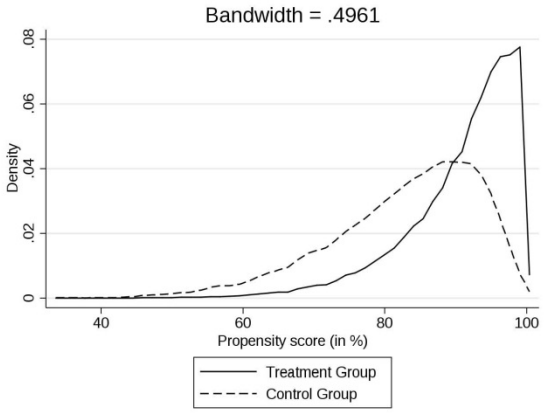
(b) Histogram, 10% treatment share



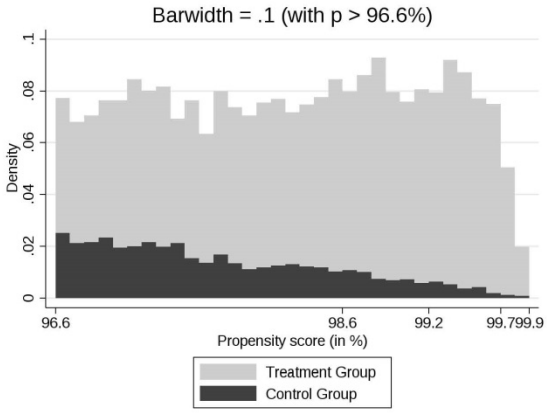
(c) Density, 50% treatment share



(d) Histogram, 50% treatment share



(e) Density, 90% treatment share



(f) Histogram, 90% treatment share

Note: See note below Figure 3.

5.3 Effect heterogeneity

The DGPs described above imply effect homogeneity, as, by definition, the pseudo-treatment has a zero effect for all individuals. This feature, which is unrealistic in many applications, favours particular procedures dealing with support problems. For example, dropping observations does not introduce any systematic large sample bias in such a setting, because Assumption D holds by construction. Therefore, we add two different types of effect heterogeneity. In the first case, effect heterogeneity depends directly on the propensity score. In the second case, effect heterogeneity depends on the propensity score as well on the binary 'support variable' *S*.

Practically, effect heterogeneity is implemented such that the outcome variables of the pseudo-treated observations are modified, because all observed outcomes stem from the non-treated group and should thus well resemble the properties of non-treated units. The effect heterogeneity observed in the population guides the implementation of the DGPs: To uncover it, first, the conditional expectations of the outcomes in the population, $E[Y|p(X) = p(x), D = 1]$, are estimated using a Nadaraya-Watson estimator. Using the same estimation procedure in the two strata defined by the values of *S*, estimates for $E[Y|p(X) = p(x), S = 1, D = 1]$ and $E[Y|p(X) = p(x), S = 0, D = 1]$ are obtained as well (see Figures B.1 and B.2 in Online Appendix B for details). In both cases, considerable effect heterogeneity is found for all three outcomes, especially in the relevant regions with high values of the propensity score. In a second step, for the treated observations residuals are computed as $\xi_i = Y_i - E[Y|p(X) = p(x_i), D = 1]$ and $\epsilon_i = Y_i - E[Y|p(X) = p(x_i), S = s_i, D = 1]$. Third, the restricted pseudo treatment outcome $\tilde{Y}_i$ is generated by adding residuals randomly drawn from these residual distributions to the estimated conditional expectation of the pseudo-treated,

$$\tilde{Y}_i = E[Y|p(X) = p(x_i), D = 1] + \xi_j, \text{ or } \tilde{Y}_i = E[Y|p(X) = p(x_i), S = s_i, D = 1] + \epsilon_j.$$

Note that although individual treatment effects become heterogeneous, we adjust the outcomes such that the average effects in the sample are unaffected and the distribution of Y_i is respected. Please find a detailed description of the distributional adjustments of $\tilde{Y}_i$ in Online Appendix B.

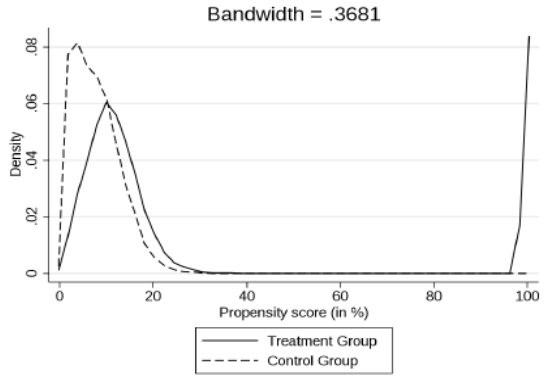
5.4 Additional support restrictions in the DGP

Having introduced the model specifications and effect heterogeneity, the next step is to introduce support problems that are even more biting than those of the baseline specification are. Therefore, there will be additional data generating processes in which the assignment process of the pseudo treatment is made further dependent on *S*. Since *S* is already a good predictor of receiving a voucher for technicians and is related to the three outcome variables (see Tables A.1 and A.2 in Online Appendix A) and their heterogeneity, *S*-based support problems have the potential to 'hurt' estimators.

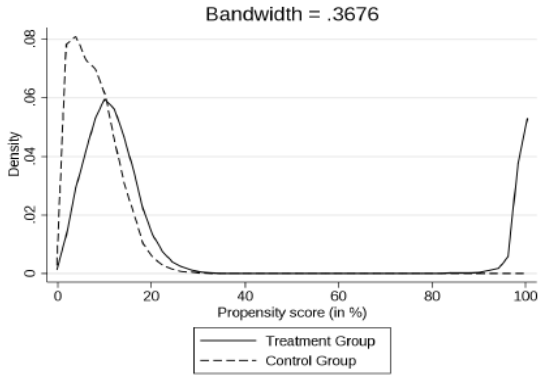
The procedure additionally removing support has several steps: First, the share $P(S = 1)$ in the simulation sample is changed as $P(S = 1)' = \delta P(D = 1)$.[5] $N_S = N \cdot P(S = 1)'$ non-treated observations are randomly drawn from the subpopulation with $S = 1$. Another $(N - N_S)$ non-treated individuals are drawn from the subpopulation with $S = 0$, such that the sample size N is maintained. Here, we choose $\delta = 0.21$ for VMSW and $\delta = 0.66$ for VTEC reflecting the corresponding fractions in the population.

[5] One way to introduce common support problems would be to set the pseudo-treatment equal to one for each observation where S_i equals one. This would cause problems when $\mathrm{P}(D = 1) < \mathrm{P}(S = 1)$, because it would be impossible to maintain the fraction of pseudo treated (10%, 50%, and 90%) and to introduce the common support problem at the same time. Further, controlling the number of observations that are affected by support problems could be an important tuning parameter. Therefore, $P(S = 1)$ is restricted in the first place. $P(S = 1)'$ is defined relative to the probability to be pseudo treated.

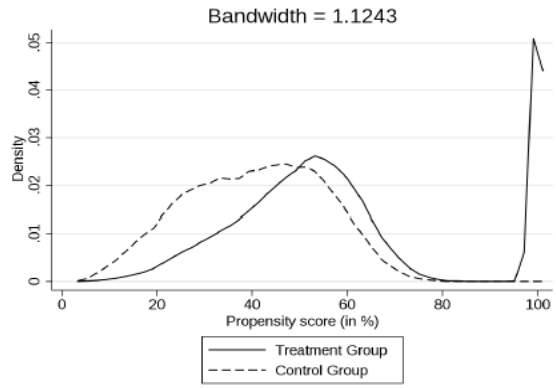
Figure 5: Kernel density estimates of the propensity score in the simulation samples with reduction of support for the treatment 'award of VMSW'



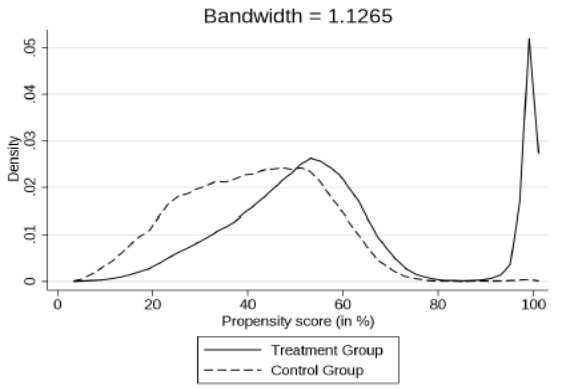
(a) 10% treatment share, Φ = 0



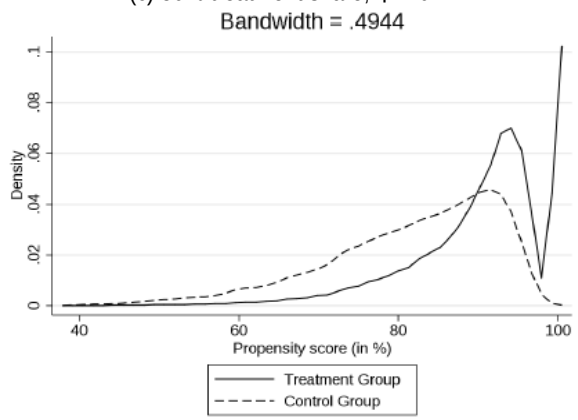
(b) 10% treatment share, Φ = 0.01



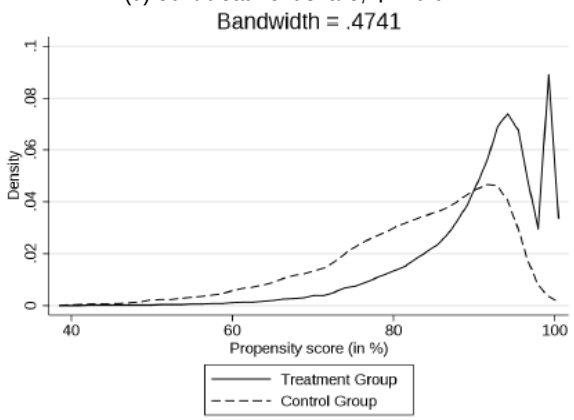
(c) 50% treatment share, Φ = 0



(d) 50% treatment share, Φ = 0.01



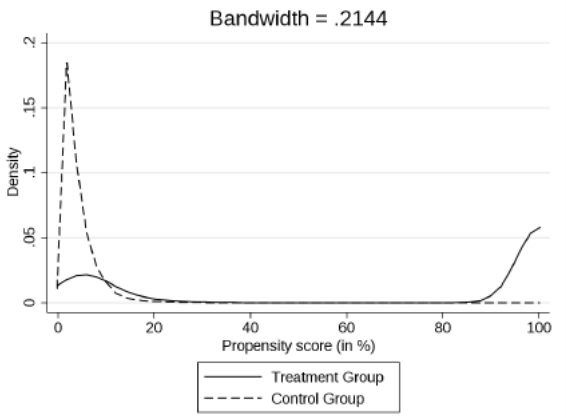
(e) 90% treatment share, Φ = 0



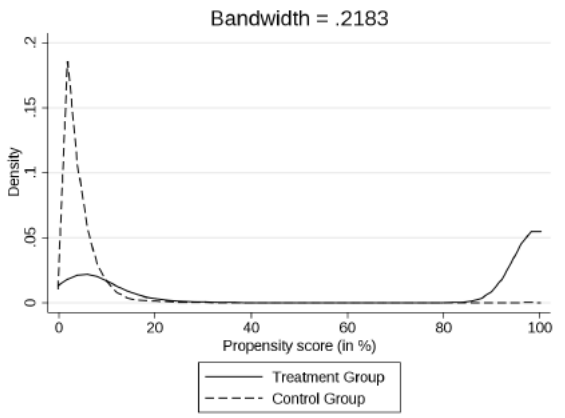
(f) 90% treatment share, Φ = 0.01

Note: We report densities of the propensity score for DGPs with restricted support. Bandwidth selections for the Gaussian kernels are based Silverman's rule of thumb. All graphs are calculated using 500,000 simulated observations. Results are based on the population propensity score.

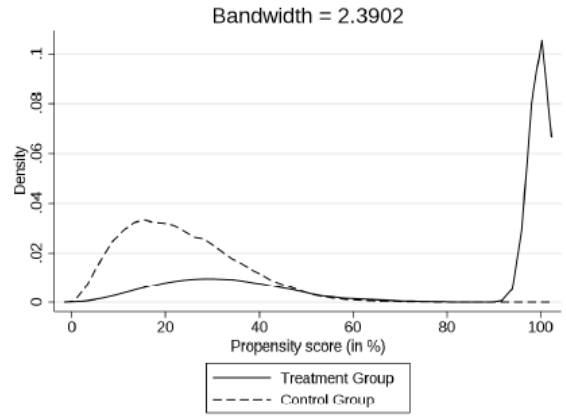
Figure 6: Kernel density estimates of the propensity score in the simulation samples with reduction of support for the treatment 'award of VTEC'



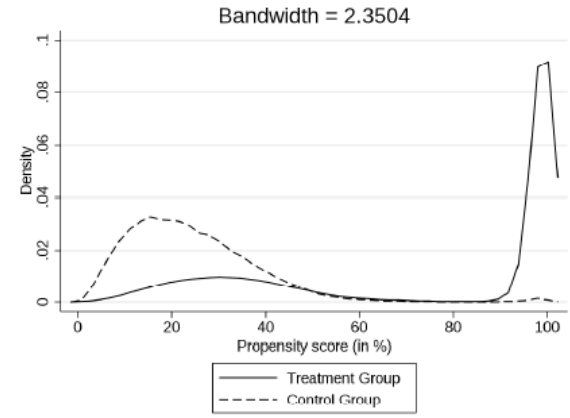
(a) 10% treatment share, Φ = 0



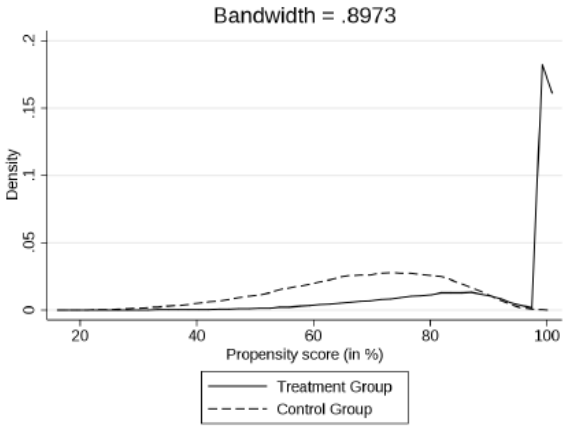
(b) 10% treatment share, Φ = 0.01



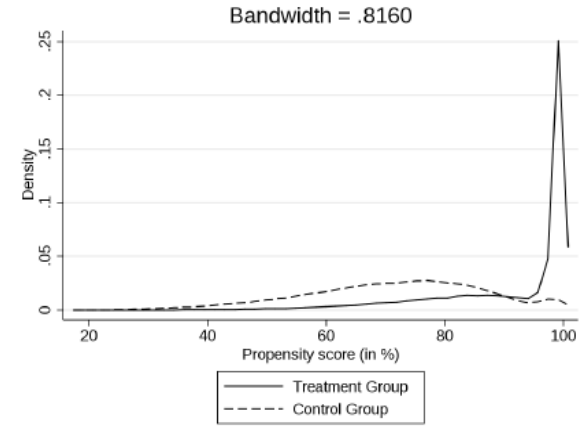
(c) 50% treatment share, Φ = 0



(d) 50% treatment share, Φ = 0.01



(e) 90% treatment share, Φ = 0



(f) 90% treatment share, Φ = 0.01

Note: See note below Figure 5.

Next, the DGPs are restricted such that $P(D=1|S=1)' = 1-\phi$, where $\phi \geq 0$ is some small number. The pseudo treatment dummy d_i is generated according to,

$$d_i = (1-s_i)\cdot 1\{x_i\hat{\beta} + a' + r_i > 0\} + s_i \cdot 1\{x_i\hat{\beta} + b + r_i > 0\},$$

with $r_i \sim N(0,1)$. The desired $P(D=1|S=1)'$ is achieved by adjusting the parameter b. The treatment shares from the baseline specifications (10%, 50%, 90%) are maintained by adjusting the parameter a'. Choosing $\phi = 0$ incorporates support problems in the population. Choosing $\phi > 0$ being a small positive number might not introduce support problems in the population, but can lead to serious support problems in finite samples. Figures 5 and 6 show the densities of the propensity score by treatment status for $\phi = 0$ and $\phi = 0.01$. The reason to choose these two values of ϕ is motivated by the expectation that they lead to interesting types of support violations. The densities of the treated subpopulation show fat tails close to $p(x) = 1$ in all specifications.

5.5 Practical implementation of estimators

5.5.1 Specifications and tuning parameters

Below, results for parametric, IPW, and matching estimators are presented. OLS regressions are used for continuous and discrete non-binary outcome variables, while probit estimators are used for binary outcome variables. In order to increase their flexibility, separate models are estimated in the treated and the non-treated subsample. *E(Y|X=x, D=0)* is estimated parametrically controlling for the full set of confounding variables *X* in the subsample of non-treated (see Tables A.1 and A.2 in Online Appendix A). This estimated model is then used to compute predicted values of *Y* at the values of *X* observed in the treated subsample. Finally, these predictions are averaged and subtracted from the mean of the outcomes in the treated subsample to obtain an estimate for γ.

IPW estimators can at least be traced back to Horvitz and Thompson (1952). They have the advantage of being easy to implement and that they may be asymptotically efficient (i.e., Hirano, Imbens, and Ridder, 2003). We use the implementation advocated by Busso, Di-Nardo, and McCrary (2014a, 'specification IPW2'), which ensures that the weights add-up to one:

$$\hat{\gamma} = \frac{1}{\sum_{i=1}^{N} d_i} \sum_{i=1}^{N} d_i y_i - \sum_{i=1}^{N} \widehat{\omega}_i y_i, \qquad \widehat{\omega}_i = (1 - d_i) \frac{\frac{\hat{p}(x_i)}{1 - \hat{p}(x_i)}}{\sum_{i=1}^{N} \frac{(1 - d_i)\hat{p}(x_i)}{1 - \hat{p}(x_i)}}. \tag{1}$$

The estimated propensity score $\hat{p}(x)$ is obtained from a probit model described below.

Finally, we investigate the radius-matching estimator suggested in Lechner, Miquel, and Wunsch (2011). This version combines the features of radius matching with regression adjustments. Moreover, this estimator allows matching on additional control variables directly, e.g. using the so-called Mahalanobis distance. A detailed description of this estimator can be found in Huber, Lechner, and Steinmayr (2012). According to the study of Huber, Lechner, and Wunsch (2013), it has good finite sample properties. Following Lechner and Wunsch (2013), we match on the variables gender and pre-treatment earnings in the Mahalanobis distance. Following Huber, Lechner, and Steinmayr (2012) we define the radius to be three times the distance between the propensity scores of the largest one-to-one match.

5.5.2 Specification of covariates

All estimators used depend on some parametric parts. For the fully parametric estimators we specify the functional form of the dependence of the conditional expectation of the observed outcome with treatment and confounding variables. For IPW and radius matching, the propensity score model is specified. For the latter, we could of course base the estimation on the true specification of the selection model (note that it is only 'true' for the case without

explicit support restriction). However, having no or only thin support introduces another problem when estimating the sample propensity score. Some variables could become (almost) perfect predictors, like the support variables S_{i1}, S_{i2}, and S_{i3}, for example. In practice, the exact propensity score is unknown. Therefore, it is most likely that researchers would not include the support variables in the specification of the propensity score in real applications. Not considering the support variables could introduce an additional bias due to misspecification. In order to disentangle the bias from support problems and from model misspecifications, we follow three different approaches. First, the sample propensity score based on the correct model specification is used. This approach is only used in DGPs without (almost) perfect predictors, i.e. DGPs without support restrictions. Second, a sample propensity score is used omitting the support variables (i.e. the three interaction terms are omitted, but the main effects they are made of are kept). Third, we match on the population propensity score, which was estimated in the full sample.[6]

For the parametric specifications, we do not use the propensity score, but control for X_i directly. These estimators are based on two model specifications. There is a correctly specified model controlling for all confounding variables. Furthermore, there is a misspecified model, which does not control for the support variables.

5.6 Number of replications

The number of replications is a trade-off between computation time and simulation noise. The simulation noise decreases with the number of replications and increases with the variance of the simulated objects. Because the latter is reduced by half when the sample size is doubled,

[6] This specification is not used for IPW with (almost) perfect predictors, because the weights $\widehat{\omega}_i$ from (1) would go to infinity.

the numbers of replications are chosen in proportion to the sample size, i.e. such that $N \times R = 8{,}000{,}000$. Consequently, there are $R = 16{,}000$ replications in the small sample, $R = 4{,}000$ replications in the medium-sized sample, and $R = 1{,}000$ replications in the large sample.

6 Results

6.1 General remarks

There are results for 252 different DGPs, 44 procedures to deal with support problems, three outcomes, three estimators, and up to three different model specifications.[7] It is not possible to report (and understand!) all these results without reducing their dimensionality. Therefore, linear regressions are used as summary measures. In these regressions, the dependent variables consist of measures of the quality of the estimators, like the root mean squared error (RMSE), the absolute bias, and the standard errors. The independent variables reflect different features of the DGPs (treatment shares, types of effect heterogeneity, sample size, and type of voucher), of the model specifications, of the estimators, and of the various rules to tackle the support problem (see Table 2 for the list of all procedures used). Controls for model specifications, treatment shares, and types of effect heterogeneity are interacted with each other (but not with the dummies for the different procedures to handle support problems). Further, separate regressions are reported for different types of vouchers, estimators, sample sizes, and types of support reductions as, a priori, considerable heterogeneity is expected with respect to those features.

[7] The DGPs vary by sample size, treatment share, type of treatment (*VMSW* or *VTEC*), type of effect heterogeneity, and type of support restrictions. Overall, there are 88,704 different results.

Table 2: Procedures handling support problems

Procedure	Description	Rule	Assumption	References
	Drop treated based fixed value of propensity score			
A	Drop no observations			
WA	Drop non-treated if weights high ($\widehat{\omega} \geq 0.04$)		D	Imbens (2004)
B1	Drop if .1 < p(x) < .9			CHIM (2009)
B2	Drop if p(x) < .9			
WBx	Bx & WA			
	Drop treated based on density of propensity score			
C1	Drop treated if f(p(x)) < q_2	1	D	HIST (1998), HIT (1997), Smith, Todd (2005)
C2	Drop treated if f(p(x)) < q_{10}	1	D	
WCx	Cx & WA	1	D	
	Drop treated based on lack of non-treated neighbours in terms of radius of propensity score			
D1	Drop treated if no close match in 0.01 radius (u=0.01)	2	D	Grzybowski et al. (2003), Vincent et al. (2002)
D2	Drop treated if no close match in 0.1 radius (u=0.1)	2	D	
WDx	Dx & WA	2	D	
	Drop treated based on upper limits of distribution of propensity score among non-treated			
E1	Drop treated above highest non-treated p-score	3	D	Dehejia, Wahba (1999)
E2	Drop treated above 99% highest non-treated p-score	3	D	
E3	Drop treated above 95% highest non-treated p-score	3	D	HLW (2013)
WEx	Ex & WA	3	D	
	Procedures that extrapolate into lack of support region			
F1	E1 & estimate $Y_{1\|N}$ - $Y_{0\|N}$ at highest non-treated p-score	3	E	
F2	E2 & estimate $Y_{1\|N}$ - $Y_{0\|N}$ at 99% highest non-treated p-score	3	E	
F3	E3 & estimate $Y_{1\|N}$ - $Y_{0\|N}$ at 95%highest non-treated p-score	3	E	
WFx	Fx & WA	3	E	
G1	E1 + lin. approx. of $Y_{1\|N}$ - $Y_{0\|N}$ above highest non-treated p-score	3	F	
G2	E2 + lin. approx. of $Y_{1\|N}$ - $Y_{0\|N}$ above 99% highest non-treated p-s.	3	F	
G3	E3 + lin. approx. of $Y_{1\|N}$ - $Y_{0\|N}$ above 95%highest non-treated p-s.	3	F	
WGx	Gx & WA	3	F	
H1	Estimate Y0|N at highest non-treated p-score	3	B	
H2	Estimate Y0|N at 99% highest non-treated p-score	3	B	
H3	Estimate Y0|N at 95% highest non-treated p-score	3	B	
WHx	Hx & WA	3	B	
I1	Lin. approx. of $Y_{0\|N}$ above the highest non-treated p-score	3	C	
I2	Lin. approx. of $Y_{0\|N}$ above the 99% highest non-treated p-score	3	C	
I3	Lin. approx. of $Y_{0\|N}$ above the 95%highest non-treated p-score	3	C	
WIx	Ix & WA	3	C	

Note: WA indicates that non-treated observations with $\widehat{\omega} \geq 0.04$ are dropped (see Section 5.5). This procedure is used in combination with other procedures (two-step procedure). CHIM: Crump, Hotz, Imbens, Mitnik; HIST: Heckman, Ichimura, Smith, Todd; HIT: Heckman, Ichimura, Todd; HLW: Huber, Lechner, Wunsch.

In the regressions, the three different outcome variables (earnings, months employed, employment) are pooled. Since they are measured on different scales, they are normalized by their standard deviation in the baseline specification (see Online Appendix C for details). Therefore, for a given estimation procedure, the regression coefficients of the binary control variables indicate by how many standard deviations the performance measure changes if this dummy turns on in comparison to the omitted category.

The next section reports regression results for the case without additional support restrictions, followed by the case with restricted support, and selected detailed results.

6.2 Regression results for DGPs not restricting common support

The DGPs for which the treatment is based on the *award of VMSW* are expected not to be subject to severe support problems, because of their low degree of selectivity (see Table 1). Although, the *award of VTEC* is more selective, even in this case, the population propensity scores exceed the level 0.9 only when the treatment share is 90%. Even then, it remains below one (see Figures 3 and 4) so that the DGPs considered in this section show no asymptotic support problems. Nevertheless, issues of thin support may still be relevant.

Tables 3 and 4 report the results from the regressions of the normalized RMSE on different procedures to handle support problems (the reference category is ignoring the problem, i.e. Procedure A). We report results for selected procedures only. The results for the complete set of common support procedures (incl. the standard errors of the estimated coefficients and R^2s), as well as for other performance measures (normalized absolute bias, standard deviation) can be found in Online Appendix D (Tables D.1 to D.6).

Table 3: Effect of support-adjustment procedures on normalized RMSE in DGPs without support restrictions for the treatment 'award of VMSW'

Sample size		500			2000			8000	
Support	Paramet.	IPW	Match.	Paramet.	IPW	Match.	Paramet.	IPW	Match.
procedures	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
No adjustment (A: reference)									
WA	.002	-.012***	.003**	.005	-.017**	.004	0	0	0
Drop treated based on fixed value of propensity score									
B2	-.007*	-.0148***	-.007***	.220***	.187***	.188***	.701***	.640***	.611***
WB2	-.007*	-.0148***	-.007***	.220***	.187***	.188***	.701***	.640***	.611***
Drop treated based on density of propensity score									
C2	.017***	.019***	.011***	.001	0	.002	0	0	0
WC2	.018***	.007	.014***	.007	-.017**	.005	0	0	0
Drop treated based on lack of non-treated neighbours in terms of radius of propensity score									
D1	-.016***	.007***	-.025***	-.001	0	-.001	0	0	-.001
WD1	-.016***	-.006	-.025***	.001	-.017**	0	0	0	-.001
D2	-.001	0	-.002	0	0	0	0	0	0
WD2	0	-.012***	.001	.005	-.017**	.003	0	0	0
Drop treated based on upper limits of distribution of propensity score among non-treated									
E1	-.013***	.003	-.019***	-.013	.001	-.013**	-.003	-.001	-.004
WE1	-.013***	-.009**	-. 020***	-.015	-.018**	-.016**	-.003	-.001	-.004
E2	-.010**	-.025***	-.018***	-.006	-.021***	-.004	.028	-.001	.019
WE2	-.009*	-.025***	-.018***	-.008	-.026***	-.005	.028	-.001	.019
E3	.037***	-.004	.017***	.070***	.027**	.058***	.242***	.182***	.186***
WE3	.038***	-.004	.018***	.072***	.028**	.059***	.242***	.182***	.186***
Procedures that extrapolate into lack of support region									
F2	.093***	.082***	.033***	.068***	.069***	.028***	.055**	.055***	.045***
WF2	.100***	.086***	.036***	.078***	.072***	.033***	.055**	.055***	.045***
G2	.042***	.014***	-.002	.032***	.014*	.018***	.020	.004	.013
WG2	.047***	.015***	-.001	.036***	.013*	.019***	.020	.004	.013
H2	.114***	.084***	.037***	.071***	.061***	.026***	.065***	.050***	.041***
WH2	.124***	.090***	.041***	.082***	.065***	.031***	.065***	.050***	.041***
I2	.037***	.005	-.009***	.022**	-.003	.001	.022	-.003	.010
WI2	.042***	.007	-.007**	.026***	-.004	.001	.022	-.003	.010
# of obs.	792	1,188	1,188	2,376	3,564	3,564	2,376	3,564	3,564

Note: OLS. The dependent variable is the normalized RMSE. The covariates contain a full set of dummy variables for the different procedures to handle support problems. The reference category is to drop no observations (Procedure A, a full description of the different procedures is given in Table 2). Further control variables are the tuning parameters of the different DGPs in a fully interacted way (see description in main text). Only selected coefficients for the different procedures are reported in this table. A complete set of results, including standard errors and R^2s are shown in Online Appendix D. ***, **,* indicate significance at the 1-, 5-, and 10-percent levels, respectively (based on robust standard errors). There are fewer observations for the small *N*, because only the case of 50% treatment shares is considered. The same holds for the parametric estimations, because only 2 specifications are considered (instead of 3, see Section 5.5.2).

Table 4: Effect of support-adjustment procedures on normalized RMSE in DGPs without support reduction for treatment 'award of VTEC'

Sample size		500			2000			8000	
Support	Parametr.	IPW	Match.	Parametr.	IPW	Match.	Parametr.	IPW	Match.
procedures	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
No adjustment (A: reference)									
WA	.014**	-.097***	.037***	.018*	-.092***	.018***	.001	-.037**	.003
Drop treated based on fixed value of propensity score									
B2	-.038***	-.112***	-.069***	.163***	.036	.118***	.576***	.374***	.461***
WB2	-.038***	-.112***	-.069***	.163***	.036	.118***	.576***	.374***	.461***
Drop treated based on density of propensity score									
C2	-.007	.0145	-.036***	.002	0	.003	0	0	0
WC2	.006	-.089***	-.006	.020**	-.091***	.021***	.001	-.037**	.003
Drop treated based on lack of non-treated neighbours in terms of radius of propensity score									
D1	-.047***	.017	-.113***	-.011	.001	-.020***	-.001	0	-.004
WD1	-.045***	-.095***	-.112***	-.023**	-.102***	-.029***	0	-.037**	-.001
D2	-.002	0	-.004	-.001	0	0	0	0	0
WD2	.006	-.099***	.021***	.017*	-.092***	.018***	.001	-.037**	.003
Drop treated based on upper limits of distribution of propensity score among non-treated									
E1	-.039***	.004	-.082***	-.039***	0	-.048***	-.016	-.003	-.023
WE1	-.046***	-.101***	-.096***	-.039***	-.106***	-.055***	-.018	-.043***	-.026
E2	-.039***	-.123***	-.098***	-.019	-.097***	-.050***	.073**	-.040**	.020
WE2	-.037***	-.132***	-.098***	-.014	-.121***	-.050***	.075**	-.040**	.021
E3	.032	-.106***	-.059***	.132***	-.013	.053**	.489***	.312***	.363***
WE3	.040	-.101***	-.054***	.140***	-.007	.059***	.490***	.313***	.364***
Procedures that extrapolate into lack of support region									
F2	.160***	.091***	.030***	.143***	.091***	.063***	.133***	.072***	.071***
WF2	.209***	.110***	.052***	.183***	.093***	.090***	.136***	.073***	.073***
G2	.091***	-.010	-.026***	.078***	-.012	.008	.085***	-.012	.022
WG2	.125***	.002	-.014*	.100***	-.019	.021***	.086***	-.012	.023
H2	.160***	.073***	.019**	.144***	.071***	.055***	.176***	.082***	.092***
WH2	.215***	.095***	.043***	.184***	.075***	.082***	.180***	.084***	.094***
I2	.048***	-.059***	-.066***	.044***	-.055***	-.028***	.067**	-.052***	-.009
WI2	.067***	-.057***	-.061***	.061***	-.065***	-.019***	.068**	-.051***	-.008
# of obs.	792	1,188	1,188	2,376	3,564	3,564	2,376	3,564	3,564

Note: See note below Table 3.

Both tables show for most procedures that if there are reductions of the normalized RMSE (in comparison to the omitted category of dropping no observations), they most likely occur in the smallest samples. Interestingly, it appears that almost all procedures either have no effect

on the normalized absolute bias of the estimators (see Tables D.3 and D.5 in Online Appendix D), or increase it (somewhat). However, such increases are usually (over-) compensated by a reduction in the variability of the estimators (see Tables D.4 and D.6 in Online Appendix D). Accordingly, the performance of the estimators can be improved, even in DGPs where the support conditions are not violated in the population. This is in line with the arguments of Busso, DiNardo, and McCrary (2014a), Crump, Hotz, Imbens, and Mitnik (2009), and Kahn and Tamer (2010) suggesting that thin support issues lead to a loss of precision. Finally, note that for the treatment *award of VTEC*, with a strong selection into treatment, the potential performance improvements are considerably larger than for the treatment *award of VMSW*. Next, the performance of the single procedures is discussed in more detail.

Dropping observations with high importance (*WA*) improves only the performance of IPW estimators. However, when this approach is combined with other procedures, then these (joint) procedures (beginning with *W*) have the potential to improve the performance of the single procedure with which it is combined. This is consistent with the findings of Huber, Lechner, and Wunsch (2013).

The procedure of Crump, Hotz, Imbens, and Mitnik (2009) is applied in two different specifications. First, we drop all observations with a propensity score below 0.1 and above 0.9 (*B1*; see Online Appendix D). Second, only observations with a propensity score above 0.9 (*B2*) are dropped. Although *B1* and *B2* seem to work for the small samples, for the larger sample they lead to biases large enough to dominate the RMSE.

The results for the procedure dropping treated observations with a low marginal density (*C*) are not encouraging either. While there appears to be some possibility of improvements on the performance of estimators if selectivity is strong enough (Table 4), in the case of weak selectivity (Table 3) the small sample performance deteriorates.

Procedure *D* drops treated observations in which the distance of the closest one-to-one match is below u (Grzybowski et al., 2003, Vincent et al., 2002), while Procedure *E* drops treated observations with a propensity score above a cut-off value $\bar{p}$ (Dehejia and Wahba, 1999). Generally, *D* and *E* improve the performances of the estimators equally well. Unlike most other procedures, only in very rare cases do these procedures hurt the performances of the estimators in terms of normalized RMSE. The largest improvement (originating from the standard deviations) can most often be obtained from *E*. This procedure works better than *D*, particularly in larger samples, when $\bar{p}$ is either specified as being the maximum or the value of the 99%-quantile of the propensity score in the non-treated sample. In the smallest sample, Procedure *D* appears to have slightly better properties than *E*. In most cases, *D* works best with $u = 0.01$ for parametric and matching estimators. Using D with $u = 0.01$ or $u = 0.1$ performs about equally well for IPW estimators. Both procedures improve somewhat when combined with *W*.

All procedures that aim to estimate the conditional treatment effect $\gamma_N = Y_{1|N} - Y_{0|N}$ or $Y_{0|N}$ off support (*F*, *G*, *H*, *I*) do not perform well. In most cases, there are no performance improvements. However, if any improvements show up, than they are smaller than for the other procedures discussed above.

6.3 Regression results for DGPs with reductions of common support

Next, DGPs are considered for which the support with respect to *S* is restricted in a way that causes (serious) support problems. To do so, $P(D = 1|S = 1)' = 1 - \phi$ is changed. If $\phi = 0$, then even asymptotically there is no common support. When $0 < \phi < 1$, there is common support in the population, but it may be thin when ϕ is small. Remember, the share of

observations with $S = 1$ is much larger for individuals awarded with *VTEC* than with *VMSW*.[8] Accordingly, the treatment *award of VTEC* is considerably more affected by the support reductions than the treatment *award of VMSW*.

Figure 7 shows the performance of parametric, IPW, and matching estimators when no observations are dropped (*A*). In this figure, $P(D = 1|S = 1)'$ is gradually increased based on the grid $\{0.9, 0.91, \dots, 1\}$. To be precise, the figures show the coefficients of the indicator variables for the different points of the grid based on the same regressions that have been reported before but pooled for the two treatments. We find an increase in the average normalized RMSE between 0.1 and 1 standard deviation if $P(D = 1|S = 1)' = 0.9$ in comparison to the specifications with no support restrictions. However, if $P(D = 1|S = 1)' = 1$ the average normalized RMSE can be increased by up to eight standard deviations. Support reductions have the strongest impact on the matching estimator. Interestingly, matching estimators appear to have on average lower normalized absolute biases than parametric and IPW estimators. This is in line with the findings of Busso, DiNardo, and McCrary (2014b), who report that the bias of matching estimators is less affected by overlap problems than the bias of IPW estimators. The normalized absolute bias of matching estimators exceed those of parametric and IPW estimators only under very strong support reductions. On the other hand, the average normalized standard deviation of matching estimators becomes very large under strong support restrictions. This is the main reason for the bad performance of matching in the specifications with strong support reductions. For parametric and IPW estimators the performance in terms of average normalized absolute bias and average normalized standard deviation is more balanced in this situation. However, the average normalized RMSE, average

[8] From Section 3.4 we get $P(S = 1)' = \delta P(D = 1)$ with $\delta = 0.21$ for VMSW and $\delta = 0.66$ for VTEC.

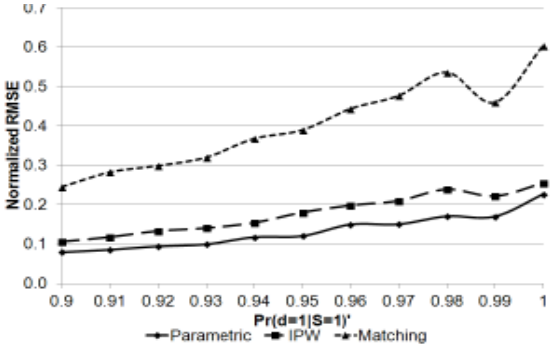
normalized absolute bias, and average normalized standard deviation of these estimators may increase by up to one standard deviation when support reductions are substantial. The performance of these estimators is almost linearly decreasing when the support is reduced.

Tables 5 and 6 report the results from regression for the normalized RMSE for the case of such support reductions. For the sake of computation time, only specifications with $\phi = 0$ and $\phi = 0.01$ are included, because these two scenarios lead to the most serious support problems. Thus, the specification of the OLS regressions is similar to the one used in the previous section. The only difference is that an additional dummy variable for the case $P(D = 1|S = 1)' = 0.99$ is included. As before, Tables 5 and 6 report only a subset of the results. The complete set of results can be found in Online Appendix E (Tables E.1 to E.6).

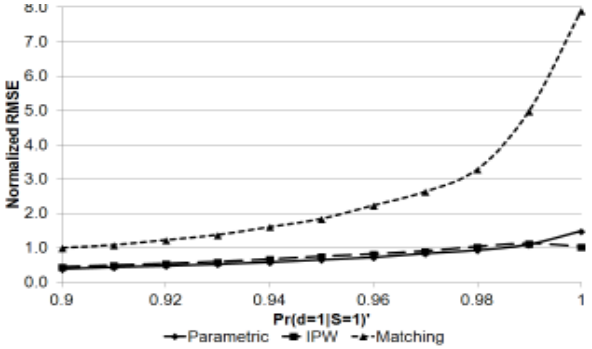
Not surprisingly, the performance improvements for the different estimators are much more pronounced than for the DGPs without support restrictions. Furthermore, parametric, IPW, and matching estimators have very different properties in the DGPs with restricted support (see Tables 5 and 6 as well as Tables E.1 and E.2 in Online Appendix E). The biggest improvements are observed for matching estimators for which all procedures handling support problems work well. However, note that various procedures improve the performance of parametric and IPW estimators as well: their normalized RMSE improves by up to 0.7 standard deviations.

As before, Procedure *WA* used alone only improves the performance of IPW estimators, but when combined with other procedures such two-step procedures may improve the performance of all estimators.

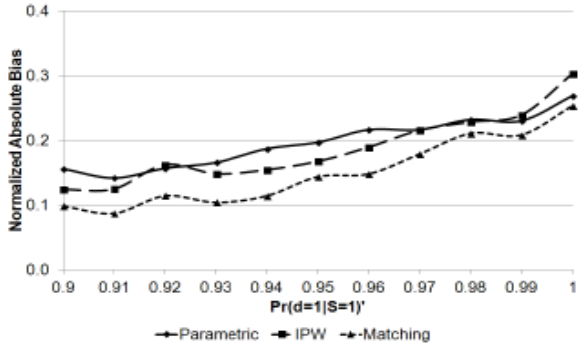
Figure 7: Simulation of the performance of different estimators under different degrees of support reduction when no observations are dropped



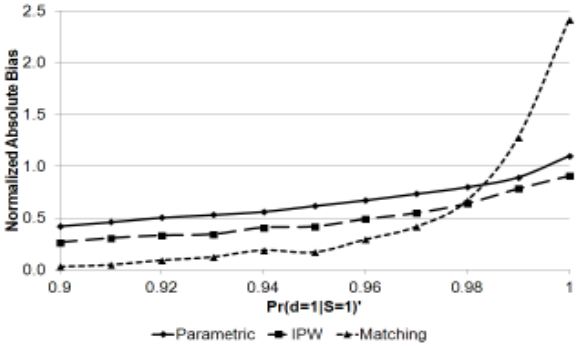
(a) Average normalized RMSE with the treatment being award of VMSW



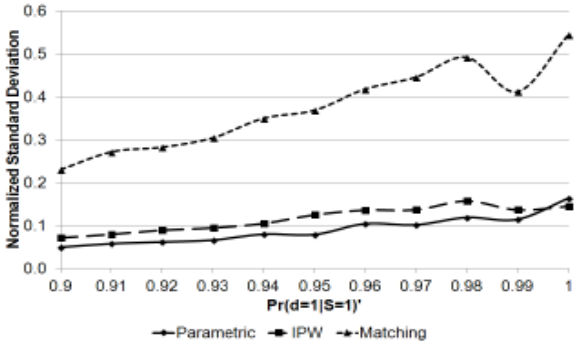
(b) Average normalized RMSE with the treatment being award of VTEC



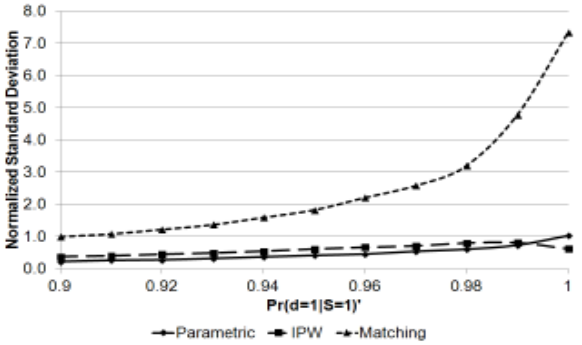
(c) Average normalized absolute bias with the treatment being award of VMSW



(d) Average normalized absolute bias with the treatment being award of VTEC



(e) Average normalized standard deviation with the treatment being award of VMSW



(f) Average normalized standard deviation with the treatment being award of VTEC

Note: Performance of different estimators for support reductions. *P(d = 1|S = 1)'* is varied on the grid {0.9, 0.91,...,1}. Dummy variables are generated for each value of *P(d = 1|S = 1)'* . OLS is run with these dummy variables. The omitted category is 'no support restrictions'. The outcome variables are the normalized performance measures RMSE, absolute bias, and standard deviation, respectively. Separate regressions are estimated by estimator, sample size, and type of voucher. Estimated performance measures reported are averaged over the regressions by sample size and type of voucher, in order to reduce the dimensionality of our results. Covariates are the share of treated, type of effect heterogeneity, and type of outcome variable (fully interacted). Only results for Procedure A are reported (no observations are dropped; see Table 2). Only estimators based on misspecified models are included (in order to avoid problems with perfect predictors). The average conditional standard deviations of parametric estimators are 275, of IPW estimators 308, and of matching estimators 326.

Table 5: Effect of support-adjustment procedures on normalized RMSE in DGPs with support restrictions for the treatment 'award of VMSW'

Sample size		500			2000			8000	
Support	Parametr.	IPW	Match.	Parametr.	IPW	Match.	Parametr.	IPW	Match.
procedures	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
No adjustment (A: reference)									
WA	.019	-.096***	.070	.018	-.081***	.155	0	-.009	4.590
Drop treated based on fixed value of propensity score									
B2	-.110***	-.109***	-1.624***	.112***	.0198	-3.709***	.563***	.394***	-7.292***
WB2	-.110***	-.109***	-1.624***	.112***	.0198	-3.709***	.563***	.394***	-7.292***
Drop treated based on density of propensity score									
C2	-.091***	.001	.002	.003	.001	.005	0	0	0
WC1	-.001	-.099***	.039	.018	-.081***	.155	0	-.009	4.590
Drop treated based on lack of non-treated neighbours in terms of radius of propensity score									
D1	-.124***	.009	-1.641***	-.023	-.015	-3.768***	.015	-.051	-7.524***
WD1	-.124***	-.095***	-1.648***	-.050***	-.110***	-3.903***	.015	-.061*	-7.675***
D2	-.012	-.003	-1.493***	0	-.001	-3.516***	.002	-.009	-7.112***
WD2	-.016	-.103***	-1.501***	.018	-.082***	-3.611***	.003	-.018	-7.184***
Drop treated based on upper limits of distribution of propensity score among non-treated									
E1	-.109***	.001	-1.625***	-.067***	-.017	-3.875***	.009	-.071**	-7.764***
WE1	-.121***	-.100***	-1.641***	-.075***	-.114***	-3.919***	.009	-.084***	-7.901***
E2	-.116***	-.113***	-1.635***	-.0396	-.102***	-3.910***	.177***	-.021	-7.910***
WE2	-.114***	-.127***	-1.636***	-.0421	-.115***	-3.913***	.177***	-.021	-7.910***
E3	-.064***	-.119***	-1.593***	.043	-.052	-3.840***	.376***	.201***	-7.730***
WE3	-.060**	-.115***	-1.591***	.047	-.050	-3.838***	.376***	.201***	-7.729***
Procedures that extrapolate into lack of support region									
F2	.331***	.399***	-1.181***	.293***	.332***	-3.514***	.500***	.505***	-7.466***
WF2	.373***	.418***	-1.168***	.316***	.326***	-3.517***	.502***	.507***	-7.470***
G2	.149***	.142***	-1.404***	.106***	.063***	-3.732***	.110***	.039	-7.856***
WG2	.199***	.177***	-1.388***	.121***	.069***	-3.723***	.110***	.039	-7.851***
H2	.233***	.308***	-1.324***	.173***	.210***	-3.621***	.341***	.323***	-7.534***
WH2	.262***	.314***	-1.318***	.196***	.204***	-3.624***	.343***	.324***	-7.551***
I2	.041***	.042***	-1.571***	.033**	-.008	-3.844***	.051	-.015	-7.956***
WI2	.059***	.047***	-1.568***	.043***	-.008	-3.842***	.051	-.015	-7.953***
# of obs.	792	792	1,584	2,376	2,376	4,752	2,376	2,376	4,752

Note: OLS. Dependent variable is the normalized RMSE. The covariates contain a full set of dummies for the different procedures handling support problems. The reference category is to drop no observations (Procedure A). Further covariates are the tuning parameters of the different DGPs fully interacted (see description in main text). Only selected coefficients for the different procedures are reported in this table. Find a complete set of results, including standard errors and R^2s in Online Appendix E. ***, **,* indicate significance at the 1-, 5-, and 10-percent levels, respectively (based on robust standard errors).

Table 6: Effect of support-adjustment procedures on normalized RMSE in DGPs with support restrictions for the treatment 'award of VTEC'

Sample Size		500			2000			8000	
Support	Parametr.	IPW	Match.	Parametr.	IPW	Match.	Parametr.	IPW	Match.
procedures	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
No adjustment (A: reference)									
WA	.257***	-.654***	3.642*	.130	-.393	.797	.0291	-.173	7.955
Drop treated based on fixed value of propensity score									
B2	-.702***	-.629***	-6.017***	-.538**	-.649**	-22.42***	-.501**	-.770***	-19.57***
WB2	-.677***	-.697***	-5.955***	-.538**	-.650**	-22.42***	-.501**	-.770***	-19.57***
Drop treated based on density of propensity score									
C2	-.048	-.013	-.031	.011	.001	.018	0	0	0
WC1	.245***	-.659***	3.598*	.13	-.393	.797	.029	-.173	7.955
Drop treated based on lack of non-treated neighbours in terms of radius of propensity score									
D1	-.594***	-.005	-5.327***	-.124	-.008	-22.80***	-.182	-.186	-19.93***
WD1	-.733***	-.650***	-6.113***	-.396*	-.495**	-23.32***	-.157	-.368*	-20.23***
D2	-.038	-.007	-4.112**	-.003	.031	-22.01***	-.0357	-.027	-18.51***
WD2	-.527***	-.708***	-5.721***	.127	-.362	-21.68***	-.007	-.200	-18.15***
Drop treated based on upper limits of distribution of propensity score among non-treated									
E1	-.576***	-.018	-5.554***	-.336	-.0379	-23.10***	-.334*	-.279	-20.28***
WE1	-.746***	-.670***	-6.117***	-.466*	-.505**	-23.45***	-.379**	-.552***	-20.75***
E2	-.712***	-.581***	-6.069***	-.447*	-.458*	-23.61***	-.507**	-.626***	-21.86***
WE2	-.739***	-.701***	-6.109***	-.495**	-.537**	-23.67***	-.518***	-.636***	-21.88***
E3	-.683***	-.704***	-6.042***	-.445**	-.533**	-23.63***	-.531***	-.661***	-21.90***
WE3	-.654***	-.689***	-6.026***	-.442**	-.534**	-23.63***	-.531***	-.661***	-21.90***
Procedures that extrapolate into lack of support region									
F2	1.697***	1.577***	-2.647*	6.216***	5.944***	-18.39***	3.607***	3.284***	-18.34***
WF2	2.146***	1.899***	-2.912*	6.223***	5.881***	-18.42***	3.630***	3.297***	-18.28***
G2	1.499***	1.275***	-3.888**	2.043***	1.800***	-21.49***	1.167***	.898***	-20.51***
WG2	3.515***	3.089***	-2.975*	2.226***	1.947***	-21.34***	1.201***	.927***	-20.48***
H2	1.247***	1.184***	-4.270***	1.724***	1.562***	-21.85***	2.390***	2.086***	-19.50***
WH2	1.195***	1.053***	-4.412***	1.748***	1.524***	-21.87***	2.403***	2.091***	-19.48***
I2	.144***	.0717	-5.494***	.188	.0178	-23.20***	.279	.0155	-21.33***
WI2	.446***	.338***	-5.353***	.263	.0749	-23.12***	.304	.0371	-21.31***
# of obs.	792	792	1,584	2,376	2,376	4,752	2,376	2,376	4,752

Note: See note below Table 5.

Procedure *B* strongly reduces the normalized standard deviation of all estimators (the parametric estimators in the largest sample are an exception; see Tables E.5 and E.6 in Online Appendix E). However, as the normalized absolute bias increases (see Tables E.3 and E.4 in

Online Appendix E), the overall effect on the normalized RMSE is ambiguous. It is only for the matching estimators that this procedure always leads to such performance improvements. In contrast, Procedure *C* has again only little impact on the performance of the different estimators. With few exceptions, *C* does neither harm nor improve their performance.

Procedures *D* and *E* have the largest positive affect on the performance of the estimators when there are strong support problems. They improve the normalized standard deviation of all estimators especially in the smaller samples. These improvements are largest for matching estimators. For the *award of VTEC* they can exceed 20 standard deviations. *D* and *E* also reduce the normalized RMSE, normalized absolute bias, especially when the treatment is *award of VTEC*. In some specifications, the normalised RMSE can be improved by up to 20 and the absolute bias by up to 5 standard deviations. In rare cases when the treatment is *award of VMSW*, *E* increases the normalized absolute bias of the parametric and IPW estimators,[9] while *D* does not affect them. However, *E* leads to larger improvements on the normalized standard deviations of the estimators than *D* in these specifications. Both procedures work better when combined with *W*.

All procedures aiming to estimate $\gamma_{ATET|N}$ or $Y_{0|N}$ off support (*F, G, H, I*) work only for matching estimators, if they work at all. However, even for matching estimators they do not outperform procedures *D* and *E*, which are much easier to implement.

6.4 Detailed results

The regression results allowed a condensed view at the average performances. Next, we investigate whether those findings appear to show heterogeneity in relevant dimensions, like

[9] This appears only very rarely when the threshold $\overline{p}$ is specified at the highest propensity score value of the non-treated subpopulation.

model specification, type of estimator, outcome variable, sample size, treatment share, types of effect heterogeneity, and different support restrictions. Here, the results are only summarised. The details can be found in Tables F.1-F.7 in Online Appendix F. The results are based on eight different performance measures. The first one is the relative RMSE, which is the additional RMSE measured relative to the RMSE of the procedure with the minimum normalized RMSE for the DGPs under consideration. Then, there are the normalized performance measures absolute bias and standard deviation. Skewness and kurtosis are presented as well because a non-normal distribution of the estimators may lead to inference problems. Furthermore, the average numbers of dropped treated observations are reported. Additionally, we show the number of dropped observations of support ($S_i = 1$). Normally, for each group of common support procedures the results of only two procedures (with and without *W*) are reported.[10] However, the full set of results is reported for *D* and *E*, because the previous results suggested that these procedures outperformed the rest.

The dimension to begin with is the model specification (see Table F.1 in Online Appendix F). These results are only based on simulations without support restrictions.[11] As expected, there is no strong bias for the *award of VMSW* because this treatment exhibits only a low degree of selectivity (when there are no additional artificial support restrictions). The lowest normalized RMSE occurs when no observations are dropped in the case when the population propensity score is used.[12] Only when the misspecified model is used, dropping observations may lead to small improvements in terms of normalized RMSE. In this specification Procedure *H1*, which

[10] The detailed results for the other procedures are available from the authors upon request.

[11] We consider only DGPs without support restrictions, because we do not use all specifications when the support is restricted in order to avoid potential problems with (almost) perfect predictors.

[12] We consider only IPW and matching estimators when the population propensity score is used.

extrapolates $Y_{0|N}$ at the highest non-treated propensity score level, would be optimal in terms of normalized RMSE.[13] For the *award of VTEC*, the results indicate that the normalized RMSE may be potentially improved by several procedures, with *WE1* performing best. This is not surprising as this treatment is subject to stronger selectivity.

From now on, DGPs with restricted support are included in the analysis as well, but specifications with the population propensity score or the true model specifications are excluded. Next, heterogeneity with respect to estimators is (re-)considered. Generally, dropping observations may improve the normalized RMSE, absolute bias, and standard deviation (see Table F.2 in Online Appendix F), with *WE2* performing best on average. The largest performance improvement can be obtained for matching estimators. For the *award of VMSW*, the normalized RMSE is improved by up to 24% when Procedure *WE1* (similarly for *WE2*) is applied (in comparison to dropping no observations, *A*). For the *award of VTEC*, the normalized RMSE may be improved by up to 275% when *WE2* is applied. In this case, even the normalized RMSE of parametric estimators improves up to 26% (*WE2*). This suggests that support thickness and overlap are important even when parametric estimators are applied. The normalized RMSE of IPW estimators improve by up to 34% (*WE2*) in this case as well. Matching estimators have the highest normalized standard deviation if no observations are dropped (A). In addition, Procedure *E* improves skewness and kurtosis of all three estimators.[14] This improvement is most relevant for IPW when the treatment is *award of VTEC*.

With respect to the different types of outcome variables, Procedure *E* works on average best for discrete and binary outcome variables (in terms of normalized RMSE of the different

[13] This is surprising, because Procedure *H* performed not well in the results presented in the previous sections.

[14] This statement holds also for the other types of heterogeneity investigated in this section.

estimators; *WE2* is best on average; see Table F.3 in Online Appendix F). For the semi-continuous outcome variables, the linear approximation of the potential outcome of the treated off support under non-treatment $Y_{0|N}$ (Procedure *I*) has on average the best performance on normalized RMSE of the estimators, mainly due to improvements in the normalized absolute bias.[15] *E* has a smaller standard deviation, but a larger absolute bias than *I*.

With regard to all sample sizes, Procedure *E* performs best on average in terms of normalized RMSE (see Table F.4 in Online Appendix F). In most cases the two-step procedures, dropping control observations with high weights in the first place, works best (in particular *WE2*).

With regard to the treatment shares, Procedure *I* works best for small treatment shares (10%-treated) and Procedure *E* works better for 50%- and 90%- treated (see Table F.5 in Online Appendix F). In contrast to the previous results, when the treatment share is 10% Procedure *I* has a smaller standard deviation but a larger absolute bias than Procedure *E*. Overall, for the low treatment shares (10% and 50%) Procedures *D*, *E*, and *I* are almost equally good.

With regard to effect homogeneity, dropping observations does not introduce an asymptotic bias, but might still decrease the normalized standard deviation (see Table F.6 in Online Appendix F), at least for the 'local' semiparametric estimators like matching and IPW that allow for effect heterogeneity. Indeed, all procedures with a high number of dropped observations perform well under effect homogeneity. However, the best procedure in terms of normalized RMSE is not the one that drops the most treated observations (*B2*) but the one that

[15] This is surprising, because Procedure *I* was not among the procedures performing well in the previous sections. However, also in the detailed results, Procedure *I* is (if at all) only slightly better than *D* and *E*. Both procedures work also well (or better) in combination with *W*. The last statement holds also for the following findings even if not mentioned explicitly.

drops the most observations with $S_i = 1$ (*E3*). Under effect heterogeneity, the procedures, which drop fewer observations, have better performances (e.g. *WE1* and *WE2*).

Finally, there is the degree of support restrictions (see Table F.7 in Online Appendix F) that may be relevant for differential performances of the different procedures. However, here Procedure *WE* has the best normalized RMSE in all specifications, only the optimal cut-off values $\bar{p}$ varies.

7 Conclusions

This paper studies the performance of different parametric and semiparametric estimators adjusting observable characteristics in no or thin support situations and the performance of remedies suggested in the literature. The performance is evaluated in a Monte Carlo Study using simulation designs based on real data. Therefore, the many different data generating processes investigated should be close to what is encountered in empirical applications, in particular in the context of program evaluation.

Our findings suggest that almost all procedures proposed in the literature to mitigate support problems have the potential to improve the performance of the estimators investigated, in particular when support problems become severe. Although the largest improvements can be achieved for matching estimators, parametric estimators benefit as well. However, not surprisingly, some procedures are more effective than others are.

The results obtained for different degrees of support problems, estimators, specifications, and other features of the data generating process, suggest that procedures based on trimming observations in the treated group with propensity scores larger the maximum value (or the 99% quantile) of the propensity score in the non-treated group consistently outperform the

other procedures considered. Furthermore, when these procedures are combined with further trimming non-treated observations that receive a 'too-large' weight, additional improvements were observed. When support problems are mild, the gains usually come from an improvement in the precision of the estimators. When support problems are strong, biases are generally reduced and precision increases. Therefore, we recommend using these methods in applied work independent of the type of estimator used.

Clearly, due to the empirical design of the Monte Carlo approach used, the results of this study should be valid for similar programme evaluation studies based on large administrative data bases. Furthermore, many features of the data generating processes have also been varied. Thus, we are tempted to claim that our simulation designs contain many different cases relevant in practise and have external validity beyond such programme evaluation studies. Whether this claim can be confirmed in another Empirical Monte Carlo study based on a completely different applied field remains speculative and deserves further research.

References

Abadie, A., and G. W. Imbens (2011): "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business and Economic Statistics*, 29 (1), 1–11.

Biewen, M., B. Fitzenberger, A. Osikominu, and M. Paul (2014): "The Effectiveness of Public Sponsored Training Revisited: The Importance of Data and Methodological Choices," *Journal of Labor Economics*, forthcoming.

Busso, M., J. DiNardo, and J. McCrary (2014a): "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects," *The Journal of Business and Economic Statistics*, forthcoming.

Busso, M., J. DiNardo, and J. McCrary (2014b): "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators," *Review of Economics and Statistics*, forthcoming.

Card, D., J. Kluve, and A. Weber (2010): "Active Labour Market Policy Evaluations: A Meta-Analysis," *The Economic Journal*, 120 (548), 452–477.

Crump, R. K., J. V. Hotz, G. W. Imbens, and O. A. Mitnik (2009): "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96 (1), 187–199.

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.

Doerr, A., B. Fitzenberger, T. Kruppe, M. Paul, and A. Strittmatter (2014): "Employment and Earnings Effects of Awarding Training Vouchers," Working Paper.

Grzybowski, M., E. A. Clements, L. Parsons, R. Welch, A. T. Tintinalli, M. A. Ross, and R. J. Zalenski (2003): "Mortality Benefit of Immediate Revascularization of Acute ST-Segment Elevation Myocardial Infarction in Patients with Contraindications to Thrombolytic Therapy: A Propensity Analysis," *The Journal of the American Medical Association*, 290 (14), 1891–1898.

Heckman, J. J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66 (5), 1017–1098.

Heckman, J. J., H. Ichimura, and P. E. Todd (1998): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64 (4), 605–654.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71 (4), 1161– 1189.

Ho, D., K. Imai, G. King, and E. Stuart (2007): "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis*, 15 (3), 199–236.

Horvitz, D. G., and D. J. Thompson (1952): "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47 (260), 663–685.

Huber, M., M. Lechner, and A. Steinmayr (2012): "Radius Matching on the Propensity Score with Bias Adjustment: Finite Sample Behaviour, Tuning Parameters and Software Implementation," SEPS Economics Working Paper Series 2012-26, University of St. Gallen.

Huber, M., M. Lechner, and C. Wunsch (2013): "The Performance of Estimators Based on the Propensity Score," *Journal of Econometrics*, 175 (1), 1–21.

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

Imbens, G. W., and K. Kalyanaraman (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies*, 79 (3), 933–959.

Imbens, G. W., and T. Lemieux (2008): "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142 (2), 615–635.

Khan, S., and E. Tamer (2010): "Irregular Identification, Support Conditions and Inverse Weight Estimation," *Econometrica*, 78 (6), 2021-2042.

Lechner, M. (2008): "A Note on the Common Support Problem in Applied Evaluation Studies," *Annales d'Economie et de Statistique*, 91-92, 217–234.

Lechner, M., R. Miquel, and C. Wunsch (2011): "Long-run Effects of Public Sector Sponsored Training," *Journal of the European Economic Association*, 9 (4), 742– 784.

Lechner, M., and C. Wunsch (2013): "Sensitivity of Matching-Based Program Evaluations to the Availability of Control Variables," *Labour Economics*, 21, 111–121.

Rinne, U., A. Uhlendorff, and Z. Zhao (2013): "Vouchers and Caseworkers in Training Programs for the Unemployed," *Empirical Economics*, 45 (3), 1089-1127.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70 (1), 41–55.

Rubin, D. B. (1973): "Matching to Remove Bias in Observational Studies," *Biometrics*, 29 (1), 159–183.

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

Vincent, J. L., J.-F. Baron, K. Reinhart, L. Gattinoni, L. Thijs, A. Webb, A. Meier-Hellmann, G. Nollet, and D. Peres-Bota (2002): "Anemia and Blood Transfusion in Critically Ill Patients," *The Journal of The American Medical Association*, 288 (12), 1499–1507.