Warsaw School of Economics–SGH
Institute of Econometrics
Department of Applied Econometrics

ISSN 2084-4573

Department of Applied Econometrics Working Papers
Warsaw School of Economics–SGH
ul. Madalinskiego 6/8
02-513 Warszawa, Poland

Working Paper No. 2-12

The Oaxaca-Blinder unexplained component as a treatment effects estimator

Tymon Słoczyński
Warsaw School of Economics–SGH, Poland

This paper is available at the Warsaw School of Economics
Department of Applied Econometrics website at: http://www.sgh.waw.pl/instytuty/zes/wp/

The Oaxaca–Blinder Unexplained Component as a Treatment Effects Estimator*

TYMON SŁOCZYŃSKI

Department of Economics I, Warsaw School of Economics, ul. Madalińskiego 6/8 p. 228, 02-513 Warszawa, Poland

(e-mail: tymon.sloczynski@gmail.com)

Abstract: In this paper I use the National Supported Work (NSW) data to examine the validity of the Oaxaca–Blinder unexplained component as an estimator of the population average treatment effect on the treated (PATT). Precisely, I utilize dataset and variable selections used in previous studies of the NSW data to compare the performance of the Oaxaca–Blinder unexplained component with methods based on the propensity score (Dehejia and Wahba, 1999) and bias-corrected matching estimators (Abadie and Imbens, 2011). I show that in both cases the Oaxaca–Blinder unexplained component performs superior compared to the previously analyzed estimators provided that common support is imposed.

JEL classification numbers: C21, J24

Keywords: Decomposition methods; Manpower training; Treatment effects.

* I would like to acknowledge financial support for this research, provided through a grant from the Warsaw School of Economics (03/BMN/25/11) and a 'Weź stypendium – dla rozwoju' scholarship, funded by the European Social Fund and administered by the Warsaw School of Economics.

I. Introduction

Recent contributions of Barsky *et al.* (2002), Melly (2006), and Fortin *et al.* (2011) have noted that the Oaxaca–Blinder decomposition, a popular method used in empirical labour economics to study intergroup wage differentials,[1] provides a consistent estimator of the population average treatment effect on the treated (PATT). Precisely, applied researchers in labour economics have often used the Oaxaca–Blinder decomposition to estimate two components of an intergroup wage differential: a component attributable to differences in group composition (the explained component) and a component attributable to net effects of group membership (the unexplained component). It is the unexplained component in the most basic version of the Oaxaca–Blinder decomposition which constitutes a consistent estimator of the PATT. Importantly, Kline (2011) has recently shown that this method is equivalent to a propensity score reweighting estimator based on a linear model for the treatment odds, and satisfies therefore a valuable 'double robustness' property (see Robins *et al.*, 1994).

In this paper I examine the validity of the Oaxaca–Blinder unexplained component as an estimator of the PATT using the National Supported Work (NSW) data, analyzed originally by LaLonde (1986) and subsequently by Heckman and Hotz (1989), Dehejia and Wahba (1999, 2002), Smith and Todd (2001, 2005), Becker and Ichino (2002), Angrist and Pischke (2009), Porro and Iacus (2009), Abadie and Imbens (2011), Kline (2011), and others. Following Dehejia and Wahba (1999), these studies have typically shown good performance of various matching estimators and methods based on the propensity score (see Angrist and Pischke, 2009, for a different view). In a recent paper, Kline (2011) has seminally applied the Oaxaca–Blinder unexplained component to these data, though he has only used a single

[1] See Blinder (1973) and Oaxaca (1973) for seminal contributions and Fortin *et al.* (2011) for a comprehensive survey.

nonexperimental control dataset and a single selection of control variables, and he has compared his result to a relatively small number of alternative estimates.

In the present paper I provide much broader a picture of the performance of the Oaxaca–Blinder unexplained component as an estimator of the PATT. I utilize dataset and variable selections used by Dehejia and Wahba (1999) and Abadie and Imbens (2011) in their studies of the NSW data to compare the performance of the Oaxaca–Blinder unexplained component with various methods based on the propensity score (Dehejia and Wahba, 1999) and bias-corrected matching estimators (Abadie and Imbens, 2011). I also examine whether nonexperimental estimates of the PATT based on the Oaxaca–Blinder unexplained component can be brought closer to the experimental benchmark by improving overlap. Precisely, I test this alternative estimator as well as the benchmark linear regression using full nonexperimental control datasets, datasets obtained after discarding all the nontreated individuals whose estimated propensity score is less than the minimum or greater than the maximum estimated propensity score for the treated individuals (i.e. imposing common support), as well as datasets trimmed according to the rule of thumb proposed by Crump *et al.* (2009). Moreover, I also compare these results with estimates based on stratification (on the propensity score) and a simple combination of stratification and linear regression, i.e. two standard methods based on the propensity score. Although this latter approach has recently been described by Imbens and Wooldridge (2009, p. 41) as 'one of the more attractive estimators [of average treatment effects] in practice', the present paper suggests that it may still perform poorer than the Oaxaca–Blinder unexplained component. The Oaxaca–Blinder decomposition is shown to perform superior compared to various methods based on the propensity score (Dehejia and Wahba, 1999) and bias-corrected matching estimators (Abadie and Imbens, 2011) provided that common support is imposed.

The remainder of this paper is organized as follows. In Section II, I review the treatment effects framework. In Section III, I use various nonexperimental estimators to reanalyze the National Supported Work (NSW) data. Finally, I conclude and review my findings in Section IV.

II. The Treatment Effects Framework

The exposition here is standard and borrows notation from Imbens and Wooldridge (2009). Let me therefore consider a population of N individuals, indexed by $i = 1, \dots, N$, who are divided into two disjoint subsets (groups), 0 and 1. Individuals in group 1 are exposed to regime that is called *treatment*, while individuals in group 0 are exposed to regime that is called *control*. There are N_1 individuals in group 1 and N_0 individuals in group 0 ($N_0 + N_1 = N$). To indicate group membership, the binary variable W_i is used, and $W_i = 0$ ($W_i = 1$) if individual i belongs to group 0 (group 1). A column vector of covariates, X_i, is also observed for each individual i.

Crucial for this framework, however, is the notion of potential outcomes. It is assumed there exist two potential outcomes for each individual i, the treated outcome $Y_i(1)$ and the nontreated outcome $Y_i(0)$, and although both of them are potentially observable, exactly one of them is eventually realized. It is the group membership of each individual i which causes one of the potential outcomes to become observable ($Y_i(0)$ if individual i belongs to group 0 and $Y_i(1)$ if individual i belongs to group 1) and the other potential outcome to become counterfactual. The realized outcome is denoted by Y_i. Consequently, $Y_i = Y_i(W_i) = Y_i(0)(1 - W_i) + Y_i(1)W_i$.

The main interest in the treatment effects framework lies in determining causal effects of treatment. Such an effect, for each individual i, is defined as the difference between the

treated outcome of this individual and her nontreated outcome, $Y_i(1) - Y_i(0)$. In general, such individual treatment effects are averaged over appropriate (sub)populations of interest. The average over the subpopulation of treated individuals is called the population average treatment effect on the treated (PATT):

$$\tau_{PATT} = \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1] \tag{1}$$

Alternatively, one may wish to average individual treatment effects over the whole population (including both treated and nontreated individuals) to obtain the population average treatment effect (PATE):

$$\tau_{PATE} = \mathbb{E}[Y_i(1) - Y_i(0)] \tag{2}$$

Let me now use the example of PATT to comment further on how to determine different average treatment effects. Since we wish to average individual treatment effects over the subpopulation of treated individuals, we observe the treated outcome for each individual of interest. At the same time, we do not observe the nontreated outcome for any of the individuals of interest, so we have to estimate these missing outcomes. A naïve solution is to use the average realized outcome of nontreated individuals as a prediction of what treated individuals would have received, on average, had they not received treatment (see, e.g., Cobb-Clark and Crossley, 2003). However, such a naïve estimator is biased if selection to treatment is present:

$$\begin{aligned}
&\mathbb{E}[Y_i|W_i = 1] - \mathbb{E}[Y_i|W_i = 0] = \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] \\
&= \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] \\
&+ \{\mathbb{E}[Y_i(0)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 1]\} \\
&= \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1] + \{\mathbb{E}[Y_i(0)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0]\} \\
&= \tau_{PATT} + \{\mathbb{E}[Y_i(0)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0]\}
\end{aligned} \tag{3}$$

Precisely, the difference in realized outcomes of treated and nontreated individuals is equal to the sum of PATT and selection bias, i.e. the extent to which the nontreated outcomes of treated and nontreated individuals are, on average, different. Since there is no reason whatsoever to expect selection bias not to appear in observational studies, such a naïve estimator can certainly be regarded as useless.

What follows, identification and estimation of average treatment effects must proceed differently. There are generally two main strands in the treatment effects literature, often referred to as selection on observables and selection on unobservables (a good survey of both has recently been provided by Imbens and Wooldridge, 2009), and this division is based on identifying assumptions which differ between these strands. This paper – and all the analyses of the NSW data in general – is only concerned with selection on observables, a strand whose main assumptions are typically referred to as unconfoundedness and overlap. Under unconfoundedness, it is assumed there do not exist such unobserved individual characteristics which would be associated both with the potential outcomes and the treatment status. Consequently:

$$W_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) | X_i \tag{4}$$

Under overlap, on the other hand, it is assumed there do not exist such (sets of) values of the control variables which would perfectly predict either of the treatment statuses:

$$0 < \mathrm{pr}(W_i = 1 | X_i = x) < 1, \text{ for all } x \tag{5}$$

Under the assumptions of unconfoundedness and overlap both the PATT and the PATE are identified (see Imbens and Wooldridge, 2009, pp. 26–27). Many estimators have been used to estimate both average treatment effects under these assumptions, including regression methods, methods based on the propensity score, matching on covariates, and various combinations of these estimators. A very good survey is provided, again, by Imbens and Wooldridge (2009), while several recent contributions (Barsky *et al.*, 2002; Melly, 2006;

Fortin *et al.*, 2011) have also noted that the Oaxaca–Blinder decomposition, a popular tool used by labour economists to study intergroup wage differentials, provides a consistent estimator of the population average treatment effect on the treated (PATT) as well. Precisely, let the model for outcomes be linear and let the regression coefficients be flexible, i.e. possibly different for the treated individuals and the nontreated individuals:

$$Y_i = X_i\beta_1 + \upsilon_{1i} \text{ if } W_i = 1;\ Y_i = X_i\beta_0 + \upsilon_{0i} \text{ if } W_i = 0 \tag{6}$$

where $\mathbb{E}[\upsilon_{1i}|X_i] = \mathbb{E}[\upsilon_{0i}|X_i] = 0$. What follows:

$$\begin{aligned}
\mathbb{E}[Y_i|W_i = 1] - \mathbb{E}[Y_i|W_i = 0] &= \mathbb{E}[\mathbb{E}(Y_i|X_i, W_i = 1)|W_i = 1] - \mathbb{E}[\mathbb{E}(Y_i|X_i, W_i = 0)|W_i = 0] \\
&= (\mathbb{E}[X_i|W_i = 1]\beta_1 + \mathbb{E}[\upsilon_{1i}|W_i = 1]) - (\mathbb{E}[X_i|W_i = 0]\beta_0 + \mathbb{E}[\upsilon_{0i}|W_i = 0]) \\
&= \mathbb{E}[X_i|W_i = 1]\beta_1 - \mathbb{E}[X_i|W_i = 0]\beta_0 \\
&= \mathbb{E}[X_i|W_i = 1]\beta_1 - \mathbb{E}[X_i|W_i = 0]\beta_0 + \mathbb{E}[X_i|W_i = 1]\beta_0 - \mathbb{E}[X_i|W_i = 1]\beta_0 \\
&= \mathbb{E}[X_i|W_i = 1](\beta_1 - \beta_0) + (\mathbb{E}[X_i|W_i = 1] - \mathbb{E}[X_i|W_i = 0])\beta_0 \\
&= \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1] + \{\mathbb{E}[Y_i(0)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0]\} \\
&= \tau_{PATT} + \{\mathbb{E}[Y_i(0)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0]\}
\end{aligned} \tag{7}$$

In other words, any intergroup differential in outcomes can be decomposed into the net effect of treatment group membership (the PATT) and a component attributable to differences in group composition (selection bias). These two components have typically been referred to as the unexplained component and the explained component, respectively.

The result in (7) follows from Barsky *et al.* (2002), Melly (2006), and Fortin *et al.* (2011). Recently, Kline (2011) has shown that this estimator is not only consistent, but also 'doubly robust' (see Robins *et al.*, 1994), since it is equivalent to a propensity score

reweighting estimator based on a linear model for the treatment odds. The next section examines the validity of this estimator using the National Supported Work (NSW) data.[2]

III. An Application of the Oaxaca–Blinder Unexplained Component to the NSW Data

The National Supported Work (NSW) data

The National Supported Work (NSW) Demonstration was a U.S. employment program implemented in the mid-1970s to provide work experience to disadvantaged workers. Unlike many similar programs, the NSW assigned treatment (participation in the program) to individuals on random, so the pool of potential participants was randomly divided into an experimental group and a control group, thus allowing for a straightforward, unbiased estimation of average treatment effects (see LaLonde, 1986, and Smith and Todd, 2005, for detailed descriptions of the NSW).

In a seminal paper, LaLonde (1986) tested the validity of various nonexperimental estimators in a novel way. He discarded the original control group from the NSW data, and created six alternative nonexperimental control datasets using standard surveys of the U.S. working population, the Panel Study of Income Dynamics (PSID) and the Current Population Survey (CPS). His approach was based on an ingenious idea: a valid nonexperimental estimator should be able to closely replicate the experimental estimate of the average treatment effect, while using a nonexperimental control group instead of the original one. LaLonde (1986) concluded that the nonexperimental estimators he used were typically unable to replicate the experimental results, while it would also have been practically impossible to

[2] A useful Stata program to perform the Oaxaca–Blinder decomposition has been provided by Jann (2008). It is used throughout the present paper.

distinguish between differently performing estimators without his knowledge of the experimental benchmark.

In the present paper I use a specific version of the NSW data which was first used by Dehejia and Wahba (1999). Table 1 presents descriptive statistics for all the subsamples used in the analysis. The first two columns present means and standard deviations of the outcome variable (earnings in 1978) and all the standard control variables for the experimental group and the original control group, respectively. The columns 3–5 present the statistics for the three nonexperimental control groups (PSID-1, PSID-2, and PSID-3) created from the Panel Study of Income Dynamics (PSID) data. The columns 6–8 present the statistics for the three nonexperimental control groups (CPS-1, CPS-2, and CPS-3) created from the Current Population Survey (CPS) data. Clearly visible are substantial disparities in means of control and outcome variables between the NSW experimental and control group and the nonexperimental control groups. It is precisely these disparities that hinder nonexperimental replication of the experimental estimate of the average treatment effect. This estimate is equal to 6349.14 – 4554.80 ≈ 1794. What is also apparent, while the experimental estimate of the average treatment effect can be considered substantial (please note that this effect is equal to ca. 40% of mean earnings of the nontreated individuals in 1978), randomization was visibly successful in equalizing mean pre-treatment characteristics in the experimental group and the original control group. In other words, it is only post-treatment outcomes – with the notable exception of No degree, i.e. a dummy variable that captures high school dropouts – which substantially differ between these groups.

TABLE 1

Sample means of outcome and control variables for the NSW and control datasets

	NSW							
	Treated	*Control*	*PSID-1*	*PSID-2*	*PSID-3*	*CPS-1*	*CPS-2*	*CPS-3*
Number of observations	185	260	2,490	253	128	15,992	2,369	429
Outcome variable								
Earnings '78	6,349	4,555	21,554	9,996	5,279	14,847	10,171	6,984
	(7,867)	(5,484)	(15,555)	(11,184)	(7,763)	(9,647)	(8,852)	(7,294)
Control variables								
Age	25.82	25.05	34.85	36.09	38.26	33.23	28.25	28.03
	(7.16)	(7.06)	(10.44)	(12.08)	(12.89)	(11.05)	(11.70)	(10.79)
Education	10.35	10.09	12.12	10.77	10.30	12.03	11.24	10.24
	(2.01)	(1.61)	(3.08)	(3.18)	(3.18)	(2.87)	(2.58)	(2.86)
No degree	0.71	0.83	0.31	0.49	0.51	0.30	0.45	0.60
	(0.46)	(0.37)	(0.46)	(0.50)	(0.50)	(0.46)	(0.50)	(0.49)
Black	0.84	0.83	0.25	0.39	0.45	0.07	0.11	0.20
	(0.36)	(0.38)	(0.43)	(0.49)	(0.50)	(0.26)	(0.32)	(0.40)
Hispanic	0.06	0.11	0.03	0.07	0.12	0.07	0.08	0.14
	(0.24)	(0.31)	(0.18)	(0.25)	(0.32)	(0.26)	(0.28)	(0.35)
Married	0.19	0.15	0.87	0.74	0.70	0.71	0.46	0.51
	(0.39)	(0.36)	(0.34)	(0.44)	(0.46)	(0.45)	(0.50)	(0.50)
'Earnings '74'	2,096	2,107	19,429	11,027	5,567	14,017	8,728	5,619
	(4,887)	(5,688)	(13,407)	(10,815)	(7,255)	(9,570)	(8,968)	(6,789)
'Nonemployed '74'	0.71	0.75	0.09	0.23	0.41	0.12	0.21	0.26
	(0.46)	(0.43)	(0.28)	(0.42)	(0.49)	(0.32)	(0.41)	(0.44)
Earnings '75	1,532	1,267	19,063	7,569	2,611	13,651	7,397	2,466
	(3,219)	(3,103)	(13,597)	(9,042)	(5,572)	(9,270)	(8,112)	(3,292)
Nonemployed '75	0.60	0.68	0.10	0.34	0.61	0.11	0.18	0.31
	(0.49)	(0.47)	(0.30)	(0.47)	(0.49)	(0.31)	(0.38)	(0.46)

Notes: Standard deviations are in parentheses. Earnings are in 1978 dollars. Education = number of years of schooling; No degree = 1 if no high school degree, 0 otherwise.

Following LaLonde (1986), the NSW data were subsequently analyzed by many researchers, including Heckman and Hotz (1989), Dehejia and Wahba (1999, 2002), Smith and Todd (2001, 2005), Becker and Ichino (2002), Angrist and Pischke (2009), Porro and Iacus (2009), Abadie and Imbens (2011), and Kline (2011). In a famous contribution, Dehejia and Wahba (1999) closely replicated the experimental estimate of the average treatment effect using various methods based on the propensity score (see Smith and Todd, 2005, for an important critique of this paper). Recently, Abadie and Imbens (2011) have tested a new class of bias-corrected matching estimators using these data. In two subsequent subsections, I provide a reanalysis of these two studies.

A reanalysis of Dehejia and Wahba (1999)

In an important paper, Dehejia and Wahba (1999) employed various methods based on the propensity score and all the six nonexperimental control datasets (PSID-1, PSID-2, PSID-3, CPS-1, CPS-2, and CPS-3), and replicated the experimental estimate of the average treatment effect relatively closely. The authors used regression on control variables, regression on a quadratic in the (estimated) propensity score, stratification on the propensity score, a combination of stratification on the propensity score and regression, nearest neighbour matching on the propensity score, and a combination of nearest neighbour matching on the propensity score and regression (weighted least squares regression with weights on the nontreated individuals equal to the number of times they were matched to a treated individual). These estimators were designed by Dehejia and Wahba (1999) to estimate the population average treatment effect on the treated (PATT), and it has become standard in the subsequent literature to focus on this estimand, although we could as well focus, for example, on the PATE.

In their analysis, Dehejia and Wahba (1999) employed three different selections of control variables, each of them matched to one, two or three nonexperimental control datasets. For PSID-1, they used Age, Age squared, Education, Education squared, Married, No degree, Black, Hispanic, 'Earnings '74', 'Earnings '74' squared, Earnings '75, Earnings '75 squared, and the product of Black and 'Nonemployed '74'.[3] For PSID-2 and PSID-3, they also decided to include 'Nonemployed '74' and Nonemployed '75, but did not use the product of Black and 'Nonemployed '74'. For CPS-1, CPS-2, and CPS-3 – as compared with the latter variable selection – they also used Age cubed and the product of Education and 'Earnings '74', but on

[3] The usage of quotation marks stems from the fact that 'Earnings '74' and 'Nonemployed '74' denote real earnings and nonemployment in months 13–24 prior to random assignment. While several studies refer to these variables as if these months overlapped with calendar year 1974, they actually overlap with 1974 or 1975, dependent on being randomized earlier or later in the experiment. See Smith and Todd (2005) for a discussion.

the other hand decided to include neither ‘Earnings ’74’ squared, nor Earnings ’75 squared. To make the subsequent comparison of the Oaxaca–Blinder estimates of the PATT with the results reported by Dehejia and Wahba (1999) fully adequate, I employ exactly the same sets of control variables throughout this subsection.

Table 2 presents root mean square deviations (RMSDs) from the experimental benchmark which were obtained using various nonexperimental estimators, including methods based on the propensity score employed by Dehejia and Wahba (1999). I have been unable to replicate their results, so I report RMSDs computed on the basis of estimates in Table 3 in Dehejia and Wahba (1999, p. 1059). Clearly visible is a substantial gain in performance in favour of estimators based on stratification on the propensity score. Nearest neighbour matching (with and without regression adjustment) performed slightly poorer, and regression on a quadratic in the propensity score performed three to four times poorer than stratification-based estimators. It also seems that stratification on the propensity score may be visibly enhanced by using within-strata regression adjustment to eliminate remaining within-strata differences in control variables. Such an estimator, combining stratification on the propensity score and regression, performed best among all the methods used by Dehejia and Wahba (1999), and its RMSD was equal to 289 (1978 dollars). It is precisely this estimator which has been recommended by Imbens and Wooldridge (2009, p. 41) in their recent survey of the econometrics of program evaluation. Table 2 also presents RMSDs for new nonexperimental estimates of the PATT contributed in this paper.

TABLE 2

A comparison of Dehejia and Wahba (1999) with various estimates of the PATT based on the Oaxaca–Blinder unexplained component or linear regression

	Improving overlap?	*RMSD*
Dehejia and Wahba (1999):		
Mean difference	No, full sample	8,779
Regression on a quadratic in the score	Common support	1,218
Stratification (on the score)	Common support	378
Stratification (on the score) and regression	Common support	289
Nearest neighbour matching	Common support	538
Nearest neighbour matching and regression	Common support	521
New estimates:		
Regression	No, full sample	1,007
Oaxaca–Blinder	No, full sample	414
Stratification (on the score)	No, full sample	3,299
Stratification (on the score) and regression	No, full sample	1,188
Regression	Common support	1,230
Oaxaca–Blinder	Common support	281
Stratification (on the score)	Common support	1,661
Stratification (on the score) and regression	Common support	1,136
Regression	Crump *et al.* (2009)	1,983
Oaxaca–Blinder	Crump *et al.* (2009)	1,640
Stratification (on the score)	Crump *et al.* (2009)	1,959
Stratification (on the score) and regression	Crump *et al.* (2009)	2,410

Notes: Root mean square deviations (RMSDs) are calculated for each estimator using nonexperimental estimates for each control dataset as one observation, and the experimental estimate as the benchmark value. They are in 1978 dollars. Propensity scores are estimated using a logit model. Each stratification-based estimator uses two strata of equal width. Common support refers to discarding all the nontreated individuals whose estimated propensity score is less than the minimum or greater than the maximum estimated propensity score for the treated individuals. Crump *et al.* (2009) refers to discarding all the individuals whose estimated propensity score is less than 0.1 or greater than 0.9. Detailed estimation results for each nonexperimental control dataset are presented in Appendix Table 1.

Among new results presented in Table 2, unambiguous is great performance of the Oaxaca–Blinder decomposition. Independent of the method of improving overlap, including no improvement at all, the Oaxaca–Blinder unexplained component guarantees closer matches than linear regression and both stratification-based estimators. While Angrist and Pischke (2009, pp. 69–70) have recently recommended using linear regression to study treatment effects and suggested that 'the differences between regression and matching estimates are unlikely to be of major empirical importance', it seems that it can actually be reasonable to expect major differences between simple linear regression and the Oaxaca–Blinder

decomposition. Clearly, the Oaxaca–Blinder decomposition *is* a version of linear regression which relaxes the restrictive assumption of treatment effects homogeneity, and such a result may indeed suggest that it is this assumption (and not linearity of the conditional expectation function) which can make linear regression problematic.

Importantly, the full-sample estimates of the PATT based on the Oaxaca–Blinder decomposition provide a closer replication of the experimental benchmark than both nearest neighbour matching estimators considered by Dehejia and Wahba (1999). While the RMSD for the Oaxaca–Blinder unexplained component is equal to 414, both nearest neighbour matching estimators performed slightly poorer, and their RMSDs exceeded 500. It seems, however, that the precision of the Oaxaca–Blinder unexplained component can be further enhanced when a first-stage estimation of the propensity score is used to screen the sample at hand and improve overlap between the treated and nontreated subsamples. Discarding all the nontreated individuals whose estimated propensity score is less than the minimum or greater than the maximum estimated propensity score for the treated individuals (i.e. imposing common support) clearly improved the RMSD of the Oaxaca–Blinder unexplained component. Its RMSD reached as little as 281, and provided therefore a slightly closer match with the experimental benchmark than any of the famous results in Dehejia and Wahba (1999), including the combination of stratification on the propensity score and regression.

While such an approach to improving overlap seems to have resulted in closer replication of the experimental benchmark, a strikingly different conclusion emerges for the rule of thumb proposed by Crump *et al.* (2009). These authors have recently developed a systematic approach to select subsamples which diminish sensitivity to the choice of specification, and concluded that the optimal rule can typically be approximated by a simple rule of thumb to discard all the individuals whose estimated propensity score is less than 0.1 or greater than 0.9. Such a rule has recently been applied to the NSW data by Angrist and

Pischke (2009), and the rule often improved the precision of linear regression and rarely worsened it. On the other hand, the present paper provides much richer a set of estimates, and can be safely used to conclude that the rule of thumb proposed by Crump *et al.* (2009) can actually bias the resulting estimates to a significant extent. Clearly, bias reduction has *not* been the main goal in Crump *et al.* (2009). It is, however, important to note that less sensitivity to the choice of specification can have a profound (negative) impact on bias reduction.

A reanalysis of Abadie and Imbens (2011)

In a recent paper, Abadie and Imbens (2011) have employed a new class of bias-adjusted matching estimators as well as simple nearest neighbour matching estimators (both on covariates and the estimated propensity score) and the PSID-1 dataset, and replicated the experimental estimate of the average treatment effect relatively closely, with the exception of simple matching estimators with a large number of matches. They have also compared these estimates with several regression-based approaches, and found that propensity score reweighting estimators performed remarkably well (please note that the Oaxaca–Blinder decomposition *is* actually a propensity score reweighting estimator; see Kline, 2011). In their analysis, Abadie and Imbens (2011) have consistently employed a single set of control variables, including Age, Education, Married, Black, Hispanic, 'Earnings '74', Earnings '75, 'Nonemployed '74' and Nonemployed '75. Again, I employ the same selection of control variables throughout this subsection to make the subsequent comparison of the Oaxaca–Blinder estimates with the results obtained by Abadie and Imbens (2011) fully adequate.

Table 3 presents RMSDs from the experimental benchmark, calculated both for the estimates reported in Table 2 in Abadie and Imbens (2011, p. 6) and for the new estimates

contributed in the present paper. Abadie and Imbens (2011) have clearly shown that their regression-based bias correction can substantially improve the performance of matching estimators. RMSDs of bias-corrected matching estimators are in the order of 700–800, a considerable improvement. Interestingly, however, matching without bias-correction provides a slightly better replication of the experimental benchmark when the number of matches is very small ($M = 1$ and $M = 4$). Clearly, the quality of matches deteriorates as the number of matches grows, and this hampers the precision of matching without bias-correction. On the other hand, bias-correction is shown to guarantee that matching estimation is quite robust to the number of matches. See Table 2 in Abadie and Imbens (2011, p. 6) for details.

TABLE 3

A comparison of Abadie and Imbens (2011) with various estimates of the PATT based on the Oaxaca–Blinder unexplained component or linear regression

	Improving overlap?	*Estimate*	*(SE)*	*RMSD*
Abadie and Imbens (2011):				
Matching on covariates	No, full sample			7.675
Bias-adjusted matching on covariates	No, full sample			0.705
Matching on the score	No, full sample			1.817
Bias-adjusted matching on the score	No, full sample			0.772
New estimates:				
Regression	No, full sample	0.115	(0.832)	1.679
Oaxaca–Blinder	No, full sample	0.843	(0.906)	0.951
Stratification (on the score)	No, full sample	-3.065	(1.274)	4.859
Stratification (on the score) and regression	No, full sample	0.983	(1.167)	0.811
Regression	Common support	1.062	(0.897)	0.732
Oaxaca–Blinder	Common support	2.133	(0.964)	0.339
Stratification (on the score)	Common support	-1.020	(1.282)	2.814
Stratification (on the score) and regression	Common support	1.257	(1.174)	0.537
Regression	Crump *et al.* (2009)	-0.585	(1.077)	2.379
Oaxaca–Blinder	Crump *et al.* (2009)	0.617	(1.416)	1.177
Stratification (on the score)	Crump *et al.* (2009)	-0.181	(1.538)	1.975
Stratification (on the score) and regression	Crump *et al.* (2009)	-0.844	(1.536)	2.638

Notes: Root mean square deviations (RMSDs) are calculated for each estimator utilized by Abadie and Imbens (2011) using nonexperimental estimates for each number of matches (1, 4, 16, 64, and 2490) as one observation, and the experimental estimate as the benchmark value. They are in thousands of 1978 dollars. For the new estimates, they are equal to their absolute deviation from the benchmark. Propensity scores are estimated using a logit model. Each stratification-based estimator uses two strata of equal width. Common support refers to discarding all the nontreated individuals whose estimated propensity score is less than the minimum or greater than the maximum estimated propensity score for the treated individuals. Crump *et al.* (2009) refers to discarding all the individuals whose estimated propensity score is less than 0.1 or greater than 0.9.

Again, new estimates reported in Table 3 suggest that the Oaxaca–Blinder unexplained component provides a robust alternative to methods based on the propensity score and matching estimators. The full-sample Oaxaca–Blinder estimate of the PATT deviates ca. 950 dollars from the experimental benchmark, an acceptable result which is slightly poorer than the performance of bias-corrected matching estimators. On the other hand, imposing common support is shown to improve the precision of the Oaxaca–Blinder unexplained component again, and the resulting absolute deviation from the benchmark is as small as ca. 350 dollars; consequently, the Oaxaca–Blinder decomposition is again shown to provide the closest match among all the considered estimators.

Moreover, other results reported in the previous subsection seem to be robust to variable selection and hold in the current reanalysis as well. First, linear regression does a consistently worse job than the Oaxaca–Blinder unexplained component in replicating the experimental benchmark, since absolute deviations are typically twice as large for regression estimates. Second, imposing common support – i.e. discarding all the nontreated individuals whose estimated propensity score is less than the minimum or greater than the maximum estimated propensity score for the treated individuals – improves the precision of all the estimators considered, both regression-based (linear regression and the Oaxaca–Blinder decomposition) and based on the propensity score (stratification and the combination of stratification and linear regression). Third, the rule of thumb proposed by Crump *et al.* (2009) typically worsens the precision of the considered estimators to a significant extent, and all the absolute deviations are larger than 1,000 dollars (typically in the order of 2,000 dollars and more). Interestingly, however, the only estimator whose deviation can still be considered (relatively) acceptable is the Oaxaca–Blinder unexplained component, with the absolute deviation in the order of 1,200 dollars. Still, it seems to be a consistently bad idea to impose the rule of thumb proposed by Crump *et al.* (2009) if one aims at bias reduction.

IV. Summary and Conclusions

In this paper I have used the NSW data (see, e.g., LaLonde, 1986; Dehejia and Wahba, 1999; Smith and Todd, 2005) to examine the validity of the Oaxaca–Blinder decomposition as an estimator of the population average treatment effect on the treated (PATT). I have utilized the same dataset and variable selections which were used by Dehejia and Wahba (1999) and Abadie and Imbens (2011) in their studies of the NSW data to make my comparison of the performance of the Oaxaca–Blinder unexplained component with methods based on the propensity score (Dehejia and Wahba, 1999) and bias-corrected matching estimators (Abadie and Imbens, 2011) fully adequate. I have been able to show that in both cases the Oaxaca–Blinder unexplained component performs superior compared to the previously analyzed estimators provided that overlap is improved by imposing common support.

There are several general conclusions which emerge from the reanalysis which has been carried out in this paper. First, the Oaxaca–Blinder unexplained component seems to provide a valuable alternative to matching estimators and standard methods based on the propensity score. This is compatible with a recent contribution of Kline (2011) who has shown that the Oaxaca–Blinder decomposition *is* a specific version of a propensity score reweighting estimator and has provided a seminal application of this method to the NSW data. Second, while the Oaxaca–Blinder decomposition is also a flexible linear regression estimator, it has been shown to consistently outperform the most basic linear regression model. Thus, if linear regression is not enough robust an estimator, it is the treatment effects homogeneity assumption which is most problematic, not linearity itself. Third, various methods of improving overlap have different consequences for bias reduction. While discarding all the nontreated individuals whose estimated propensity score is less than the minimum or greater than the maximum estimated propensity score for the treated individuals

seems to be a highly recommendable choice (and it has indeed often been applied in practice; see, e.g., Dehejia and Wahba, 1999), the rule of thumb proposed by Crump *et al.* (2009) seems not to be working well for bias reduction; it is not inconsistent, however, with the original contribution of these authors.

APPENDIX TABLE 1

New estimates of the PATT using Dehejia and Wahba (1999) variable selections

	Improving overlap?	*PSID-1*	*PSID-2*	*PSID-3*	*CPS-1*	*CPS-2*	*CPS-3*	*RMSD*
Regression	No, full sample	217	677	787	1,567	916	1,075	1,007
		(810)	(1,000)	(1,057)	(631)	(664)	(734)	
Oaxaca–Blinder	No, full sample	1,733	2,345	2,450	1,669	1,274	1,843	414
		(879)	(1,088)	(1,253)	(660)	(743)	(922)	
Stratification (on the score)	No, full sample	-2,933	-587	1,208	-3,729	-391	502	3,299
		(1,011)	(1,169)	(1,246)	(663)	(793)	(969)	
Stratification (on the score) and regression	No, full sample	797	-649	840	1,755	1,543	1,067	1,188
		(1,043)	(1,269)	(1,378)	(764)	(868)	(989)	
Regression	Common support	-4	374	471	1,572	946	648	1,230
		-896	(960)	(1,122)	(638)	(664)	(726)	
Oaxaca–Blinder	Common support	1,478	1,971	1,714	1,900	1,691	1,232	281
		(965)	(1,122)	(1,225)	(692)	(800)	(973)	
Stratification (on the score)	Common support	-769	-231	922	312	541	630	1,661
		(966)	(1,204)	(1,346)	(665)	(793)	(970)	
Stratification (on the score) and regression	Common support	820	-455	765	1,909	1,543	1,023	1,136
		(993)	(1,302)	(1,475)	(751)	(862)	(989)	
Regression	Crump *et al.* (2009)	-1,395	-408	-349	1,695	356	417	1,983
		(1,204)	(1,140)	(1,292)	(743)	(742)	(852)	
Oaxaca–Blinder	Crump *et al.* (2009)	-844	40	-248	2,195	1,230	590	1,640
		(1,471)	(1,390)	(1,588)	(934)	(953)	(1,118)	
Stratification (on the score)	Crump *et al.* (2009)	-530	-1,585	-25	2,371	842	508	1,959
		(1,184)	(1,254)	(1,436)	(838)	(861)	(1,022)	
Stratification (on the score) and regression	Crump *et al.* (2009)	-882	-2,047	-1,427	2,471	937	626	2,410
		(1,326)	(1,373)	(1,530)	(896)	(910)	(979)	

Notes: Standard errors are in parentheses. Root mean square deviations (RMSDs) are calculated for each estimator using nonexperimental estimates for each control dataset as one observation, and the experimental estimate as the benchmark value. They are in 1978 dollars. Propensity scores are estimated using a logit model. Each stratification-based estimator uses two strata of equal width. Common support refers to discarding all the nontreated individuals whose estimated propensity score is less than the minimum or greater than the maximum estimated propensity score for the treated individuals. Crump *et al.* (2009) refers to discarding all the individuals whose estimated propensity score is less than 0.1 or greater than 0.9.

References

Abadie, A. and Imbens, G. W. (2011). 'Bias-corrected matching estimators for average treatment effects', *Journal of Business and Economic Statistics*, Vol. 29, pp. 1–11.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Princeton–Oxford.

Barsky, R., Bound, J., Charles, K. K. and Lupton, J. P. (2002). 'Accounting for the black-white wealth gap: a nonparametric approach', *Journal of the American Statistical Association*, Vol. 97, pp. 663–673.

Becker, S. O. and Ichino, A. (2002). 'Estimation of average treatment effects based on propensity scores', *Stata Journal*, Vol. 2, pp. 358–377.

Blinder, A. S. (1973). 'Wage discrimination: reduced form and structural estimates', *Journal of Human Resources*, Vol. 8, pp. 436–455.

Cobb-Clark, D. A. and Crossley, T. (2003). 'Econometrics for evaluations: an introduction to recent developments', *Economic Record*, Vol. 79, pp. 491–511.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2009). 'Dealing with limited overlap in estimation of average treatment effects', *Biometrika*, Vol. 96, pp. 187–199.

Dehejia, R. H. and Wahba, S. (1999). 'Causal effects in nonexperimental studies: reevaluating the evaluation of training programs', *Journal of the American Statistical Association*, Vol. 94, pp. 1053–1062.

Dehejia, R. H. and Wahba, S. (2002). 'Propensity score-matching methods for nonexperimental causal studies', *Review of Economics and Statistics*, Vol. 84, pp. 151–161.

Fortin, N., Lemieux, T. and Firpo, S. (2011). 'Decomposition methods in economics', in Ashenfelter O. and Card D. (eds), *Handbook of Labor Economics*, Vol. 4A, Elsevier, San Diego–Amsterdam, pp. 1–102.

Heckman, J. J. and Hotz, V. J. (1989). 'Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training', *Journal of the American Statistical Association*, Vol. 84, pp. 862–874.

Imbens, G. W. and Wooldridge, J. M. (2009). 'Recent developments in the econometrics of program evaluation', *Journal of Economic Literature*, Vol. 47, pp. 5–86.

Jann, B. (2008). 'The Blinder-Oaxaca decomposition for linear regression models', *Stata Journal*, Vol. 8, pp. 453–479.

Kline, P. (2011). 'Oaxaca-Blinder as a reweighting estimator', *American Economic Review*, Vol. 101, pp. 532–537.

LaLonde, R. J. (1986). 'Evaluating the econometric evaluations of training programs with experimental data', *American Economic Review*, Vol. 76, pp. 604–620.

Melly, B. (2006). 'Applied quantile regression', PhD dissertation, University of St. Gallen.

Oaxaca, R. (1973). 'Male-female wage differentials in urban labor markets', *International Economic Review*, Vol. 14, pp. 693–709.

Porro, G. and Iacus, S. M. (2009). 'Random recursive partitioning: a matching method for the estimation of the average treatment effect', *Journal of Applied Econometrics*, Vol. 24, pp. 163–185.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). 'Estimation of regression coefficients when some regressors are not always observed', *Journal of the American Statistical Association*, Vol. 89, pp. 846–866.

Smith, J. A. and Todd, P. E. (2001). 'Reconciling conflicting evidence on the performance of propensity-score matching methods', *American Economic Review*, Vol. 91, pp. 112–118.

Smith, J. A. and Todd, P. E. (2005). 'Does matching overcome LaLonde's critique of nonexperimental estimators?', *Journal of Econometrics*, Vol. 125, pp. 305–353.