Blinder-Oaxaca as a Reweighting Estimator[*]

Patrick Kline
UC Berkeley / NBER
pkline@econ.berkeley.edu

June, 2010

Abstract

The traditional regression-based estimator of Alan Blinder (1973) and Ronald Oaxaca (1973) constitutes a reweighting estimator based upon a linear model for the conditional odds of being treated. As such it enjoys the status of a "doubly robust" estimator of counterfactuals as in Robins, Rotnizky, and Zhao (1994) and Egel, Graham, and Pinto (2009) – estimation is consistent if *either* the propensity score assumption or the model for outcomes is correct. To illustrate the method, the Blinder-Oaxaca estimator is applied to LaLonde's (1986) study of the National Supported Work program where it is found to replicate experimental impacts more closely than competing approaches.

JEL codes: C1, C14.
Keywords: Counterfactuals, Blinder-Oaxaca, Propensity Score Reweighting, Double Robustness.

[*] Department of Economics, 508-1 Evans Hall #3880, Berkeley, CA 94720. I thank David Card, Justin McCrary, Bryan Graham, and Andres Santos for useful comments.

A large applied econometrics literature focuses on the use of reweighting methods for estimation of treatments effects and missing data problems.[1] Though much has been made of the efficiency properties of semi-parametric versions of reweighting estimators (Hirano, Imbens, and Ridder, 2004), in practice virtually all applications involve use of a parametric propensity score.[2] The purpose of this note is to point out that the traditional regression based estimator of Alan Blinder (1973) and Ronald Oaxaca (1973) constitutes a reweighting estimator based upon a linear model for the conditional odds of being treated – a functional form which emerges, for example, from an assignment model with a latent log-logistic error.[3] As such it enjoys the status of a "doubly robust" estimator of counterfactuals as in Robins, Rotnizky, and Zhao (1994) and Egel, Graham, and Pinto (2009) – estimation is consistent if *either* the propensity score assumption or the model for outcomes is correct. To illustrate the method, the Blinder-Oaxaca estimator is applied to LaLonde's (1986) study of the National Supported Work program where it is found to replicate experimental impacts more closely than competing approaches.

1 The Blinder-Oaxaca Estimator

Consider a population of individuals falling into two groups indexed by $D_i \in \{0,1\}$. We will refer to observations with $D_i = 1$ as the treatment group and those with $D_i = 0$ as the controls. A relevant example comes from LaLonde (1986) who studies a treatment group consisting of workers participating in a job training program and a corresponding set of controls composed of workers in the same cities known not to be participating.

Let X_i be a $K \times 1$ vector of covariates (which we assume includes an intercept) and Y_i some outcome of interest such as earnings. I assume throughout that $E[X_i X_i']$ is finite and invertible. We begin by indexing the potential outcomes associated with treatment as follows:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

where Y_i^1 is the outcome individual i would experience if treated and Y_i^0 is the outcome he would experience in the absence of treatment.

The Blinder-Oaxaca (B-O) approach is predicated on a model for the potential outcomes of the form:

$$Y_i^d = X_i' \beta^d + \varepsilon_i^d \tag{1}$$

$$E\left[\varepsilon_i^d | X_i, D_i\right] = 0 \;\text{ for }\; d \in \{0,1\} \tag{2}$$

Hence, one merely needs to obtain estimates of (β^1, β^0) in order to compute counterfactual means among covariate groups. Natural estimators of these coefficients come from linear regression in the two populations indexed by D_i.

Suppose in particular that we are interested in the counterfactual mean outcomes the treatment group would have experienced in the absence of treatment which we denote as:

$$\mu_0^1 \equiv E\left[Y_i^0 | D_i = 1\right]$$

[1] Imbens (2004) provides a review.

[2] See Dinardo, Fortin, and Lemieux (1996) for a prominent example.

[3] A special case of this equivalence was worked out in Dinardo (2002).

According to the model in (1) and (2):

$$\begin{aligned}\mu_0^1 &= E\left[X_i'\beta^0|D_i=1\right] \\ &= E\left[X_i|D_i=1\right]'\beta^0 \\ &= E\left[X_i|D_i=1\right]'E\left[X_iX_i'|D_i=0\right]^{-1}E\left[X_iY_i|D_i=0\right] \\ &\equiv \delta^{BO}\end{aligned}$$

where in the second line I have used the fact that β_0 is identified by the population regression of Y_i on X_i among members of the control group. When each of the moments in δ^{BO} is replaced by its sample analogue one obtains the B-O estimate of the counterfactual mean, which by standard arguments can be shown to be consistent for the parameter of interest.

2 Reweighting Estimators

A popular alternative to the Blinder-Oaxaca approach is to assume that the potential outcomes are conditionally independent of treatment given covariates or that:

$$\left(Y_i^1, Y_i^0\right) \perp\!\!\!\perp D_i|X_i \tag{3}$$

This restriction, which was popularized by Rosenbaum and Rubin (1983), amounts to assuming that treatment status (D_i) was assigned randomly conditional on covariates. Note that the parametric B-O model would satisfy this condition were we to strengthen the mean independence assumption (2) to encompass full conditional independence of the errors.[4] However (3) is usually considered less restrictive than the B-O assumptions since it is agnostic about the dependence of the potential outcomes on the covariates. It is instructive then to consider the population moments that identify μ_0^1 using only the nonparametric restrictions inherent in (3).

We must first make an additional assumption that ensures identification:

$$p\left(X_i\right) < 1 \tag{4}$$

where $p\left(X_i\right) = P\left(D_i=1|X_i\right)$ is the propensity score. This "common support" assumption merely states that suitable controls can be found for every treated unit, which allows us to state the following useful result:

Lemma 1 *If* (4) *holds then:*

$$\frac{dF\left(X_i|D_i=1\right)}{dF\left(X_i|D_i=0\right)} = \frac{p\left(X_i\right)}{1-p\left(X_i\right)}\frac{1-\pi}{\pi}$$

where $\pi = P\left(D_i=1\right)$ *and* $dF\left(X_i|D_i\right) = P\left(X_i|D_i\right)$.

Proof. Direct application of Bayes' rule. ■

We now use this Lemma to prove the following well-known result justifying the use of propensity score reweighting estimators:

[4] This would be equivalent to assuming in addition to (2) that $E\left[g\left(\varepsilon_i^d\right)|X_i,D_i\right] = E\left[g\left(\varepsilon_i^d\right)|X_i\right]$ for any continuous function $g(.)$ vanishing outside a finite interval and for $d \in \{0,1\}$. See e.g. Theorem 1.17 in Chapter V of Feller (1966).

Proposition 1 *If* (3) *and* (4) *hold then:*

$$\begin{aligned} \mu_0^1 &= E\left[w\left(X_i\right)Y_i|D_i=0\right] \qquad (5) \\ w\left(X_i\right) &= \frac{p\left(X_i\right)}{1-p\left(X_i\right)}\frac{1-\pi}{\pi} \end{aligned}$$

Proof.

$$\begin{aligned} E\left[w\left(X_i\right)Y_i|D_i=0\right] &= E\left[w\left(X_i\right)Y_i^0|D_i=0\right] \\ &= E\left[w\left(X_i\right)E\left[Y_i^0|X_i\right]|D_i=0\right] \\ &= \int E\left[Y_i^0|X_i\right]w\left(X_i\right)dF\left(X_i|D_i=0\right) \\ &= \int E\left[Y_i^0|X_i\right]dF\left(X_i|D_i=1\right) \\ &= E\left[Y_i^0|D_i=1\right] \end{aligned}$$

where the second line makes use of (3) and the fourth makes use of Lemma 1. ■

Thus, a weighted average of the control outcomes, with weights proportional to the conditional odds of treatment, identifies the counterfactual mean of the treated population. A large literature considers using sample analogues of (5) for estimation of μ_0^1, where $p\left(X_i\right)$ is replaced by some parametric or nonparametric estimator.[5]

3 Equivalence

Let us now return to the parametric B-O estimand δ^{BO}. That this quantity has an interpretation as a weighted average of the control outcomes is self-evident. Here I show that these weights have a particularly simple interpretation given the common support assumption (4). The following result is useful in developing that interpretation:

Lemma 2 *If* (4) *holds then:*

$$E\left[X_i|D_i=1\right] = E\left[X_i\frac{p\left(X_i\right)}{1-p\left(X_i\right)}\frac{1-\pi}{\pi}|D_i=0\right]$$

Proof. Application of Lemma 1 to definition of $E\left[X_i|D_i=1\right]$. ■

Note that this Lemma, which states that any covariate moment in the treated population has a representation as a propensity score reweighted moment in the control population, does not depend on (3). Armed with this relationship we may now show that the B-O estimand is equivalent to a propensity score reweighted average of the control outcomes given a linear model for the odds of treatment:

Proposition 2 *If* (4) *holds then:*

$$\begin{aligned} \delta^{BO} &= E\left[\widetilde{w}\left(X_i\right)Y_i|D_i=0\right] \\ \widetilde{w}\left(X_i\right) &= X_i'E\left[X_iX_i'|D_i=0\right]^{-1}E\left[X_i\frac{p\left(X_i\right)}{1-p\left(X_i\right)}\frac{1-\pi}{\pi}|D_i=0\right] \end{aligned}$$

[5] See Rosenbaum and Rubin (1983), Rosenbaum (1987), Dinardo, Fortin, and Lemieux (1996), Hirano, Imbens, and Ridder (2003), and Imbens (2004).

Proof. By Lemma 2, $\delta^{BO} = E\left[X_i \frac{p(X_i)}{1-p(X_i)} \frac{1-\pi}{\pi} | D_i = 0\right]' E\left[X_i X_i' | D_i = 0\right]^{-1} E\left[X_i Y_i | D_i = 0\right]$ ■

Note that the B-O weights $\widetilde{w}(X_i)$ are simply the normalized projection of the true treatment odds $\frac{p(X_i)}{1-p(X_i)}$ onto the column space of X_i – i.e. they are the predicted values from an (infeasible) population regression of $w(X_i)$ on X_i. Hence, the B-O specification provides a minimum mean squared error approximation to the true nonparametric weights $w(X_i)$.

Of course if the true odds of treatment are actually linear in X_i then $\widetilde{w}(X_i) = w(X_i)$ and Proposition 1 implies the B-O estimand will identify μ_0^1 even if the model for the outcomes is misspecified provided that (3) and (4) hold.[6] This functional form for the treatment odds arises naturally from an assignment model of the form:

$$D_i = 1\left[X_i'\delta + v_i > 0\right]$$

where $1[.]$ is an indicator for whether the condition in brackets is true and the assignment error v_i is an *iid* draw from a standardized log-logistic distribution.

Conversely, if the model for the outcomes in (1) and (2) is correct, the B-O estimand will identify μ_0^1 even if the common support condition (4) fails and/or the implicit model for the propensity score is incorrect. Hence the estimator is "doubly robust" as in Robins, Rotnizky, and Zhao (1994) and Egel, Graham, and Pinto (2009) as it identifies the parameter of interest under two independent sets of assumptions.

A Remark on Misspecification

The double robustness property offers little comfort to the applied econometrician who suspects any propensity score model, like any model for the conditional mean, to provide only a rough approximation to the data generating process. Note from Propositions 1 and 2 that the population bias in the B-O approximation may be written:

$$\mu_0^1 - \delta^{BO} = E\left[\left(w(X_i) - \widetilde{w}(X_i)\right) Y_i | D_i = 0\right]$$

Though the B-O weights may yield specification errors at particular values of X_i, those errors will only induce bias if they are correlated with outcomes in the control sample.[7] If, for instance, $\widetilde{w}(X_i) = w(X_i) + \xi_i$ where ξ_i is a random specification error obeying $E\left[\xi_i Y_i | D_i = 0\right] = 0$ then the B-O estimator will retain consistency. By Proposition 2 however we know that $\widetilde{w}(X_i)$ is the projection of $w(X_i)$ onto X_i implying that any specification errors ξ_i must at least be orthogonal to X_i. If Y_i^0 is approximately linear in the control sample (as is assumed in 1) then this is enough to also guarantee orthogonality with respect to Y_i.

By contrast, the standard practice of estimating a logistic propensity score via maximum likelihood can be shown to impose:

$$E\left[\left(p(X_i) - \widetilde{p}(X_i)\right) X_i\right] = 0$$

where $\widetilde{p}(X_i)$ is the logistic approximation to the propensity score. This expression cannot be manipulated to yield obvious restrictions on the implied logistic weights. An important question then is whether, in the absence of prior knowledge of the propensity score, approximations ought to be sought with respect to the propensity score or the weights themselves. The B-O approach follows the latter approach, the logistic reweighting model the former. Which approach removes more bias in a misspecified environment will depend on the true data generating process.

[6] This is to be contrasted with the standard practice of using a parametric logit model for the propensity score which assumes the odds of treatment take the form $\exp(X_i'\gamma)$ for some coefficient vector γ.

[7] Both sets of weights can be shown to have mean one which implies $E\left[w(X_i) - \widetilde{w}(X_i) | D_i = 0\right] = 0$.

4 Sample Properties

Thus far we have focused on the properties of the population moments defining the Blinder-Oaxaca estimator. It turns out that the sample moments have some interesting properties as well. Define $N_1 = \sum_i D_i$, and $\boldsymbol{X} = [\mathbf{1}, \boldsymbol{x_2}, ..., \boldsymbol{x_K}]$ where $\mathbf{1}$ is an $N \times 1$ vector of ones, and the elements of $\{\boldsymbol{x_2}, ..., \boldsymbol{x_K}\}$ are $N \times 1$ covariate vectors. Then we may write the B-O estimate of the counterfactual mean in matrix notation as:

$$\begin{aligned} \hat{\mu}_0^1 &= \frac{1}{N_1}\boldsymbol{D'X}\left(\boldsymbol{X'WX}\right)^{-1}\boldsymbol{X'WY} \\ &= \frac{1}{N_1}\boldsymbol{D'HY} \end{aligned} \tag{6}$$

where $\boldsymbol{Y}$ is the $N \times 1$ vector of outcomes, $\boldsymbol{D}$ is an $N \times 1$ vector whose elements consist of D_i, and $\boldsymbol{W}$ is a diagonal $N \times N$ weighting matrix taking values of one for control observations and zeros otherwise. The $N \times N$ matrix $\boldsymbol{H}$ is a generalization of the conventional "hat" matrix associated with OLS (Hoaglin and Welch, 1978). Each row $\mathbf{h}_i'$ of this matrix provides a set of weights $\widehat{\omega}_{ij}$ that sum to one across columns $\left(\sum_j \widehat{\omega}_{ij} = 1\right)$.[8] The inner product $\mathbf{h}_i'\mathbf{Y}$ is a weighted average of the outcomes that yields a prediction $X_i'\widehat{\beta}_0$ of the conditional untreated mean. Premultiplying the hat matrix by $\frac{1}{N_1}\boldsymbol{D'}$ averages these weights across across treated observations and hence yields an estimate $\widehat{\mu}_0^1$ of the average counterfactual outcome in the treated sample. A few properties of the averaged weights $\widehat{\omega}_j = \frac{1}{N_1}\sum_i D_i\widehat{\omega}_{ij} = \frac{1}{N_1}\boldsymbol{D'H}$ are notable:

1. The weights are zero for treated observations.

2. The weights sum to one.

3. Some of the weights may be negative. This occurs when the treatment odds implied by the linear model are negative.

Like conventional propensity score weights, B-O weights can be thought of as reweighting the controls to match the covariate distribution of the treated units. Note that for any covariate $\boldsymbol{x_j}$ in $\boldsymbol{X}$ we have by the properties of projection matrices that:

$$\frac{1}{N_1}\boldsymbol{D'Hx_j} = \frac{1}{N_1}\boldsymbol{D'x_j}$$

In words, the reweighted mean of every control covariate exactly equals its mean value among the treated sample. Hence the weights embodied in the Blinder-Oaxaca approach ensure exact balance of moments included in the regression model as in the recent paper by Egel, Graham, and Pinto (2009).

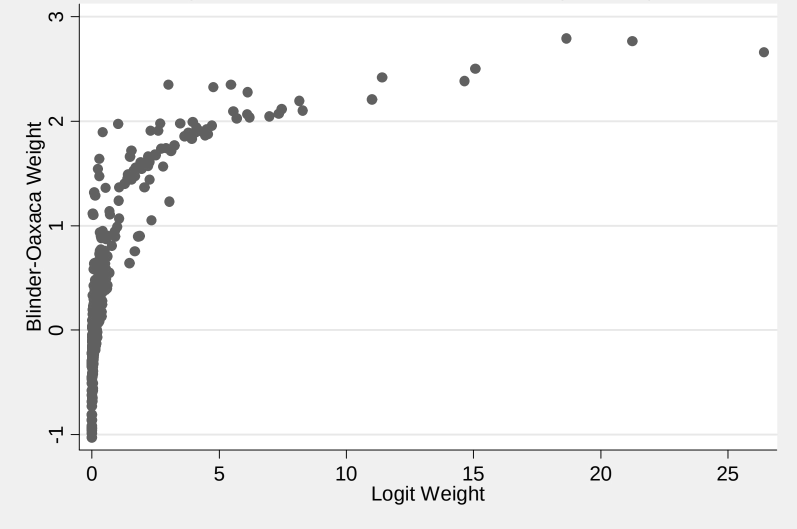
[8] See Appendix for proof.

5 Application

To illustrate use of the Blinder-Oaxaca estimator we revisit LaLonde's (1986) classic analysis of the National Supported Work (NSW) program using observational controls from the Current Population Survey. Attention is confined to a sample of men studied by Dehejia and Wahba (1999) with valid earnings data in both 1974 and 1975 who were present either in the NSW experimental sample or in Lalonde's "CPS-3" control group which consists of the poor and recently unemployed.[9] Because these data have been studied many times, I omit summary statistics which are reported elsewhere.[10] Three estimators: OLS, B-O, and reweighting based upon a logistic propensity score are contrasted; each using the set of demographic controls considered in Dehejia and Wahba (1999) along with 1974 and 1975 earnings.

Figure 1 plots a scatter of the renormalized B-O weights (the elements of $\boldsymbol{D'H}$) against the weights $\frac{\widehat{p}(X_i)}{1-\widehat{p}(X_i)}\frac{1-\widehat{\pi}}{\widehat{\pi}}$ derived from a propensity score reweighting estimator where $\widehat{p}(X_i)$ are predicted probabilities from a logit model estimated by Maximum Likelihood and $\widehat{\pi}$ is the fraction of treated observations.

Figure 1: Blinder-Oaxaca vs. Logit Weights



Unsurprisingly, the relationship between the two sets of weights is approximately logarithmic. However the B-O weights are often negative, a sign the implicit log-logistic propensity score model is likely misspecified. Of course the logistic model, despite yielding predictions in the unit interval, may also be misspecified. Ultimately, interest centers not on whether a propensity score model is literally correct, but the quality of approximation that can be provided to the true counterfactual μ_0^1.

Table 1 assesses this question empirically by comparing treatment effect estimates generated by

[9] See Smith and Todd (2005) for a detailed discussion of the implications of these sample restrictions.

[10] See for example Dehejia and Wahba (1999), Smith and Todd (2005), and Angrist and Pischke (2009).

each estimator using the observational CPS-3 controls and the experimental NSW controls.[11]

Table 1 - Estimated Impact of NSW on Men's 1978 Earnings

Estimator/Control Group	CPS-3	NSW
Raw Difference	−$635	$1794
	(677)	(671)
OLS	$1369	$1676
	(739)	(677)
Logistic Reweighting*	$1440	$1808
	(863)	(705)
Blinder-Oaxaca	$1701	$1785
	(841)	(677)
Sample Size	614	445

Note: Heteroscedasticity robust standard errors in parentheses.

*Reweighting standard errors computed from 1000 bootstrap replications.

Clearly covariate adjustments of virtually any sort help to remove bias in the observational sample. However, the B-O estimator yields observational impacts closest to those found in the experimental sample. This suggests the assumption of near linearity of untreated earnings in covariates provides a better approximation to the data generating process than the implicit assumptions of the workhorse logistic reweighting estimator or of simple OLS. Also of note is that the B-O estimator yields slightly smaller standard error estimates than logistic reweighting, even in the experimental sample.

6 Conclusion

The regression based Blinder-Oaxaca estimator of counterfactual means is equivalent to a propensity score reweighting estimator modeling the odds of treatment as a linear function of the covariates. This is be to contrasted with the standard practice in the applied literature of modeling the propensity score via a logit or probit link and using the estimated parameters to form estimates of the odds of treatment. The latter approach can be thought of as indirectly approximating the unknown odds via a different set of basis functions, albeit a set that imposes the side constraint that the odds are nonnegative. Whether, in the presence of misspecification, the imposition of this side constraint yields a better approximation to the counterfactual of interest is an empirical question and will depend on the data generating process.

Despite its allowance of negative weights, the Blinder-Oaxaca estimator has several features to commend it. It is easy to implement and allows for straightforward computation of standard errors and regression diagnostics. It is consistent if either the linear model for the potential outcomes or the implicit log-logistic model for the propensity score are correct. And if the outcome model is correctly specified and the errors are homoscedastic Blinder-Oaxaca is parametrically efficient even if the model for the treatment odds is incorrect or if the common support condition is violated. Finally, unlike standard reweighting estimators, the B-O weights yield exact covariate balance and are finite sample unbiased for the counterfactual under proper specification of the outcome equation.

[11] The Blinder-Oaxaca treatment effect estimate simply subtracts $\hat{\mu}_0^1$ from the mean sample outcome of treated units.

Appendix

Proof that $\sum_j \widehat{\omega}_{ij} = 1$:

From (6) we may write

$$\omega_{ij} = X_i' \boldsymbol{(X'WX)^{-1}} X_j \left(1 - D_j\right)$$

and

$$\sum_j \omega_{ij} = X_i' \boldsymbol{(X'WX)^{-1}} \boldsymbol{X}' \boldsymbol{W1}$$

It follows that:

$$\begin{aligned} \sum_j \omega_{ij} &= tr\left(X_i' \boldsymbol{(X'WX)^{-1}} \boldsymbol{X}' \boldsymbol{W1}\right) \\ &= tr\left(\boldsymbol{(X'WX)^{-1}} \boldsymbol{X}' \boldsymbol{W1} X_i'\right) \\ &= tr\left(\boldsymbol{(X'WX)^{-1}} \boldsymbol{X}' \boldsymbol{WJ}\right) \\ &= tr\left(\boldsymbol{B}\right) \end{aligned}$$

Note that $\boldsymbol{(X'WX)^{-1}} \boldsymbol{X}' \boldsymbol{WJ}$ is a multivariate regression of each element of X_i' on $\boldsymbol{X}$ in the control sample. Because i is fixed this yields a $K \times K$ coefficient matrix $\boldsymbol{B} = \left[\begin{array}{c} X_i' \\ \boldsymbol{0} \end{array}\right]$. By assumption the constant term was ordered first in $\boldsymbol{X}$, hence $tr\left(\boldsymbol{B}\right) = 1$.

References

1. Angrist, Joshua and Steven Pischke. 2009. *Mostly Harmless Econometrics.* Princeton: Princeton University Press.

2. Blinder, Alan. 1973. "Wage Discrimination: Reduced Form and Structural Estimates." *Journal of Human Resources* 8(4): 436-455.

3. Dehejia, Rajeev and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(448): 1053-1062.

4. Dinardo, John. 2002. "Propensity Score Reweighting and Changes in Wage Distributions." Mimeo.

5. DiNardo, John, Nicole Fortin, and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." *Econometrica* 64(5):1001-1044.

6. Egel, Daniel, Bryan Graham, and Cristine Pinto. 2009. "Efficient Estimation of Data Combination Problems by the Method of Auxiliary-to-Study Tilting." Mimeo.

7. Feller, William. 1966. *An Introduction to Probability Theory and Its Applications.* Volume II New York: John Wiley & Sons.

8. Heckman, James and Richard Robb. 1984. "Alternative Methods for Evaluating the Impact of Interventions." in J. Heckman and B. Singer (eds.) *Longitudinal Analysis of Labor Market Data* Cambridge, U.K.: Cambridge University Press.

9. Hoaglin, David and Roy Welsch. 1978. "The Hat Matrix in Regression and ANOVA." *The American Statistician* 32(1): 17-22.

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

11. Hirano, Keisuke, Guido Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71(4): 1161-1189.

12. LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76(4):604-620.

13. Oaxaca, Ronald. 1973. "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review* 14(3): 693-709.

14. Robins, James, Andrea Rotnizky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89(427): 846-866.

15. Rosenbaum, Paul. 1987. "Model-Based Direct Adjustment." *Journal of the American Statistical Association* 82(398):387-394.

16. Rosenbaum, Paul and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41-45.

17. Smith, Jeffrey and Petra Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125(1):305-353.