Matching using Semiparametric Propensity Scores

Gregory Kordas
Department of Economics
University of Pennsylvania
kordas@ssc.upenn.edu

Steven F. Lehrer
SPS and Department of Economics
Queen's University
lehrers@post.queensu.ca

October 2004
First Version: December 2002

Abstract

This paper considers the application of semiparametric methods to estimate propensity scores or probabilities of program participation, which are in central in program evaluation methods. To evaluate the practical benefits we use data from the NSW experiment, CPS and PSID. We compare treatment effect and evaluation bias estimates using propensity scores estimated from parametric logit, semiparametric single index and semiparametric binary quantile regression models. We find that the quantile regression model exhibits the smallest average absolute bias error. Our results suggest that it is important to account for very general forms of heterogeneity in (semiparametric) estimation of the propensity score.

JEL codes: C14, C81, C99, H53, I38

*We are grateful to Mianna Plesca and seminar participants at the 2004 NASM of the Econometric Society meetings, 2003 CEA meetings, 2003 iHEA World Congress, Concordia University, Florida State University, Lehigh University, McGill University, Queens University, Simon Fraser University, SUNY Albany, UNC-Greensboro, University of South Florida and the Wharton School, University of Pennsylvania for comments and suggestions which have helped to improve this paper. We are responsible for all errors.

1 Introduction

An increasing body of evidence has found that there is significant diversity and heterogeneity in response to a given policy. Heckman (2001) argues that this has profound consequences for economic theory and for economic practice. In particular, accounting for heterogeneity may improve the performance of non-experimental estimators. In this paper, we introduce and evaluate the performance of a semiparametric propensity score matching estimation strategy that explicitly accounts for heterogeneity in response across observed covariates along the conditional willingness to participate in the treatment intervention distribution.

Matching estimators evaluate the effects of a treatment intervention by comparing outcomes such as wages, employment, fertility or mortality for treated persons to those of similar persons in a comparison group. The use of the propensity score as a basis for matching treated and untreated individuals (and thus for evaluating the magnitude of treatment effects) is becoming increasingly common in clinical medicine, demographic and economic research. The propensity score is defined as the conditional probability of being treated given the individual's covariates and requires the assumption of selection on observables.[1]

Existing studies use parametric estimators of binary response models, such as the probit and logit which imposes strong distributional assumptions on the underlying data. In particular, the dangers of misspecification may be severe if the error terms are not independent and identically distributed from their known parametric distributions.[2] Kordas (2004) outlines the benefits of using Manski's (1975, 1985) binary regression quantiles to provide consistent estimates of the conditional probability at different points of the distribution. This estimator avoids the distributional restrictions embedded in the parametric approach and has the advantage that it is robust and can accommodate heteroskedasticity of very general form. This property is extremely valuable in our setting as the estimator can accommodate problems of heterogeneity, self-selection and misclassification.

Todd (1999) presents the only other study that we are aware of that considers matching using semiparametrically estimated propensity scores. She considers matching using the estimated probabilities from both the semiparametric least squares estimator of Ichimura (1993) and the quasi maximum likelihood estimator of Klein and Spady (1993).[3] Her Monte Carlo study demonstrates that the gains from semiparametric procedures relative to parametric alternatives are greatest when either the systematic component of the model is misspecified or when the error distribution is highly asymmetric.

Our approach offers several additional benefits for empirical researchers. First, this estimator does not require the researcher to select higher order or interaction terms to ensure balancing of covariates across the treatment and control groups. While recent work in economics (Dehejia and Wahba, 2002) has proposed the use of balancing tests to determine if additional higher order or interaction terms should be included in the estimates of the propensity scores but does not provide guidance on precisely which of these terms should be included.[4] Second, quantile treatment effects are simple to calculate permitting an examination of the average treatment effect on the treated at different points along the probability of participation distribution.

2 Framework

Let $D_i = 1$ indicate if person i received treatment and $D_i = 0$ if not. Let Y_{1i} and Y_{0i} are the outcome of interest if person i received treatment or did not respectively. Evaluators are often interested in the average effect of the treatment on the treated ($ATT_{D=1}(X)$), which can be defined conditional on some characteristics X as $ATT_{D=1}(X) = E(Y_1 - Y_0|X, D = 1) = E(Y_1|X, D = 1) - E(Y_0|X, D = 1)$. In the absence of a randomized control group, matching methods allow for the construction of a comparison group for the treated under the assumption that conditional on observed covariates assignment to treatment is random. In many empirical

applications, the dimension of these observed characteristics is high and Rosenbaum and Rubin (1983) suggest using a scalar measure $P(X)$, where

$$P(X) = Pr(D_i = 1|X). \tag{1}$$

is the propensity score.[5]

Implementation involves two steps. First, equation 1 is estimated to calculate the propensity score. Second, a matching algorithm is used to construct the matched comparison group for the treated. Algorithms differ in the weights they place on individuals in the comparison group.

2.1 Econometric Methods

Let D^* denote the latent *propensity to participate* index. D^* measures the (indirect) latent utility differential between receiving and not receiving the treatment. We consider three estimation approaches, a parametric logit, the semiparametric single-index model of Klein and Spady (1993), and Manski's (1975, 1985) binary regression quantiles. The logit and single-index estimators maximize the familiar log-likelihood of the binary response model,

$$\ell(\beta) = n^{-1} \sum_{i=1}^{n} D_i \log(P_i) + (1 - D_i) \log(1 - P_i), \tag{2}$$

differing from each other only in the specification of $P_i \equiv \Pr(Y_i = 1|X_i'\beta)$. The homoskedastic logit model is given by

$$P_i^L = L(X_i'\beta/\sigma) \equiv \frac{1}{1 + \exp(-X_i'\beta/\sigma)}. \tag{3}$$

In this model, the covariates exert a pure location effect on the latent utility differential. To make the model more flexible researchers often include higher order and interaction terms which are selected based on balancing tests results (i.e. Hotelling T^2 tests for differences in means) which determine whether a covariate adds information on the selection process conditional on the propensity score.

The semiparametric procedures avoid the distributional and other restrictions embedded in the parametric specification of $\Pr(D_i = 1|X_i)$. First, we consider the semiparametric *single-index* model of Klein and Spady (1993). To define this estimator, let $K(\cdot)$ be a kernel function (in this paper we use a Gaussian kernel function), h_n be a bandwidth parameter that converges to zero as n becomes large, and define the leave-i-out estimate of $\Pr(Y = 1|X_i'\beta)$ by

$$\tilde{P}_i = \sum_{j \neq i} D_j K\left(\frac{X_i'\beta - X_j'\beta}{h_n}\right) \Bigg/ \sum_{j \neq i} K\left(\frac{X_i'\beta - X_j'\beta}{h_n}\right).$$

To avoid anomalous behavior at the tails, let ε be a small positive number and define

$$P_i^{SI} = \max\{\varepsilon, \min\{1 - \varepsilon, \tilde{P}_i\}\}. \tag{4}$$

Klein and Spady (1993) have shown that, under some additional regularity conditions, if $nh_n^6 \to \infty$ and $nh_n^8 \to 0$ as $n \to \infty$, the estimator converges at the parametric root-n rate to an asymptotically normal random variable, and achieves the semiparametric efficiency bound.

Our second semiparametric estimation strategy is Manski's (1975, 1985) binary regression quantiles. We assume the following linear quantile specification of D_i^*

$$Q_{D_i^*}(q|X_i) = X_i'\alpha(q), \qquad q \in (0, 1). \tag{5}$$

where $\alpha(q)$ is the coefficient vector of the q-th conditional quantile. Using the equivariance property of quantile functions with respect to monotonic transformations, we can write the conditional quantile function of $D_i = 1\{D_i^* \geq 0\}$ as (See Kordas (2004) equation 6)

$$Q_{D_i}(q|X_i) \equiv Q_{1\{D_i^* \geq 0\}}(q|X_i) = 1\{Q_{D_i^*}(q|X_i) \geq 0\} = 1\{X_i'\alpha(q) \geq 0\}. \tag{6}$$

This estimator is the binary response analogue to the linear quantile regression estimator introduced by Koenker and Bassett (1978) and offers a robust and efficient semiparametric alternative to commonly used parametric models. From an empirical point of view, their main

advantage is their ability to model very general forms of population heterogeneity by allowing the coefficient vector to vary across the conditional quantiles of the dependent variable.

Estimates of the scaled coefficients $\alpha(q)$ such that $||\alpha(q) = 1||$, are obtained by solving the quantile regression problem

$$\alpha(q) = \text{argmin}_{a:||a||=1} \left\{ S_N(a) = N^{-1} \sum_{i=1}^{N} \rho_q(D_i - 1\{X_i'a \geq 0\}) \right\}, \tag{7}$$

where $\rho_q(u) = (q - 1\{u < 0\}) \cdot u$, and $S_N(\cdot)$ is the *score function*. The score function is a multimodal step function so optimization is performed using the simulated annealing algorithm of Goffe et al. (1994). The discontinuities of the score function also affect the asymptotic behavior of the estimators that have been shown to converge at the slow $N^{1/3}$ rate to a non-gaussian random variable (Kim and Pollard, 1990). To overcome these problems Horowitz (1992) smoothed the median score function and derived a smoothed median estimator that is asymptotically normally distributed. Kordas (2004) extended these results to show joint asymptotic normality of families of smoothed binary quantile estimates and showed how these smoothed estimates may be optimally combined for efficient estimation. Note the single index and binary regression quantile models are not nested.

Since our focus is on estimating propensity scores only unsmoothed estimates will be computed here. The probabilities are computed by noting that the quantile regression model in (6) implies that if an individual's q-th conditional quantile $X_i'\alpha(q)$ is (approximately) equal to zero, his conditional probability of receiving treatment is (approximately) equal to $1 - q$, i.e.,

$$\Pr(D_i = 1 | X'\alpha(q) = 0) = 1 - q. \tag{8}$$

Given estimates of $\alpha(q)$ over a grid $\theta = \{q_1, q_2, \cdots, q_M | q_1 < q_2 <, \cdots < q_M\}$ of quantiles, Kordas (2004) shows how this equation may be used to derive semiparametric *interval* probability estimates. Let

$$\hat{q}_i = \text{argmin}_{q \in \theta} \{q : X_i'\alpha(q) \geq 0\} \tag{9}$$

be the smallest quantile in the grid for which i's index function is positive. Then an interval estimate of the conditional probability $P_{i,1|X_i} \equiv Pr(D_i = 1|X_i)$ is given by

$$\hat{P}_{i,1|X_i(\theta)} = [1 - \hat{q}_i, 1 - \hat{q}_{-1,i}), \tag{10}$$

where $\hat{q}_{-1,i}$ denotes the quantile immediately preceding $\hat{q}_i$. In our application, $\theta = \{0.05, 0.10, \cdots, 0.95\}$, so, for example, if i's quantile indices are negative for quantiles below 0.70 and are positive for quantiles 0.70 and above, $\hat{q}_i = 0.70$, and $\hat{P}_{i,1|X_i(\theta)} = [0.30, 0.35)$.

2.2 Matching Algorithm

Since the estimated choice probabilities from binary regression quantiles are discrete (interval probabilities) the average treatment effect on the treated ($ATT_{D=1}(X)$) is calculated using stratification matching. At each probability interval, we compute the difference in average outcomes of treated and controls, providing an estimate of the quantile treatment effect ($ATT_{D=1}(X)^q$), $q = 1, 2, ..., Q$

$$ATT_{D=1}(X)^q = \frac{\sum_{i \in L_q} Y_{1i}}{N_q^1} - \frac{\sum_{j \in L_q} Y_{0j}}{N_q^0} \tag{11}$$

where N_q^1 and N_q^0 number of treated and untreated individuals at quantile q respectively. The average treatment effect on the treated is computed using a weighted (by the number of treated) average of these quantile treatment effects as

$$ATT_{D=1}(X) = \sum_{q=1}^{Q} ATT_{D=1}(X)^q * \frac{\sum_{i \in N_q^1} D_i}{\sum_{i \in N_1} D_i} \tag{12}$$

where Q is the total number of quantiles estimated and N_1 is the total number of treated individuals that are matched. Assuming independence of outcomes across units, the variance of $ATT_{D=1}(X)$ is given by

$$Var(ATT_{D=1}(X)) = \frac{1}{N_1} \left\{ Var(Y_{1i}) + \sum_{q=1}^{Q} \frac{N_q^1}{N_1} * \frac{N_q^1}{N_q^0} Var(Y_{0j}) \right\} \tag{13}$$

Bootstrapped standard errors could also be calculated and are presented as we are matching on the estimated and not the actual propensity score.[6] Since the estimated participation probabilities from the logit and Klein and Spady (1993) models are continuous, we consider a larger variety of matching algorithms that are described in Smith and Todd (2004).

3 Returns to the NSW Job Training Program

3.1 Data

Following LaLonde (1986), numerous studies have examined whether econometric (non-experimental) estimators recover impacts on post-intervention outcomes that are similar to those produced from a randomized experiment such as the National Supported Work Demonstration program (NSW).[7] Experimental treated units are combined with non-experimental comparison units drawn from two national survey datasets; CPS and PSID. Treatment effect estimates obtained using econometric estimators are then compared with the benchmark results from the experiment. This exercise presents challenges for non-experimental estimator since there are substantial differences in demographic and economic characteristics between individuals in the CPS, PSID and the experimental samples.

Studies evaluating propensity score matching with this data initially found that these methods were able to replicate experimental treatment effects (Dehejia and Wahba (1999)). More recent evidence calls these findings in question (Smith and Todd (2004)) and indicate that accounting for permanent unobserved heterogeneity does lower the estimated bias with propensity score matching estimators. We follow Smith and Todd (2004) by considering three alternative experimental samples from the NSW data (LaLonde's full sample, the Dehejia and Wahba extract, an extract containing subjects assigned in the first four months of the program) in addition to the survey data.[8]

3.2 Propensity Score Estimates

In practice, researchers focus on the quality of matches obtained from propensity score estimation and specifications are selected with the best balance between matched individuals. In our application, we consider two alternative specifications for the binary response model i) based on Dehejia and Wahba (1999,2002) that includes higher order and interaction terms to satisfy balancing tests and ii) omitting higher order and interaction terms since binary regression quantiles have the desirable property of being robust to heteroskedasticity of general forms. Note, a slightly different set of higher order and interaction terms are used for specifications with the CPS and PSID samples.[9]

This empirical approach used in the propensity score matching literature to create specification one differs dramatically from traditional diagnostics used to assess binary response models which are considered with parameter estimates that are used in calculating propensity scores. Likelihood ratio tests between the logit and heteroskedastic logit strongly reject the null hypothesis of a homoskedastic residual.[10] Thus, the parametric homoskedastic binary response models are misspecified.

A graphical examination of the normalized quantile, logit and single index model coefficient estimates across quantiles indicate disagreements between binary regression quantile estimates and the other approaches only appear at the higher quantiles as the parametric and single index models under predict the probability of participation. The general pattern of over and under prediction by the logit models could also be used to provide further evidence of the restrictiveness of the parametric model which tend to extrapolate the behavior of individuals near the mean to individuals that belong in the tails of the propensity to participate distribution. In general, for treatment effects and evaluation bias estimates the results are similar from propensity scores estimated using the Klein and Spady (1993) and logit model so we will focus our discussion on binary regression quantiles.[11]

3.3 Treatment Effects

Table 1 presents estimates for both specifications of the causal effect of the NSW Work Demonstration on earnings in calender year 1978 based on stratification matching with binary regression quantile propensity scores. The rows differ solely in the number of bins that are employed excluding the lowest probability bin in the top panel. While the experimental impact is captured within a 95% bootstrapped confidence interval, the estimates are extremely accurate for each experimental sample matched with the PSID (even numbered columns). Notice that the results with twenty bins are practically identical between specifications 1 and 2. Further, the results do not appear to be sensitive to the number of bins used to stratify the sample match. The results improve with fewer bins for some subsamples but weaken for others such as column 2 and 6 whose estimated magnitude decreases by approximately 67% when only five bins are used.

Table 2 presents Hotelling T^2 tests for differences in means (i.e. balance) within each quintile probability interval. Each entry lists the number of covariates which failed the test at the 5% level.[12] Notice that there are significant failures at the lowest probability interval capturing the dissimilarities between the experimental and CPS non-experimental samples. The bottom panel of Table 1 demonstrates a dramatic reduction in the estimated magnitude of the treatment effect for the columns matched with CPS sample when the lowest probability interval bin is included.[13]

In Figure 1, we graph quantile treatment effects for the sample that corresponds to specification 2 and column 6 of Table 1. With balanced covariates the quantile treatment effects have a clear interpretation. Notice that the training program had a negative impact for those subjects in the middle quantiles who tend to be either blacks or Hispanics that had low earnings in 1974 but high earnings in 1975 if the control group was drawn from the PSID. Similarly, the largest gains from NSW were achieved at the extreme quantiles containing individuals with low

earnings in 1975.

3.4 Evaluation Bias Estimates

An alternative approach to evaluate non-experimental estimates is to match the randomized out control group with the non-experimental samples. As neither group has received the intervention the difference in earnings between matched individuals from each experimental control group and non-experimental sample should be zero. Any deviation is evaluation bias. Table 3 presents direct estimates of the bias using stratification matching with binary regression quantile propensity scores. Notice that with the exception of column 4 of specification one, the bias is of the order of a few hundred dollars and is less than 15% of the experimental treatment impact in columns 2, 3, 5 and 6 respectively. Unlike the treatment effect estimates, including the lowest probability quintile has little effect on the bias since fewer individuals are assigned to this probability interval.

The evaluation bias increases by approximately $200 in column 1- 3 when the higher order and interaction terms are omitted from the estimating equation. Column 6 continues to exhibit low bias whereas column 5's bias is also reduced in absolute value. Surprisingly we find a high degree of bias in the Dehejia and Wahba samples, columns which exhibit low bias with parametric propensity scores.

To uncover an explanation as to why the evaluation bias calculated using semiparametric propensity scores exceeded the estimate obtained using parametric propensity scores in column 4 of Table 4 we conducted a more detailed examination of how the estimated bias differs across quantiles. Figure 2 presents a graph of the quantile bias effects at each interval for both parametric and semiparametric propensity scores for specification 1. Notice that in almost all quantiles the semiparametric procedure exhibits lower bias. The results in Table 3 (and) present a number of treated individuals weighted average of these quintile biases and suggest

that the lower bias for the Dehejia and Wahba subsample is based in part on having the larger biases across quantiles cancel out.

To provide additional guidance for empirical researchers on the performance of propensity score matching algorithms we compare the average absolute bias error of our matching algorithm with a variety of different matching algorithms based on parametric propensity scores.[14] For each matched outcome we calculate the absolute bias error

$$Bias\ error = |Y_{1i} - \hat{E}(Y_{0i}|P(X_i), D_i = 0)|$$

where $\hat{E}(Y_{0i}|P(X_i), D_i = 0)$ is calculated by the algorithm under investigation. The average absolute bias error is calculate by dividing the sum of these bias errors by the number of individuals in the treatment group who were successfully matched. Table 4 reports summary statistics on the average absolute bias error.[15]

Notice that with one exception, the smallest average absolute bias error is attained using stratification matching with propensity scores calculated by binary regression quantiles. In general, bias error estimates obtained by stratification matching procedures are smaller than the nonparametric and distance metric algorithms. In general when using parametric propensity scores algorithms that use a larger distance produce smooth results; whereas narrow intervals produce larger bias errors on average. In part, this occurs since fewer individuals have matches as the distance shrinks. The results from specification 1 find that Kernel and local linear matching estimator exhibit significantly less bias error than nearest neighbor or caliper matching algorithms. Overall, it appears that using 20 bins produces estimates with the smallest mean squared error. The results suggest that adding the lowest probability quantile to the stratification matching algorithm increases bias up to an average of \$500 and \$670 per treated participant for specification 1 and 2 respectively.

The increased average size of the bias error from parametric procedures ranges from slightly more than \$55.00 to approximately \$5200 for specification 1. As a percentage of the estimated

treatment impact this range is equivalent 6.2% to 586.9%. For specification 2, stratification matching using parametric propensity scores does exhibit smaller bias error for column 3.[16] Of the remaining columns, the size of the average bias error ranges from \$91 to \$6250 or 10.3% to 706% of the experimental treatment impact per matched treated individual. While the semiparametric procedure yielded the smallest average absolute bias error in 11 of the 12 columns in Table 4, the number is still large relative to the experimental impact. This casts doubt as to whether all observables were included in the estimation of the propensity score and is a potential cause for concern for empirical researchers interested in using these methods. After all, the matched individuals using one nearest neighbor matching intuitively should be most alike, yet the average absolute bias error is extremely large.

Stratification matching with parametric and binary regression quantile propensity scores yield similar average absolute bias error but wildly different treatment effects. In general if one excludes the lowest probability bin (0.0-0.05%) the procedures rarely placed individuals within the same interval. This is demonstrated by examining the scarcity of individuals lying on the prime diagonal of Table 5 and the large number of individuals residing in the off diagonal elements. This table presents information on the horizontal rows of which bin the semiparametric procedure assigns and the columns provide the bins that the parametric procedure assigns. Notice that ignoring the lowest probability quantile, approximately 30% of all the observations fall in the same probability bin for the two methods. For all 12 subsamples the similarities range between 22%-43%.

The improved performance of stratification matching relative to other algorithms with parametric propensity scores is due in part to the balancing tests which occur over large intervals. Replicating these tests in smaller bins (1%) indicate substantial failures with each sample. In contrast, using smaller bins constructed by binary regression quantiles we found little evidence of increased failures in balancing tests.

Finally, we compared binary regression quantiles (Table 4) to the Klein and Spady (1993) estimator. From a heterogeneity perspective the major distinction between these estimators is that the former allows for very general forms while the latter is more limited only allowing heterogeneity through the index. The results using specification two are provided in Appendix Table 4. Notice again that the stratification matching estimates with binary regression quantiles are substantially lower in absolute bias error for nearly all columns.

4 Conclusions

In situations with non-experimental data matching methods provide a means to estimate program impacts when the variables determining assignment to treatment are observed and the support of treatment and comparison groups overlap. Since the use of the propensity score as a basis for estimating treatment effects is becoming increasingly common in research in a variety of disciplines researchers should test for possible model misspecification and if present, consider the methods described above to improve inference and reduce concerns regarding the specific higher order and interaction terms to include when estimating the treatment program participation equation.

Notes

[1]The assumption of selection on observables requires that conditioning on the observed variables the assignment to treatment is random. Propensity score matching (Rosenbaum and Rubin (1983)) reduce the dimensionality of having to match participants and non participants on the set of conditioning variables (X) by matching solely on the basis estimated propensity scores ($P(X)$).

[2]For example, Horowitz (1993) demonstrates that misspecification is likely to be severe under heteroskedasticity and bimodality.

[3]These methods estimate a conditional mean and overcome the distributional restrictions embedded in the parametric approach but allows for only limited forms of heterogeneity. Note, we also consider the Klein and Spady (1993) estimator.

[4]See Dehejia (2004) and Smith and Todd (2004) for a demonstration of the difficulties in choosing the appropriate specification of higher order and interaction terms.

[5]This estimator assumes i) $E(Y_0|P(X), D = 1) = E(Y_0|P(X), D = 1)$ ($0 < \Pr(D = 1|X) < 1$) and ii) .

[6]Notice that if a quantile contains numerous treated units and few controls it will increase the variance $ATT_{D=1}(X)$. Quantiles with few treated and many controls work in an opposite manner but receive little weight in the calculation of $ATT_{D=1}(X)$.

[7]Our results uses the same data as in LaLonde (1986), Heckman and Hotz (1989), Dehejia and Wahba (1999, 2002, 2004), Abadie and Imbens (2002) and Smith and Todd (2002, 2004) among others, making it easier to draw comparisions.

[8]See Appendix Table 1 or Smith and Todd (2002) for further information on these datasets.

[9]The selection of variables to include in the estimation of the propensity score is very important since even small changes in the estimated probabilities can dramatically affect the magnitude of treatment effects in the matching stage and cause a substantial difference in the

amount of bias present in the matching estimator. See Heckman, Ichimura, Smith and Todd (1998), Smith and Todd (2002) or Dehejia and Wahba (2004) for a discussion. Finally it is worth noting that while Dehejia and Wahba (1999, 2002) did not find evidence that the treatment effect estimated was sensitive to the inclusion of these terms, they stress the importance of variable selection to ensure that the balancing hypothesis is satisfied.

[10]This assumption is rejected below the 5% level for all columns and both specifications with the exception of column 5 in specification 1 which is rejected at the 15% level.

[11]For several samples, we experimented with different bandwiths with Klein and Spady (1993) and the results were robust. The full set of results are available from the authors by request.

[12]The results do not change significantly if we report the 10% or 20% level.

[13]Note these individuals are generally not included in the parametric matching procedures due to trimming conditions as in Dehejia and Wahba (1999). See Appendix Table 2 for estimates corresponding to logit propensity scores

[14]We also considered Abadie and Imbens (2004) matching procedure with homoskedastic and heteroskedastic weighting matrices and either one or four individuals matched. The results indicated larger absolute bias error than local linear matching with a bandwidth of 0.01.

[15]For the parametric propensity score we match on the estimated propensity score with the exception of kernel and local linear matching estimators where we match on the odds ratio due to the choice based nature of the sample. Heckman and Todd (1995) demonstrate that matching methods can be applied with the odds ratio to gain consistent estimates when the sample is choice based. Note failure to account for choice based samples should not affect nearest neighbor or stratification point estimates. Finally, following Smith and Todd (2004) we investigated the sensitivity of our results to alternative seeds for several of these algorithms. There were no major shifts in the magnitude of the absolute bias error.

[16]This column requires further study as it exhibited a significant balancing test failures.

References

[1] Abadie, A. and G. Imbens, "Large Sample Properties of Matching Estimators for Average Treatment Effects," *mimeo*, University of California at Berkeley (2004).

[2] Dehejia, R., "Practical Propensity Score Matching: A Reply to Smith and Todd," forthcoming in *Journal of Econometrics*, (2004).

[3] Dehejia, R., and S. Wahba, "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 94:448 (1999), 1053-1062.

[4] Dehejia, R., and S. Wahba, "Propensity Score Matching Methods for Non-Experimental Causal Studies," *Review of Economics and Statistics* 84:1 (2002), 151-161.

[5] Goffe, W., G. D. Ferrier and J. Rogers, "Global Optimization of Statistical Functions with Simulated Annealing," *Journal of Econometrics* 60:1/2 (1994), 65–101.

[6] Heckman, J. J., "Heterogeneity, Microdata and the Evaluation of Public Policy: Nobel Lecture," J*ournal of Political Economy* 109:4 (2001), 673-748.

[7] Heckman, J. J., H. Ichimura, J. Smith and P. Todd, "Characterizing Selection Bias using Experimental Data," *Econometrica* 66:5 (1998), 1017-1098.

[8] Heckman, J. J., H. Ichimura and P. Todd, "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65:2 (1998), 261-294.

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

[10] Heckman, J. J. and P. Todd, "Adapting Propensity Score Matching and Selection Models to Choice-based Samples," *mimeo*, University of Chicago (1995).

[11] Heckman, J. J., and J. V. Hotz, "Choosing Among Alternative Non-Experimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," J*ournal of the American Statistical Association* 84:408 (1989), 862-880.

[12] Horowitz, J., "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60:3 (1992), 505-531.

[13] Horowitz, J., "Semiparametric and Nonparametric Estimation of Quantal Response Models," in *Handbook of Statistics*, Vol 11 (1993), Maddala, GS and Vinod, HD (eds.), North-Holland.

[14] Ichimura, H., "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58:1-2 (1993), 71-120.

[15] Kim, J. and D. Pollard, "Cube Root Asymptotics," *Annals of Statistics* 18 (1990), 191–219.

[16] Klein, R. and R. Spady, "An Efficient Semiparametric Estimator for Discrete Choice Models," *Econometrica* 61:2 (1993), 387-422.

[17] Koenker, R. and G. Bassett, Jr., "Regression Quantiles," *Econometrica* 46:1 (1978), 33-50.

[18] Kordas, G., "Smoothed Binary Regression Quantiles," forthcoming in *Journal of Applied Econometrics* (2004).

[19] LaLonde, R., "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review* 76:4 (1986), 604-620.

[20] Manski, C. F., "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator." *Journal of Econometrics* 27: (1975), 313-333.

[21] Manski, C. F., "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics* 3: (1975), 205-228.

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

[23] Smith, J., and P. Todd, "Rejoinder to Does Matching Overcome LaLonde's Critique of Non-Experimental Estimators" forthcoming in *Journal of Econometrics* (2004).

[24] Smith, J., and P. Todd, "Does Matching Overcome LaLonde's Critique of Non-Experimental Estimators," forthcoming in *Journal of Econometrics* (2002).

[25] Todd, P., "Local Linear Approaches to Program Evaluation using a Semiparametric Propensity Score", *mimeo*, University of Pennsylvania (2002).

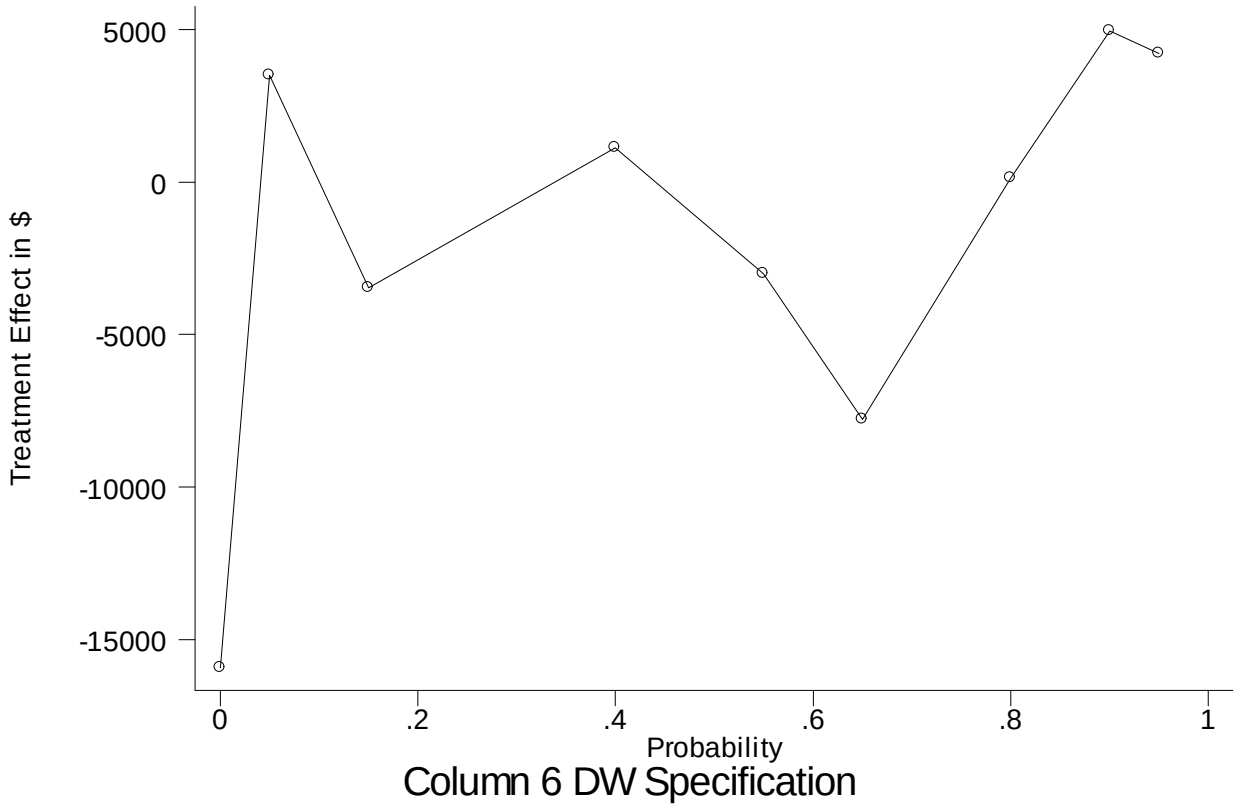

Figure 1: Quantile Treatment Effects

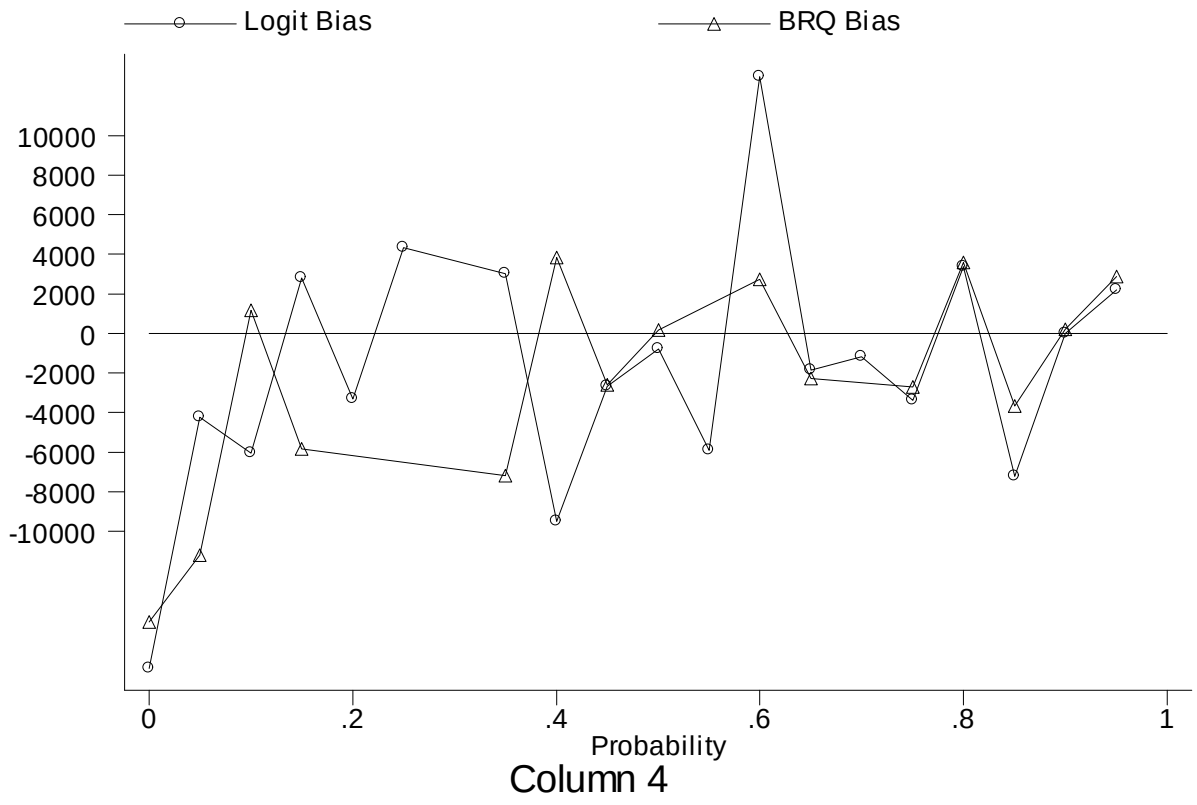

Figure 2: Estimated Bias Parametric and Semiparametric Estimates Column 3 of Table 4

Table 1: Treatment Effects Estimates with Binary Regression Quantile Propensity Scores using Stratification Matching

	SPECIFICATION ONE						**SPECIFICATION TWO**					
	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Experiment Impact	886.32	886.32	1794.34	1794.34	2748.49	2748.49	886.32	886.32	1794.34	1794.34	2748.49	2748.49
20 Bins	85.44 (618.77)	665.55 (900.59)	1925.19 (678.11)	1390.79 (675.37)	2779.06 (1047.0)	2697.48 (1624.8)	96.62 (587.43)	652.04 (874.25)	2039.67 (831.12)	1406.82 (705.95)	2556.02 (1196.2)	2418.79 (1619.5)
10 Bins	114.13 (508.81)	773.91 (886.37)	1391.97 (935.18)	1206.23 (684.84)	2051.61 (1242.6)	2922.78 (1329.7)	-201.02 (584.42)	-179.87 (1234.5)	2145.43 (866.29)	1441.15 (674.73)	2750.53 (1202.6)	2989.07 (1326.6)
5 Bins	221.76 (667.10)	126.80 (945.53)	2400.79 (961.73)	1258.74 (669.76)	2635.53 (1903.0)	963.14 (1933.2)	-34.71 (593.42)	-568.94 (1102.3)	2205.51 (873.32)	1436.19 (668.59)	2640.13 (1226.4)	2852.39 (1143.1)
Including Lowest Probability Bin												
20 Bins	-627.70 (574.16)	476.12 (841.31)	223.27 (833.20)	1108.07 (681.35)	592.24 (1127.3)	1834.56 (1635.9)	-881.44 (535.94)	389.32 (874.25)	684.29 (736.18)	1014.49 (710.99)	1070.34 (1057.5)	1664.49 (1509.2)
10 Bins	-850.68 (649.83)	516.66 (837.60)	253.79 (804.42)	720.96 (669.84)	-116.96 (1092.5)	1769.43 (1337.1)	-1267.4 (550.35)	-561.83 (1284.1)	470.36 (779.58)	825.42 (702.81)	621.01 (1068.0)	1816.20 (1291.6)
5 Bins	-1157.1 (622.46)	-838.17 (854.67)	73.13 (891.22)	488.72 (735.58)	-668.71 (1075.1)	-391.72 (1903.0)	-1447.6 (510.77)	-1455.2 (1036.5)	-206.31 (820.30)	577.36 (700.04)	156.02 (1042.2)	1274.53 (1153.3)

Note: Bootstrapped standard errors in parentheses. 1000 Bootstrap replications.

Table 2: Balancing Test Results

	SPECIFICATION ONE						**SPECIFICATION TWO**					
Quantile	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
0↔0.05	11	2	14	4	12	4	7	3	10	3	10	3
0.05↔0.10	0	0	0	0	1	0	0	0	3	0	0	0
0.10↔0.15	0	0	0	0	0	0	0	0	0	1	0	0
0.15↔0.20	0	0	1	0	0	2	0	0	3	0	0	0
0.20↔0.25	0	0	0	0	0	0	0	2	0	0	0	0
0.25↔0.30	0	2	0	0	0	0	0	0	0	0	0	0
0.30↔0.35	0	2	0	0	0	0	2	0	2	0	0	0
0.35↔0.40	3	1	0	0	0	0	0	0	0	0	1	0
0.40↔0.45	0	0	0	0	0	1	0	0	0	0	3	0
0.45↔0.50	0	0	1	0	0	0	0	0	2	0	0	0
0.50↔0.55	1	0	0	1	0	0	0	0	0	0	0	0
0.55↔0.60	0	0	0	0	0	0	1	0	6	0	0	0
0.60↔0.65	0	0	0	0	0	0	0	0	0	0	0	0
0.65↔0.70	0	0	0	0	0	0	4	0	0	0	2	0
0.70↔0.75	0	1	0	0	0	0	2	1	1	0	2	0
0.75↔0.80	0	2	0	0	0	0	0	0	0	0	0	0
0.80↔0.85	0	0	0	0	0	1	0	0	0	0	0	0
0.85↔0.90	0	0	0	0	0	0	1	0	1	0	0	0
0.90↔0.95	0	0	0	0	0	0	0	0	0	0	0	0
0.95↔1.00	0	0	0	0	0	0	0	0	0	0	0	0

Note: Number of unbalanced covariates at the 5% level reported.

Table 3: Evaluation Bias Estimates with Binary Regression Quantile Propensity Scores using Stratification Matching

	SPECIFICATION ONE						**SPECIFICATION TWO**					
	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Experiment Impact	886.32	886.32	1794.34	1794.34	2748.49	2748.49	886.32	886.32	1794.34	1794.34	2748.49	2748.49
20 Bins	-336.54 (480.81)	-139.93 (810.83)	161.28 (741.64)	1673.97 (746.29)	-589.35 (461.54)	515.52 (641.74)	436.09 (640.18)	1916.99 (704.56)	-639.12 (680.63)	153.73 (972.87)	2556.02 (1196.2)	2418.79 (1619.5)
10 Bins	-335.03 (594.49)	241.27 (853.37)	384.81 (771.63)	1465.84 (812.25)	-908.48 (447.11)	-903.74 (23.12)	642.51 (640.09)	1648.89 (755.05)	-536.74 (716.03)	674.81 (995.20)	2750.53 (1202.6)	2989.07 (1326.6)
5 Bins	-405.13 (556.55)	-735.40 (861.99)	1138.15 (781.37)	1666.38 (649.94)	-983.01 (413.25)	-1195.69 (1124.9)	732.39 (652.39)	1586.93 (590.43)	39.20 (805.45)	555.40 (777.46)	2640.13 (1226.4)	2852.39 (1143.1)
Including Lowest Probability Bin												
20 Bins	-319.93 (482.99)	-26.65 (872.13)	271.95 (701.25)	1486.15 (767.14)	340.51 (817.92)	-214.48 (1000.8)	-1387.62 (458.99)	292.88 (666.71)	-17.52 (589.98)	1630.89 (716.22)	-1442.96 (641.02)	-614.39 (1008.24)
10 Bins	-1215.47 (560.22)	-91.68 (833.26)	-573.74 (624.82)	863.00 (831.63)	-1371.57 (713.60)	376.71 (911.09)	-1648.51 (430.82)	-1150.33 (1153.4)	-418.66 (635.01)	896.64 (719.66)	-1727.59 (690.56)	-506.14 (981.67)
5 Bins	-1568.71 (556.55)	-1359.71 (845.89)	-650.74 (678.53)	744.53 (698.07)	-1843.34 (669.55)	-829.00 (976.36)	-2003.59 (423.29)	-1988.88 (1101.5)	-948.47 (621.08)	729.50 (636.01)	-2077.87 (725.08)	-914.14 (866.48)

Note: Bootstrapped standard errors in parentheses. 1000 Bootstrap replications.

Table 4: Average Absolute Bias Error

Matching Algoritm	SPECIFICATION ONE						SPECIFICATION TWO					
	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Nearest Neighbor 1 W/O Common Support	6531.57 (5746.5)	6515.78 (8011.1)	5043.09 (5375.8)	5231.05 (5190.2)	5956.38 (6444.2)	5834.49 (6005.6)	6495.40 (5719.1)	6065.43 (6199.8)	6073.86 (5848.0)	5027.43 (6518.4)	6200.70 (6322.6)	5657.73 (5847.6)
Nearest Neighbor 10 W/O Common Support	6540.42 (5840.8)	6515.78 (8011.1)	5141.34 (5426.0)	5231.05 (5190.2)	6100.74 (6488.0)	5834.49 (6005.6)	6466.16 (5740.2)	6065.43 (6199.8)	5953.66 (5963.3)	5027.43 (6518.4)	6298.87 (6309.0)	5657.73 (5847.6)
Nearest Neighbor 1 W. Common Support	6544.23 (5779.5)	6515.78 (8011.1)	4996.16 (5450.1)	5231.05 (5190.2)	5983.31 (6502.3)	5834.49 (6005.6)	6561.55 (5785.3)	6065.43 (6199.8)	5867.80 (5834.5)	5027.43 (6518.4)	6500.32 (6476.5)	5657.73 (5847.6)
Nearest Neighbor 10 W. Common Support	6597.81 (5820.5)	6515.78 (8011.1)	5109.14 (5587.4)	5231.05 (5190.2)	6077.74 (6564.0)	5834.49 (6005.6)	6551.98 (5769.2)	6065.43 (6199.8)	5989.54 (5978.8)	5027.43 (6518.4)	6077.74 (6564.0)	5657.73 (5847.6)
Kernel (Bandwidth 0.04)	5155.23 (3738.3)	5146.12 (4086.8)	4536.32 (3878.1)	4746.40 (4204.6)	5329.49 (4378.1)	5051.71 (4588.3)	5711.01 (4405.8)	7105.71 (5955.1)	4819.80 (4509.0)	6596.85 (5487.0)	5415.97 (4940.4)	6055.22 (5164.6)
Kernel (Bandwidth 0.01)	4987.96 (3839.0)	5216.04 (4335.5)	4307.99 (3962.5)	4788.36 (4339.8)	5034.23 (4671.2)	5300.18 (4720.5)	5664.64 (4572.7)	6553.06 (8027.5)	4841.42 (4720.5)	6012.27 (5287.2)	5484.68 (5198.4)	6209.80 (5337.7)
Local Linear (Bandwidth 0.04)	4675.54 (3345.0)	4845.76 (3749.6)	4134.49 (3524.6)	4298.21 (3636.9)	4670.90 (3857.7)	4871.93 (4038.7)	4754.06 (3459.7)	5051.76 (3891.8)	4254.59 (3546.9)	4427.15 (3908.4)	4831.42 (3790.4)	4900.41 (4223.0)
Local Linear (Bandwidth 0.01)	4705.60 (3379.1)	4995.25 (4105.4)	4145.24 (3498.1)	4353.05 (3895.5)	4743.96 (3883.0)	4813.54 (4356.2)	4680.02 (3464.6)	5059.29 (4100.5)	4150.35 (3613.5)	4460.52 (3917.0)	4743.96 (3883.0)	5046.30 (4247.1)
Caliper (0.01)	6505.52 (5758.3)	6600.18 (8138.6)	4978.08 (5365.1)	4945.54 (5098.1)	5928.61 (6492.5)	5529.58 (5879.6)	6791.76 (5844.4)	6340.82 (6428.5)	6128.93 (5760.1)	5168.24 (6612.4)	6357.82 (6499.1)	5769.70 (6022.4)
Caliper (0.001)	6860.53 (5981.8)	6175.00 (6132.1)	4942.18 (5057.3)	4434.86 (4838.2)	6841.91 (6621.5)	5976.01 (6788.6)	6619.27 (5791.0)	7073.58 (7189.0)	6448.08 (6387.0)	5683.33 (8489.1)	6380.06 (6218.1)	5092.21 (5162.5)
Caliper (0.0001)	6079.16 (5766.4)	6166.09 (6749.7)	5268.85 (5375.1)	4349.57 (5403.5)	5904.92 (6546.1)	6359.16 (8353.7)	6812.13 (6073.7)	8286.05 (10196.)	5788.78 (5185.5)	8622.76 (14083.)	7946.84 (6395.2)	4429.00 (5506.3)
Stratification 20 Bins Logit	2131.75 (1879.6)	2272.02 (1884.1)	1798.77 (1825.0)	3206.00 (2961.3)	2953.57 (2382.5)	3519.43 (3140.3)	2360.50 (1879.6)	2356.05 (2890.4)	1798.77 (1825.0)	3206.00 (2961.3)	2953.57 (2382.5)	3519.43 (3140.3)
Stratification 20 Bins BRQ	2054.54 (2192.5)	2210.44 (2312.4)	1741.31 (1833.8)	2864.02 (1997.1)	2929.14 (2032.3)	2234.41 (2063.6)	2451.74 (2633.4)	2585.78 (1968.3)	2372.92 (1611.9)	2596.23 (2174.3)	2511.21 (2099.7)	3133.87 (3444.3)
Stratification 10 Bins BRQ	1768.95 (2364.5)	2113.56 (2833.4)	2654.56 (2412.1)	2445.20 (2844.3)	2716.92 (2137.9)	2713.09 (2757.8)	2262.78 (2626.1)	1356.82 (2577.7)	2834.35 (2043.8)	2414.72 (2954.8)	2675.21 (2954.8)	3536.80 (3624.5)
Stratification 20 Bins BRQ No Zeros	1669.03 (1116.5)	2064.17 (2100.5)	1352.69 (2306.5)	2726.98 (1549.5)	2412.47 (1468.4)	2612.59 (1917.5)	1770.34 (1878.5)	2028.09 (1557.5)	1981.78 (1210.4)	2376.78 (1281.4)	1874.07 (1437.1)	2480.27 (1578.6)
Stratification 20 Bins Logit No Zeros	1971.97 (1218.8)	2119.36 (2352.0)	1421.64 (1164.9)	2991.63 (2430.7)	2458.90 (2052.7)	3137.16 (2219.0)	1971.97 (1218.8)	2119.36 (2352.0)	1421.64 (1164.9)	2991.63 (2430.7)	2458.90 (2052.7)	3137.16 (2219.0)

Note: Standard deviation in parentheses

Table 5: Number of Individuals Assigned to a Bin by Parametric and Semiparametric Estimates Using PSID and Early Random Assignment Experimental Sample via Specification 2

Logit Bins -> BRQ Bins ↓	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Total
[0-0.05%)	2178	58	17	5	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2266
[.05-0.1%)	42	42	23	21	19	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	153
[0.1-0.15%)	0	1	2	2	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	10
[0.15-0.2%)	0	0	0	5	6	4	3	1	1	0	0	0	0	0	0	0	0	0	0	0	20
[0.2-0.25%)	1	1	0	1	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	5
[0.25-0.3%)	0	0	0	0	2	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
[0.3-0.35%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[0.35-0.4%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
[0.4-0.45%)	0	1	2	0	2	1	4	0	4	5	2	2	6	0	0	0	0	0	0	0	27
[0.45-0.5%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[0.5-0.55%)	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2
[0.55-0.6%)	0	0	1	2	1	1	5	2	1	2	4	5	7	4	7	2	1	1	0	0	45
[0.6-0.65%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[0.65-0.7%)	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	2	0	1	0	0	6
[0.7-0.75%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[0.75-0.8%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[0.8-0.85%)	0	0	0	0	0	0	0	0	2	3	3	5	10	7	10	2	5	10	3	1	61
[0.85-0.9%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[0.9-0.95%)	0	0	0	0	0	0	0	0	0	0	0	1	2	5	5	3	2	5	12	32	62
[0.95-1.0%)	0	0	0	0	0	0	1	0	2	0	1	0	0	1	1	7	20	8	14	24	79
Total	2221	103	45	36	42	12	13	6	12	10	11	18	20	14	23	16	28	24	29	57	2740

Appendix Table 1: Summary Statistics

Sample	LaLonde Treated	LaLonde Controls	DW Treated	DW Controls	Early Assignment Treated	Early Assignment Control	CPS	PSID
Sample Size	297	425	185	260	108	142	15992	2490
Age	24.626 (6.686)	24.447 (6.590)	25.816 (7.155)	25.054 (7.058)	25.370 (6.251)	26.014 (7.108)	33.225 (11.045)	34.851 (10.441)
Years of Education	10.380 1.818)	10.188 (1.619)	10.346 (2.011)	10.088 (1.614)	10.491 (1.643)	10.275 (1.572)	12.028 (2.871)	12.117 (3.082)
Hispanic	0.094	0.113	0.059	0.108	0.074	0.113	0.072	0.032
Black	0.801	0.80	0.843	0.827	0.824	0.817	0.074	0.251
Married	0.168	0.158	0.189	0.154	0.204	0.190	0.712	0.866
Dropout	0.731	0.814	0.708	0.835	0.713	0.803	0.296	0.305
Zero Earnings in 1974	0.441	.461	0.708	0.75	0.50	0.542	0.120	0.086
Zero Earnings in 1975	0.374	0.419	0.60	0.685	0.324	0.472	0.109	0.100
Real Earnings in 1974	3571.00 (5773.13)	3672.49 (6521.53)	2095.57 (4886.62)	2107.03 (5687.91)	3589.64 (5970.74)	3857.94 (7254.27)	14016.8 (9569.80)	19428.8 (13406.9)
Real Earnings in 1975	3066.10 (4874.89)	3026.68 (5201.25)	1532.06 (3219.25)	1266.91 (3102.98)	2596.03 (3871.68)	2276.96 (3919.28)	13650.8 (9270.40)	19063.3 (13596.9)
Real Earnings in 1978	5976.35 (6923.80)	5090.05 (5718.09)	6349.14 (7867.40)	4554.80 (5483.84)	7357.41 (9027.18)	4608.92 (6031.96)	14846.66 (9647.39)	21553.9 (15555.4)

Note: Standard Deviation in Parentheses

Appendix Table 2: Treatment Effect Estimates with Parametric Propensity Scores using Stratification Matching

	SPECIFICATION ONE						SPECIFICATION TWO					
	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Experiment Impact	886.32	886.32	1794.34	1794.34	2748.49	2748.49	886.32	886.32	1794.34	1794.34	2748.49	2748.49
20 Bins Excluding 0-0.05	-277.54 (603.86)	-401.81 (788.77)	1327.11 (860.30)	1631.55 (1157.9)	1386.91 (1108.3)	2242.87 (1265.9)	-373.80 (556.89)	18.44 (706.37)	1361.09 (794.66)	1940.58 (911.70)	2437.87 (1063.0)	1380.82 (1303.5)
20 Bins no exclusion	-1067.65 (596.06)	-611.86 (759.50)	14.76 (816.72)	1151.39 (1094.7)	211.18 (959.90)	1441.67 (1351.0)	-1312.4 (544.38)	-301.82 (707.88)	171.88 (776.22)	1354.01 (947.75)	591.07 (1060.0)	465.81 (1237.7)
20 Bins DW exclusion	-819.79 (582.53)	-632.00 (800.09)	1096.57 (800.63)	1394.41 (1060.4)	1255.21 (1032.5)	1297.61 (1125.5)	-803.78 (489.33)	-268.06 (688.45)	902.33 (731.83)	1513.39 (920.17)	1634.18 (938.87)	694.29 (1245.2)

Note: Bootstrapped standard errors in parentheses. 1000 Bootstrap replications. DW exclusion drops all individuals in the treatment group with estimated propensity scores above the maximum propensity score in the control group and drops all control individuals whose estimated propensity score is less than the minimum propensity score of the treatment group.

Appendix Table 3: Evaluation Bias Estimates with Parametric Propensity Scores using Stratification Matching

	SPECIFICATION ONE						**SPECIFICATION TWO**					
	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Experiment Impact	886.32	886.32	1794.34	1794.34	2748.49	2748.49	886.32	886.32	1794.34	1794.34	2748.49	2748.49
20 Bins	-1318.41	-1011.45	62.14	726.59	-1216.63	-302.93	-1303.93	-685.21	-399.41	679.02	-1275.2	-1862.5
	(530.93)	(696.50)	(673.98)	(779.96)	(769.63)	(911.30)	(427.43)	(594.36)	(547.29)	(760.47)	(634.42)	(1108.8)
10 Bins	-1148.28	-1070.33	-124.51	84.97	-1135.18	-825.08	-1248.07	-985.08	-229.11	248.66	-1129.4	-1156.1
	(539.09)	(783.31)	(675.83)	(926.07)	(849.69)	(952.01)	(437.63)	(634.63)	(566.49)	(1056.9)	(641.74)	(876.66)
Including Lowest Probability Bin												
20 Bins	-1749.99	-1269.00	-411.11	455.02	-1851.27	-765.03	-1890.1	-1091.06	-912.84	270.58	-1939.3	-2286.0
	(517.42)	(695.99)	(644.94)	(770.67)	(737.83)	(947.28)	(399.59)	(600.09)	(528.85)	(805.11)	(630.75)	(1097.8)
10 Bins	-2004.45	-1593.87	-977.55	-505.34	-2282.38	-1471.57	-2051.4	-1421.89	-1275.7	-375.59	-2526.2	-2186.9
	(491.63)	(741.56)	(643.09)	(911.10)	(762.06)	(863.60)	(398.61)	(650.42)	(511.59)	(1057.2)	(627.09)	(934.91)

Note: Bootstrapped standard errors in parentheses. 1000 Bootstrap replications.

Table 4: Average Absolute Bias Error using Klein Spady (1993) Propensity Scores.

Matching Algoritm	SPECIFICATION ONE						SPECIFICATION TWO					
	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Nearest Neighbor 1 W/O Common Support	6352.60 (6193.4)	5877.75 (5394.6)	5428.18 (5478.5)	7102.12 (5732.3)	6299.20 (5745.2)	4992.30 (4360.3)	6982.26 (8108.0)	6982.26 (8108.0)	6073.86 (5848.0)	4996.93 (5405.9)	6646.79 (5585.2)	6165.67 (6501.4)
Nearest Neighbor 10 W/O Common Support	5022.36 (3667.6)	4915.66 (4005.2)	4395.05 (3769.7)	4372.99 (4070.6)	5034.94 (4037.6)	4728.47 (5530.1)	6242.61 (6630.9)	6242.61 (6630.9)	5953.66 (5963.3)	4371.97 (4313.1)	4953.87 (4190.9)	5179.72 (4200.4)
Nearest Neighbor 1 W. Common Support	6352.60 (6193.4)	5877.75 (5394.6)	5428.18 (5478.5)	7102.12 (5732.3)	6299.20 (5745.2)	4992.30 (4360.3)	6982.26 (8108.0)	6281.00 (6519.6)	5783.91 (5562.5)	5329.23 (5852.4)	6641.07 (5604.8)	6165.67 (6501.4)
Nearest Neighbor 10 W. Common Support	5022.36 (3667.6)	4915.66 (4005.2)	4395.05 (3769.7)	4372.99 (4070.6)	5034.94 (4037.6)	4728.47 (5530.1)	5389.74 (3573.7)	5169.50 (4489.5)	4446.48 (3550.9)	4533.35 (4581.9)	4968.01 (4192.8)	5139.80 (4767.8)
Kernel (Bandwidth 0.04)	5042.46 (3757.7)	5053.83 (3821.2)	4647.82 (3784.4)	4372.99 (4070.6)	5592.54 (4275.5)	5858.21 (5163.0)	5288.00 (3749.6)	5178.34 (4240.0)	4791.99 (3773.9)	4835.59 (4899.4)	5366.73 (4400.9)	5659.19 (4835.2)
Kernel (Bandwidth 0.01)	4864.39 (3807.3)	5066.20 (3940.8)	4646.35 (4162.6)	4907.52 (4606.9)	4941.11 (4265.6)	4926.83 (4817.3)	5165.07 (3568.3)	5313.46 (4579.4)	4505.29 (3787.2)	4418.51 (4599.3)	5010.83 (4338.8)	5139.80 (4767.8)
Local Linear (Bandwidth 0.04)	4643.84 (3361.2)	4779.31 (3443.1)	4235.01 (3484.8)	4649.24 (4030.1)	4847.18 (3730.9)	4891.54 (3996.1)	4963.27 (3390.6)	5087.11 (4164.5)	4376.77 (3385.1)	4385.57 (4278.7)	4816.22 (3781.0)	4911.10 (4017.3)
Local Linear (Bandwidth 0.01)	4619.84 (3401.0)	4852.26 (3615.3)	4294.89 (3468.0)	4772.71 (4037.1)	4768.45 (3804.2)	5026.13 (4125.3)	4950.94 (3414.4)	6220.03 (6567.4)	4400.88 (3298.2)	4443.02 (4215.7)	4700.37 (3845.8)	5012.75 (4074.9)
Caliper (0.01)	6339.56 (6261.6)	6689.82 (6748.7)	5446.49 (5212.1)	6330.32 (5532.0)	6277.75 (5884.8)	5193.95 (5978.1)	7073.26 (6439.0)	6902.02 (8119.1)	6231.49 (5726.5)	5313.34 (5674.4)	6887.10 (5729.9)	6278.50 (6495.2)
Caliper (0.001)	7036.05 (6396.3)	7554.11 (6420.4)	6094.01 (5941.0)	6304.42 (5244.3)	6740.51 (5551.1)	5599.46 (5952.7)	8043.27 (6573.9)	9112.63 (7070.8)	7641.16 (6630.4)	6489.76 (6698.5)	7614.11 (5101.5)	8719.40 (7685.5)
Caliper (0.0001)	8504.31 (7520.5)	8995.17 (6439.2)	6926.92 (5921.7)	8826.38 (6777.3)	6456.99 (5892.1)	7676.29 (7393.1)	9112.63 (7070.8)	10875.3 (12640.)	6495.55 (5913.0)	8338.97 (8341.5)	8000.81 (5487.2)	11143.0 (11077.)
Stratification 20 Bins KS	2520.56 (2569.6)	2762.02 (2109.9)	2289.02 (2310.8)	1786.94 (2714.3)	3269.07 (2592.6)	3511.51 (5567.7)	2849.80 (2053.8)	2718.37 (2728.4)	2310.46 (2244.7)	2078.51 (2811.4)	3022.43 (3408.2)	3475.25 (3674.8)
Stratification 20 Bins BRQ	2054.54 (2192.5)	2210.44 (2312.4)	1741.31 (1833.8)	2864.02 (1997.1)	2929.14 (2032.3)	2234.41 (2063.6)	2451.74 (2633.4)	2585.78 (1968.3)	2372.92 (1611.9)	2596.23 (2174.3)	2511.21 (2099.7)	3133.87 (3444.3)
Stratification 10 Bins BRQ	1768.95 (2364.5)	2113.56 (2833.4)	2654.56 (2412.1)	2445.20 (2844.3)	2716.92 (2137.9)	2713.09 (2757.8)	2262.78 (2626.1)	1356.82 (2577.7)	2834.35 (2043.8)	2414.72 (2954.8)	2675.21 (2954.8)	3536.80 (3624.5)
Stratification 20 Bins BRQ No Zeros	1669.03 (1116.5)	2064.17 (2100.5)	1352.69 (2306.5)	2726.98 (1549.5)	2412.47 (1468.4)	2612.59 (1917.5)	1770.34 (1878.5)	2028.09 (1557.5)	1981.78 (1210.4)	2376.78 (1281.4)	1874.07 (1437.1)	2480.27 (1578.6)
Stratification 20 Bins KS No Zeros	1743.21 (1541.6)	2586.14 (1623.0)	1910.55 (1575.8)	1606.58 (2155.5)	2344.50 (1665.9)	2842.86 (4660.3)	2309.40 (1341.8)	2464.88 (2136.3)	1784.02 (1543.9)	1761.06 (1685.3)	2358.11 (3214.4)	2941.83 (2420.8)

Note: Standard deviation in parentheses