Sensitivity to Exogeneity Assumptions in Program Evaluation

By GUIDO W. IMBENS*

In many empirical studies of the effect of social programs researchers assume that, conditional on a set of observed covariates, assignment to the treatment is exogenous or unconfounded (aka selection on observables). Often this assumption is not realistic, and researchers are concerned about the robustness of their results to departures from it. One approach (e.g., Charles Manski, 1990) is to entirely drop the exogeneity assumption and investigate what can be learned about treatment effects without it. With unbounded outcomes, and in the absence of alternative identifying assumptions, there are no restrictions on the set of possible values for average treatment effects. This does not mean, however, that all evaluations are equally sensitive to departures from the exogeneity assumption. In this paper I explore an alternative approach, developed by Paul Rosenbaum and Donald Rubin (1983), where the assumption of exogeneity is explicitly relaxed by allowing for a limited amount of correlation between treatment and unobserved components of the outcomes.

The starting point of the sensitivity analysis is the assumption that the exogeneity assumption is satisfied only conditional on an additional unobserved covariate. Making assumptions about the effect of the unobserved covariate on the outcome and its correlation with the treatment, I trace out the set of possible values for the treatment effect of interest. By considering a sufficiently large set of possible correlations with outcomes and treatment, one can recover the bounds on the treatment effect derived by Manski (1990). The approach here, in the spirit of Rosenbaum and Rubin (1983) and Rosenbaum (1995), is to allow only a limited amount of correlation and to judge the sensitivity of average treatment-effect estimates to such correlations. There are two novel features of the proposed analysis. First, rather than formulate the sensitivity in terms of coefficients on the unobserved covariate, the sensitivity results are presented in terms of partial R^2 values, which may be easier to interpret. Second, the partial R^2 values of the unobserved covariates are compared to those for the observed covariates in order to facilitate judgments regarding the plausibility of values necessary to substantially change results obtained under exogeneity.

The proposed sensitivity analysis is conceptually related to the practice of assessing sensitivity of estimates by comparisons with results obtained by discarding one or more observed covariates (James Heckman and V. Joseph Hotz, 1989; Rajeev Dehejia and Sadek Wahba, 1999; Jeffrey Smith and Petra Todd, 2001). The attraction of the sensitivity analysis is that it is more directly relevant: one is not interested in what would have happened in the absence of covariates actually observed, but in biases that are the result from not observing all relevant covariates.

I. Setup

The interest is in the effect of a program or treatment evaluated on the basis of data from a population of units, individuals, or firms, some of whom were exposed to the active treatment and the remainder of whom were exposed to the control treatment. Let $W_i \in \{0, 1\}$ be the indicator for the treatment, where $i = 1, \ldots, N$ labels the units in the population. Following the potential-outcome notation popularized for causal inference in observational studies by Donald Rubin (1974), let $Y_i(w)$ for $w = 0, 1$ denote the outcome for unit i if treatment w is applied. For unit i, I observe the treatment indicator W_i and the outcome corresponding to the treatment actually received, $Y_i = Y_i(W_i)$. In addition a vector of pretreatment variables or covariates $\mathbf{X}_i$ is observed.

The starting point is the following unconfoundedness (exogeneity) assumption:

* Department of Economics and Department of Agricultural and Resource Economics, University of California, 549 Evans Hall, Berkeley, CA 94720-3880 and NBER (e-mail: imbens@econ.berkeley.edu). I am grateful for comments and suggestions by Petra Todd and for financial support under grant SES 0136789 by the NSF.

$$Y_i(0),\ Y_i(1) \perp W_i|\mathbf{X}_i.$$

Combined with the assumption that the probability of receiving the treatment, given covariates, is bounded away from 0 and 1, this assumption implies that one can estimate the average effect of the treatment on the outcome by first estimating the average effect with subpopulations homogenous in the covariates, through the following equality:

$$\tau(x) \equiv E[Y(1) - Y(0)|X = x]$$

$$= E[Y|X = x,\ W = 1]$$

$$- E[Y|X = x,\ W = 0]$$

followed by averaging this over the distribution of covariates,

$$\tau \equiv E[Y(1) - Y(0)] = E[\tau(X)].$$

For a recent survey of methods for implementing such estimators, see Imbens (2002). In many cases this is an attractive starting point. Even if the exogeneity assumption is controversial, it is often sensible to follow a discussion of the summary statistics, which would include a simple comparison of averages for treated and control outcomes, by a discussion of this comparison adjusted for differences in covariates. Following such an analysis one may wish to go further and consider alternative approaches such as instrumental-variables analyses (e.g., Heckman and Richard Robb, 1985; Joshua Angrist et al., 1996) or bounds analyses (Manski, 1990; John Pepper, 2003) that allow selection into the treatment to be partially or wholly determined by potential outcomes. Here we discuss an alternative approach, due to Rosenbaum and Rubin (1983).

In the sensitivity analysis, the unconfoundedness assumption is weakened to require independence of the potential outcomes and the treatment indicator only after conditioning on one additional, unobserved, covariate U_i:

$$(1) \qquad Y_i(0),\ Y_i(1) \perp W_i|X_i,\ U_i.$$

This assumption is without loss of generality, and one can recover the bounds by appropriate choices for the conditional distribution of potential outcomes and assignment given the unobserved and observed covariates. Moreover, without loss of generality one can choose U_i to be independent of X_i. To reduce the set of possible average treatment effects, one therefore has to restrict the set of distributions for the assignment given U and X and the set of distributions for the potential outcomes given U and X.

II. Implementation

The first key point is that a parametric model is postulated. For expositional reasons, a simple parametric model is used here. It can be generalized in many ways, but often most of the insights of a sensitivity analysis can be obtained with relatively simple models.

First, the marginal distribution of U is postulated to be binomial, with

$$U \sim \mathcal{B}(1,\ 1/2)$$

so that $\Pr(U = 1) = \Pr(U = 0) = 1/2$. Second, it is assumed that the distribution of W given U and $\mathbf{X}$ follows a logistic distribution:

$$\Pr(W = 1|\mathbf{X},\ U) = \frac{\exp(\boldsymbol{\gamma}'\mathbf{X} + \alpha U)}{1 + \exp(\boldsymbol{\gamma}'\mathbf{X} + \alpha U)}.$$

Third, it is assumed that the distribution of $Y(w)$ given U and $\mathbf{X}$ is normal with a constant treatment effect τ:

$$Y(w)|\mathbf{X},\ U \sim \mathcal{N}(\tau w + \boldsymbol{\beta}'\mathbf{X} + \delta U,\ \sigma^2).$$

One way to recover the standard estimates based on the unconfoundedness or exogeneity assumption is to fix $\alpha = \delta = 0$ and estimate the remaining parameters by maximum likelihood. The sensitivity analysis corresponds to choosing alternative values for (α, δ) and calculating the maximum-likelihood estimate for τ. Note that no attempt is made to recover α and δ from the data. Although these parameters may be identified given the distributional and functional form assumptions, their identification is very weak. Specifically, their identification is not nonparametric, as without functional form and distributional assumptions one cannot reject the unconfoundedness assumption.

Specifically, let $L(\tau, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \alpha, \delta)$ be the logarithm of the likelihood function given a sample $(Y_i, W_i, \mathbf{X}_i)$, $i = 1, \ldots, N$, obtained by integrating out the missing potential outcomes and the unobserved covariate:

$$L(\tau, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \alpha, \delta) = \sum_{i=1}^{N} \ln\Bigg[\frac{1}{2}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \times \exp\left(-\frac{1}{2\sigma^2}(Y_i - \tau W_i - \boldsymbol{\beta}'\mathbf{X}_i)^2\right) \times \frac{(\exp(\boldsymbol{\gamma}'\mathbf{X}_i))^{W_i}}{1 + \exp(\boldsymbol{\gamma}'\mathbf{X}_i)} + \frac{1}{2}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \times \exp\left(-\frac{1}{2\sigma^2}(Y_i - \tau W_i - \boldsymbol{\beta}'\mathbf{X}_i - \delta)^2\right) \times \frac{(\exp(\boldsymbol{\gamma}'\mathbf{X}_i + \alpha))^{W_i}}{1 + \exp(\boldsymbol{\gamma}'\mathbf{X}_i + \alpha)}\Bigg].$$

For fixed (α, δ), the remaining parameters will be estimated by maximum likelihood. Thus, one can express the maximum-likelihood estimate for τ fixed (α, δ) as $\hat{\tau}(\alpha, \delta)$. It is this representation of the estimated average treatment effect in terms of the sensitivity parameters α and δ that is the focus of the sensitivity analysis. By considering a range of plausible values for α and δ, a range of average treatment effects consistent with these values is obtained.

In order to carry out this analysis it is crucial to specify a range of plausible values for these sensitivity parameters. This is difficult because the sensitivity parameters α and δ are not always easy to interpret. I therefore transform the sensitivity parameters into more easily interpretable quantities. This transformation involves separating the variation in outcomes and treatment assignment into variation explained by the observed covariates, the unobserved covariates, and the remainder. I then consider the amount of variation that is explained by the unobserved covariate relative to the amount not explained by the observed covariates.

Formally, let $R^2_Y(\alpha, \delta)$ be the share of the variation in Y explained by $\mathbf{X}$, W, and U:

$$R^2_Y(\alpha, \delta) = 1 - \hat{\sigma}^2(\alpha, \delta)/\Sigma_Y$$

where $\Sigma_Y = \Sigma_i (Y_i - \bar{Y})^2/N$ is the variance of Y. The proportion of the previously unexplained variation in Y that is explained by the unobserved covariate U is

$$R^2_{Y,\text{par}}(\alpha, \delta) = \frac{R^2_Y(\alpha, \delta) - R^2_Y(0, 0)}{1 - R^2_Y(0, 0)} = \frac{\hat{\sigma}^2(0, 0) - \hat{\sigma}^2(\alpha, \delta)}{\hat{\sigma}^2(0, 0)}.$$

For the treatment indicator regression, there is no natural R^2. Instead I use the explained variation in the latent index in a latent-index representation. Under the logit model, the latent-index error term has a logistic distribution with variance $\pi^2/3$. Let $\boldsymbol{\Sigma}_\mathbf{X}$ be the sample covariance matrix of the observed covariates $\mathbf{X}$ (omitting the constant term). The explained variation is the variation in $\mathbf{X}'\boldsymbol{\gamma} + U\alpha$ which, because of the independence of $\mathbf{X}$ and U, is equal to $\boldsymbol{\gamma}'\boldsymbol{\Sigma}'_\mathbf{X}\boldsymbol{\gamma} + \alpha^2/4$. Hence the implicit R^2 is

$$R^2_W(\alpha, \delta) = \frac{\hat{\boldsymbol{\gamma}}(\alpha, \delta)'\boldsymbol{\Sigma}_X\hat{\boldsymbol{\gamma}}(\alpha, \delta) + \alpha^2/4}{\hat{\boldsymbol{\gamma}}(\alpha, \delta)'\boldsymbol{\Sigma}_\mathbf{X}\hat{\boldsymbol{\gamma}}(\alpha, \delta) + \alpha^2/4 + \pi^2/3}$$

and the partial R^2 used in the sensitivity analysis is

$$R^2_{W,\text{par}}(\alpha, \delta) = \frac{R^2_W(\alpha, \delta) - R^2_W(0, 0)}{1 - R^2_W(0, 0)}.$$

To present the results, I will construct pairs of values of $(R^2_{W,\text{par}}(\alpha, \delta), R^2_{Y,\text{par}}(\alpha, \delta))$ for such pairs (α, δ) so that the implied average treatment effect $\hat{\tau}(\alpha, \delta)$ changes by some preset amount. If the set of all such values does not include reasonable values of the partial R^2 values, then the sign of the estimated average treatment effect under unconfoundedness is judged to be robust. To judge whether a particular pair of values is reasonable, I will compare them to pairs of partial R^2 values corresponding to observed covariates.

III. An Application

In this section I discuss an application of the sensitivity analysis to the evaluation of labor-market programs. In nonexperimental evaluation of labor-market programs researchers are often concerned about biases arising from differences in motivation between those volunteering for a program and those who do not. Strong

TABLE 1—MEAN AND STANDARD DEVIATIONS, LALONDE DATA

Variable	Experimental data: Treatment (N = 185)	Experimental data: Control (N = 260)	PSID control (N = 2,490)	Restricted control (N = 242)
Married	0.19	0.15	0.87	0.78
	(0.39)	(0.36)	(0.34)	(0.42)
Age	25.82	25.05	34.85	38.61
	(7.16)	(7.06)	(10.44)	(11.45)
Black	0.84	0.83	0.25	0.27
	(0.36)	(0.38)	(0.43)	(0.44)
Hispanic	0.06	0.11	0.03	0.04
	(0.24)	(0.31)	(0.18)	(0.20)
Education	10.35	10.09	12.12	11.37
	(2.01)	(1.61)	(3.08)	(3.40)
Earnings, 1974	2.10	2.11	19.43	0.77
	(4.89)	(5.69)	(13.41)	(1.40)
Unemployed, 1974	0.71	0.75	0.09	0.71
	(0.46)	(0.43)	(0.28)	(0.46)
Earnings, 1975	1.53	1.27	19.06	0.65
	(3.22)	(3.10)	(13.60)	(1.33)
Unemployed, 1975	0.60	0.68	0.10	0.75
	(0.49)	(0.47)	(0.30)	(0.43)
Earnings, 1978	6.35	4.55	21.55	3.45
	(7.87)	(5.48)	(15.56)	(7.43)

motivation to enroll in a job-training program may lead to more favorable outcomes in either regime. Hence, omitting such individual characteristics may lead to biases in average treatment effects estimated under the assumption of exogenous treatment assignment. With such specific interpretations for the unobserved covariates, one may be able to provide ranges of reasonable values for the partial R^2 values for the unobserved covariates. Such ranges may be based on such considerations as whether it is reasonable that motivation has more explanatory power for future earnings than last period's earnings, or than a multi-period earnings history. Similarly, one may consider whether motivation has more explanatory power for selection into the program than education or the presence of children.

The application of the sensitivity analysis is to data from a job-training program first analyzed by Robert Lalonde (1986) and subsequently by Heckman and Hotz (1989), Dehejia and Wahba (1999), and Smith and Todd (2001). I use both the experimental data and the nonexperimental sample from the Panel Study of Income Dynamics (PSID). The experimental estimate of the average treatment effect is \$1,672 (SE = 626). Using the experimental data, I present the set of partial R^2 values that would be required to change the value of the implied average treatment effect by more than \$1,000. (Alternatively one could present the set of partial R^2 values that would be required to change the sign of the average treatment effect. In order to make the graphs comparable across samples, I do the former.) The sensitivity analyses are carried out for four comparisons: (i) trainees versus experimental controls, (ii) trainees versus the full set of PSID controls, (iii) trainees versus the full set of PSID controls with outcome defined as change in earnings, (iv) a restricted set of trainees and PSID controls where individuals with earnings exceeding \$5,000 in earnings in 1974 or 1975 are dropped. (This restricted sample includes most of the experimental sample [148 out of 185], but less than 10 percent of the PSID control sample [242 out of 2,490].)

Table 1 presents summary statistics for the trainees and the three control groups (experimental controls, all PSID controls, and restricted PSID controls). There are nine covariates (married, age, indicators for black and Hispanic, education, earnings in 1974 and 1975, and indicators for positive earnings in 1974 and 1975), and the outcome (earnings in 1978). Note the large differences in background characteristics between the trainees and the PSID sample. This is what makes drawing causal inferences from comparisons between the PSID sample and the trainee group a potentially tenuous task.

In Figures 1–4 the sensitivity analyses are presented for the four comparisons. In each figure, the solid curve is the set of partial R^2 values that corresponds to an average treatment effect different from the estimate under exogeneity by \$1,000. In addition the partial R^2 values for each of the nine observed covariates are represented by "+" signs. Finally the partial R^2 value for the two variables corresponding to the most recent lag of earnings is represented by a "○," and the partial R^2 value for all pre-program earnings by a "*."

Consider Figure 1 for the experimental comparison. The curve describes how strongly an unobserved binary covariate would have to be correlated with outcomes and the treatment indicator to change the average treatment effect by \$1,000. For example, it would require an unobserved covariate explaining 20 percent of the variation in treatment assignment and 2

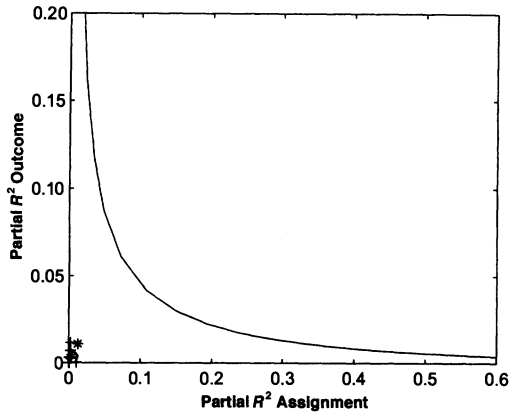

FIGURE 1. LALONDE EXPERIMENTAL SAMPLE

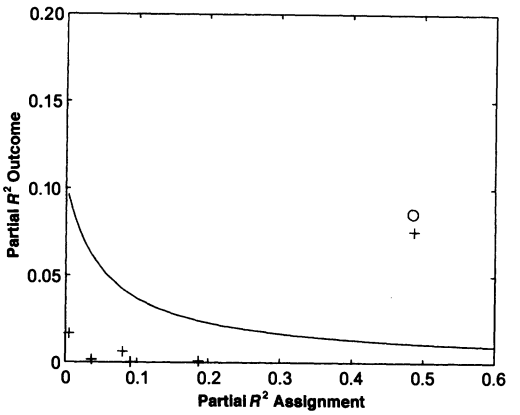

FIGURE 3. LALONDE NONEXPERIMENTAL GAIN SAMPLE

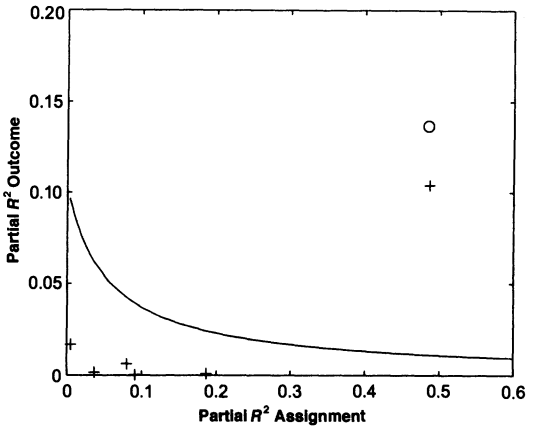

FIGURE 2. LALONDE NONEXPERIMENTAL SAMPLE

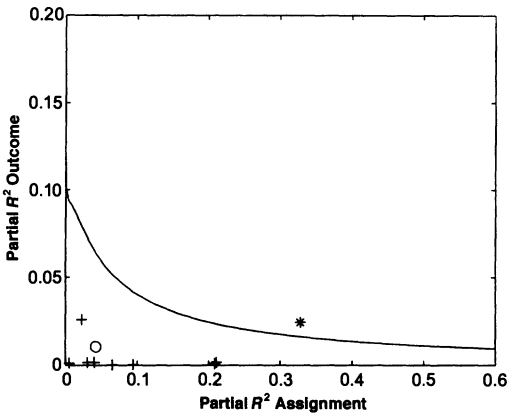

FIGURE 4. LALONDE RESTRICTED SAMPLE

percent of the variation in 1978 earnings not explained by the observed covariates to change the average treatment effect by $1,000. However, consider the explanatory power of the observed covariates. Even the combined power of all four earnings variables is very little, only 1.3 percent and 1.1 percent for the variation in treatment assignment and outcome, respectively. In this case, any unobserved covariate that would be sufficiently strongly correlated with treatment assignment and 1978 earnings to change the sign of the average treatment effect would have to be much more important than 1974 and 1975 earnings combined. The implication is that in this case the results are very robust to violations of the unconfoundedness assumption.

Next, consider the sensitivity analysis for the nonexperimental data. Here, an unobserved covariate explaining 25 percent of the variation in the treatment assignment and 2 percent of the variation in the outcome not explained by the observed covariates could lead to a bias sufficient to change the average treatment effect by $1,000. To judge whether that is a substantial amount of explanatory power, let us turn to observed covariates. The covariate "1975-earnings" explains about 48 percent of the variation in treatment assignments and about 10 percent of the variation in 1978 earnings not explained by the other covariates. In fact if one looks at the combination of 1975 earnings and the indicator for positive 1975 earnings, one sees that these explain 48 percent and 14 percent of the variation in treatment assignment and outcome (the "o" in Fig. 2). Hence, the explanatory power of the unobserved covariate could be substantially

less than that of 1975 earnings and still change the estimate of the average treatment effect by more than $1,000. In this case the sensitivity analysis suggests that the results based on the PSID controls are very sensitive to the unconfoundedness assumption. (In this figure the "*" is not visible, as it is at the point (0.72, 0.48): combined the preprogram earnings explain a very large proportion of the variation in treatment assignment and the outcome.) I also carry out the sensitivity analysis using the change in earnings as the outcome variable, following the argument in Heckman et al. (1997) that this can improve robustness. Changing the outcome lowers the partial R^2 values of the observed covariates but does not change the curve of R^2 values required for changing the average treatment effect by $1,000. Figure 3 shows that the estimated average treatment effect is still very sensitive to the exogeneity assumption.

Finally, consider the restricted nonexperimental sample. Restricting the sample to those with earnings below $5,000 in both 1974 and 1975 ensures that the PSID and experimental groups are much more similar in terms of covariates. The results from this sensitivity analysis are presented in Figure 4. The set of partial R^2 values required to change the average treatment effect by at least $1,000 does not differ much from that for the full nonexperimental sample. However, the explanatory power of the observed covariates is much reduced. Now in order to change the estimated average treatment effect by more than $1,000 the unobserved covariate would have to be much more powerful than the two 1975 earnings variables combined (the "○"), although it would not have to be as powerful as the four earnings variables together (the "*"). The conclusion is that restricting the sample in a way that makes the two treatment groups more homogenous can lead to more robust results.

IV. Conclusion

In this paper I extend the sensitivity analysis developed by Rosenbaum and Rubin (1983) and apply it to the evaluation of a job-training program previously analyzed by Lalonde (1986). The results based on the experimental data are shown to be much more robust than those based on the nonexperimental data. Using a restricted subset of the nonexperimental data improves the robustness considerably. The sensitivity analyses appear to be useful tools complementing analyses relying exclusively on unconfoundedness assumptions as well as bounds analyses that avoid such assumptions altogether.

REFERENCES

Angrist, Joshua; Imbens, Guido and Rubin, Donald. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, June 1996, *91*(434), pp. 444–72.

Dehejia, Rajeev and Wahba, Sadek. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*, December 1999, *94*(448), pp. 1053–62.

Heckman, James and Hotz, V. Joseph. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association*, December 1989, *84*(408), pp. 862–74.

Heckman, James; Ichimura, Hidehiko and Todd, Petra. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies*, October 1997, *64*(4), pp. 605–54.

Heckman, James and Robb, Richard. "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer, eds., *Longitudinal analysis of labor market data*. New York: Cambridge University Press, 1985, pp. 156–245.

Imbens, Guido. "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review." Unpublished manuscript, University of California–Berkeley, December 2002.

Lalonde, Robert. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*, September 1986, *76*(4), pp. 604–20.

Manski, Charles. "Nonparametric Bounds on Treatment Effects." *American Economic Review*, May 1990 (*Papers and Proceedings*), *80*(2), pp. 319–23.

Pepper, John. "Using Experiments to Evaluate Performance Standard: What Do Welfare-to-Work Demonstrations Reveal to Welfare Reformers?" *Journal of Human Resources*, 2003 (forthcoming).

Rosenbaum, Paul. *Observational studies*. New York: Springer-Verlag, 1995.

Rosenbaum, Paul and Rubin, Donald. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome." *Journal of the Royal Statistical Society, Series B*, 1983, *45*(2), pp. 212–18.

Rubin, Donald. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, 1974, *66*(5), pp. 688–701.

Smith, Jeffrey and Todd, Petra. "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods." *American Economic Review*, May 2001 (*Papers and Proceedings*), *91*(2), pp. 112–18.