RWI : Discussion Papers

Boris Augurzky and Jochen Kluve

Assessing the performance of matching algorithms when selection into treatment is strong

No. 21

RWI
ESSEN

Rheinisch-Westfälisches Institut für Wirtschaftsforschung

Board of Directors:
Prof. Dr. Christoph M. Schmidt, Ph.D. (President),
Prof. Dr. Thomas K. Bauer
Prof. Dr. Wim Kösters

Governing Board:
Dr. Eberhard Heinke (Chairman);
Dr. Dietmar Kuhnt, Dr. Henning Osthues-Albrecht, Reinhold Schulte (Vice Chairmen);
Prof. Dr.-Ing. Dieter Ameling, Manfred Breuer, Christoph Dänzer-Vanotti, Dr. Hans Georg Fabritius, Prof. Dr. Harald B. Giesel, Karl-Heinz Herlitschke, Dr. Thomas Köster, Hartmut Krebs, Tillmann Neinhaus, Dr. Günter Sandermann, Dr. Gerd Willamowski

Advisory Board:
Prof. David Card, Ph.D., Prof. Dr. Clemens Fuest, Prof. Dr. Walter Krämer, Prof. Dr. Michael Lechner, Prof. Dr. Till Requate, Prof. Nina Smith, Ph.D., Prof. Dr. Harald Uhlig, Prof. Dr. Josef Zweimüller

Honorary Members of RWI Essen
Heinrich Frommknecht, Prof. Dr. Paul Klemmer

RWI : Discussion Papers
No. 21

Published by Rheinisch-Westfälisches Institut für Wirtschaftsforschung, Hohenzollernstrasse 1/3, D-45128 Essen, Phone +49 (0) 201/81 49-0

All rights reserved. Essen, Germany, 2004

Editor: Prof. Dr. Christoph M. Schmidt, Ph.D.
ISSN 1612-3565 – ISBN 3-936454-35-3

The working papers published in the Series constitute work in progress circulated to stimulate discussion and critical comments. Views expressed represent exclusively the authors' own opinions and do not necessarily reflect those of the RWI Essen.

RWI : Discussion Papers
No. 21

Boris Augurzky and Jochen Kluve

Assessing the performance of matching algorithms when selection into treatment is strong

RWI
ESSEN

Bibliografische Information Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

ISSN 1612-3565
ISBN 3-936454-35-3

Assessing the performance of matching algorithms when selection into treatment is strong

Boris Augurzky and Jochen Kluve*

Abstract

This paper investigates the method of matching regarding two crucial implementation choices, the distance measure and the type of algorithm. We implement optimal full matching – a fully efficient algorithm – and present a framework for statistical inference. The implementation uses data from the NLSY79 to study the effect of college education on earnings. We find that decisions regarding the matching algorithm depend on the structure of the data: In the case of strong selection into treatment and treatment effect heterogeneity a full matching seems preferable. If heterogeneity is weak, pair matching suffices.

JEL Classification: C14, C61

Keywords: Matching algorithms, optimal full matching, selection into treatment

---

*Both RWI Essen and IZA-Bonn. All correspondence to Boris Augurzky, Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI), Hohenzollernstraße 1-3, D-45128 Essen, Germany; Fax: ++49 201 81 49 200; e-mail: augurzky@rwi-essen.de. – This version: August 31st, 2004. We thank Martin Biewen, Michael Lechner, and Jeffrey Smith for valuable comments, and the Bureau of Labor Statistics and the Center for Human Resource Research, Ohio State University, for their quick and valuable aid concerning the NLSY data. We are especially grateful to Christoph M. Schmidt for many useful discussions. Thanks go to Maria Delgado Vazquez for assistance in text processing. All remaining errors are our own. Financial support by the Friedrich-Ebert-Stiftung is gratefully acknowledged.

1 Introduction

Much research in applied econometrics is concerned with the assessment of causal effects of policy interventions, or, generically, the effects of a specific "treatment". In the absence of experimental data, which is largely the case, alternative identification strategies have to be found. In recent years the statistical technique of matching has found widespread attention and has become a particularly popular tool for the evaluation of treatments in observational studies. Complementing the fundamental and continuing work on matching in the statistics literature (see e.g. Cochran 1965, Rubin 1974, 1977, Rosenbaum and Rubin 1983, 1984, Rosenbaum 1995, Rubin and Thomas 1996, Ming and Rosenbaum 2000, etc.), the econometrics literature has discussed matching methods extensively in both empirical and theoretical work.[2]

Matching is an intuitively appealing technique for assessing causal effects because of its main feature of mimicking a randomized experiment *ex post*. The technique is valid if, in statistics parlance, there exists only "overt bias" between treatment and control groups, i.e. - in econometric terms - "selection is on observables". Common wordings for this central assumption are "unconfoundedness" (see e.g. Imbens 2004), "ignorable treatment assignment" (Rosenbaum and Rubin 1983), and "conditional independence assumption" (Lechner 1999). In implementing matching estimators, however, many decisions – some of them *ad hoc* in nature - have to be made, and frequently it is not entirely clear if and how treatment effect estimates will be affected by these decisions. Recent research has addressed a variety of questions in this regard, such as (i) efficiency issues, (ii) the general applicability of matching methods, and (iii) a set of more specific issues regarding practical implementation.

First, for instance, Angrist and Hahn (2004) study efficiency comparisons of covariate matching with propensity score matching and show that the former may be more efficient in finite samples than the latter. They also suggest that propensity score matching is to be preferred when cell sizes are small, the explanatory power of the covariates is low, or the treatment probability is close to 0 or 1.

[2]Cf., for instance, Hahn (1998) and Hirano, Imbens and Ridder (2003) for efficiency issues. Angrist and Krueger (1999) and Heckman, LaLonde and Smith (1999) contain overviews of matching estimators in the labor economics context. Applications using propensity score matching techniques are manifold, cf. for instance Lechner (1999), Dehejia and Wahba (1999), Heckman, Ichimura and Todd (1997), and - for applications using exact covariate matching - Angrist (1998), Kluve, Lehmann and Schmidt (1999). In particular, cf. articles in a recent symposium on the econometrics of matching in *The Review of Economics and Statistics* (2004, Vol. 86, No. 1, pp. 1-194). As part of the symposium, Imbens (2004) provides a comprehensive review of the assumptions necessary for consistent estimation under matching, alternative estimands, efficiency issues, alternative methods developed for matching, variance estimation issues, testing the plausibility of the identifying assumption, and of the applications and simulation studies to date.

Second, the general issue regarding the capacity of matching estimators to produce the "true" effect - which can be analyzed e.g. by comparing the performance of matching against known experimental results, a technique going back to LaLonde (1986) - is discussed at length in the debate between Dehejia/Wahba and Smith/Todd (cf. Dehejia and Wahba 1999, 2002, Smith and Todd 2004a, 2004b, and Dehejia 2004). It is clear from the debate that matching is a useful econometric tool for policy evaluation, but does not represent a general solution to the evaluation problem under each and every circumstance. Whereas the question remains whether such a general issue could have been resolved at all using a relatively small and uninformative data set, the debate covers many essential pros and cons regarding matching methods and, moreover, gives important guidance regarding their implementation: For instance, Dehejia and Wahba (1999) discuss proper specification of the propensity score, and Smith and Todd (2004a) elucidate a whole set of different matching estimators, such as nearest neighbor matching, kernel and local linear matching, and difference-in-differences matching.

Third, Zhao (2004) compares propensity score matching methods with covariate matching estimators by discussing data requirements and studying small-sample properties through Monte Carlo experiments. The paper finds that propensity score matching performs well when correlations between covariates and the participation indicator are high, but does not perform well relative to other matching estimators when sample size is too small. Zhao (2004) also discusses different matching metrics, finding that Mahalanobis matching is relatively robust under different settings.

A simulation study by Gu and Rosenbaum (1993) concludes that matching using the propensity score distance is better at producing balanced samples than matching using the Mahalanobis metric if the overt bias is large and there are many covariates (i.e. matching is "more difficult", cf. Gu and Rosenbaum 1993). The study also discusses matching algorithms and matching "structure": The algorithms considered are nearest neighbor matching, which is a so-called "greedy" algorithm, implying that it will not minimize total distance within matched pairs, and an "optimal" matching minimizing that total distance.[3] Optimal matching, which is better than greedy matching at producing close matches *per definitionem*, in the simulation turns out to produce marginally up to noticeably better matches, though it is no better at producing covariate balance in matched samples. The matching "structures"

[3]In general, cf. Rosenbaum (1995, p211), a "greedy" algorithm is an algorithm that divides a large decision problem into a series of simpler decisions each of which is handled optimally, and makes those decisions one at a time without reconsidering early decisions as later ones are made. Hence, greedy algorithms do solve a small class of problems optimally, but the matching problem is not a member of that class (in terms of total distance minimization). See our discussion in section 3.

compared by Gu and Rosenbaum (1993) regard 1-k matching, i.e. one treated unit is matched to k controls, and full matching, i.e. a treated unit may have one or more controls or a control may have one or more treated units, using all observations in the sample. The study juxtaposes optimal 1-k matching and optimal full matching and finds that optimal full matching is often much better.

In general, matching estimation is characterized by the type of algorithm and the distance measure chosen. The choice of matching algorithm and the distance measure depends on data characteristics and involves an inherent trade-off between bias and variance. The variance of the estimator depends on the uniformity of the matched sample: When uniformity is highest, variance is lowest. Furthermore, two kinds of biases arise, a bias due to lack of balance of the covariates and a bias due to loss of treated individuals after matching.

Two data features determine the relevance of these estimation problems. First, the strength of selection into treatment and second, the degree of treatment effect heterogeneity. In case of weak selection into treatment matching algorithms will in general achieve high uniformity of matched samples, and sample variance is not very important. In case of homogeneous treatment effects there is no bias due to loss of treated units, whereas in the case of heterogeneous effects each treated unit carries information on the treatment effect and loss of treated units might increase the estimation bias. In all cases, however, bias due to imbalance of covariates cannot be neglected. Depending on these data characteristics and the bias-variance trade-off a matching technique has to be chosen.

In this paper, we investigate sensitivity of treatment effect estimates regarding the choice of matching technique, i.e. the choice of distance measure and type of algorithm. We analyze these issues using data from the National Longitudinal Survey of Youths 1979 (NLSY) to study the effect of college education on labor market earnings. The data exhibit strong selection into treatment, aggravating matching at the top and bottom of the propensity score distribution.

The distance measures we discuss are propensity score, index score, and Mahalanobis distance. We implement, to our knowledge for the first time in the context of labor econometrics, an optimal full matching, and compare it with greedy full matching and greedy pair matching. So far, optimal full matching has not received much attention in the applied literature, perhaps due to the fact that fully efficient matching methods are considered computationally cumbersome such that other methods have prevailed, as observed by Imbens (2004).

The paper is structured as follows. The next section presents the methodological

framework of the matching approach. Section 3 outlines the matching algorithms. Section 4 describes the data and section 5 presents empirical results. Section 6 concludes.

2 The Matching Approach

Interest lies in estimating the effect of a binary treatment, e.g. participation in labor market training or holding a bachelor's degree, on a corresponding outcome variable (response), such as labor market performance expressed through employment probability or earnings.

Let Y_i^1 denote the potential response of individual i being exposed to the treatment and Y_i^0 the potential response if i receives no treatment. Furthermore, let T_i be a binary variable indicating treatment status. Then, $Y_i = T_i Y_i^1 + (1 - T_i) Y_i^0$ gives the observed outcome. This framework has become known as the *potential outcome approach to causality* (cf. Rubin 1974, 1977, Holland 1986, Kluve 2004). To identify the causal effect of treatment, 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 for each individual, Y_i^0 and Y_i^1, corresponding to treatment states $T_i = 0$ and $T_i = 1$, respectively. There are no further potential outcomes depending on the treatment 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 given by $\delta_i = Y_i^1 - Y_i^0$ and is never observable since either Y_i^1 or Y_i^0 is missing at the unit level. Still, the essential conceptual point is that each individual has two potential outcomes associated with herself. As individual treatment effects are never observable, interest usually lies in an appropriate summary measure. Two parameters have received particular interest in the literature, the *average treatment effect for the population* and *the average treatment effect on the treated*. Cf. Imbens (2004) for further discussion of these parameters and alternative estimands.

The average treatment effect for the population is given by

$$\tau_P = \mathbb{E}(\delta_i) = E(Y_i^1 - Y_i^0). \tag{1}$$

It is generally not identified from observational data since $\mathbb{E}(Y^1)$ is not observed for the subpopulation with $T_i = 0$ and $\mathbb{E}(Y^0)$ is not observed for the subpopulation with $T_i = 1$. Alternatively one might focus on the average effect of treatment on

the treated individuals given by

$$\tau_T = \mathbb{E}(\delta_i|T_i = 1) = \mathbb{E}(Y_i^1|T_i = 1) - \mathbb{E}(Y_i^0|T_i = 1). \tag{2}$$

Again, while the first expectation $\mathbb{E}(Y_i^1|T_i = 1)$ can be identified for the treatment group subsample, the counterfactual expectation $\mathbb{E}(Y_i^0|T_i = 1)$ is not identifiable without invoking further assumptions.

The subsequent formal setup follows Rosenbaum (1995) and first focuses on the ideal case of a randomized experiment, and then considers the case of nonexperimental data. Assume that N units under observation are being stratified into S strata on the basis of their covariates X_i. Let T_{si} indicate whether unit i in stratum s, $s = 1, ..., S$, is randomly assigned to treatment ($T_{si} = 1$) or not ($T_{si} = 0$). Each stratum s comprises n_s units, $m_s = \sum_{i=1}^{n_s} T_{si}$ treated and $n_s - m_s$ controls.[4] Furthermore, let $\mathbf{T}_s$ be the vector of $(T_{s1}, ..., T_{sn_s})'$ and $\mathbf{T}$ the vector of $(\mathbf{T}_1', ..., \mathbf{T}_S')'$. Let the random variable Y_{si}^1 be the outcome of unit i in stratum s after treatment and $\mathbf{Y}^1$ be the N-tuple of Y_{si}^1 arranged in the same order as $\mathbf{T}$. Y_{si}^0 and $\mathbf{Y}^0$ denote outcomes without treatment. If $Y_{si}^1 = Y_{si}^0$ the treatment has no effect on unit si. Under the null hypothesis of no treatment effect the responses are fixed, denoted y_{si}, and the only random variable left is $\mathbf{T}$.

The mean stratum effect Δ_s is estimated as the difference in the mean outcomes of the treated units and their controls in stratum s

$$\begin{aligned}
\hat{\Delta}_s &= \frac{1}{m_s}\mathbf{T}_s'\mathbf{y}_s - \frac{1}{n_s - m_s}(\mathbf{1} - \mathbf{T}_s)'\mathbf{y}_s \\
&= \frac{1}{m_s}\mathbf{T}_s'\mathbf{y}_s - \frac{n_s}{n_s - m_s}\bar{y}_s + \frac{1}{n_s - m_s}\mathbf{T}_s'\mathbf{y}_s \\
&= \frac{n_s}{m_s(n_s - m_s)}(\mathbf{T}_s'\mathbf{y}_s - m_s\bar{y}_s)
\end{aligned} \tag{3}$$

for all $s = 1, ..., S$, where $\mathbf{1}$ is a suitable vector of ones and $\bar{y}_s = \frac{1}{n_s}\mathbf{1}'\mathbf{y}_s$ denotes the mean over the y_{si} in stratum s. The overall mean effect τ is a weighted average of the stratum effects Δ_s, estimated by

$$\hat{\tau} = \sum_{s=1}^{S} \omega_s \hat{\Delta}_s, \tag{4}$$

where ω_s are positive stratum weights summing to one: $\sum_{s=1}^{S} \omega_s = 1$. $\hat{\tau}$ identifies the *average effect of treatment on the treated* (2) if the stratum weights ω_s are proportional to m_s since τ_T is the expectation conditional on the subsample of treated units. $\hat{\tau}$ identifies the *average treatment effect for the population* τ_P (1) if

[4] In this study either one treated unit will be matched to one or more controls or one control to more than one treated. Thus, m_s will either be equal to 1 or equal to $n_s - 1$.

the stratum weights are proportional to n_s.[5] In this case, each individual would have the same weight.[6]

Under the null hypothesis of no treatment effect $Y_{si}^1 = Y_{si}^0$ for all si and $\mathbb{E}\hat{\Delta}_s = 0$ for all s, $\mathbb{E}\hat{\tau} = 0$, and the variances of $\hat{\Delta}_s$ and $\hat{\tau}$ are

$$\sigma_s^2 = Var(\hat{\Delta}_s) = \frac{n_s}{(n_s-1)^2}\sum_{i=1}^{n_s}(y_{si}-\bar{y}_s)^2, \tag{5}$$

$$Var(\hat{\tau}) = \sum_{s=1}^{S}\omega_s^2\sigma_s^2. \tag{6}$$

The formula (5) is derived in appendix B, see also Rosenbaum (1995, pp. 29,54). The stratum differences $\hat{\Delta}_s$ are mutually independent and their variances differ across strata. Under very mild assumptions asymptotic normality of $\hat{\tau}$ is established for $S \to \infty$ (see appendix C). Statistical inference will be based on large sample theory exploiting the moments of the relevant test statistics.[7]

In contrast to the randomized experiment, in an observational study the distribution of the assignment vector $\mathbf{T}$ is unknown because individuals themselves decide whether to participate in treatment or not. If the treatment and control group differ prior to treatment in ways that matter for the outcome under study an observational study is *biased*. An *overt bias* is one that is produced by observable covariates $\mathbf{X}$ and that, in general, can be controlled using adjustments such as matching. Assuming that there is only overt bias, i.e. the potential outcomes are conditionally independent of treatment assignment given the covariates, matching on $\mathbf{X}$ mimics *ex post* a randomized experiment in each stratum defined by $\mathbf{X}$. Thus, given $\mathbf{X}$ the formalism for the randomized experiment outlined above can be applied. Alas, whenever $\mathbf{X}$ is of high dimension exact matching will, in all likelihood, be impossible. Alternatively, Rosenbaum and Rubin (1983) suggest to match on the one-dimensional *propensity score*, i.e. the probability to participate in treatment given $\mathbf{X}$, $p(\mathbf{x}) = \mathbb{P}(T = 1|\mathbf{X} = \mathbf{x})$, where $\mathbb{P}$ denotes probability. They show that if matching on $\mathbf{X}$ removes overt bias, then matching on $p(\mathbf{X})$ will do so, too.

[5] In particular, in the case of heterogeneous treatment effects the matching estimator as a weighted average of individual effects builds on a more intuitive weighting scheme than OLS. Angrist and Krueger (1999) and Angrist (1998) show how in the case of heterogeneous treatment effects a saturated linear model estimated by OLS weights the individual effects by the individual variances of the treatment indicator. In contrast, matching weights the individual effects by the probability to participate in treatment.

[6] Sample weights will also be taken into account in order to identify the US population parameters.

[7] Alternatively, it could rest on an exact permutation test. Calculating all feasible permutations of zeroes and ones of the vector $\mathbf{T}$ and counting how often the test statistic of the permuted data exceeds the sample test statistic (4) would produce exact p-values. Though, for a large number of strata such a test would exceed computer power by far. Good (1994) provides a practical guide to permutation tests and resampling methods in general.

3 The Matching Algorithm

Two main decisions have to be made when implementing a matching algorithm: First, the distance measure between treated and untreated individuals expressed in their covariates has to be defined, and second, the specific type of matching algorithm has to be chosen. We will consider these decisions in turn.

Distance Measures

The distance d between a treated and an untreated individual depends on their observable characteristics $\mathbf{x}^t$ and $\mathbf{x}^c$, respectively, and can be expressed as

$$d = d(\mathbf{x}^t, \mathbf{x}^c). \tag{7}$$

There are various common ways how this distance can be defined. Examples are

- the difference in propensity scores,
- the difference in the linear index of the probit model when estimating the propensity score p, $\Phi^{-1}(p)$, where Φ is the cumulative normal density function, and
- a weighted Euclidean distance, the Mahalanobis metric. The pooled covariance matrix V of the covariates $\mathbf{x}$ serves as weights.[8]

Hence, we have

$$d(\mathbf{x}^t, \mathbf{x}^c) = \begin{cases} (\mathbf{x}^t - \mathbf{x}^c)' V^{-1} (\mathbf{x}^t - \mathbf{x}^c) & \text{Mahalanobis metric,} \\ |p(\mathbf{x}^t) - p(\mathbf{x}^c)| & \text{Propensity score difference,} \\ |\Phi^{-1}(p(\mathbf{x}^t)) - \Phi^{-1}(p(\mathbf{x}^c))| & \text{Index score difference.} \end{cases} \tag{8}$$

A practical way to restrict matching to the close vicinity of each individual and, moreover, to substantially accelerate the speed of matching algorithms is the introduction of a caliper width $\varepsilon > 0$ outside which matching is not allowed, see e.g. Rosenbaum and Rubin (1985). Additionally, matching might be restricted to certain

[8] Matching using the Mahalanobis distance is discussed in Rubin (1980). A comparison, on the basis of a simulation study, of three distance measures – Mahalanobis, propensity score distance, and Mahalanobis distance within propensity score calipers – is provided in Gu and Rosenbaum (1993). Furthermore, propensity score calipers are discussed in Rosenbaum and Rubin (1985: 3) and Rosenbaum (1989: 3.4).

subsamples of the data, e.g. to subsamples defined by some pre-specified covariates x_k, $k \in K$. Therefore,

$$d^*(\mathbf{x}^t, \mathbf{x}^c) = \begin{cases} \infty & \text{if} \quad d(\mathbf{x}^t, \mathbf{x}^c) > \varepsilon \quad \text{or} \quad \exists k \in K : x_k^t \neq x_k^c \\ d(\mathbf{x}^t, \mathbf{x}^c) & \text{otherwise.} \end{cases} \tag{9}$$

Two different caliper widths ε will be compared, a narrow and a broad one. For propensity score matching, the narrow one will be set equal to 0.05 while the broad one will be 0.10. For index score matching, the respective numbers are 0.30 and 0.60. They are chosen such that both matching on the propensity score and on the index employ an approximately equal number of treated units. Broad calipers allow for matching of more individuals at the expense of a potentially less favorable balance of covariates. Narrow calipers generate closer similarity of matched units but several individuals at the top and bottom end of the propensity score scale are dropped. In case of the Mahalanobis distance calipers will be defined by the propensity or the index score as in Gu and Rosenbaum (1993).

Note that there is a bijective mapping between the propensity score p and the index $\Phi^{-1}(p)$. Basically, it should not matter which of them is used in the matching algorithm. However, the mapping is not linear since $p = 0$ is mapped to $-\infty$ and $p = 1$ to $+\infty$. Thus, differences in distances between two units on the unit interval do not necessarily remain constant under the mapping: Suppose that a treated unit with a propensity score of 0.95 might have to "choose" between two potential matches from the pool of untreated, one with propensity score of 0.90 and one with 0.98. Since the second one is closer to the treated it would be used for matching. In terms of the index, the treated unit is situated at 1.64, the first potential match at 1.28, the second at 2.05. This time, the treated unit will be matched to the other control, which is now closer. In other words, close to the boundaries of the unit interval matching on the propensity score might yield different strata than matching on the index.

Algorithms

The second decision regarding the matching procedure is how to minimize the distance between treated and untreated units. An appropriate algorithm is characterized by two main features, the uniformity of the stratification it produces, and the degree of distance minimization it can potentially attain. First, in a uniform stratification treated units are uniformly distributed across a large number of strata. Pair matching is the ideal uniform stratification since there are as many strata as

there are treated units. Augurzky and Schmidt (2000) propose a *variance inflation factor* for measuring uniformity.[9] The disadvantage, however, is that many treated units either would have to be matched to quite distant untreated units or would have to be dropped, both of which might produce an estimation bias. Therefore, in this paper, we investigate pair matching and full matching. Full matching distributes all treated units with finite distance across strata and thus produces strata that consist of more than one treated unit.

Second, among all possible stratifications there is one stratification achieving a minimum total distance. A so-called *optimal* algorithm searches for exactly the minimum total distance stratification. By contrast, other algorithms generally do not attain the minimum total distance: a *greedy* algorithm randomly selects a treated unit and matches it to the closest untreated unit available (in terms of the specified distance measure). The matched control unit is then removed from the control reservoir and the next treated unit is again matched to the nearest untreated unit from the remaining pool of controls. Hence, treated units looking for matches "late" in the algorithm may not find suitable controls, as the "sought-after" controls (i.e. usually those with high propensity scores) are no longer available. A treated unit is removed if it does not find a control unit with finite distance. After the last treated unit has been assigned a match the algorithm stops in case of pair matching. All remaining treated and untreated units are dropped. The final shape of the matched sample after greedy pair matching is therefore a set of 1-1 matches, and both a certain number of unmatched treated and unmatched untreated units, both of which are not considered in treatment effect calculation.

In the case of *full* matching the algorithm continues at the point where the sample of 1-1 matches has been produced, since the full matching algorithm aims at utilizing *all* treated and untreated units. The next step is to distribute the remaining unmatched untreated units to the existing 1-1 strata. The procedure is as follows: From the existent set of 1-1 strata one stratum is drawn randomly and assigned the closest untreated unit from the remaining pool, turning the stratum into a 1-2 stratum (unless no untreated unit with finite distance is available, which would leave the stratum in 1-1 shape). This procedure is repeated for all 1-1 strata, potentially turning all of them into 1-2 strata. If, at the end of this step, there remain unmatched untreated units, the procedure is repeated for all 1-2 strata, potentially turning them into 1-3 strata, etc., until all untreated units (with finite distance) are distributed. The intermediate shape of the matched sample is then a set of strata,

[9] Suppose all estimated stratum treatment effects have the same variance. Then "variance inflation" due to unfavorable stratification can be measured using $\frac{1}{(\sum_{s=1}^{S} m_s)^2} \sum_{s=1}^{S} \frac{m_s^2}{(1-1/n_s)^2}$, which is then compared to the benchmark stratification $m_s = 1$ and $n_s = 2$ for all s: $4/\sum_{s=1}^{S} m_s$.

with each strata in a 1-k shape.

In a final step of the full matching algorithm, the remaining treated units – those who did not find a match in the first step of pair matching – must now be assigned to appropriate strata. Logically, these can only be the 1-1 strata. The algorithm randomly selects one 1-1 stratum and assigns the closest treated unit, producing a 2-1 stratum. Then the next 1-1 stratum is assigned the closest treated unit, etc. If, at the end of this step, there are still treated units left, the procedure continues with constructing 3-1 strata, etc. The resulting shape of the full matching algorithm is a stratification with k-1, 1-1, and 1-k strata. Only those observations with infinite distance to all potential matching partners will be discarded.

Although this full matching algorithm attempts to minimize the *total distance* between treated units and their controls it will, in general, not attain the minimum. Rosenbaum (1991) shows how a greedy algorithm (which could be greedy pair or greedy full matching) might produce a stratification with a total distance arbitrarily worse than the optimal. A further unpleasant side effect is that results are different each time the algorithm is used because of the initial random order of observations. Optimal full matching circumvents these shortcomings. It attains the overall minimum in that it works backwards and rearranges already matched units if some specific treated unit turns out to be a better (closer) match with a control unit previously matched to another treated unit. In such a case, the first match is broken up, the second match is assigned, and the corresponding treated unit from the first match is again available for matching. Optimal full matching can easily be transformed into a *minimum cost flow problem*[10] (Rosenbaum, 1991).

In the application, three algorithms will be compared, a *greedy pair matching*, a *greedy full matching*, and an *optimal full matching*. Note that all stratifications consist of non-overlapping strata, i.e. no unit will be member of two different strata. Dehejia and Wahba (2002), for instance, also suggest an algorithm where controls are allowed to be used more than once in a matching algorithm with replacement. However, their algorithm generally produces overlapping strata. This makes statistical inference as outlined above more difficult due to stochastic dependencies across strata.

[10]Bertsekas (1991) discusses *linear network optimization* and minimum cost flow problems.

4 The Data

The data are taken from the *National Longitudinal Survey of Youth 1979* (NLSY) administered by the US Bureau of Labor Statistics. The NLSY is a sample of 12,686 youths first interviewed in 1979 when they were aged between 14 and 22 and re-interviewed annually until 1994. A detailed description of the data is given by the NLS Handbook (1997) and the NLSY79 User's Guide (1997). Annual data on hourly wages until 1994 are extracted for men.[11] Oversampling of Non-whites and economically disadvantaged Whites suggests the use of sample weights pertaining to 1979 in order to identify the population parameters.

The treatment period is the time it takes to achieve the bachelor's degree after graduating from high school. The treated individuals are those who obtained the degree and left college immediately thereafter, i.e. who did not continue college and eventually dropped out before achieving a higher degree. Controls are drawn from the pool of individuals with only a high school diploma who never attended college. High school dropouts and individuals with a *general educational development* (GED) are removed from the sample.

The year in which a respondent received the high school diploma marks the beginning of the treatment phase of those who went to college. In turn, the year in which he received his bachelor's degree marks the end. A treated and a control person should ideally have finished high school in the same year and at the same age. The control then starts to work and gain labor market experience while the treated is allowed to either go to college straight away, interrupt college for a while, or even start to work a certain time before finally attending college. Note that the estimation strategy pursued here does not identify the *return to education* but the *effect of the college degree* on earnings, which also includes indirect effects on labor market experience.

The outcome measure is the hourly rate of pay inflated to 1996 dollars using the US consumer price index and transformed into logarithms. For presentation of the results, the estimate $\hat{\tau}$ will be retransformed to $\exp(\hat{\tau}) - 1$.[12] The effect of college education is evaluated during the first ten years after graduation. Socioeconomic

[11] The sample is restricted to men because of their higher labor market participation compared to women.

[12] To eliminate outliers, all values below \$1 are set equal to \$1 and maximum or minimum wages of observations whose wages oscillate enormously across years are removed as well. For example, an hourly wage of \$5 in one year, \$1000 in the second, and again \$5 in the third seems more likely to reflect inconsistencies in the calculation of the hourly wage by the NLSY than real fundamental economic changes which is why \$1000 would be removed. See e.g. the NLSY79 User's Handbook (1997: p. 266): "... the calculation procedure [...] produces, at times, extremely low and extremely high pay rate values."

background variables, information about the high school career, and ability measures play an important role in the decision to attend college. They are used to estimate the propensity score.

Individuals of the same race, the same age +/- one year, and the same year of obtaining their high school degree +/- one year are permitted to be matched. In other words, K in equation (9) consists of the covariates *race*, *age* ± 1 and *high school graduation year* ± 1. This guarantees that treated and untreated individuals within a stratum share a similar economic environment at the beginning of the treatment phase. Exact matches on *age* and *the year of the high school diploma* would be preferable, but would substantially reduce the number of potential controls.

The pool of potential controls for each treated unit comprises all untreated units with finite distance. If some potential control for a given year after college graduation shows a missing value in hourly wage he is removed from the pool for this year. Ten years after college graduation will be examined and each year will be stratified separately such that individuals who are removed in some year due to missing wage information may still be available in other years.

5 Results

In evaluating the performance of the matching algorithms, we will focus on (i) the variance of the matching estimates and uniformity of stratification and (ii) potential biases as given by the balance of covariates after matching and the systematic loss of treated units.

The estimation of the propensity and index score is done using a probit model, results are presented in appendix A. The probit model, on the one hand, achieves to successfully separate college and high school graduates, i.e. selection into treatment is strong. This fact, on the other hand, aggravates matching at the boundaries, and observations at the top and bottom end of the propensity score distribution will have difficulties finding matching partners.[13] Full matching algorithms therefore might produce a stratification with a low degree of uniformity.

Table 1 compares the absolute frequencies of treated and untreated individuals for given propensity score and index score intervals. At the top and bottom of the propensity score scale there are more individuals than at the top and bottom of the index score scale. This is because the index score stretches the unit interval of the

[13]Note that a classic ordinary least squares model would linearly interpolate between the extremes, which is not necessarily superior.

propensity score to plus and minus infinity. Hence, matching on the index will drop several low-score untreated individuals. At the top of the distribution, the situation is comparable but less pronounced.

Estimation results for treatment effects are reported in tables 2 to 5 for four different distance measures: Propensity score distance, Mahalanobis distance with propensity score calipers, Index score distance, and Mahalanobis distance with index score calipers. Results for the first, third, fifth, seventh and ninth year after college graduation are shown. The first column of each table indicates the number of years after college. The second and third columns for the full matching algorithms report estimates of τ_T and τ_P. The estimates indicate an upward trend over time. While there is no effect $\hat{\tau}_T$ in the first year after college it rises up to 35% in the ninth year.[14]

For greedy pair matching, the two treatment effect estimates coincide. A supplementary column (7) reports the standard deviations induced by the initial random order of treated units in the greedy algorithm. They are calculated for $\hat{\tau}$ as well as for its standard error. The simulation errors for greedy full matching are omitted because they are negligibly low. The greedy algorithms are repeated 20 times. Columns (4) and (5) display for the full algorithms the number of strata, of treated, and of untreated individuals used for stratification. For pair matching, all three numbers coincide. The number of individuals and strata diminishes continuously from the first to the ninth year because many individuals, especially younger ones, are not in the sample for the whole nine-year period after college. The last column of the full matching algorithms reports the mean and maximum number of treated units in strata that consist of more than one treated. Large numbers typically increase the standard errors.

Table 6 describes the balancing properties of the matching algorithms. Since there are numerous covariates and ten stratifications reflecting the ten years after college, some aggregate measures of balance are introduced to facilitate assessment. The detailed results are reported in appendix D. The first column of table 6 shows the average reduction of differences in the variables over all years after college between treated and controls.[15] "0" means no reduction, "100" total reduction. Since the matching algorithms face severe problems in balancing the variable *born in south* (see appendix D) column (2) reports the average bias reduction disregarding this variable.

[14]Further investigation shows that part of the increase can be explained by the fact that college graduates accumulate experience more quickly after leaving college than high school graduates. However, interaction between labor market experience and schooling does not appear to be existent.

[15]Each year after college is weighted by the number of strata.

Columns (3) and (4) report average bias reductions for the presumably most important single variables *math scores* and *parents' education*. *Math scores* exhibit the highest t-value in the probit estimation (appendix A). Also, they are important determinants of wages as documented in other studies (Blackburn and Neumark, 1993, or Murnane, Willett, and Levy, 1995). *Parents' education* exhibits the second largest t-value. Finally, columns (5) display the propensity score difference between treated individuals before and after matching. A negative sign points to a systematic loss of treated units at the top of the propensity score distribution and, thus, to a possible bias in the estimates if the treatment effect is heterogeneous.

Sensitivity Analysis (i): Type of algorithm

The greedy full matching algorithm achieves to produce a more favorable, i.e. a more uniform, stratification than the optimal full matching. This is expressed by the mean and maximum number of units in strata consisting of more than one treated which is smaller for greedy matching. The number of strata is slightly larger in the greedy case, especially when calipers are broad. This pattern is more pronounced for index score matching although the estimates do not differ strongly. As noted in Gu and Rosenbaum (1993), this might be because greedy and full matching use the same individuals even though the specific stratification differs.

Surprisingly, overall balance is somewhat superior for greedy full matching, too. The main reason is that the optimal matching faces severe problems in balancing the variable *born in south*. Yet, notice that optimal matching tends to balance *math scores* better. Disregarding *born in south*, balancing success is more or less equal.[16] This finding is in line with Gu and Rosenbaum (1993) who observe that in terms of balance, optimal matching seems to have no advantage over greedy matching.

Greedy pair matching produces approximately the same number of strata as greedy full matching, i.e. the effective sample size is constant across algorithms. Nevertheless, standard errors are smaller for pair matching. This is because it produces the highest degree of uniformity. In case of pair matching, there is no reason to distinguish between $\hat{\tau}_T$ and $\hat{\tau}_P$ for two reasons. First, there is only one weighting scheme for pair matching and, second, since the majority of treated and untreated units are not matched, identification of the respective population parameters is doubtful anyway. These doubts are substantiated considering $\Delta\hat{p}$ in table 6. As

[16] A weakly significant interaction between *parents' education* and *born in south* has been included in the probit estimation, but improvements were not attained; other interactions were statistically insignificant. Moreover, exact matching on *born in south* reduced the matched sample size by roughly 20%, though, the number of strata did not diminish much; estimates of the treatment effects increased slightly.

expected, although pair matching produces the highest degree of uniformity and the most favorable balance, the loss of treated units with high propensity scores is dramatic. Yet, this systematic loss does not lead to very different estimates compared to full matching except for the case of the Mahalanobis distance. Thus, the results do not point to strong heterogeneity in the treatment effects. In such a case pair matching seems to be a superior strategy.

Sensitivity analysis (ii): Distance measure

If calipers are broader, more strata are produced because there are less units with infinite distances. The difference in the number of strata is more pronounced when the Mahalanobis distance is used. Nonetheless, estimates do not differ systematically and standard errors are not lower for the case of broad calipers because the larger number of strata is offset by a substantially reduced uniformity across strata, especially for optimal full matching. It is not offset for pair matching, where standard errors do decrease. For the full matching algorithms, percent bias reduction is larger for narrow calipers. For pair matching, the discrepancy is negligible. However, once *born in south* is disregarded, narrow and broad calipers produce an almost equal overall balance. Considering the systematic loss of treated units, a clear distinction can be made. For narrow calipers $\Delta\hat{p}$ is more negative than for broad calipers. This is because treated individuals with a high propensity score have more difficulties in finding a control with an equally high score and, thus, more treated units are dropped.

Estimation results for τ_T and τ_P do not differ strongly for different distance measures. The Mahalanobis case tends to supply more strata, though based on the same number of treated and untreated units. This observation is especially evident for index score matching. As a result, standard errors tend to be lower in the Mahalanobis case.

The most striking difference between the results based on propensity and index score distance is the number of controls used for stratification. Index score matching drops numerous untreated units, which is consistent with findings reported for table 1. For instance, in the first year, index matching utilizes over 200 controls less than propensity score matching. Because of that, it is unclear whether index matching really identifies τ_P. However, a clear distinction between estimates can hardly be established except for the fact that standard errors of $\hat{\tau}_P$ are slightly lower for index matching. With regard to balance there seems to be no discrepancy worth mentioning.

Finally, note that τ_P appears to be lower than τ_T in almost all specifications. While the difference is not statistically significant, however, this might be weak evidence in favor of heterogeneous effects. Results for τ_P are more or less of the same magnitude as results of the greedy *pair* matching.

6 Conclusion

The implementation of matching estimators to estimate treatment effects is characterized by the type of algorithm and the distance measure. The choice of matching algorithm and the distance measure depends on data characteristics and involves an inherent trade-off between bias and variance. The variance of the estimator is related to the uniformity of stratification. If there are only strata consisting of one treated individual, then uniformity is highest and variance lowest. Furthermore, two kinds of biases arise: A bias due to lack of balance of the covariates and a bias due to loss of treated individuals after matching.

Two data characteristics determine the relevance of these estimation problems: (i) the strength of selection into treatment and (ii) the strength of heterogeneity of the treatment effect. In case of weak selection into treatment all algorithms will achieve high uniformity of stratification. Sample variance is not very important. In case of homogeneous treatment effects there is no bias due to loss of treated units. Since each treated has the same effect there is no need to keep all treated individuals. If the effect is heterogeneous, however, each treated unit carries individual information on the treatment effect and loss of treated might increase the estimation bias. In all cases, however, bias due to lack of balance of covariates cannot be neglected and balance is best achieved if caliper widths are narrow.

Depending on the data characteristics and the bias-variance trade-off a matching technique has to be chosen in terms of type of algorithm and distance measure. This paper has addressed the sensitivity of matching estimates with respect to these decisions, using data from the NLSY79 to estimate the effect of a college degree on labor market earnings. The data exhibit strong selection into treatment, i.e. bias in relevant covariates prior to treatment is large and matching becomes a serious challenge.

The distance measures we have considered are propensity score, index score, and Mahalanobis distance. Our results show that choice of distance measure within the caliper appears less important than achieving uniformity of stratification. We have implemented an optimal full matching algorithm, and have compared it with

greedy pair matching and greedy full matching. Although the greedy full algorithm presented in this paper performs well, it is not its "greed" but its uniformity that drives the good results. An optimal procedure with restrictions on the number of treated per stratum seems to be a better alternative. Moreover, an optimal algorithm does not depend on the random initial order of treated units. Therefore, we recommend to use optimal, i.e. minimum distance algorithms with restrictions on the number of treated per stratum. See also Ming and Rosenbaum (2000) for a related discussion. Given a distance or caliper width, such restrictions allow to cover the whole range of matching strategies, from optimal pair to optimal full matching and everything in between.

If the data exhibit strong selection and treatment effect heterogeneity the algorithm should allow more than one treated per stratum and caliper widths should not be too narrow. In this paper, heterogeneity does not seem to be very strong and therefore pair matching produces good results. Alternatively, under fairly strict restrictions on the number of treated per stratum, optimal full matching would perform equally well at least, and would be preferable in the presence of strong heterogeneity.

Appendix A: Estimation of the Propensity Score

Table 7 displays the results of the estimation of the propensity score using a probit model. The model includes several covariates that drive selection into college, specifically socioeconomic background and high school career. Furthermore, it comprises two ability variables, scores on *math* and *auto and shop information* tests (adjusted for *age*).[17] The first tend to capture academic and the second non-academic abilities, see also the classification in Blackburn and Neumark (1995). *Parents' education* is the mean of the father's and mother's education, it takes on mother's education if father's education is missing and vice versa. *Parents' occupational status* is a binary variable indicating the social status of parents' occupation – high or low – and is given by the mean of mother's and father's status. Again, it takes on the father's status if the mother's is missing and vice versa. Except for *Hispanic* all variables are statistically significant at conventional levels. *Family income* is excluded due to many missing observations.

Apparently, selection into college is fairly strong, as has been found by other studies, too. Ashenfelter and Rouse (1998) report that (observed and unobserved) family background explains about 60% of the variance in schooling attainment and Murnane, Willett, and Levy (1995) assert that math test scores are a strong predictor of subsequent educational attainment.

Appendix B: Derivation of the Stratum Variance

This appendix derives the stratum variance formula given in equation (5). Starting with equation (3)

$$\hat{\Delta}_s = \frac{n_s}{m_s(n_s - m_s)}(\mathbf{T}'_s\mathbf{y}_s - m_s\bar{y}_s), \tag{10}$$

the variance of $\bar{\Delta}_s$ is

$$Var(\hat{\Delta}_s) = Var\left(\frac{n_s}{m_s(n_s - m_s)}\mathbf{T}'_s\mathbf{y}_s\right) = \frac{n_s^2}{m_s^2(n_s - m_s)^2}\mathbf{y}'_s Var(\mathbf{T}_s)\mathbf{y}_s. \tag{11}$$

[17] The NLSY provides ten ability measures, the *Armed Services Vocational Aptitude Battery* scores. Since respondents participated in the tests at different ages the scores are adjusted by regressing the raw scores on age dummies and using the residuals subsequently as explanatory variables, analogous to Blackburn and Neumark (1993).

Note that y_{si} is no random variable under the null hypothesis. Based on the distribution of $\mathbf{T}_s$ the covariance matrix is

$$
\begin{aligned}
Var(\mathbf{T}_s) &= \begin{pmatrix} \frac{m_s}{n_s}(1-\frac{m_s}{n_s}) & & \frac{m_s}{n_s}(\frac{m_s-1}{n_s-1}-\frac{m_s}{n_s}) \\ & \ddots & \\ \frac{m_s}{n_s}(\frac{m_s-1}{n_s-1}-\frac{m_s}{n_s}) & & \frac{m_s}{n_s}(1-\frac{m_s}{n_s}) \end{pmatrix} \\
&= \frac{m_s}{n_s}\frac{n_s-m_s}{n_s}\begin{pmatrix} 1 & & -\frac{1}{n_s-1} \\ & \ddots & \\ -\frac{1}{n_s-1} & & 1 \end{pmatrix}
\end{aligned} \tag{12}
$$

and

$$
\begin{aligned}
&(y_{s1},\ldots,y_{sn_s})\begin{pmatrix} 1 & & -\frac{1}{n_s-1} \\ & \ddots & \\ -\frac{1}{n_s-1} & & 1 \end{pmatrix}\begin{pmatrix} y_{s1} \\ \vdots \\ y_{sn_s} \end{pmatrix} \\
&= (y_{s1},\ldots,y_{sn_s})\begin{pmatrix} y_{s1}-\frac{1}{n_s-1}(y_{s2}+\ldots+y_{sn_s}) \\ \vdots \\ y_{sn_s}-\frac{1}{n_s-1}(y_{s1}+\ldots y_{sn_{s-1}}) \end{pmatrix} \\
&= \frac{n_s}{n_s-1}(y_{s1},\ldots,y_{sn_s})\begin{pmatrix} y_{s1}-\bar{y}_s \\ \vdots \\ y_{sn_s}-\bar{y}_s \end{pmatrix} \\
&= \frac{n_s}{n_s-1}(y_{s1}(y_{s1}-\bar{y}_s)+\ldots+y_{sn_s}(y_{sn_s}-\bar{y}_s)) \\
&= \frac{n_s}{n_s-1}\left[\sum_{i=1}^{n_s}(y_{si}-\bar{y}_s)^2+\bar{y}_s\underbrace{((y_{s1}-\bar{y}_s)+\ldots+(y_{sn_s}-\bar{y}_s))}_{=0}\right].
\end{aligned} \tag{13}
$$

Thus,

$$
\begin{aligned}
Var(\hat{\Delta}_s) &= \frac{n_s^2}{m_s(n_s-m_s)^2}\frac{m_s}{n_s}\frac{n_s-m_s}{n_s}\frac{n_s}{n_s-1}\sum_{i=1}^{n_s}(y_{si}-\bar{y}_s)^2 \\
&= \frac{n_s}{m_s(n_s-m_s)(n_s-1)}\sum_{i=1}^{n_s}(y_{si}-\bar{y}_s)^2.
\end{aligned} \tag{14}
$$

In the special case $m_s = 1$ or $m_s = n_s - 1$

$$
Var(\hat{\Delta}_s) = \frac{n_s}{(n_s-1)^2}\sum_{i=1}^{n_s}(y_{si}-\bar{y}_s)^2. \tag{15}
$$

Appendix C: Asymptotic Normality of $\hat{\tau}$

The estimator of the treatment effect $\hat{\tau}$ is the weighted sum of S independent strata effects $\hat{\Delta}_s$

$$\hat{\tau} = \sum_{s=1}^{S} \omega_s \hat{\Delta}_s. \tag{16}$$

To establish asymptotic normality of $\hat{\tau}$ the different variances of the $\hat{\Delta}_s$ have to be taken into account, i.e. Lindeberg condition has to be satisfied. Rewrite the problem and let

$$X_s = \omega_s \hat{\Delta}_s = \frac{\omega_s n_s}{m_s(n_s - m_s)} (\mathbf{T}'_s \mathbf{y}_s - m_s \bar{y}_s) \tag{17}$$

with $\mathbb{E} X_s = 0$ and variance σ_s^2 according to equation (6). The Lindeberg condition requires that for all $\varepsilon > 0$

$$\frac{1}{\sum_{s=1}^{S} \sigma_s^2} \sum_{s=1}^{S} \mathbb{E}\left(X_s^2 \, \mathbf{1}\left(X_s^2 \geq \varepsilon \sum_{s=1}^{S} \sigma_s^2 \right) \right) \quad \longrightarrow \quad 0 \tag{18}$$

as $S \longrightarrow \infty$, where $\mathbf{1}(...)$ is one if its argument is true and zero otherwise. In optimal full matching only the two cases $m_s = 1$ and $m_s = n_s - 1$ are relevant. The subsequent probabilities are necessary to compute the expectation of (18).
For $m_s = 1$[18]

$$\mathbb{P}\left(X_s = \frac{\omega_s n_s}{n_s - 1} (y_{si} - \bar{y}_s) \right) = \frac{1}{n_s} \qquad \forall i = 1, .., n_s. \tag{19}$$

For $m_s = n_s - 1$

$$\mathbb{P}\left(X_s = \frac{\omega_s n_s}{n_s - 1} (\bar{y}_s - y_{si}) \right) = \frac{1}{n_s} \qquad \forall i = 1, .., n_s. \tag{20}$$

Hence[19],

$$\mathbb{P}\left(X_s^2 = \left(\frac{\omega_s n_s}{n_s - 1} \right)^2 (y_{si} - \bar{y}_s)^2 \right) = \frac{1}{n_s} \qquad \forall i = 1, .., n_s. \tag{21}$$

It follows that

$$\begin{aligned} &\frac{1}{\sum_{s=1}^{S} \sigma_s^2} \sum_{s=1}^{S} \mathbb{E}\left(X_s^2 \, \mathbf{1}\left(X_s^2 \geq \varepsilon \sum_{s=1}^{S} \sigma_s^2 \right) \right) \qquad (22) \\ &= \frac{1}{\sum_{s=1}^{S} \sigma_s^2} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} \left(\frac{\omega_s n_s}{n_s - 1} \right)^2 (y_{si} - \bar{y}_s)^2 \cdot \mathbf{1}\left(\left(\frac{\omega_s n_s}{n_s - 1} \right)^2 (y_{si} - \bar{y}_s)^2 \geq \varepsilon \sum_{s=1}^{S} \sigma_s^2 \right) \\ &\leq \frac{1}{\sum_{s=1}^{S} \sigma_s^2} \sum_{s=1}^{S} k_s M_s \cdot \mathbf{1}\left(k_s M_s \geq \varepsilon \sum_{s=1}^{S} \sigma_s^2 \right), \end{aligned}$$

[18] If there are ties, i.e. there is an $i \neq j$ such that $y_{si} = y_{sj}$, $\mathbb{P}\left(X_s = \frac{\omega_s n_s}{n_s - 1} (y_{si} - \bar{y}_s) \right)$ is at least $\frac{2}{n_s}$. However, counting each i separately for all $i = 1, ..., n_s$ as in the subsequent steps and giving each the same probability $\frac{1}{n_s}$ does not cause problems.

[19] X_s^2 not being bijective ties can emerge again, see previous footnote.

where $k_s = \left(\frac{\omega_s n_s}{n_s - 1}\right)^2$ and $M_s = \max_{i=1,\ldots,n_s}(y_{si} - \bar{y}_s)^2$. Furthermore, if M^S is the maximum over all M_s, $M^S = \max_{s \leq S} M_s$, and if

$$\frac{M^S}{\sum_{s=1}^{S} \sigma_s^2} \longrightarrow 0 \tag{23}$$

for $S \longrightarrow \infty$, condition (18) is satisfied. Assumption (23) is rather harmless because y_{si} being wages M^S even remains finite.

Appendix D: Balance of Covariates

Tables 8 to 11 display the balancing properties for all covariates and for all specifications. They show the means of covariates by treatment status before and after matching. After matching, weighted averages over all stratifications of the ten years after college are reported. The weights correspond to the number of strata in each year. The means are compared by a conventional t-test under the assumption of equal variances in both groups. A "1" indicates that the means are not significantly different. Fractions are due to averaging. Moreover, the reduction of the bias in covariates is shown as a percentage for each variable and as an average over all variables. Since the full matching algorithms face severe problems in balancing the variable *born in south*, the last row displays the average over all variables when it is excluded.

References

Angrist, J.D. (1998), 'Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants', *Econometrica*, 66, 249-288.

Angrist, J.D. and A.B. Krueger (1999), 'Empirical Strategies in Labor Economics', in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, North-Holland, Amsterdam.

Angrist, J.D. and J. Hahn (2004), 'When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects', *The Review of Economics and Statistics*, 86, 58-72.

Ashenfelter, O. and C. Rouse (1998), 'Income, Schooling and Ability: Evidence from a New Sample of Identical Twins', *Quarterly Journal of Economics*, 113, 253-284.

Augurzky, B. and C.M. Schmidt (2000), 'The Propensity Score – A Means to an End', *IZA Discussion paper*, No. 271, Bonn.

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

Blackburn, M.L. and D.B. Neumark (1993), 'Omitted-Ability Bias and the Increase in the Return to Schooling', *Journal of Labor Economics*, 11, 521-544.

Blackburn, M.L. and D.B. Neumark (1995), 'Are OLS Estimates of the Return to Schooling Biased Downward? Another Look', *Review of Economics and Statistics*, 77, 217-230.

Cochran, W.G. (1965), 'The Planning of Observational Studies of Human Populations (with discussion)', *Journal of the Royal Statistical Society Series A*, 128, 234-266.

Dehejia, R. and S. Wahba (1999), 'Causal Effects in Nonexperimental Studies:

Reevaluating the Evaluation of Training Programs', *Journal of the American Statistical Association*, 94, 1053-1062.

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

Dehejia, R. (2004), 'Practical Propensity Score Matching: A Reply to Smith and Todd', *Journal of Econometrics*, forthcoming.

Good, P. (1994), *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, Springer Series in Statistics, New York.

Gu, X.S. and P.R. Rosenbaum (1993), 'Comparison of multivariate matching methods: Structures, distances and algorithms', *Journal of Computational and Graphical Statistics*, 2, 405-420.

Hahn, J. (1998), 'On the Role of the Propensity Score in the Efficient Semiparametric Estimation of Average Treatment Effects', *Econometrica*, 66, 315-332.

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

Heckman, J.J., R.J. Lalonde and J. Smith (1999), 'The Economics and Econometrics of Active Labor Market Programs', in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, North-Holland, Amsterdam.

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

Holland, P.W. (1986), 'Statistics and Causal Inference (with discussion)', *Journal of the American Statistical Association*, 81, 945-970.

Imbens, G.W. (2004), 'Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review', *Review of Economic and Statistics*, 86, 4-29.

Kluve, J. (2004), 'On the Role of Counterfactuals in Inferring Causal Effects', *Foundations of Science*, 9, 65-101.

Kluve, J., H. Lehmann and C.M. Schmidt (1999), 'Active Labor Market Policies: Human Capital Enhancement, Stigmatization, or Benefit Churning?', *Journal of Comparative Economics*, 27, 61-89.

LaLonde, R.J. (1986), 'Evaluating the Econometric Evaluations of Training Programs with Experimental Data', *American Economic Review*, 76, 604-620.

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

NLS Handbook (1997), U.S. Department of Labor, Bureau of Labor Statistics.

NLSY79 User's Guide (1997), Center for Human Resource Research: The Ohio State University.

Ming, K. and P.R. Rosenbaum (2000), 'Substantial Gains in Bias Reduction from Matching with a Variable Number of Controls, *Biometrics*, 56, 118-124.

Murnane, R.J., J.B. Willett and F. Levy (1995), 'The Growing Importance of Cognitive Skills in Wage Determination', *Review of Economics and Statistics*, 77, 251-266.

Rosenbaum, P.R. (1989), 'Optimal Matching for Observational Studies', *Journal of the American Statistical Association*, 84, 1024-1032.

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

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

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

Rosenbaum, P.R. and D.B. Rubin (1984), 'Reducing Bias in Observational Studies using Subclassification on the Propensity Score', *Journal of the American Statistical Association*, 79, 516-524.

Rosenbaum, P.R. and D.B. Rubin (1985), 'Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score', *The American Statistician*, 39, 33-38.

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

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

Rubin, D.B. (1980), 'Bias Reduction Using Mahalanobis Metric Matching', *Biometrics*, 36, 293-298.

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

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

Smith, J.A. and P.E. Todd (2004a), 'Does Matching overcome LaLonde's critique of Nonexperimental estimators?', *Journal of Econometrics*, forthcoming.

Smith, J.A. and P.E. Todd (2004b), 'Rejoinder', *Journal of Econometrics*, forthcoming.

Zhao, Z. (2004), 'Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence', *The Review of Economics and Statistics*, 86, 91-107.

Table 1: **Distribution of Estimated Propensity and Index Score.**

Estimated Prop. score	Untreated	Treated	Estimated Index score	Untreated	Treated
[0.0 , 0.1)	946	29	[−4.70 , −3.94)	11	0
[0.1 , 0.2)	150	23	[−3.94 , −3.18)	74	0
[0.2 , 0.3)	80	21	[−3.18 , −2.42)	285	2
[0.3 , 0.4)	56	20	[−2.42 , −1.66)	407	9
[0.4 , 0.5)	29	21	[−1.66 , −0.90)	298	37
[0.5 , 0.6)	33	35	[−0.90 , −0.14)	175	54
[0.6 , 0.7)	15	34	[−0.14 , +0.62)	64	96
[0.7 , 0.8)	20	60	[+0.62 , +1.38)	24	139
[0.8 , 0.9)	9	79	[+1.38 , +2.14)	4	86
[0.9 , 1.0]	4	128	[+2.14 , +2.90]	0	27
Mean score	0.11	0.67		-1.77	0.61
Observations	1342	450		1342	450

Comparison of the number of treated and untreated individuals by propensity score and index score intervals.

Table 2: **Estimated Treatment Effects. Propensity score distance.**

	Optimal Full Matching					Greedy Full Matching					Greedy Pair Matching		
(1)	(2)	(3)	(4)	(5)	(6)	(2)	(3)	(4)	(5)	(6)	(2)	(7)	(8)
			S	T	Mean			S	T	Mean		Simul.	S
Year	$\hat{\tau}_T$	$\hat{\tau}_P$		C	Max	$\hat{\tau}_T$	$\hat{\tau}_P$		C	Max	Effect	Error	Error
Narrow Caliper													
1	-0.007	-0.026	150	287	4.9	0.028	0.027	152	286	4.0	-0.011	0.021	151
	(0.088)	(0.079)		898	27.0	(0.088)	(0.077)		898	27.0	(0.057)	(0.004)	1.34
3	0.188***	0.211**	149	250	4.0	0.175**	0.156**	149	250	3.2	0.188***	0.028	148
	(0.076)	(0.091)		880	16.0	(0.075)	(0.079)		880	15.0	(0.066)	(0.003)	0.97
5	0.201***	0.019	137	230	4.0	0.234***	0.097	137	229	3.2	0.217***	0.026	138
	(0.072)	(0.083)		893	14.0	(0.073)	(0.083)		893	11.0	(0.077)	(0.004)	1.02
7	0.260***	0.216**	123	197	3.7	0.276***	0.222***	123	197	3.2	0.275***	0.032	123
	(0.091)	(0.100)		789	14.0	(0.086)	(0.089)		789	11.0	(0.081)	(0.004)	0.54
9	0.355***	0.244**	93	151	3.9	0.299***	0.234**	92	150	3.1	0.314***	0.059	92
	(0.130)	(0.117)		578	12.0	(0.113)	(0.104)		578	9.0	(0.119)	(0.010)	0.78
Broad Caliper													
1	-0.020	0.059	159	333	5.2	0.006	0.151**	163	332	4.2	-0.022	0.026	164
	(0.090)	(0.090)		1122	28.0	(0.082)	(0.082)		1122	21.0	(0.054)	(0.004)	1.46
3	0.158	0.253**	158	308	5.1	0.164*	0.204***	163	307	4.1	0.206***	0.031	162
	(0.117)	(0.110)		1116	34.0	(0.101)	(0.086)		1116	24.0	(0.067)	(0.004)	1.26
5	0.179**	0.054	148	285	5.2	0.192***	0.205***	150	283	4.0	0.221***	0.035	152
	(0.089)	(0.090)		1084	32.0	(0.075)	(0.081)		1084	20.0	(0.072)	(0.005)	1.56
7	0.228**	0.219**	130	242	4.5	0.256***	0.269***	134	242	3.8	0.278***	0.044	134
	(0.112)	(0.101)		959	32.0	(0.103)	(0.088)		959	21.0	(0.076)	(0.004)	1.10
9	0.362***	0.271***	99	188	4.7	0.361***	0.297***	101	188	3.9	0.327***	0.057	102
	(0.152)	(0.117)		730	27.0	(0.107)	(0.104)		730	16.0	(0.113)	(0.008)	0.85

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. The weighting schemes take account of the NLSY sample weights. The first columns of the full matching algorithms show estimates of the mean effect of treatment on the treated while the second show the mean effect on the population. For greedy pair matching, simulation standard deviations are additionally reported in an own column. Columns denoted by S, T, and C display the number of strata, of treated, and of control units, respectively. For pair matching all three numbers are equal, simulation standard deviations for S are reported. Finally, columns titled "Mean" and "Max" show the mean and maximum number of treated units in strata that comprise more than one treated, respectively.

Table 3: **Estimated Treatment Effects. Mahalanobis distance, p.score calipers.**

	Optimal Full Matching					Greedy Full Matching					Greedy Pair Matching		
(1)	(2)	(3)	(4)	(5)	(6)	(2)	(3)	(4)	(5)	(6)	(2)	(7)	(8)
			S	T	Mean			S	T	Mean		Simul.	S
Year	$\hat{\tau}_T$	$\hat{\tau}_P$		C	Max	$\hat{\tau}_T$	$\hat{\tau}_P$		C	Max	Effect	Error	Error
Narrow Caliper													
1	0.023	0.025	153	287	4.9	0.019	0.010	151	286	3.9	-0.027	0.019	152
	(0.090)	(0.077)		898	27.0	(0.087)	(0.074)		898	27.0	(0.059)	(0.003)	1.07
3	0.214***	0.209***	151	250	3.9	0.172***	0.151**	151	249	3.5	0.143**	0.020	150
	(0.078)	(0.087)		880	17.0	(0.071)	(0.075)		880	16.0	(0.061)	(0.003)	1.14
5	0.219***	0.119	139	230	3.9	0.225***	0.094	137	229	3.4	0.162**	0.030	138
	(0.075)	(0.090)		893	14.0	(0.075)	(0.082)		893	12.0	(0.073)	(0.004)	1.19
7	0.298***	0.258***	123	197	4.9	0.302***	0.238***	122	196	3.4	0.216***	0.028	123
	(0.093)	(0.097)		789	15.0	(0.090)	(0.090)		789	12.0	(0.074)	(0.005)	0.85
9	0.318***	0.286***	91	151	4.2	0.304***	0.253***	92	151	3.2	0.249***	0.041	92
	(0.121)	(0.118)		578	14.0	(0.104)	(0.109)		578	10.0	(0.097)	(0.007)	0.97
Broad Caliper													
1	0.028	0.232***	167	333	5.4	0.002	0.140*	165	333	4.1	-0.003	0.021	166
	(0.086)	(0.097)		1122	25.0	(0.078)	(0.079)		1122	20.0	(0.056)	(0.006)	1.70
3	0.177*	0.248***	166	308	4.9	0.162*	0.196***	168	308	3.9	0.165***	0.037	166
	(0.111)	(0.099)		1116	29.0	(0.094)	(0.082)		1116	21.0	(0.057)	(0.006)	1.62
5	0.174**	0.243***	156	285	4.9	0.216***	0.196***	158	285	4.0	0.142**	0.025	156
	(0.085)	(0.101)		1084	27.0	(0.075)	(0.081)		1084	14.0	(0.069)	(0.004)	0.95
7	0.261***	0.309***	138	242	5.0	0.275***	0.283***	137	242	3.6	0.232***	0.036	137
	(0.107)	(0.104)		959	27.0	(0.086)	(0.085)		959	17.0	(0.064)	(0.004)	0.78
9	0.372***	0.352***	104	188	4.7	0.324***	0.292***	103	187	3.8	0.263***	0.043	103
	(0.142)	(0.125)		730	22.0	(0.114)	(0.104)		730	13.0	(0.090)	(0.007)	1.42

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. The weighting schemes take account of the NLSY sample weights. The first columns of the full matching algorithms show estimates of the mean effect of treatment on the treated while the second show the mean effect on the population. For greedy pair matching, simulation standard deviations are additionally reported in an own column. Columns denoted by S, T, and C display the number of strata, of treated, and of control units, respectively. For pair matching all three numbers are equal, simulation standard deviations for S are reported. Finally, columns titled "Mean" and "Max" show the mean and maximum number of treated units in strata that comprise more than one treated, respectively.

Table 4: **Estimated Treatment Effects. Index score distance.**

	Optimal Full Matching					Greedy Full Matching					Greedy Pair Matching		
(1)	(2)	(3)	(4)	(5)	(6)	(2)	(3)	(4)	(5)	(6)	(2)	(7)	(8)
			S	T	Mean			S	T	Mean		Simul.	S
Year	$\hat{\tau}_T$	$\hat{\tau}_P$		C	Max	$\hat{\tau}_T$	$\hat{\tau}_P$		C	Max	Effect	Error	Error
Narrow Caliper													
1	0.001	0.010	159	300	4.5	0.033	0.024	166	299	3.8	-0.008	0.017	165
	(0.079)	(0.062)		669	19.0	(0.088)	(0.062)		669	19.0	(0.055)	(0.005)	1.65
3	0.199***	0.205***	160	268	4.1	0.230***	0.186***	167	268	3.2	0.232***	0.030	166
	(0.076)	(0.073)		692	18.0	(0.071)	(0.067)		692	15.0	(0.067)	(0.006)	1.67
5	0.202***	0.093	148	249	4.4	0.205***	0.098	151	249	3.5	0.216***	0.033	153
	(0.072)	(0.071)		712	15.0	(0.070)	(0.070)		712	12.0	(0.072)	(0.004)	1.48
7	0.284***	0.249***	130	214	3.8	0.278***	0.246***	135	214	3.3	0.264***	0.034	136
	(0.090)	(0.079)		635	15.0	(0.079)	(0.078)		635	12.0	(0.075)	(0.004)	1.09
9	0.354***	0.239***	99	165	3.9	0.339***	0.262***	102	165	3.0	0.310***	0.040	102
	(0.127)	(0.101)		465	14.0	(0.104)	(0.097)		465	11.0	(0.111)	(0.007)	1.06
Broad Caliper													
1	-0.017	0.019	167	340	5.1	0.077	0.053	189	339	3.9	0.021	0.025	190
	(0.087)	(0.072)		878	25.0	(0.079)	(0.064)		878	18.0	(0.053)	(0.003)	1.31
3	0.162	0.200**	164	322	5.2	0.192**	0.174***	185	322	3.7	0.235***	0.029	184
	(0.116)	(0.087)		865	34.0	(0.090)	(0.068)		865	18.0	(0.064)	(0.006)	2.26
5	0.192**	0.072	153	301	5.4	0.205***	0.113*	173	300	3.8	0.236***	0.024	175
	(0.092)	(0.078)		874	32.0	(0.063)	(0.069)		874	13.0	(0.066)	(0.003)	2.92
7	0.248**	0.245***	134	258	4.9	0.310***	0.276***	151	257	3.4	0.290***	0.044	153
	(0.111)	(0.091)		788	31.0	(0.088)	(0.082)		788	12.0	(0.071)	(0.004)	1.80
9	0.356***	0.253***	104	195	4.8	0.366***	0.274***	118	194	3.2	0.341***	0.051	116
	(0.152)	(0.105)		621	27.0	(0.109)	(0.093)		621	9.0	(0.107)	(0.008)	1.55

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. The weighting schemes take account of the NLSY sample weights. The first columns of the full matching algorithms show estimates of the mean effect of treatment on the treated while the second show the mean effect on the population. For greedy pair matching, simulation standard deviations are additionally reported in an own column. Columns denoted by S, T, and C display the number of strata, of treated, and of control units, respectively. For pair matching all three numbers are equal, simulation standard deviations for S are reported. Finally, columns titled "Mean" and "Max" show the mean and maximum number of treated units in strata that comprise more than one treated, respectively.

Table 5: **Estimated Treatment Effects. Mahalanobis distance, index score calipers.**

	Optimal Full Matching					Greedy Full Matching					Greedy Pair Matching		
(1)	(2)	(3)	(4)	(5)	(6)	(2)	(3)	(4)	(5)	(6)	(2)	(7)	(8)
			S	T	Mean			S	T	Mean		Simul.	S
Year	$\hat{\tau}_T$	$\hat{\tau}_P$		C	Max	$\hat{\tau}_T$	$\hat{\tau}_P$		C	Max	Effect	Error	Error
Narrow Caliper													
1	0.046	0.035	173	300	4.8	0.027	0.024	172	300	3.7	-0.014	0.030	172
	(0.083)	(0.063)		669	19.0	(0.081)	(0.060)		669	19.0	(0.054)	(0.005)	1.89
3	0.212***	0.169***	173	268	4.3	0.178***	0.159***	171	268	3.2	0.154**	0.032	172
	(0.074)	(0.067)		692	18.0	(0.067)	(0.062)		692	16.0	(0.058)	(0.004)	1.40
5	0.180***	0.133*	158	249	4.1	0.177***	0.093	158	249	3.2	0.126*	0.027	157
	(0.069)	(0.075)		712	14.0	(0.061)	(0.067)		712	10.0	(0.065)	(0.004)	1.62
7	0.290***	0.257***	141	214	4.3	0.280***	0.244***	140	214	3.1	0.215***	0.037	140
	(0.087)	(0.081)		635	15.0	(0.076)	(0.076)		635	9.0	(0.066)	(0.006)	1.24
9	0.312***	0.282***	106	165	4.0	0.280***	0.252***	104	165	3.1	0.220**	0.042	105
	(0.109)	(0.103)		465	14.0	(0.100)	(0.096)		465	11.0	(0.087)	(0.007)	0.92
Broad Caliper													
1	0.036	0.055	195	340	5.1	0.005	0.040	207	340	3.5	0.023	0.027	202
	(0.074)	(0.068)		878	19.0	(0.068)	(0.062)		878	15.0	(0.053)	(0.003)	2.15
3	0.165*	0.147**	193	322	5.0	0.168**	0.154**	195	322	3.5	0.151***	0.032	198
	(0.092)	(0.072)		865	19.0	(0.087)	(0.065)		865	16.0	(0.052)	(0.003)	1.78
5	0.211***	0.141*	179	301	4.6	0.207***	0.130**	186	301	3.2	0.188***	0.028	185
	(0.075)	(0.079)		874	18.0	(0.059)	(0.068)		874	10.0	(0.062)	(0.003)	1.87
7	0.316***	0.295***	158	258	4.8	0.260***	0.259***	163	254	3.1	0.228***	0.037	162
	(0.098)	(0.091)		788	16.0	(0.072)	(0.077)		788	9.0	(0.066)	(0.006)	1.84
9	0.302***	0.264***	121	195	4.5	0.358***	0.276***	125	195	3.1	0.199**	0.034	123
	(0.118)	(0.104)		621	13.0	(0.095)	(0.093)		621	7.0	(0.089)	(0.007)	1.44

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. The weighting schemes take account of the NLSY sample weights. The first columns of the full matching algorithms show estimates of the mean effect of treatment on the treated while the second show the mean effect on the population. For greedy pair matching, simulation standard deviations are additionally reported in an own column. Columns denoted by S, T, and C display the number of strata, of treated, and of control units, respectively. For pair matching all three numbers are equal, simulation standard deviations for S are reported. Finally, columns titled "Mean" and "Max" show the mean and maximum number of treated units in strata that comprise more than one treated, respectively.

Table 6: **Balance of Covariates, Aggregate Measures.**

			Optimal Full					Greedy Full					Greedy Pair				
Caliper based on	Dis-tance	Caliper Width ε	(1) Mean	(2) -1	(3) Math	(4) Educ	(5) $\Delta\hat{p}$	(1) Mean	(2) -1	(3) Math	(4) Educ	(5) $\Delta\hat{p}$	(1) Mean	(2) -1	(3) Math	(4) Educ	(5) $\Delta\hat{p}$
p.score	p.score	narrow	82	90	98	89	-.09	84	89	98	88	-.09	85	90	97	93	-.24
p.score	p.score	broad	76	89	98	85	-.05	78	87	95	81	-.05	87	91	98	91	-.22
p.score	Mahal	narrow	81	88	96	85	-.09	84	88	98	87	-.09	86	90	97	91	-.24
p.score	Mahal	broad	74	87	98	77	-.05	79	87	94	79	-.05	87	90	97	87	-.22
index	p.score	narrow	81	92	98	91	-.09	85	91	98	88	-.09	87	91	98	92	-.23
index	p.score	broad	75	90	98	84	-.04	83	90	91	79	-.04	89	91	97	87	-.20
index	Mahal	narrow	80	89	98	91	-.09	86	91	97	95	-.09	88	91	98	93	-.22
index	Mahal	broad	80	88	93	77	-.04	85	88	88	77	-.04	85	87	93	85	-.18

The first three columns specify the sensitivity parameters. The first column of each matching algorithm represents mean overall percent bias reduction, the second is the mean reduction when the variable *born in south* is disregarded. The third and fourth display bias reduction in math scores and in parents' education, respectively. The fifth column reports the difference in mean propensity scores between treated units before and after matching.

Table 7: **Probit Estimation Results.**

Variables	Mean	Coeff.	t-value	P-value
Black	0.263	0.274	1.971	0.049
Hispanic	0.091	0.256	1.443	0.149
Math test scores	-0.442	0.098	15.384	0.000
Auto and shop test scores	4.911	-0.018	-2.857	0.004
Attended private school	0.052	0.432	2.397	0.017
Ever expelled or suspended from school	0.272	-0.536	-4.314	0.000
High school curriculum: college preparatory	0.288	0.972	6.392	0.000
High school curriculum: general program	0.509	0.358	2.439	0.015
Parents' education	11.185	0.154	6.857	0.000
Parents' occup. status high when resp. was 14	0.129	0.432	2.361	0.018
Number of siblings	3.600	-0.065	-2.796	0.005
Born in the south	0.365	0.346	3.333	0.001
Constant	1.000	-3.142	-9.750	0.000
Observations	1792			
$\chi^2(12)$	1046.4			
Overall p-value	0.000			
Pseudo R^2	0.518			

All variables with "yes/no" answers are dummy variables with 1 for "yes" and 0 for "no". The Pseudo R^2 reports the the likelihood ratio index, i.e. $1 - L_1/L_0$, where L_1 is the log likelihood of the full model and L_0 is the log likelihood of the "constant-only" model.

Table 8: **Balance of Covariates. Propensity score distance.**

	Initially			Optimal Full				Greedy Full				Greedy Pair			
	C	T	t	C	T	t	%	C	T	t	%	C	T	t	%
Narrow Caliper															
Propensity score	0.11	0.67	0	0.57	0.58	1.00	99	0.57	0.58	1.00	99	0.43	0.43	1.00	100
Index score	-1.77	0.61	0	0.17	0.21	1.00	99	0.16	0.21	1.00	98	-0.28	-0.27	1.00	100
Black	0.30	0.16	0	0.15	0.15	1.00	100	0.15	0.15	1.00	100	0.23	0.23	1.00	100
Hispanic	0.10	0.07	1	0.04	0.04	1.00	100	0.04	0.04	1.00	100	0.06	0.06	1.00	100
Age	17.50	17.65	1	17.89	17.85	1.00	70	17.90	17.85	1.00	63	17.82	17.79	1.00	75
Year of high school diploma	79.37	78.74	0	78.57	78.54	1.00	94	78.56	78.54	1.00	95	78.64	78.66	1.00	95
Math test scores	-3.95	10.02	0	8.35	8.36	1.00	98	8.20	8.35	1.00	98	5.25	5.02	1.00	97
Auto+shop test scores	3.88	7.98	0	7.95	7.78	1.00	94	8.02	7.79	1.00	94	7.62	6.54	1.00	73
Attended private school	0.03	0.12	0	0.11	0.09	1.00	80	0.11	0.09	1.00	79	0.08	0.08	1.00	83
Expelled or susp. from school	0.33	0.10	0	0.14	0.11	0.71	88	0.16	0.11	0.71	82	0.15	0.14	1.00	94
Curriculum: college prepar.	0.16	0.67	0	0.60	0.57	1.00	93	0.62	0.57	1.00	91	0.49	0.46	1.00	94
Curriculum: general	0.59	0.28	0	0.33	0.37	1.00	86	0.31	0.37	1.00	82	0.41	0.44	1.00	88
Highest grades of parents	10.50	13.21	0	12.44	12.59	1.00	89	12.40	12.60	0.55	88	11.93	12.03	1.00	93
Occupation parents high	0.08	0.29	0	0.24	0.22	1.00	94	0.24	0.23	1.00	92	0.18	0.18	1.00	94
Number of siblings	3.92	2.64	0	2.80	2.75	1.00	90	2.79	2.75	1.00	91	2.93	2.82	1.00	88
Born in south	0.38	0.33	1	0.25	0.30	0.88	-27	0.27	0.30	1.00	18	0.31	0.34	1.00	17
Mean percent bias reduction							82				84				85
– *born in south* excluded							90				89				90
Broad Caliper															
Propensity score	0.11	0.67	0	0.61	0.62	1.00	98	0.59	0.62	1.00	95	0.43	0.45	1.00	98
Index score	-1.77	0.61	0	0.27	0.37	1.00	96	0.20	0.37	0.31	93	-0.25	-0.22	1.00	98
Black	0.30	0.16	0	0.17	0.17	1.00	100	0.17	0.17	1.00	100	0.24	0.24	1.00	100
Hispanic	0.10	0.07	1	0.04	0.04	1.00	100	0.04	0.04	1.00	100	0.08	0.08	1.00	100
Age	17.50	17.65	1	17.91	17.89	1.00	73	17.97	17.89	1.00	45	17.83	17.82	1.00	79
Year of high school diploma	79.37	78.74	0	78.57	78.50	1.00	89	78.53	78.50	1.00	93	78.63	78.63	1.00	94
Math test scores	-3.95	10.02	0	8.97	9.04	1.00	98	8.34	9.05	0.95	95	5.33	5.24	1.00	98
Auto+shop test scores	3.88	7.98	0	7.14	7.63	1.00	88	7.20	7.65	1.00	89	7.34	6.44	1.00	78
Attended private school	0.03	0.12	0	0.10	0.10	1.00	78	0.11	0.10	1.00	78	0.08	0.08	1.00	83
Expelled or susp. from school	0.33	0.10	0	0.13	0.11	0.83	89	0.14	0.11	0.71	86	0.14	0.15	1.00	93
Curriculum: college prepar.	0.16	0.67	0	0.65	0.63	1.00	94	0.65	0.63	1.00	94	0.50	0.47	1.00	94
Curriculum: general	0.59	0.28	0	0.29	0.32	1.00	88	0.28	0.32	1.00	84	0.40	0.43	1.00	90
Highest grades of parents	10.50	13.21	0	12.36	12.76	0.37	85	12.25	12.76	0.37	81	11.88	12.10	1.00	91
Occupation parents high	0.08	0.29	0	0.23	0.24	1.00	91	0.24	0.24	1.00	93	0.18	0.17	1.00	93
Number of siblings	3.92	2.64	0	2.79	2.76	1.00	90	2.77	2.76	1.00	91	2.93	2.82	1.00	88
Born in south	0.38	0.33	1	0.23	0.32	0.11	-99	0.25	0.32	0.76	-37	0.32	0.34	1.00	38
Mean percent bias reduction							76				78				87
– *born in south* excluded							89				87				91

For reasons of parsimony, weighted averages over all ten years after college are shown. Weights correspond to the number of strata in each year after college. C denotes the control units while T represents treated units, t indicates whether a t-test accepts balance of covariates ($t = 1$), and % denotes percent bias reduction. The last two rows of each panel report the simple average over all single percent bias reductions excluding those of the propensity and index score.

Table 9: **Balance of Covariates. Mahalanobis distance, p.score calipers.**

	Initially			Optimal Full				Greedy Full				Greedy Pair			
	C	T	t	C	T	t	%	C	T	t	%	C	T	t	%
Narrow Caliper															
Propensity score	0.11	0.67	0	0.57	0.58	1.00	99	0.57	0.58	1.00	98	0.43	0.43	1.00	99
Index score	-1.77	0.61	0	0.17	0.21	1.00	98	0.15	0.21	1.00	98	-0.30	-0.27	1.00	99
Black	0.30	0.16	0	0.15	0.15	1.00	100	0.15	0.15	1.00	100	0.23	0.23	1.00	100
Hispanic	0.10	0.07	1	0.04	0.04	1.00	100	0.04	0.04	1.00	100	0.06	0.06	1.00	100
Age	17.50	17.65	1	17.89	17.85	1.00	69	17.92	17.84	1.00	49	17.82	17.77	1.00	63
Year of high school diploma	79.37	78.74	0	78.57	78.54	1.00	94	78.55	78.54	1.00	96	78.63	78.68	1.00	91
Math test scores	-3.95	10.02	0	8.83	8.36	1.00	96	8.19	8.38	1.00	98	5.31	5.04	1.00	97
Auto+shop test scores	3.88	7.98	0	8.44	7.78	1.00	84	7.99	7.77	1.00	93	7.82	6.56	0.98	69
Attended private school	0.03	0.12	0	0.09	0.09	1.00	80	0.11	0.09	1.00	78	0.07	0.08	1.00	83
Expelled or susp. from school	0.33	0.10	0	0.15	0.11	0.71	86	0.15	0.11	0.71	83	0.16	0.14	1.00	93
Curriculum: college prepar.	0.16	0.67	0	0.62	0.57	1.00	89	0.61	0.57	1.00	91	0.48	0.46	1.00	95
Curriculum: general	0.59	0.28	0	0.32	0.37	0.88	83	0.32	0.37	1.00	83	0.44	0.44	1.00	94
Highest grades of parents	10.50	13.21	0	12.23	12.59	0.37	85	12.36	12.59	0.45	87	11.82	12.04	1.00	91
Occupation parents high	0.08	0.29	0	0.20	0.22	1.00	89	0.24	0.22	1.00	93	0.17	0.17	1.00	94
Number of siblings	3.92	2.64	0	2.88	2.75	0.78	87	2.85	2.75	1.00	88	2.77	2.82	1.00	93
Born in south	0.38	0.33	1	0.25	0.30	0.88	-10	0.27	0.30	0.88	34	0.31	0.33	1.00	35
Mean percent bias reduction							81				84				86
– *born in south* excluded							88				88				90
Broad Caliper															
Propensity score	0.11	0.67	0	0.60	0.62	1.00	95	0.58	0.62	0.88	93	0.42	0.45	1.00	95
Index score	-1.77	0.61	0	0.21	0.36	0.89	93	0.16	0.36	0.13	91	-0.33	-0.20	1.00	95
Black	0.30	0.16	0	0.17	0.17	1.00	100	0.17	0.17	1.00	100	0.24	0.24	1.00	100
Hispanic	0.10	0.07	1	0.04	0.04	1.00	100	0.04	0.04	1.00	100	0.08	0.08	1.00	100
Age	17.50	17.65	1	17.95	17.89	1.00	61	17.97	17.89	1.00	45	17.85	17.82	1.00	68
Year of high school diploma	79.37	78.74	0	78.57	78.50	1.00	88	78.53	78.50	1.00	94	78.63	78.62	1.00	93
Math test scores	-3.95	10.02	0	8.94	9.04	1.00	98	8.24	9.04	0.95	94	4.96	5.36	1.00	97
Auto+shop test scores	3.88	7.98	0	7.73	7.63	1.00	95	7.52	7.63	1.00	96	7.32	6.48	1.00	79
Attended private school	0.03	0.12	0	0.08	0.10	1.00	78	0.11	0.10	0.88	77	0.07	0.08	1.00	82
Expelled or susp. from school	0.33	0.10	0	0.13	0.11	0.83	90	0.14	0.11	0.71	87	0.14	0.15	1.00	94
Curriculum: college prepar.	0.16	0.67	0	0.68	0.63	1.00	90	0.64	0.63	1.00	95	0.47	0.48	1.00	96
Curriculum: general	0.59	0.28	0	0.26	0.32	0.76	79	0.29	0.32	1.00	88	0.44	0.43	1.00	93
Highest grades of parents	10.50	13.21	0	12.14	12.76	0.17	77	12.18	12.76	0.36	79	11.75	12.11	0.90	87
Occupation parents high	0.08	0.29	0	0.20	0.24	0.80	80	0.23	0.24	1.00	91	0.17	0.18	1.00	93
Number of siblings	3.92	2.64	0	2.89	2.76	0.89	89	2.80	2.76	1.00	91	2.82	2.81	1.00	93
Born in south	0.38	0.33	1	0.23	0.32	0.00	-92	0.25	0.32	0.77	-35	0.32	0.34	1.00	45
Mean percent bias reduction							74				79				87
– *born in south* excluded							87				87				90

For reasons of parsimony, weighted averages over all ten years after college are shown. Weights correspond to the number of strata in each year after college. C denotes the control units while T represents treated units, t indicates whether a t-test accepts balance of covariates ($t = 1$), and % denotes percent bias reduction. The last two rows of each panel report the simple average over all single percent bias reductions excluding those of the propensity and index score.

Table 10: **Balance of Covariates: Index score distance.**

	Initially			Optimal Full				Greedy Full				Greedy Pair			
	C	T	t	C	T	t	%	C	T	t	%	C	T	t	%
Narrow Caliper															
Propensity score	0.11	0.67	0	0.57	0.58	1.00	99	0.56	0.58	1.00	97	0.43	0.44	1.00	98
Index score	-1.77	0.61	0	0.18	0.21	1.00	99	0.14	0.21	1.00	97	-0.26	-0.23	1.00	99
Black	0.30	0.16	0	0.16	0.16	1.00	100	0.16	0.16	1.00	100	0.24	0.24	1.00	100
Hispanic	0.10	0.07	1	0.05	0.05	1.00	100	0.05	0.05	1.00	100	0.08	0.08	1.00	100
Age	17.50	17.65	1	17.90	17.89	1.00	82	17.92	17.89	1.00	74	17.84	17.83	1.00	79
Year of high school diploma	79.37	78.74	0	78.56	78.50	1.00	91	78.57	78.50	1.00	89	78.62	78.61	1.00	94
Math test scores	-3.95	10.02	0	8.34	8.27	1.00	98	8.09	8.27	1.00	98	5.26	5.25	1.00	98
Auto+shop test scores	3.88	7.98	0	7.73	7.58	1.00	92	7.73	7.58	1.00	92	7.40	6.47	1.00	76
Attended private school	0.03	0.12	0	0.10	0.10	1.00	84	0.11	0.10	1.00	82	0.08	0.08	0.99	83
Expelled or susp. from school	0.33	0.10	0	0.13	0.12	0.95	90	0.14	0.12	0.83	90	0.14	0.15	1.00	93
Curriculum: college prepar.	0.16	0.67	0	0.61	0.58	1.00	93	0.61	0.58	1.00	93	0.50	0.47	1.00	93
Curriculum: general	0.59	0.28	0	0.32	0.36	0.88	85	0.32	0.36	1.00	85	0.40	0.44	1.00	88
Highest grades of parents	10.50	13.21	0	12.32	12.53	0.88	91	12.23	12.53	0.45	88	11.88	12.06	1.00	92
Occupation parents high	0.08	0.29	0	0.23	0.23	1.00	95	0.23	0.23	1.00	95	0.18	0.18	1.00	94
Number of siblings	3.92	2.64	0	2.80	2.77	1.00	92	2.79	2.77	1.00	91	2.92	2.84	1.00	91
Born in south	0.38	0.33	1	0.24	0.31	0.58	-59	0.28	0.31	1.00	18	0.33	0.35	1.00	41
Mean percent bias reduction							81				85				87
– *born in south* excluded							92				91				91
Broad Caliper															
Propensity score	0.11	0.67	0	0.61	0.63	1.00	96	0.56	0.63	0.06	87	0.43	0.47	0.98	92
Index score	-1.77	0.61	0	0.28	0.38	1.00	96	0.13	0.38	0.00	89	-0.25	-0.13	1.00	95
Black	0.30	0.16	0	0.17	0.17	1.00	100	0.17	0.17	1.00	100	0.23	0.23	1.00	100
Hispanic	0.10	0.07	1	0.06	0.06	1.00	100	0.06	0.06	1.00	100	0.09	0.09	1.00	100
Age	17.50	17.65	1	17.92	17.91	1.00	78	17.96	17.91	1.00	56	17.81	17.81	1.00	82
Year of high school diploma	79.37	78.74	0	78.55	78.47	1.00	87	78.53	78.48	1.00	92	78.64	78.61	1.00	94
Math test scores	-3.95	10.02	0	9.09	9.19	1.00	98	7.95	9.19	0.55	91	5.48	5.93	1.00	97
Auto+shop test scores	3.88	7.98	0	7.34	7.71	1.00	91	7.84	7.72	1.00	97	7.74	6.75	1.00	76
Attended private school	0.03	0.12	0	0.10	0.10	1.00	76	0.10	0.10	1.00	86	0.08	0.08	0.99	85
Expelled or susp. from school	0.33	0.10	0	0.12	0.11	0.83	88	0.13	0.11	0.76	91	0.14	0.15	1.00	94
Curriculum: college prepar.	0.16	0.67	0	0.66	0.63	1.00	94	0.63	0.63	1.00	97	0.50	0.49	1.00	97
Curriculum: general	0.59	0.28	0	0.29	0.32	1.00	89	0.30	0.32	1.00	93	0.42	0.43	1.00	94
Highest grades of parents	10.50	13.21	0	12.31	12.75	0.29	84	12.18	12.76	0.25	79	11.85	12.17	0.82	87
Occupation parents high	0.08	0.29	0	0.23	0.24	1.00	93	0.23	0.24	1.00	93	0.18	0.18	1.00	94
Number of siblings	3.92	2.64	0	2.77	2.79	1.00	91	2.75	2.78	1.00	92	2.92	2.80	0.99	86
Born in south	0.38	0.33	1	0.22	0.32	0.00	-118	0.27	0.32	1.00	-5	0.33	0.34	1.00	59
Mean percent bias reduction							75				83				89
– *born in south* excluded							90				90				91

For reasons of parsimony, weighted averages over all ten years after college are shown. Weights correspond to the number of strata in each year after college. C denotes the control units while T represents treated units, t indicates whether a t-test accepts balance of covariates ($t = 1$), and % denotes percent bias reduction. The last two rows of each panel report the simple average over all single percent bias reductions excluding those of the propensity and index score.

Table 11: **Balance of Covariates. Mahalanobis distance, index score calipers.**

	Initially			Optimal Full				Greedy Full				Greedy Pair			
	C	T	t	C	T	t	%	C	T	t	%	C	T	t	%
Narrow Caliper															
Propensity score	0.11	0.67	0	0.56	0.58	1.00	97	0.55	0.58	1.00	95	0.43	0.45	1.00	96
Index score	-1.77	0.61	0	0.15	0.21	1.00	98	0.11	0.21	1.00	96	-0.27	-0.21	1.00	98
Black	0.30	0.16	0	0.16	0.16	1.00	100	0.16	0.16	1.00	100	0.23	0.23	1.00	100
Hispanic	0.10	0.07	1	0.05	0.05	1.00	100	0.05	0.05	1.00	100	0.08	0.08	1.00	100
Age	17.50	17.65	1	17.91	17.89	1.00	79	17.93	17.89	1.00	72	17.85	17.82	1.00	71
Year of high school diploma	79.37	78.74	0	78.60	78.50	1.00	84	78.55	78.50	1.00	92	78.62	78.62	1.00	95
Math test scores	-3.95	10.02	0	8.48	8.27	1.00	98	7.86	8.27	1.00	97	5.39	5.39	1.00	98
Auto+shop test scores	3.88	7.98	0	8.20	7.58	1.00	84	7.76	7.57	1.00	92	7.46	6.51	0.99	76
Attended private school	0.03	0.12	0	0.08	0.10	1.00	83	0.11	0.09	1.00	82	0.08	0.09	1.00	81
Expelled or susp. from school	0.33	0.10	0	0.13	0.12	0.83	91	0.14	0.12	0.82	90	0.15	0.15	1.00	95
Curriculum: college prepar.	0.16	0.67	0	0.63	0.58	0.88	89	0.60	0.58	1.00	95	0.46	0.47	1.00	96
Curriculum: general	0.59	0.28	0	0.30	0.36	0.88	80	0.32	0.37	1.00	86	0.45	0.44	1.00	94
Highest grades of parents	10.50	13.21	0	12.18	12.53	0.37	86	12.22	12.53	0.37	88	11.85	12.07	1.00	92
Occupation parents high	0.08	0.29	0	0.21	0.23	1.00	91	0.23	0.23	1.00	95	0.17	0.18	1.00	93
Number of siblings	3.92	2.64	0	2.89	2.77	1.00	89	2.78	2.77	1.00	93	2.74	2.83	1.00	92
Born in south	0.38	0.33	1	0.26	0.31	0.88	-29	0.28	0.31	1.00	17	0.33	0.35	1.00	48
Mean percent bias reduction							80				86				88
– *born in south* excluded							89				91				91
Broad Caliper															
Propensity score	0.11	0.67	0	0.56	0.63	0.05	87	0.53	0.63	0.00	83	0.42	0.49	0.05	87
Index score	-1.77	0.61	0	0.13	0.38	0.00	89	0.05	0.38	0.00	86	-0.29	-0.07	0.14	91
Black	0.30	0.16	0	0.17	0.17	1.00	100	0.17	0.17	1.00	100	0.22	0.22	1.00	100
Hispanic	0.10	0.07	1	0.06	0.06	1.00	100	0.06	0.06	1.00	100	0.08	0.08	1.00	100
Age	17.50	17.65	1	17.97	17.91	1.00	62	17.98	17.91	1.00	52	17.83	17.80	1.00	76
Year of high school diploma	79.37	78.74	0	78.55	78.47	1.00	88	78.51	78.47	1.00	94	78.64	78.61	1.00	94
Math test scores	-3.95	10.02	0	8.28	9.19	1.00	93	7.52	9.19	0.00	88	5.39	6.31	0.99	93
Auto+shop test scores	3.88	7.98	0	8.40	7.71	1.00	83	8.01	7.72	1.00	92	7.76	6.87	1.00	78
Attended private school	0.03	0.12	0	0.09	0.10	1.00	78	0.10	0.10	1.00	85	0.07	0.09	1.00	77
Expelled or susp. from school	0.33	0.10	0	0.12	0.11	0.88	93	0.14	0.11	0.82	90	0.12	0.15	1.00	87
Curriculum: college prepar.	0.16	0.67	0	0.64	0.63	1.00	99	0.58	0.63	0.88	90	0.43	0.50	0.96	86
Curriculum: general	0.59	0.28	0	0.30	0.32	1.00	94	0.34	0.32	1.00	93	0.48	0.42	0.98	80
Highest grades of parents	10.50	13.21	0	12.12	12.75	0.00	77	12.14	12.75	0.13	77	11.80	12.21	0.59	85
Occupation parents high	0.08	0.29	0	0.21	0.24	1.00	87	0.22	0.24	1.00	93	0.16	0.19	0.98	85
Number of siblings	3.92	2.64	0	2.87	2.79	1.00	91	2.73	2.79	1.00	93	2.83	2.81	1.00	94
Born in south	0.38	0.33	1	0.26	0.32	0.93	-30	0.29	0.32	1.00	37	0.35	0.34	1.00	59
Mean percent bias reduction							80				85				85
– *born in south* excluded							88				88				87

For reasons of parsimony, weighted averages over all ten years after college are shown. Weights correspond to the number of strata in each year after college. C denotes the control units while T represents treated units, t indicates whether a t-test accepts balance of covariates ($t = 1$), and % denotes percent bias reduction. The last two rows of each panel report the simple average over all single percent bias reductions excluding those of the propensity and index score.