Augurzky, Boris; Schmidt, Christoph M.

Working Paper

The Propensity Score: A Means to An End

IZA Discussion Papers, No. 271

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Augurzky, Boris; Schmidt, Christoph M. (2001) : The Propensity Score: A Means to An End, IZA Discussion Papers, No. 271, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/21122

DISCUSSION PAPER SERIES

IZA DP No. 271

The Propensity Score: A Means to An End

Boris Augurzky
Christoph M. Schmidt

March 2001

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

The Propensity Score: A Means to An End

Boris Augurzky
University of Heidelberg, and IZA, Bonn

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

Discussion Paper No. 271
March 2001

IZA

P.O. Box 7240
D-53072 Bonn
Germany

Tel.: +49-228-3894-0
Fax: +49-228-3894-210
Email: iza@iza.org

IZA Discussion Paper No. 271
March 2001

ABSTRACT

The Propensity Score: A Means to An End*

Propensity score matching is a prominent strategy to reduce imbalance in observational studies. However, if imbalance is considerable and the control reservoir is small, either one has to match one control to several treated units or, alternatively, discard many treated persons. The first strategy tends to increase standard errors of the estimated treatment effects while the second might produce a matched sample that is not anymore representative of the original one. As an alternative approach, this paper argues to carefully reconsider the selection equation upon which the propensity score estimates are based. Often, all available variables that rule the selection process are included into the selection equation. Yet, it would suffice to concentrate on only those exhibiting a large impact on the outcome under scrutiny, as well. This would introduce more stochastic noise making treatment and comparison group more similar. We assess the advantages and disadvantages of the latter approach in a simulation study.

JEL Classification: C14, C15,

Keywords: Estimation of the propensity score, balance of relevant covariates, simulation study

Christoph M. Schmidt
Department of Economics
University of Heidelberg
Grabengasse 14
69177 Heidelberg
Germany
Tel.: +49-6221-54 2955
Fax: +49 6221 54 3640
Email: schmidt@uni-hd.de

* Boris Augurzky gratefully acknowledges financial support by the Friedrich-Ebert-Stiftung. This research was in part supported by the Deutsche Forschungsgemeinschaft (DFG) under the research grant “Sonderforschungsbereich (SFB) 544, Control of Tropical Infectious Diseases".

1 Introduction

In contrast to a randomized experiment, in an observational study the treatment and the comparison group usually differ systematically in terms of their observable and unobservable covariates. Yet, appropriate weighting schemes may provide for a convincing evaluation strategy. In particular, balancing all observable covariates by the method of matching allows the identification of the mean effect of treatment if the remaining unobservable covariates are irrelevant. Usually, the number of covariates is high, thus making exact matching – in all likelihood – impossible. Rosenbaum & Rubin (1983) suggest to alternatively balance the one-dimensional propensity score, which is the conditional probability to participate in treatment given all relevant covariates. They show that this strategy, on average, achieves overall balance, thus circumventing the curse of dimensionality.

However, if treatment and comparison group differ to a considerable extent, i.e. if selection into treatment is remarkably strong, achieving an acceptable balance will be difficult. A *full matching* using all treated and untreated units in the sample might produce many strata consisting of one control and more than one treated unit. Generally, one would like to achieve a stratification which is more *uniform*. Uniform stratifications tend to produce smaller standard errors of the matching estimates. See, for instance, Augurzky (2000a) and Dehejia & Wahba (1998) whose matching is far from producing a uniform stratification because treated units with high propensity scores hardly find adequate controls. Alternatively, *pair matching* tends to discard the majority of treated individuals at the high end of the propensity score scale. As a result, it restricts evaluation of the treatment effect to individuals with low and medium propensity scores. If effects are different for different locations on the propensity score scale pair matching estimates will be biased.

This paper argues to carefully reconsider the selection equation upon which the propensity score estimates are based. It is common practice to include all available variables that might rule the selection process, with the objective of capturing the selection decision precisely. Yet, we will argue in this paper that, if selection turns out to be extremely strong, one should better concentrate on only those variables with a large impact on both the selection and the outcome under scrutiny. This procedure increases the random part of

the participation process – the whole approach rests on sufficient randomness being retained *after* deriving individuals' propensity score. Alas, a consistent estimation of the propensity score might require including into the selection equation variables which rule the selection process but which are excluded from or only play a minor role in the outcome equation.

In contrast to our arguments, current applied research emphasizes the importance of consistent estimation. For instance, Lechner (1999, 2000) performs and recommends several specification tests to examine whether a probit model is adequate for describing the selection decision. Augurzky (2000a) includes into the probit model several variables that might determine the selection. Heckman, Ichimura & Todd (1997: section 8) choose predictor variables to maximize the within-sample correct prediction rates. Although a thorough understanding of the selection process might in itself be an important contribution, it is not the main objective of propensity score matching for identifying the mean effect of treatment. At best, it is a side effect. What is to be achieved by propensity score matching is balance of all relevant covariates as reflected, for example, in Dehejia & Wahba's (1998) pragmatic estimation strategy concerning the selection equation.

To put it otherwise, there is a trade-off between a consistent estimation of the selection equation that probably balances irrelevant variables, too, and a pragmatic – but probably inconsistent – estimation that concentrates on balancing the relevant variables only. We assess this trade-off in a simulation study relying on the mean squared error criterion. The next section discusses matching as an evaluation strategy and, in particular, outlines the idea behind propensity score matching. Section 3 presents the data generating processes and the dimensions of the simulation study while section 4 explains the algorithm used for matching. Section 5 is dedicated to results for some interesting parameter constellations and the last section summarizes the findings and offers recommendations for applied research.

2 The Matching Approach

In this section, the framework and the idea of propensity score matching are briefly discussed. Rosenbaum (1995), Heckman, LaLonde & Smith (1999), and Schmidt (1999) provide a thorough overview of estimation strategies via matching. Let R_i^1 denote the potential response of individual i under the treatment state and R_i^0 the potential response if i receives no treatment. Furthermore, let D_i denote a binary variable indicating treatment status, thus, $R_i = D_i R_i^1 + (1 - D_i) R_i^0$ is the observed outcome. This framework has become known as the *potential outcome approach to causality* suggested by Roy (1951), Rubin (1974, 1977), and Holland (1986). It requires that the response of an individual be independent of the decisions of all other individuals. This implies that there are only two potential outcomes, namely R_i^0 and R_i^1, one for the personal state $D_i = 0$, and one for $D_i = 1$, respectively. There are no further potential outcomes depending on the assignment of any other individual. This requirement is often referred to as *stable unit treatment value assumption* (sutva, see Rubin, 1986).

The individual treatment effect is $\delta_i = R_i^1 - R_i^0$ which, however, is not observable since either R_i^1 or R_i^0 is missing. Alternatively, one might focus on the mean effect of treatment on the treated individuals

$$\mathbb{E}(\delta_i | D_i = 1) = \mathbb{E}(R_i^1 | D_i = 1) - \mathbb{E}(R_i^0 | D_i = 1). \tag{1}$$

Yet, while the first expectation $\mathbb{E}(R_i^1 | D_i = 1)$ can be identified in the subsample of the treatment group, the counterfactual expectation $\mathbb{E}(R_i^0 | D_i = 1)$ is not identifiable without invoking further assumptions.

Somehow one has to rely on the untreated units ($D_i = 0$) of the comparison group to obtain information on the counterfactual outcome of the treated in the no-treatment state. A simple replacement of $\mathbb{E}(R_i^0 | D_i = 1)$ by $\mathbb{E}(R_i^0 | D_i = 0)$ is unlikely to be the appropriate strategy, though, since treated and untreated units tend to differ considerably in their characteristics that determine the outcome if they themselves select into treatment. An ideal randomized experiment solves this problem, see Heckman (1996) or Schmidt, Baltussen & Sauerborn (1999). It generates a treatment and a control group by a randomization process ensuring exogenous selection into treatment and thus resulting, on

average, in balance of all covariates between treatment and control group, in particular those determining outcome.

In contrast, in an observational study, where self-selection into treatment is typically non-negligible, matching tries to mimic *ex post* a randomized experiment by stratifying the sample of treated and untreated units with respect to covariates X_i that rule both the selection into treatment and the outcome under study. Such a stratification eliminates selection bias provided all variables X_i are observed and balanced. In this case, each stratum would represent a separate small randomized experiment and simple differences between treated and controls would provide an unbiased estimate of the treatment effect. This technique does not require linearity, parametric, or distributional assumptions.

Formally, assume that the response R_i^0 is conditionally independent of D_i given X_i yielding $\mathbb{E}(R_i^0|X_i, D_i = 1) = \mathbb{E}(R_i^0|X_i, D_i = 0)$. Moreover, assume $\mathbb{P}(D_i = 0|X_i = x) > 0$ for all x which guarantees that, with positive probability, there are untreated units for each x. The data generating processes of the simulation presented in the next section are such that these requirements for matching will be fulfilled. The conditional mean response of the *treated* under no treatment for a given X can thus be estimated by the conditional mean response of the *untreated* under no treatment. The overall estimated mean effect is the weighted average over all stratum effects. The stratum weights are proportional to the number of treated units in the stratum in order to identify $\mathbb{E}(\delta_i|D_i = 1)$.

However, in a finite sample balancing X is difficult or even impossible if the vector of observables is of high dimension. To escape this curse of dimensionality, Rosenbaum & Rubin (1983) suggest to alternatively use the conditional probability to participate in treatment $p(x) = \mathbb{P}(D_i = 1|X_i = x)$, the *propensity score*, for purposes of stratifying the sample. They show that if R_i^0 is independent of D_i given X_i, R_i^0 and D_i are also independent given $p(X_i)$. Matching treated and untreated units with the same propensity scores and placing them into one stratum means that the decision whether to participate or not is random in such a stratum. The probability of participation in this stratum equals the propensity score. Alas, some disadvantages accompany this strategy. First, the propensity score itself has to be estimated. Second, since it is a continuous variable exact matches will hardly be achieved and a certain distance between treated and untreated

units has to be accepted nonetheless. Prominent candidates measuring the distance are the difference in propensity scores or the Mahalanobis metric (Rubin, 1980).

The Idea Behind Propensity Score Matching

Let there be three kinds of covariates X, Y, and Z characterizing individuals. Generally, both potential outcomes and the participation probability depend on all three variables. For reasons of clarity of the argument further assume that Y and Z are binary and let all considerations to follow be conditional on X. In sum, $R^0 = R^0(Y,Z)$, $R^1 = R^1(Y,Z)$, and $p = p(Y,Z)$.

There are four cells

	$Z=0$	$Z=1$
$Y=0$	n_{00}	n_{01}
$Y=1$	n_{10}	n_{11}

each comprising n_{jk} individuals, $j,k \in \{0,1\}$. For the sake of notational convenience, abbreviate cell-wise expectations as follows

$$\begin{aligned} R^1_{jk} &= \mathbb{E}(R^1|Y=j,Z=k,D=1) \\ R^0_{jk} &= \mathbb{E}(R^0|Y=j,Z=k,D=1) = \mathbb{E}(R^0|Y=j,Z=k,D=0), \end{aligned}$$

$\Delta_{jk} = R^1_{jk} - R^0_{jk}$, and p_{jk} denotes the propensity score in the corresponding cell. As a result, the mean effect Δ (conditional on X) can be written

$$\Delta = \frac{1}{n_t}\left(\Delta_{00}\, p_{00}\, n_{00} + \Delta_{01}\, p_{01}\, n_{01} + \Delta_{10}\, p_{10}\, n_{10} + \Delta_{11}\, p_{11}\, n_{11}\right), \tag{2}$$

n_t denotes the total number of treated individuals, $n_t = \sum p_{jk} n_{jk}$.

Selection on Z only. If the propensity score merely depends on Z, $p_{00} = p_{10} = p_{.0}$ and $p_{01} = p_{11} = p_{.1}$. This implies that Y can be expected to be already balanced and that cells with the same value of Z can be combined. Defining $n_{.k} = n_{0k} + n_{1k}$ and the effect in the combined cell $\Delta_{.k} = (\Delta_{0k}\, n_{0k} + \Delta_{1k}\, n_{1k})/n_{.k}$, equation (2) reduces to

$$\Delta = \frac{1}{n_t}\left(\Delta_{.0}\, p_{.0}\, n_{.0} + \Delta_{.1}\, p_{.1}\, n_{.1}\right). \tag{3}$$

The combination of cells that share the same propensity score is the very advantage of propensity score matching with regard to exact covariate matching. On the one hand, this means that individuals with different characteristics might be matched, here with different values of Y. As a result, in finite samples where Y may still be unbalanced the combined-cell-specific estimates of the treatment effect may deviate from the true value. On the other hand, combination of cells avoids that cells comprising only treated or only untreated units have to be dropped. This would give rise to both larger variance of the estimates and possibly a bias if the treatment effect is heterogeneous and the loss of cells is systematic.

Angrist & Hahn (1999) assess this bias-variance trade-off both theoretically and by means of a simulation study. They argue that the very virtue of propensity score estimation emerges when cells are finite. If cell sizes themselves increased beyond all bounds propensity score matching would not be advantageous to exact matching, see Hahn (1998).

Exclusion Restriction of Z. A symmetric special case arises if the outcome does not but the selection does vary with Z. Consequently, cells with the same value of Z could be combined even though they are subject to a different selection process, i.e. their propensity score differs. Analogously to above, it follows that $\Delta_{00} = \Delta_{01} = \Delta_{0.}$ and $\Delta_{10} = \Delta_{11} = \Delta_{1.}$, implying that imbalance of Z has no effect on the estimation of the outcome and that cells with the same value of Y can be combined without loss of information. Let $n_{k.} = n_{k0} + n_{k1}$ and $p_{k.} = (p_{k0}n_{k0} + p_{k1}n_{k1})/n_{k.}$, equation (2) can be reduced to

$$\Delta = \frac{1}{n_t} \left(\Delta_{0.}\, p_{0.}\, n_{0.} + \Delta_{1.}\, p_{1.}\, n_{1.} \right). \tag{4}$$

If both cases are fulfilled, i.e. the outcome depends on Y and the selection process is ruled by Z only, all four cells can be combined to one and Δ is just the difference between the unconditional responses of treated and untreated persons in the combined cell (merely defined by X). This point reflects the fact that solely covariates which rule both the outcome and the selection into treatment need to be balanced by matching. Consequently, the question is raised whether the propensity score depending on X and Z is the right measure to match upon or whether it might be better replaced by the marginal

propensity score depending solely on X. Matching on the latter would not unnecessarily balance Z. Therefore, one could concentrate on the balance of X. This would probably result in a more uniform stratification of the sample. That is, one control would not be matched to an overwhelmingly large number of treated persons.

In other words, omitting irrelevant variables increases randomness of the selection process and diminishes its deterministic part. For example, if selection were completely determined by certain known variables the propensity score of treated units would be 1 and that of untreated 0. Consequently, no reasonable strategy whatsoever would be able to match controls to any given treated person. In contrast, the more variables determining the selection process can be regarded as stochastic noise because their impact on the outcome variable is negligible, the more randomness will enter the process and the easier treated individuals will find adequate controls. One might equate the *Pseudo* R^2 of a probit model as reflecting the degree of the selection determination.

3 The Data Generating Processes

As above, let R_i denote the outcome of individual i, $i = 1, ..., n$, and D_i the binary treatment indicator. On average, there will be 150 treated individuals and between 300 and 900 comparison units. The latter number is variable such that finding adequate controls is more or less difficult. The outcome is a linear function of confounding covariates, $(X_1, X_2, Y_1, Y_2, Z_1, Z_2)$, an individual treatment effect δ_i, and normally distributed stochastic noise $\varepsilon_i \sim \mathcal{N}(0, 9)$

$$R_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 Y_{1i} + \beta_4 Y_{2i} + \beta_5 Z_{1i} + \beta_6 Z_{2i} + \delta_i D_i + \varepsilon_i. \tag{5}$$

The selection equation depends on the same covariates

$$D_i = \mathbf{1}[\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 Y_{1i} + \alpha_4 Y_{2i} + \alpha_5 Z_{1i} + \alpha_6 Z_{2i} + \eta_i > 0] \tag{6}$$

where $\mathbf{1}$, the indicator function, is 1 if its argument holds and zero otherwise, and $\eta \sim \mathcal{N}(0, 1)$ is standard normal.

The coefficients of the Z-variables β_5, β_6 in the outcome equation (5) are comparatively small and, likewise, the same is assumed for those of the Y-variables in the selection equa-

tion (6), α_3, α_4. This means that Y tends to be already partly balanced between treated and untreated units and, furthermore, although Z will be highly unbalanced its impact on the outcome is minor. The X-variables are the strongest predictors of both the outcome and the selection and most effort should therefore be spent on balancing them. The simulation aims at examining the relative performance of the matching estimator when the propensity score is estimated by means of a probit model including all variables (X, Y, Z) and when based on the most relevant variables X only. Furthermore, the treatment effect δ_i depends on i reflecting heterogeneity in the following manner

$$\delta_i = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 Y_{1i} + \gamma_4 Y_{2i} + \gamma_5 Z_{1i} + \gamma_6 Z_{2i}.$$

Depending on the parameter setting self-selection into treatment plays a more or less important role resulting in more or less severe imbalance of covariates. If Y and Z are of minor relevance, merely X should actively be balanced by matching on $\mathbb{P}(D = 1|X)$. However, $\mathbb{P}(D = 1|X, Y, Z)$ follows a probit specification in accordance with equation (6)

$$\mathbb{P}(D = 1|X, Y, Z) = \Phi(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 Y_1 + \alpha_4 Y_2 + \alpha_5 Z_1 + \alpha_6 Z_2),$$

where Φ is the cumulative normal density function. Thus, a probit estimation using covariates X, Y, and Z – henceforth called the *full probit* – would yield consistent estimates of individual propensity scores but matching on them would unnecessarily balance Z, as well. On the other hand, a misspecified probit estimation merely on X – henceforth called the *partial probit* – would indeed use only the most relevant variables but might yield inconsistent estimates of $\mathbb{P}(D = 1|X)$. The choice to proceed as if a probit model held might therefore be one reason for bias in estimates of the mean treatment effect.[2] In general, the functional form of $\mathbb{P}(D = 1|X) = \mathbb{E}(\mathbb{E}(D|X, Y, Z)|X)$ does not follow a probit specification

$$\mathbb{P}(D = 1|X) = \int \Phi(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 y_1 + \alpha_4 y_2 + \alpha_5 z_1 + \alpha_6 z_2)\, f_{(Y,Z)|X}(y, z)\, d(y, z)$$

where $f_{(Y,Z)|X}$ is the conditional density of (Y_1, Y_2, Z_1, Z_2) given X.[3] Another source of bias arises if the impact of Y and Z on the outcome and on the selection are not zero.

[2] Note, though, that consistent estimation of the coefficients α in the probit model are not of any interest. Furthermore, see Yatchew & Griliches (1984) for a discussion of specification errors in probit models.

[3] Since it is not easy to solve the integral analytically the true values are calculated by ways of an

In consequence, the questions of this paper are (i) whether neglecting to balance (Y, Z) produces a bias which is offset by a larger variance of the estimates of the full model, and (ii) whether the functional specification error in estimating $\mathbb{P}(D = 1|X)$ by a probit model causes severe problems. We assess the trade-off on the basis of the mean squared error criterion.

The described setup allows to perform simulations along five dimensions. First, the impact of Z on R and of Y on D may be altered. To this end, β_5, β_6 and α_3, α_4 are varied between 0 and 0.1 while the remaining α- and β-coefficients are set equal to 1, and the constant β_0 equals 0. This strategy allows an exploration of the question whether near exclusion restrictions carry the same implications as genuine exclusion restrictions. Second, the average number of comparison units in the sample is gradually increased from 300 to 900 while the average number of treated is fixed at 150 by accordingly adjusting the constant α_0. Thereby, we address the issue by how much the described trade-off is altered as more and more comparison observations become available.

Third, the deterministic part of the selection equation is successively weakened which means that all α-coefficients except for α_0 are simultaneously reduced until they reach 25% of their original value. This shows how the degree of selection determination influences the stratification results. Fourth, effects δ_i may be homogeneous or heterogeneous corresponding to whether $\gamma = (1, 0, 0, 0, 0, 0, 0)$ or $\gamma = (0.5, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25)$. The homogeneous case presents an interesting benchmark to compare the full and partial probit model. Pair matching might be an unbiased and a more efficient evaluation strategy than full matching when effects are homogeneous. Yet, in this study, the choice of the matching algorithm will not be explored.

Finally, the distribution of $(X_1, X_2, Y_1, Y_2, Z_1, Z_2)$ is varied. In a *basic model* all six variables are independently and identically (iid) standard normal implying that omission of (Y, Z) from the probit model does not bias propensity score estimates because the omitted variables are perfectly absorbed in normally distributed stochastic noise. To avoid this favorable aspect, Z will alternatively be distributed in an odd fashion. Several

auxiliary Monte Carlo simulation: 200 times adequate (Y, Z)'s are generated and $\mathbb{P}(D = 1|X, Y = y, Z = z)$ is calculated for each iteration inserting the given $(Y, Z) = (y, z)$. The mean over all iterations is an approximation to $\mathbb{P}(D = 1|X)$.

alternatives have been investigated but those maintaining independence between Z and X and reducing to an exchange of the distribution of Z have been unable to produce biased propensity score estimates.[4]

Apparently, the probit model seems quite insensitive to misspecification of the error distribution as far as the overall fit is concerned and coefficients are of no interest. Yet, as soon as independence of X and Z is abandoned omission of Z leads to heteroskedastic errors of the selection equation and to arbitrarily large biased propensity score estimates, up to estimates that are almost constant for all values of X. One specification that is presented below – called *alternative model* – defines (X_1, X_2, Y_1, Y_2) as *iid uniformly* distributed random variables with mean zero and variance one. In contrast, Z_j will follow the functional form

$$Z_j = U_j \exp(-\mu X_j), \qquad j = 1, 2, \tag{7}$$

where U_j is a uniform random variable in the unit interval and $\mu = 1.35$. In addition, Z_j is standardized to have mean zero and variance one in each iteration of the simulation. This is necessary to ensure that selection due to Z is normalized and comparable to the basic model.[5] Furthermore, interactions between Z and X are introduced into the selection equation (6) such that it becomes

$$\begin{aligned} D_i \;=\; & \mathbf{1}[\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 Y_{1i} + \alpha_4 Y_{2i} + \alpha_5 Z_{1i} + \alpha_6 Z_{2i} + \\ & \alpha_7 X_{1i} Z_{1i} + \alpha_8 X_{1i} Z_{2i} + \alpha_9 X_{2i} Z_{1i} + \alpha_{10} X_{2i} Z_{2i} + \eta_i > 0] \end{aligned}$$

Omission of Z might lead to severe misspecification problems which, however, can substantially be alleviated by adding higher order terms of X into the probit specification. The conditional expectation of Z_j given X_1, X_2 is a function of X_1, X_2

$$\mathbb{E}(Z_j | X_1, X_2) = f(X_1, X_2). \tag{8}$$

Hence, inclusion of higher order terms of (X_1, X_2) approximates a Taylor expansion of $f(X_1, X_2)$ such that, again, almost only the stochastic part of Z will be absorbed by the error term of the model. Three alternative probit models will therefore be specified to

[4] Even very asymmetric strange densities of Z failed to generate inconsistencies.

[5] If Z has high variance it will strongly determine selection. To normalize its impact with respect to the basic model the variance is required to be 1.

Table 1: **The Simulation Setup.**

	Distribution		Parameters	
Variable	Basic	Alternative*	Outcome	Selection
Constant	–	–	1	α_0 (adjusted)
X_1	$\mathcal{N}(0,1)$	$\mathcal{U}[-0.5,0.5]$	$\beta_1 = 1$	$\alpha_1 = 1$
X_2	$\mathcal{N}(0,1)$	$\mathcal{U}[-0.5,0.5]$	$\beta_2 = 1$	$\alpha_2 = 1$
X_1X_2	–	–	$\beta_{12} \in \{0,1\}$	0
Y_1	$\mathcal{N}(0,1)$	$\mathcal{U}[-0.5,0.5]$	$\beta_3 = 1$	$\alpha_3 \in \{0, 0.05, 0.10\}$
Y_2	$\mathcal{N}(0,1)$	$\mathcal{U}[-0.5,0.5]$	$\beta_4 = 1$	$\alpha_4 \in \{0, 0.05, 0.10\}$
Y_1Y_2	–	–	$\beta_{34} \in \{0,1\}$	0
Z_1	$\mathcal{N}(0,1)$	$U_1 \exp(-\mu X_1)$	$\beta_5 \in \{0, 0.05, 0.10\}$	$\alpha_5 = 1$
Z_2	$\mathcal{N}(0,1)$	$U_1 \exp(-\mu X_1)$	$\beta_6 \in \{0, 0.05, 0.10\}$	$\alpha_6 = 1$
Z_1Z_2	–	–	$\beta_{56} \in \{0,1\}$	0
X_1Z_1	–	–	0	$\alpha_7 \in \{0,1\}$
X_1Z_2	–	–	0	$\alpha_8 \in \{0,1\}$
X_2Z_1	–	–	0	$\alpha_9 \in \{0,1\}$
X_2Z_2	–	–	0	$\alpha_{10} \in \{0,1\}$
D_i	–	–	δ_i	see below
U_1	–	$\mathcal{U}[0,1]$	0	0
U_2	–	$\mathcal{U}[0,1]$	0	0
ε_i	$\mathcal{N}(0,9)$	$\mathcal{N}(0,9)$	1	1
η_i	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	1	1

Size of the control reservoir $\in \{300, 600, 900\}$
Size of the treatment group: 300
Importance of the deterministic part $\in \{0.25, 0.50, 0.75, 1.00\}$
$\delta_i = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 Y_{1i} + \gamma_4 Y_{2i} + \gamma_5 Z_{1i} + \gamma_6 Z_{2i}$ and
$\gamma \in \{(1,0,0,0,0,0,0), (0.5, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25)\}$
$\mu = 1.35$

* Furthermore, all variables are standardized to have mean zero and variance 1.

demonstrate this issue. The first model consists of linear terms in X only, the second one includes an interaction X_1X_2, and the third one further adds quadratic terms in X.

Other interesting features consider (i) whether asymmetry of the parameters $(\beta_1, \beta_2) = (0.5, 2)$ or (ii) whether interaction terms in the outcome equation as follows

$$\begin{aligned} R &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_3 Y_1 + \beta_4 Y_2 + \beta_{34} Y_1 Y_2 \\ &+ \beta_5 Z_1 + \beta_6 Z_2 + \beta_{56} Z_1 Z_2 + \varepsilon \end{aligned} \tag{9}$$

might cause additional problems. To keep the presentation of the alternative model simple only a certain parameter constellation of the basic model will be considered more closely: a medium impact of Y on R and Z on D, i.e. with coefficients $\alpha_3 = \alpha_4 = \beta_5 = \beta_6 = 0.05$, a medium size of the control reservoir (600), and a selection determination of 0.75. The setup is summarized in table 1.

4 The Matching Algorithm

Consider the basic specification retaining independence between X and Z, with Z having no impact on the outcome R, and Y none on selection D but all other α and β-coefficients are 1, and, furthermore, where there are 600 comparison units. This constellation already motivates the use of the special matching algorithm presented below. The columns under the heading *full probit* of table 2 compare the absolute frequencies of treated and untreated individuals by propensity score intervals. Obviously, the distribution is very unfavorable for matching at the boundaries. In effect, the full probit model successfully separates the treated from the untreated. Unfortunately, high predictive ability of the model implies difficulties in finding adequate controls for high propensity score treated individuals. The

Table 2: **Distribution of Treated and Untreated Individuals.**

Estimated Propensity score	Full Probit		Partial Probit	
	untreated	treated	untreated	treated
$0.0 \leq \hat{p} < 0.1$	459.58	6.80	293.39	12.41
$0.1 \leq \hat{p} < 0.2$	49.59	8.69	135.02	23.67
$0.2 \leq \hat{p} < 0.3$	29.39	9.77	75.58	24.86
$0.3 \leq \hat{p} < 0.4$	19.88	9.86	45.08	23.71
$0.4 \leq \hat{p} < 0.5$	14.29	10.74	25.49	20.54
$0.5 \leq \hat{p} < 0.6$	10.11	12.48	14.16	16.63
$0.6 \leq \hat{p} < 0.7$	7.18	13.47	6.93	12.65
$0.7 \leq \hat{p} < 0.8$	5.22	15.83	3.01	9.11
$0.8 \leq \hat{p} < 0.9$	3.04	18.90	1.01	4.98
$0.9 \leq \hat{p} \leq 1.0$	1.46	43.50	0.07	1.70
Mean propensity score	0.09	0.65	0.15	0.38
Observations	600	150	600	150

The means are averages over 100 iterations. Comparison of number of treated and untreated individuals by certain propensity score intervals.

picture improves substantially if Z (and Y) are omitted from the selection equation. Estimation results of the partial probit are presented in the last two columns of the table. Apparently, the difference in the distributions of the estimated propensity scores for treated and untreated is less extreme than in the full probit. Therefore, matching can be expected to be much easier.

After estimation of individual propensity scores a distance between treated and untreated individuals has to be defined because exact matching on the continuous score is impossible. Here a propensity score caliper approach is pursued (Cochran & Rubin, 1973). A small pool of potential controls is generated for each treated unit by excluding all untreated units whose propensity score distance to the chosen treated exceeds a certain caliper ε. Within the caliper, the distances from treated individual to potential control is defined in terms of the *Mahalanobis metric* based on variables W consisting of the estimated propensity score and all matching covariates, either (X, Y, Z) or X for the full or partial specification, respectively. It is a weighted Euclidean distance $d(w_t, w_c) = (w_t - w_c)'V^{-1}(w_t - w_c)$, where indices t, c represent the treated and the potential control units, respectively. V is the pooled covariance matrix of W which serves to norm the vectors. In sum, the distance is

$$d(w_t, w_c) = \begin{cases} \infty & \text{if } |p_t - p_c| > \varepsilon \\ (w_t - w_c)'V^{-1}(w_t - w_c) & \text{else.} \end{cases} \tag{10}$$

An infinite distance indicates that matching is forbidden.

Matching using the Mahalanobis distance is discussed in Rubin (1980). Gu & Rosenbaum (1993) perform simulations to compare three distance measures. Furthermore, propensity score calipers are discussed in Rosenbaum & Rubin (1985) and Rosenbaum (1989). Calipers help substantially reduce the number of potential controls and, thus, considerably accelerate the matching algorithm and, what is more, they prevent that too distant individuals are being matched. The critical ε is chosen such that there are enough but not too many potential controls in the vicinity of each treated which otherwise would considerably slow down the algorithm without improving results. Table 3 summarizes the choices of the critical ε. The results may depend on the choice of ε. A small ε will come with a loss of many treated (and untreated) individuals. On the other

Table 3: **Specification of Caliper Width ε.**

	Basic Model						Alternative Model		
Selection	Full Probit			Partial Probit			Full	Partial Probit	
Determ.	300	600	900	300	600	900	Probit	Order	
0.25	.030	.020	.010	.015	.010	.005	.03	1	.002
0.50	.040	.030	.020	.030	.015	.010		2	.010
0.75	.050	.040	.030	.040	.020	.010		3	.010
1.00	.060	.050	.040	.050	.025	.010			

The first column of the basic model presents the factors which the α-coefficients of the selection equation are multiplied with. The next columns headed by the size of the control reservoir display the critical ε. The first column of the alternative model shows the caliper width used in the full probit, the second shows whether no interactions (1), interactions (2), and additionally squares (3) are included in the partial probit, and the last displays ε.

hand, however, it increases similarity of the matched units.

The final decision is how to implement the chosen matching criteria, in other words, how the distances between treated units and controls is to be minimized. A stratification producing small strata is preferable in order to ensure that the distance between the units within a stratum is not too large and stratum members are very similar to each other. This yields strata with either one treated and one or more controls or one control and more than one treated unit. It turns out that strata with very high propensity scores contain more than one treated and strata with low scores consist of a large number of controls.

In this study, *optimal full matching* as proposed by Rosenbaum (1991) is implemented. It minimizes the overall distances between treated and controls in that it works backwards and rearranges already matched units if an unmatched treated would better be matched to an already used untreated. In such a case, the existing match is broken up and its treated is available for matching again.[6] The strata will be non-overlapping, i.e. individuals are not members of more than one stratum, which facilitates the calculation of variances.[7] Optimal full matching can easily be transformed into a *minimum cost flow*

[6]This is in contrast to so-called *greedy* algorithms which do not generally achieve a minimum, see Rosenbaum (1991).

[7]Statistical inference is described in Rosenbaum (1995) and adapted to this setup in Augurzky (2000a). However, non-overlapping strata are not necessary if different techniques are used, see Quade (1981) or Heckman, Ichimura & Todd (1998).

problem, a special case of *linear network optimization*.[8] Empirical applications can be found in Augurzky (2000a,b).

Matching produces different strata in terms of number of treated and controls per stratum. Some might be very extreme comprising numerous treated units and only one control. It is they who substantially increase the variance of the estimated mean effect of treatment on the treated. On the other hand, strata with one treated but countless controls will work in the opposite direction but receive less weight. Therefore, an aggregate measure assessing the uniformity of a given stratification with respect to a benchmark stratification is helpful. To this end, suppose all estimated stratum treatment effects have the same variance, the following formula measures *variance inflation* due to unfavorable stratification[9]

$$\frac{1}{(\sum_{s=1}^{S} m_s)^2} \sum_{s=1}^{S} \frac{m_s^2}{(1-1/n_s)^2}$$

where m_s indicates the number of treated units and n_s the number of all individuals in stratum $s = 1, ..., S$.

In order to make the formula meaningful it ought to be compared to a benchmark stratification which is defined as follows. Let all treated units get their own stratum with exactly one control. Therefore, redefine $\tilde{m}_{\tilde{s}} = 1$ and $\tilde{n}_{\tilde{s}} = 2$ for all $\tilde{s} = 1, ..., \tilde{S}$ with $\tilde{S} = \sum_{s=1}^{S} m_s$, yielding a variance inflation of $4/\sum_{s=1}^{S} m_s$. The ratio of the two expressions yields a *relative variance inflation factor* denoted κ^2

$$\kappa^2 = \frac{1}{4\sum_{s=1}^{S} m_s} \sum_{s=1}^{S} \frac{m_s^2}{(1-1/n_s)^2}. \tag{11}$$

For example, pair matching produces $\kappa = 1$, 1-k-matching, i.e. one treated and k controls share a common stratum, leads to $\kappa = 0.5\,(1+1/k)$, k-1-matching has $\kappa = (k+1)/(2\sqrt{k})$. Note that the benchmark stratification can in general never be achieved since all treated who are used in the optimal stratification would have to find an own control. This would only be possible if there are no high propensity score treated units or else if several high propensity score treated individuals were matched to medium score controls which is either ruled out by a caliper approach or which otherwise would compare the incomparable. As

[8] Bertsekas (1991) discusses *linear network optimization* and provides FORTRAN-algorithms for minimum cost flow problems. Furthermore, there is an *operations research* procedure called *netflow* in SAS for these kinds of problems.

[9] See Augurzky (2000a) or Rosenbaum (1995) for the deduction of the general variance formula.

such, κ incorporates neither the balance of covariates after matching nor how many treated units remain unmatched but only the uniformity of the stratification.

As outlined in the introduction, pair matching might be more efficient than full matching. What is more, if the treatment effect is homogeneous pair matching estimates are unbiased. Nevertheless, pair matching is disregarded in this study even in the case of homogeneous effects. The principal aim is to shed more light on the estimation of the propensity score when selection is strong. The homogeneous case is for illustrative purposes only and serves as a valuable benchmark.

Finally, matching should produce balance of all important covariates implying that at least their means for treated and controls be approximately equal. Therefore, to verify balance, simple t-tests of the hypothesis of equal means under equal variances are performed for each of the six variables $j = 1, ..., 6$. If the null hypothesis cannot be rejected at a 5% significance level let $t_j = 1$ and zero otherwise. Then, for an overall measure of balance, define the *aggregate balance* τ as

$$\tau = \frac{\sum_{j=1}^{6} \beta_j \, t_j}{\sum_{j=1}^{6} \beta_j}, \tag{12}$$

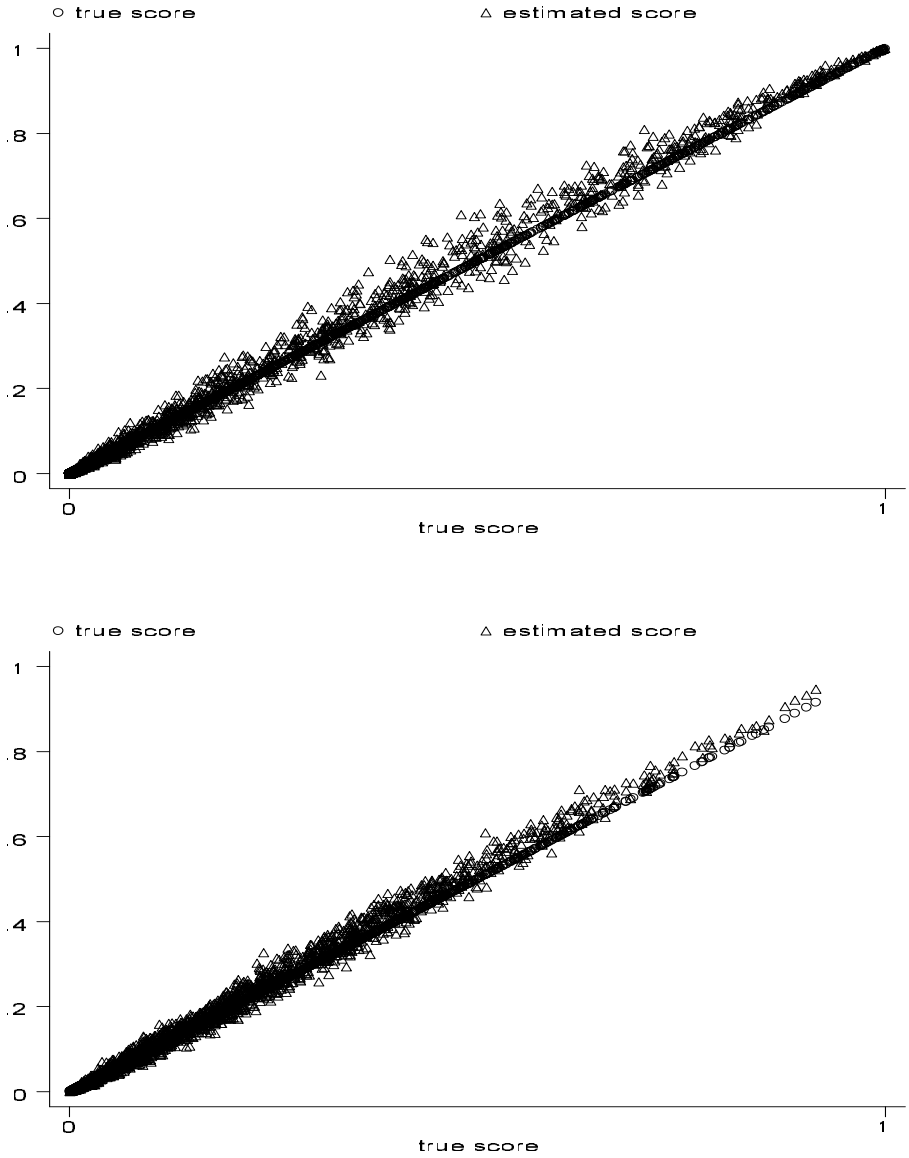
the β's being the coefficients of the outcome equation (5). Weighting by β takes into account that imbalance of the less important variables Z would cause less problems than that of X and Y.[10]

5 Results

Each simulation is performed 100 times and mean estimation results over all iterations are presented and discussed for the parameter constellations mentioned above. Variability across simulations is reflected by simulation standard errors which, however, are not presented in the tables below for reasons of clarity. Figure 1 shows propensity score estimation results of the basic model with true and estimated scores on the vertical axis and the true ones on the horizontal axis. The data are taken from the constellation where

[10]Percent bias reduction has also been examined. Yet, results were quite unsatisfactory because a negative percent bias reduction is basically unbounded. If balance before matching is already given the denominator in the formula is close to zero. On the other hand, percent bias reduction is at most +100%. Therefore, the mean reduction turned out to be rather low in each single iteration.

Figure 1: **Basic Model, Full Probit** and **Partial Probit**



True and estimated propensity scores on the vertical axis versus true scores on the horizontal. The figures represent *one* iteration of the simulation study, the full model on the top, the partial one on the bottom.

the critical coefficients of Y and Z are 0.1, with 600 untreated individuals, and with selection determination of 0.75.

Apparently, the estimates of both the full and the partial specification are unbiased. Tables 4 and 5 go into the estimation and stratification results of the full and partial model. The first three columns characterize the simulation scenario. The first column reports the values of the coefficients $\beta_5, \beta_6, \alpha_3$, and α_4, the second the size of the control reservoir,

and the third shows the factor the α-coefficients of the selection equation are multiplied with. The lower this factor the larger the randomness of the selection process and the less severe self-selection is. The next four columns report the bias and the RMSE in the homogeneous and the heterogeneous case. The remaining columns are self-explanatory.

The most striking result is that matching on the propensity score estimated by the full probit model produces almost always unbiased estimates of the mean effect of treatment on the treated while the bias of the partial probit matching rises to roughly 40% when the impacts of Y and Z are largest. Nevertheless, root mean squared errors of the latter are markedly lower when selection determination is highest. As selection determination successively increases, the full model puts an increasingly heavy burden upon treated individuals in finding appropriate controls, a fact reflected in the diminishing number of strata and the growing number of lost treated individuals. For instance, full probit matching ends with roughly 57 strata if selection determination is highest and the control reservoir is smallest. By contrast, partial probit matching still produces around 101 strata under these circumstances. That is, stratification in the latter case is more uniform as can also be seen from its lower value of κ which never surpasses 0.92 whereas full probit matching even surpasses $\kappa = 1$. However, the difference in RMSE decreases for a more extensive control reservoir.

Furthermore, the partial probit estimates are unbiased if the omitted variables Y and Z do not have an impact on the selection or outcome equation, respectively. However, there is an increasing bias if their impact increases. Note that there appears to be also a weak upward bias in the full probit model if selection determination is highest, specifically in the homogeneous case. This bias arises due to the remaining imbalance expressed by a τ of around 0.6. An additional bias of opposite direction emerges in the heterogeneous case partly offsetting the initial bias. This is because many high propensity score treated units who tend to experience a higher effect in the heterogeneous case are discarded by the matching algorithm. Furthermore, the RMSE in the heterogeneous case seems to be as large as or larger than in the homogeneous case. Yet, it is smaller for high selection determination and for large control reservoir. This finding might be explained by the additional variability of a heterogeneous δ_i. Since δ_i depends on the observable covariates, its variability diminishes as selection caused by the observables becomes more important.

Table 4: **Basic Model, Full Probit.**

Scenario			Effects				Stratification				
			Homogeneous		Heterogeneous		N^o of	Lost	κ	ΔP-	Bal.
(a)	(b)	(c)	Bias	RMSE	Bias	RMSE	strata	tr'd		score	τ
0.00	300	0.25	-0.01	0.37	-0.04	0.50	129.97	3.36	0.84	-0.02	1.00
		0.50	-0.08	0.57	-0.14	0.67	97.15	8.99	0.97	-0.05	0.96
		0.75	0.09	0.79	0.03	0.84	73.58	13.10	1.14	-0.05	0.82
		1.00	0.15	1.15	0.11	1.16	56.49	11.34	1.36	-0.04	0.60
	600	0.25	0.03	0.34	0.01	0.41	143.34	3.22	0.71	-0.03	1.00
		0.50	0.03	0.43	-0.02	0.43	115.43	9.17	0.80	-0.07	0.99
		0.75	-0.01	0.56	-0.05	0.51	89.35	12.04	0.95	-0.06	0.90
		1.00	0.05	0.88	-0.00	0.79	72.28	15.59	1.10	-0.06	0.72
	900	0.25	0.03	0.30	-0.00	0.34	141.79	4.58	0.66	-0.04	1.00
		0.50	-0.00	0.39	-0.05	0.37	122.68	9.96	0.74	-0.08	0.99
		0.75	0.01	0.46	-0.04	0.40	100.08	15.10	0.86	-0.09	0.92
		1.00	-0.01	0.61	-0.06	0.51	81.95	17.82	0.99	-0.09	0.79
0.05	300	0.25	-0.01	0.36	-0.04	0.48	128.32	4.02	0.84	-0.02	1.00
		0.50	0.09	0.52	0.05	0.58	98.90	8.18	0.98	-0.04	0.96
		0.75	0.07	0.78	0.02	0.82	73.53	12.51	1.15	-0.05	0.85
		1.00	0.16	1.15	0.11	1.15	57.15	13.50	1.34	-0.05	0.59
	600	0.25	-0.01	0.35	-0.03	0.42	142.70	3.06	0.71	-0.03	1.00
		0.50	0.02	0.42	-0.03	0.43	116.97	8.19	0.82	-0.06	0.99
		0.75	0.04	0.56	-0.02	0.52	90.68	13.99	0.94	-0.08	0.91
		1.00	-0.02	0.81	-0.08	0.71	72.17	17.17	1.09	-0.07	0.68
	900	0.25	-0.04	0.32	-0.08	0.38	141.98	4.63	0.66	-0.05	1.00
		0.50	0.04	0.38	-0.01	0.35	122.73	10.51	0.74	-0.09	0.99
		0.75	0.09	0.43	0.03	0.36	98.23	14.25	0.86	-0.09	0.92
		1.00	0.10	0.71	0.03	0.57	79.78	18.32	0.99	-0.09	0.79
0.10	300	0.25	0.03	0.38	0.01	0.51	129.60	3.43	0.85	-0.02	1.00
		0.50	0.09	0.52	0.05	0.57	96.04	8.77	0.98	-0.04	0.97
		0.75	0.09	0.80	0.03	0.82	72.43	13.22	1.15	-0.05	0.79
		1.00	0.22	0.93	0.16	0.93	58.69	14.34	1.32	-0.05	0.63
	600	0.25	-0.01	0.36	-0.04	0.45	143.86	3.26	0.71	-0.03	1.00
		0.50	-0.07	0.46	-0.11	0.46	116.51	7.92	0.81	-0.06	0.98
		0.75	0.14	0.55	0.07	0.48	90.84	13.78	0.95	-0.07	0.88
		1.00	0.03	0.69	-0.03	0.61	73.88	18.28	1.07	-0.08	0.75
	900	0.25	0.00	0.30	-0.04	0.34	140.43	4.84	0.66	-0.05	1.00
		0.50	-0.01	0.44	-0.06	0.41	121.32	10.77	0.74	-0.09	1.00
		0.75	0.09	0.48	0.02	0.40	100.14	15.97	0.86	-0.10	0.97
		1.00	0.09	0.68	0.02	0.54	80.22	18.47	0.99	-0.09	0.77

The results are averages over all 100 iterations. The first block represents the scenario: (a) value of the coefficients $\beta_5, \beta_6, \alpha_3, \alpha_4$, (b) size of control reservoir, (c) selection determination. The next block reports bias and RMSE for the homogeneous and the heterogeneous case. The last block shows stratification results: the number of strata and of lost treated units, the stratification measure κ, the difference in true propensity scores of treated units after and before matching, and the aggregate balance τ.

Table 5: **Basic Model, Partial Probit.**

Scenario			Effects				Stratification				
			Homogeneous		Heterogeneous		N° of	Lost	κ	ΔP-	Bal.
(a)	(b)	(c)	Bias	RMSE	Bias	RMSE	strata	tr'd		score	τ
0.00	300	0.25	0.01	0.40	-0.01	0.54	129.07	2.92	0.82	-0.01	0.96
		0.50	-0.05	0.49	-0.07	0.56	117.00	4.36	0.85	-0.02	0.93
		0.75	0.03	0.46	0.01	0.48	107.81	4.24	0.89	-0.01	0.93
		1.00	0.06	0.45	0.05	0.46	101.21	4.46	0.92	-0.01	0.90
	600	0.25	0.04	0.39	0.04	0.47	144.00	2.24	0.69	-0.01	0.95
		0.50	0.05	0.42	0.03	0.42	130.61	5.02	0.71	-0.03	0.94
		0.75	-0.02	0.44	-0.04	0.40	124.04	6.19	0.74	-0.02	0.94
		1.00	-0.01	0.39	-0.02	0.35	120.12	6.36	0.76	-0.02	0.92
	900	0.25	0.00	0.32	-0.01	0.37	143.61	3.16	0.63	-0.02	0.95
		0.50	0.00	0.37	-0.02	0.35	137.36	5.66	0.66	-0.03	0.95
		0.75	0.02	0.35	-0.00	0.30	130.95	8.91	0.69	-0.03	0.94
		1.00	-0.03	0.40	-0.04	0.33	125.78	11.50	0.70	-0.03	0.92
0.05	300	0.25	0.04	0.34	0.04	0.45	128.47	2.95	0.81	-0.01	0.94
		0.50	0.18	0.46	0.19	0.51	116.53	3.52	0.86	-0.01	0.91
		0.75	0.14	0.50	0.13	0.52	108.82	4.67	0.89	-0.01	0.90
		1.00	0.18	0.51	0.16	0.50	101.76	4.48	0.91	-0.01	0.87
	600	0.25	0.04	0.39	0.04	0.47	143.15	2.35	0.68	-0.01	0.94
		0.50	0.12	0.39	0.11	0.39	134.06	5.59	0.72	-0.03	0.91
		0.75	0.17	0.38	0.14	0.34	123.44	6.57	0.74	-0.02	0.91
		1.00	0.18	0.44	0.14	0.38	120.77	6.40	0.75	-0.02	0.87
	900	0.25	0.02	0.31	0.01	0.36	143.88	3.13	0.63	-0.02	0.93
		0.50	0.20	0.42	0.17	0.38	137.84	5.33	0.66	-0.03	0.93
		0.75	0.20	0.42	0.15	0.35	129.33	9.10	0.68	-0.04	0.90
		1.00	0.30	0.48	0.23	0.38	124.34	12.18	0.69	-0.03	0.89
0.10	300	0.25	0.19	0.46	0.24	0.60	129.41	2.85	0.82	-0.01	0.90
		0.50	0.32	0.50	0.34	0.55	115.86	4.19	0.85	-0.02	0.87
		0.75	0.33	0.56	0.32	0.57	107.64	4.46	0.89	-0.01	0.84
		1.00	0.42	0.65	0.41	0.64	103.05	3.53	0.91	-0.01	0.83
	600	0.25	0.14	0.38	0.15	0.46	144.53	2.49	0.68	-0.01	0.92
		0.50	0.24	0.42	0.22	0.40	132.97	4.84	0.71	-0.02	0.86
		0.75	0.41	0.56	0.35	0.49	125.63	6.26	0.74	-0.02	0.84
		1.00	0.44	0.60	0.37	0.51	121.74	6.22	0.76	-0.02	0.82
	900	0.25	0.17	0.35	0.18	0.39	142.11	3.56	0.63	-0.02	0.88
		0.50	0.29	0.48	0.25	0.43	134.63	5.81	0.66	-0.03	0.86
		0.75	0.38	0.52	0.30	0.43	131.07	9.05	0.68	-0.03	0.85
		1.00	0.41	0.57	0.31	0.44	123.81	11.76	0.70	-0.03	0.83

The results are averages over all 100 iterations. The first block represents the scenario: (a) value of the coefficients $\beta_5, \beta_6, \alpha_3, \alpha_4$, (b) size of control reservoir, (c) selection determination. The next block reports bias and RMSE for the homogeneous and the heterogeneous case. The last block shows stratification results: the number of strata and of lost treated units, the stratification measure κ, the difference in true propensity scores of treated units after and before matching, and the aggregate balance τ.

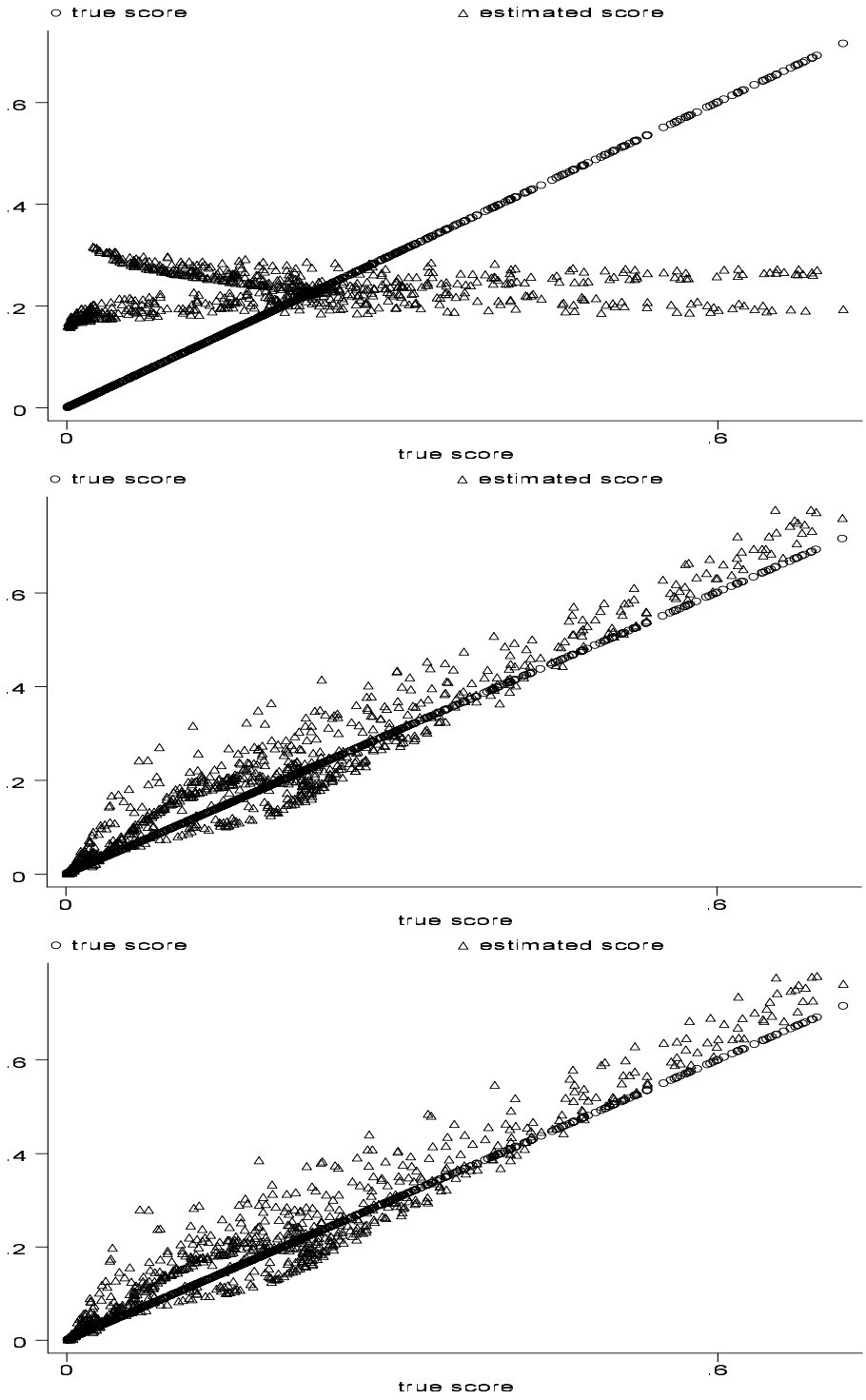
As far as balancing success is concerned, no strategy surpasses the other in all scenarios. If selection determination is weak full probit always achieves perfect balance. However, its performance diminishes quickly as selection determination is growing. On the other hand, partial probit's balancing success starts worse but does not reduce as fast as full probit's. Part of this finding is explicable by the choice of the caliper width ε. It is wider for strong selection (see table 3), hence, treated individuals might choose controls with a relatively low propensity score. For the same reason, τ deteriorates faster in the full than in the partial probit model. However, a constant ε for all scenarios would have produced a large casualty list of treated units in the full model.

In spite of non-constant ε, the full probit loses more treated units such that the relative difference in the *true full* propensity scores between treated individuals before matching and the remaining treated after matching $\Delta P.score$ is more pronounced than in the partial probit. The negative signs show that treated individuals are lost in the high end of the propensity score scale. However, while partial probit matching never exceeds 3%, full probit matching even reaches 10%. Note, however, that the number of lost treated increases with the size of the control reservoir. This counterintuitive result arises because of decreasing caliper widths, see table 3.

In sum, partial probit produces a better overall performance than full probit for the examined parameter constellations. Alas, if the coefficients of Y and Z grew above the 0.1 considered here, full probit could be expected to be the preferred strategy. Moreover, if there is no strong selection into treatment full probit matching is not at a disadvantage, in contrast, it even sometimes outperforms partial probit. Yet, strong selection as in Dehejia & Wahba (1998) or Augurzky (2000a) calls for a careful assessment of the importance of the variables included in the selection equation.

The basic model seems to be overly optimistic as far as the distributions of Y and Z are concerned. The top panel of figure 2 presents propensity score estimates under the alternative partial probit model. Apparently, it underestimates propensity scores for individuals with high $\mathbb{P}(D = 1|X)$ and overestimates for those with low scores. This is in contrast to the next two pictures which present estimated propensity scores built on probit models with higher order terms. Pictures of the full probit are not presented for they are virtually identical to those of the basic model.

Figure 2: **Alternative Model, Partial Probit With and Without Higher Order Terms**



True and estimated propensity scores on the vertical axis versus true scores on the horizontal. The figures represent *one* iteration of the simulation study. The first picture shows results when no interactions are included; the second contains interactions, the third additionally contains squares in X.

Table 6: **Mean Ranks.**

	Full probit		Partial probit	
	untreated	treated	untreated	treated
Basic Model				
True propensity score	313.59	625.75	332.14	550.78
Estimated propensity score	313.28	627.01	331.79	552.19
Alternative Model				
True propensity score	329.87	557.76	335.14	536.65
Estimated, without higher order terms	328.67	562.54	367.40	407.94
Estimated, w/ interactions	328.67	562.54	336.59	530.82
Estimated, w/ inter. & squares	328.67	562.54	335.43	535.44

The results are mean ranks in the treatment and in the comparison group. They are further averaged over all iterations.

One might ask whether the order of treated and untreated units with regard to their estimated biased propensity scores would be similar to the order of individuals in accordance with their true scores $\mathbb{P}(D = 1|X)$. In this case, treated and untreated would hardly change their ranks within the sample. As a result, stratification might be similar to that if the true scores were used for matching and the biased propensity score estimates would not be a source of bias in the matching estimates. However, as illustrated in table 6, the mean rank of treated units has diminished considerably for the alternative model with no higher order terms implying that a large number of untreated and treated units must have interchanged their ranks. Alas, once higher order terms are taken into account – particularly interaction between X_1 and X_2 – there is no difference in mean ranks worth mentioning anymore.

Table 7 presents simulation results for the alternative model in simulation scenario $(0.05, 600, 0.75)$ for the full and the partial probit. Consider first the partial probit results. Surprisingly, they are still better than the comparable ones of the full probit basic model, though worse than those of the partial probit basic model. Interactions in the *outcome equation* (9) lead to an increase of the RMSE and produce a larger bias if no higher order terms in the probit model are accounted for. Yet, this pattern disappears once they are

Table 7: **Alternative Model.**

	Effects				Stratification				
	Homogeneous		Heterogeneous		N° of	Lost	κ	ΔP-	Bal.
	Bias	RMSE	Bias	RMSE	strata	tr'd		score	τ
Full Probit									
Mahalanobis distance within calipers									
– no interactions in outcome equation									
	0.08	0.48	0.05	0.80	118.82	13.51	0.74	-0.12	0.98
– with interactions in outcome equation									
	0.11	0.53	0.11	0.87	118.82	13.51	0.74	-0.12	0.98
Propensity score distance within calipers									
– no interactions in outcome equation									
	0.08	0.51	0.06	0.84	113.16	13.51	0.75	-0.12	0.98
– with interactions in outcome equation									
	0.07	0.56	0.04	0.92	113.16	13.51	0.75	-0.12	0.98
Partial Probit									
Mahalanobis distance within calipers									
– no interactions in outcome equation									
1	0.14	0.43	0.23	0.72	134.37	0.50	0.72	-0.00	0.91
2	0.16	0.47	0.24	0.78	128.72	7.82	0.72	-0.05	0.92
3	0.19	0.49	0.30	0.81	128.16	6.58	0.73	-0.04	0.93
– with interactions in outcome equation									
1	-0.33	0.56	-0.55	0.94	134.37	0.50	0.72	-0.00	0.91
2	-0.06	0.48	-0.13	0.81	128.72	7.82	0.72	-0.05	0.92
3	-0.03	0.48	-0.07	0.80	128.16	6.58	0.73	-0.04	0.93
Propensity score distance within calipers									
– no interactions in outcome equation									
1	0.13	0.42	0.21	0.71	140.85	0.50	0.71	-0.00	0.92
2	0.12	0.49	0.17	0.81	127.86	7.82	0.73	-0.05	0.92
3	0.18	0.47	0.27	0.77	126.94	6.58	0.73	-0.04	0.92
– with interactions in outcome equation									
1	-0.80	0.95	-1.34	1.58	140.85	0.50	0.71	-0.00	0.92
2	-0.10	0.52	-0.21	0.88	127.86	7.82	0.73	-0.05	0.92
3	-0.05	0.46	-0.11	0.78	126.94	6.58	0.73	-0.04	0.92

The means are averages over all 100 iterations for scenario (0.05, 600, 0.75) of table 5. The first column refers to the partial probit model, 1: no higher order terms, 2: interactions, 3: interactions and squares. The first block reports bias and RMSE for the homogeneous and the heterogeneous case. The last block shows stratification results: the number of strata and of lost treated units, the stratification measure κ, the difference in true propensity scores of treated units after and before matching, and the aggregate balance τ. *Interactions in outcome equation* means that $(\beta_{12}, \beta_{34}, \beta_{56}) = (1, 1, 1)$.

included. Asymmetry in the coefficients $(\beta_1, \beta_2) = (0.5, 2)$ instead of $(1, 1)$ of the response equation does not at all alter the results which is why they are omitted. In contrast to the basic model, heterogeneous effects lead to substantially worse estimation results in that biases and RMSEs are markedly larger than in the homogeneous case.

These still surprisingly favorable results in spite of severe misspecifications expressed in the first picture of figure 2 might be explained by the fact that within the propensity score calipers the Mahalanobis distance, which is not misspecified, still matches the correct individuals. To explore this hypothesis all results are repeated replacing the Mahalanobis distance by the propensity score distance within calipers. The results are also shown in table 7. They are fairly similar to the previous results with one notable exception: the bias and RMSE are markedly larger in case interactions in the response model are introduced but none in the probit model.

For the sake of comparability, the table displays estimates of the full probit model, as well. The most striking result is that it achieves an almost perfect overall balance τ. This unexpected finding, however, may partly be explained by the fact that a considerable number of high propensity score treated units is lost facilitating balancing the variables of the remaining sample. As a result, the superior balance is accompanied by an unfavorably $\Delta P.score$ of 12% making the matched sample less representative. Similarly, κ is almost as small as in the partial probit model because it merely reports uniformity of the realized stratification given the number of lost treated units. Finally, using a propensity score distance within calipers does not alter the results except for slightly increased RMSE. In sum, the partial probit does not do worse than the full probit even if the partial probit model is severely misspecified. Including higher order terms into the selection equation might be a way to alleviate problems caused by omission of variables which are correlated with the included ones.

6 Conclusion

This paper investigates propensity score matching when selection into treatment is remarkably strong and thus the treatment and comparison group differ considerably in

their observable covariates. In such a scenario, matching adequate units is demanding. To alleviate this problem, we suggest to carefully reconsider the selection equation with respect to variables that might play a subordinate role in the outcome equation. Omission of these variables helps increase the randomness of the selection process and reduce the variance of the matching estimates. However, their omission from the selection equation might lead to inconsistent propensity score estimates and hence biased matching estimates. This study assesses the bias-variance trade-off in a simulation resting on the mean squared error criterion.

To this end, we presuppose existence of variables Z which strongly influence the selection decision but which, on the other hand, do not or do only weakly determine the outcome under scrutiny. For a large enough sample size, specification tests of the probit model would then recommend the inclusion of Z to consistently estimate the propensity score. Likewise, we introduce variables Y which are relevant to the outcome but irrelevant to the participation decision. Matching on a propensity score estimate based on Z and Y will balance Z at the expense of balance of the variables most relevant for both the outcome and the selection. Moreover, unnecessary effort is spent to remove small imbalance in the variable Y. In consequence, (i) some treated have to be systematically discarded from the sample because they do not find adequate controls and, (ii) more treated have to share one control, a fact that reduces uniformity of the stratification and thus increases standard errors.

In effect, the results show that matching on inconsistent estimates of the propensity score, i.e. those achieved when Z (and Y) are excluded, produces estimation results of the mean effect of treatment that are often better in terms of the RMSE than those achieved by matching on estimates that rest on all covariates relevant for the selection. This remains true even if Z shows some impact on the outcome as long as this impact is limited. DRAKE (1993) points to a similar direction in concluding that misspecifying the propensity score results in smaller biases than misspecifying the response model. Therefore, we recommend to only include variables into the selection equation that are highly significant. Variables with low significance levels are obvious candidates for exclusion even if they might play a role in the outcome equation. Moreover, if established research suggests that certain variables Z are irrelevant to the outcome under study they should solely be included into

the selection equation if there are other strong reasons for doing so.

If, nevertheless, imbalance of some variables seems to be inacceptable after matching, an additional linear regression adjustment might be pursued with presumably less cost than balancing all the remaining variables in advance. If misspecification of the propensity score seems to be inacceptable, one might additionally take account of statistically significant higher order terms of those variables included in the selection equation. A sensitivity analysis that compares partial models with the full model might be a way to assess different approaches, see e.g. Heckman, Ichimura & Todd (1997: section 13) or Augurzky, (2000b). In sum, the main criterion of success for matching remains the balance of the relevant covariates and not the proper estimation of the selection equation. This aim is easily obtained by a full probit model only if selection determination is low and/or the control reservoir is large but in several applied situations it might be better obtained by a partial model.

References

Augurzky, Boris (2000a) "Matching the Extremes – A Sensitivity Analysis Based on Real Data", unpublished manuscript, Heidelberg.

Augurzky, Boris (2000b) "Evaluating the Effect of Postsecondary Education", unpublished manuscript, Heidelberg.

Angrist, Joshua D. & Jinyong Hahn (1999) "When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects", *NBER Technical Working Paper* no. 241.

Bertsekas, Dimitri B. (1991) *Linear Network Optimization: Algorithms and Codes*, Cambridge MA, MIT Press.

Cochran, W.G. & Donald B. Rubin (1973) "Controlling Bias in Observational Studies: A Review", *Sankhyā*, Series A, 35: 417-46.

Dehejia, Rajeev H. & Sadek Wahba (1998) "Propensity Score Matching Methods for Non-Experimental Causal Studies", *NBER Working Paper*, 6829.

Drake, Christiana (1993) "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect", *Biometrics*, 49: 1231-1236.

Gu, X. Sam & Rosenbaum, Paul R. (1993) "Comparison of multivariate matching methods: Structures, distances and algorithms", *Journal of Computational and Graphical Statistics*, 2: 405-20.

Hahn, Jinyong (1998) "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects", *Econometrica*, 66: 315-31.

Heckman, James J. (1996) "Randomization As An Instrumental Variable", *Review of Economics and Statistics*, 77(2): 336-41.

Heckman, James J., Hidehiko Ichimura & Petra Todd (1997) "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program", *Review of Economic Studies*, 64: 605-54.

Heckman, James J., Hidehiko Ichimura & Petra Todd (1998) "Matching as an Econometric Evaluation Estimator: Theory and Methods", *Review of Economic Studies*, 65: 261-94.

Heckman, James J., Robert J. Lalonde & Jeffrey Smith (1999) "The Economics and Econometrics of Active Labor Market Programs", *Handbook of Labor Economics*, Chapter 31, Volume 3, edited by Orely Ashenfelter and David Card. New York, NY: North-Holland.

Holland, Paul W. (1986) "Statistics and Causal Inference (with discussion)", *Journal of the American Statistical Association*, 81: 945-70.

Lechner, Michael (1999) "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification", *Journal of Business and Economic Statistics*, 17/1: 74-90.

Lechner, Michael (2000) "An Evaluation of Public Sector Sponsored Continuous Vocational Training Programs in East Germany", *The Journal of Human Resources*, 35: 347-75.

Quade, Dana (1981) "Nonparametric Analysis of Covariance by Matching", *Biometrics*, 38: 597-611.

Rosenbaum, Paul R. & Donald B. Rubin (1983) "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70: 41-55.

Rosenbaum, Paul R. & Donald B. Rubin (1985) "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score", *The American Statistician*, 39: 33-38.

Rosenbaum, Paul R. (1989) "Optimal Matching for Observational Studies", *Journal of the American Statistical Association*, 84: 1024-32.

Rosenbaum, Paul R. (1991) "A Characterization of Optimal Designs for Observational Studies", *Journal of the Royal Statistical Association*, Series B, 53: 597-610.

Rosenbaum, Paul R. (1995) *Observational Studies*, New York: Springer Series in Statistics.

Roy, Andrew D. (1951) "Some Thoughts on the Distribution of Earnings", *Oxford Economic Papers*, 3: 135-46.

Rubin, Donald B. (1974) "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 66: 688-701.

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

Rubin, Donald B. (1980) "Bias Reduction Using Mahalanobis Metric Matching", *Biometrics*, 36: 293-98.

Rubin, Donald B. (1986) "What Ifs Have Causal Answers?", *Journal of the American Statistical Association*, 81: 961-62.

Schmidt, Christoph M. (1999) "Knowing What Works: The Case for Rigorous Program Evaluation", IZA Discussion Paper no. 77.

Schmidt, Christoph M., Rob Baltussen & Rainer Sauerborn (1999) "Evaluation of Community-Based Interventions: Group-Randomization, Limits and Alternatives", *Discussion paper series* no. 281, Department of Economics, University of Heidelberg.

Yatchew, Adonis & Zvi Griliches (1984) "Specification Error in Probit Models", *Review of Economics and Statistics*, 66: 134-39.

IZA Discussion Papers

No	Author(s)	Titel	Area	Date
181	E. Wasmer Y. Zenou	Space, Search and Efficiency	2	8/00
182	M. Fertig C. M. Schmidt	Discretionary Measures of Active Labor Market Policy: The German Employment Promotion Reform in Perspective	6	8/00
183	M. Fertig C. M. Schmidt	Aggregate-Level Migration Studies as a Tool for Forecasting Future Migration Streams	1	8/00
184	M. Corak B. Gustafsson T. Österberg	Intergenerational Influences on the Receipt of Unemployment Insurance in Canada and Sweden	3	8/00
185	H. Bonin K. F. Zimmermann	The Post-Unification German Labor Market	4	8/00
186	C. Dustmann	Temporary Migration and Economic Assimilation	1	8/00
187	T. K. Bauer M. Lofstrom K. F. Zimmermann	Immigration Policy, Assimilation of Immigrants and Natives' Sentiments towards Immigrants: Evidence from 12 OECD-Countries	1	8/00
188	A. Kapteyn A. S. Kalwij A. Zaidi	The Myth of Worksharing	5	8/00
189	W. Arulampalam	Is Unemployment Really Scarring? Effects of Unemployment Experiences on Wages	3	8/00
190	C. Dustmann I. Preston	Racial and Economic Factors in Attitudes to Immigration	1	8/00
191	G. C. Giannelli C. Monfardini	Joint Decisions on Household Membership and Human Capital Accumulation of Youths: The role of expected earnings and local markets	5	8/00
192	G. Brunello	Absolute Risk Aversion and the Returns to Education	5	8/00
193	A. Kunze	The Determination of Wages and the Gender Wage Gap: A Survey	5	8/00
194	A. Newell F. Pastore	Regional Unemployment and Industrial Restructuring in Poland	4	8/00
195	F. Büchel A. Mertens	Overeducation, Undereducation, and the Theory of Career Mobility	5	9/00

196	J. S. Earle K. Z. Sabirianova	Equilibrium Wage Arrears: A Theoretical and Empirical Analysis of Institutional Lock-In	4	9/00
197	G. A. Pfann	Options to Quit	1	9/00
198	M. Kreyenfeld C. K. Spiess G. G. Wagner	A Forgotten Issue: Distributional Effects of Day Care Subsidies in Germany	3	9/00
199	H. Entorf	Rational Migration Policy Should Tolerate Non-Zero Illegal Migration Flows: Lessons from Modelling the Market for Illegal Migration	1	9/00
200	T. Bauer G. S. Epstein I. N. Gang	What are Migration Networks?	1	9/00
201	T. J. Dohmen G. A. Pfann	Worker Separations in a Nonstationary Corporate Environment	1	9/00
202	P. Francois J. C. van Ours	Gender Wage Differentials in a Competitive Labor Market: The Household Interaction Effect	5	9/00
203	J. M. Abowd F. Kramarz D. N. Margolis T. Philippon	The Tail of Two Countries: Minimum Wages and Employment in France and the United States	5	9/00
204	G. S. Epstein	Labor Market Interactions Between Legal and Illegal Immigrants	1	10/00
205	A. L. Booth M. Francesconi J. Frank	Temporary Jobs: Stepping Stones or Dead Ends?	1	10/00
206	C. M. Schmidt R. Baltussen R. Sauerborn	The Evaluation of Community-Based Interventions: Group-Randomization, Limits and Alternatives	6	10/00
207	C. M. Schmidt	Arbeitsmarktpolitische Maßnahmen und ihre Evaluierung: eine Bestandsaufnahme	6	10/00
208	J. Hartog R. Winkelmann	Dutch Migrants in New Zealand: Did they Fare Well?	1	10/00
209	M. Barbie M. Hagedorn A. Kaul	Dynamic Effciency and Pareto Optimality in a Stochastic OLG Model with Production and Social Security	3	10/00
210	T. J. Dohmen	Housing, Mobility and Unemployment	1	11/00
211	A. van Soest M. Das X. Gong	A Structural Labour Supply Model with Nonparametric Preferences	5	11/00

212	X. Gong A. van Soest P. Zhang	Sexual Bias and Household Consumption: A Semiparametric Analysis of Engel Curves in Rural China	5	11/00
213	X. Gong A. van Soest E. Villagomez	Mobility in the Urban Labor Market: A Panel Data Analysis for Mexico	1	11/00
214	X. Gong A. van Soest	Family Structure and Female Labour Supply in Mexico City	5	11/00
215	J. Ermisch M. Francesconi	The Effect of Parents' Employment on Children's Educational Attainment	5	11/00
216	F. Büchel	The Effects of Overeducation on Productivity in Germany — The Firms' Viewpoint	5	11/00
217	J. Hansen R. Wahlberg	Occupational Gender Composition and Wages in Sweden	5	11/00
218	C. Dustmann A. van Soest	Parametric and Semiparametric Estimation in Models with Misclassified Categorical Dependent Variables	1	11/00
219	F. Kramarz T. Philippon	The Impact of Differential Payroll Tax Subsidies on Minimum Wage Employment	5	11/00
220	W. A. Cornelius E. A. Marcelli	The Changing Profile of Mexican Migrants to the United States: New Evidence from California and Mexico	1	12/00
221	C. Grund	Wages as Risk Compensation in Germany	5	12/00
222	W.P.M. Vijverberg	Betit: A Family That Nests Probit and Logit	7	12/00
223	M. Rosholm M. Svarer	Wages, Training, and Job Turnover in a Search-Matching Model	1	12/00
224	J. Schwarze	Using Panel Data on Income Satisfaction to Estimate the Equivalence Scale Elasticity	3	12/00
225	L. Modesto J. P. Thomas	An Analysis of Labour Adjustment Costs in Unionized Economies	1	12/00
226	P. A. Puhani	On the Identification of Relative Wage Rigidity Dynamics: A Proposal for a Methodology on Cross-Section Data and Empirical Evidence for Poland in Transition	4/5	12/00

227	L. Locher	Immigration from the Eastern Block and the former Soviet Union to Israel: Who is coming when?	1	12/00
228	G. Brunello S. Comi C. Lucifora	The College Wage Gap in 10 European Countries: Evidence from Two Cohorts	5	12/00
229	R. Coimbra T. Lloyd-Braga L. Modesto	Unions, Increasing Returns and Endogenous Fluctuations	1	12/00
230	L. Modesto	Should I Stay or Should I Go? Educational Choices and Earnings: An Empirical Study for Portugal	5	12/00
231	G. Saint-Paul	The Economics of Human Cloning	5	12/00
232	E. Bardasi M. Francesconi	The Effect of Non-Standard Employment on Mental Health in Britain	5	12/00
233	C. Dustmann C. M. Schmidt	The Wage Performance of Immigrant Women: Full-Time Jobs, Part-Time Jobs, and the Role of Selection	1	12/00
234	R. Rotte M. Steininger	Sozioökonomische Determinanten extremistischer Wahlerfolge in Deutschland: Das Beispiel der Europawahlen 1994 und 1999	3	12/00
235	W. Schnedler	Who gets the Reward? An Empirical Exploration of Bonus Pay and Task Characteristics	5	12/00
236	R. Hujer M. Caliendo	Evaluation of Active Labour Market Policy: Methodological Concepts and Empirical Estimates	6	12/00
237	S. Klasen I. Woolard	Surviving Unemployment without State Support: Unemployment and Household Formation in South Africa	3	12/00
238	R. Euwals A. Börsch-Supan A. Eymann	The Saving Behaviour of Two Person Households: Evidence from Dutch Panel Data	5	12/00
239	F. Andersson K. A. Konrad	Human Capital Investment and Globalization in Extortionary States	5	01/01
240	W. Koeniger	Labor and Financial Market Interactions: The Case of Labor Income Risk and Car Insurance in the UK 1969-95	5	01/01

241	W. Koeniger	Trade, Labor Market Rigidities, and Government-Financed Technological Change	2	01/01
242	G. Faggio J. Konings	Job Creation, Job Destruction and Employment Growth in Transition Countries in the 90's	4	01/01
243	E. Brainerd	Economic Reform and Mortality in the Former Soviet Union: A Study of the Suicide Epidemic in the 1990s	4	01/01
244	S. M. Fuess, Jr. M. Millea	Pay and Productivity in a Corporatist Economy: Evidence from Austria	5	01/01
245	F. Andersson K. A. Konrad	Globalization and Human Capital Formation	5	01/01
246	E. Plug W. Vijverberg	Schooling, Family Background, and Adoption: Does Family Income Matter?	5	01/01
247	E. Plug W. Vijverberg	Schooling, Family Background, and Adoption: Is it Nature or is it Nurture?	5	01/01
248	P. M. Picard E. Toulemonde	The Impact of Labor Markets on Emergence and Persistence of Regional Asymmetries	2	01/01
249	B. M. S. van Praag P. Cardoso	"Should I Pay for You or for Myself?" The Optimal Level and Composition of Retirement Benefit Systems	3	01/01
250	T. J. Hatton J. G. Williamson	Demographic and Economic Pressure on Emigration out of Africa	1	01/01
251	R. Yemtsov	Labor Markets, Inequality and Poverty in Georgia	4	01/01
252	R. Yemtsov	Inequality and Income Distribution in Georgia	4	01/01
253	R. Yemtsov	Living Standards and Economic Vulnerability in Turkey between 1987 and 1994	4	01/01
254	H. Gersbach A. Schniewind	Learning of General Equilibrium Effects and the Unemployment Trap	3	02/01
255	H. Gersbach A. Schniewind	Product Market Reforms and Unemployment in Europe	3	02/01
256	T. Boeri H. Brücker	Eastern Enlargement and EU-Labour Markets: Perceptions, Challenges and Opportunities	2	02/01

257	T. Boeri	Transition with Labour Supply	4	02/01
258	M. Rosholm K. Scott L. Husted	The Times They Are A-Changin': Organizational Change and Immigrant Employment Opportunities in Scandinavia	1	02/01
259	A. Ferrer-i-Carbonell B. M.S. van Praag	Poverty in the Russian Federation	4	02/01
260	P. Cahuc F. Postel-Vinay	Temporary Jobs, Employment Protection and Labor Market Performance	1/3	02/01
261	M. Lindahl	Home versus School Learning: A New Approach to Estimating the Effect of Class Size on Achievement	5	02/01
262	M. Lindahl	Summer Learning and the Effect of Schooling: Evidence from Sweden	5	02/01
263	N. Datta Gupta N. Smith	Children and Career Interruptions: The Family Gap in Denmark	5	02/01
264	C. Dustmann	Return Migration, Wage Differentials, and the Optimal Migration Duration	1	02/01
265	M. Rosholm M. Svarer	Structurally Dependent Competing Risks	1	02/01
266	C. Dustmann O. Kirchkamp	The Optimal Migration Duration and Activity Choice after Re-migration	1	02/01
267	A. Newell	The Distribution of Wages in Transition Countries	4	03/01
268	A. Newell B. Reilly	The Gender Pay Gap in the Transition from Communism: Some Empirical Evidence	4	03/01
269	H. Buddelmeyer	Re-employment Dynamics of Disabled Workers	3	03/01
270	B. Augurzky C. M. Schmidt	The Evaluation of Community-Based Interventions: A Monte Carlo Study	6	03/01
271	B. Augurzky C. M. Schmidt	The Propensity Score: A Means to An End	6	03/01