sociological science

Heterogeneous Causal Effects and Sample Selection Bias

Richard Breen,[a] Seongsoo Choi,[a] Anders Holm[b]

a) Yale University; b) University of Copenhagen

Abstract: The role of education in the process of socioeconomic attainment is a topic of long standing interest to sociologists and economists. Recently there has been growing interest not only in estimating the average causal effect of education on outcomes such as earnings, but also in estimating how causal effects might vary over individuals or groups. In this paper we point out one of the under-appreciated hazards of seeking to estimate heterogeneous causal effects: conventional selection bias (that is, selection on baseline differences) can easily be mistaken for heterogeneity of causal effects. This might lead us to find heterogeneous effects when the true effect is homogenous, or to wrongly estimate not only the magnitude but also the sign of heterogeneous effects. We apply a test for the robustness of heterogeneous causal effects in the face of varying degrees and patterns of selection bias, and we illustrate our arguments and our method using National Longitudinal Survey of Youth 1979 (NLSY79) data. data.

Keywords: causality; sample selection bias; heterogeneous treatment effects; propensity score; education; NLSY79

The role of education in the process of socioeconomic attainment is a topic of longstanding interest to sociologists and economists. Following the causal revolution of the past thirty years in the social sciences, there is a growing interest not only in estimating the causal effect of education on outcomes, such as earnings, but also in understanding and estimating how these effects differ among individuals (Brand and Xie 2010 provide a sociological example, Carneiro, Heckman, and Vytlacil 2011 an economic one). In this paper, drawing on earlier work by Heckman and Navarro-Lozano (2004), we point out one of the under-appreciated hazards of seeking to estimate heterogeneous causal effects: conventional selection bias (that is, selection on baseline differences) can easily be confused with heterogeneity of causal effects. One consequence is that analyses might find heterogeneous effects when the true effect is homogenous; another is that we might wrongly estimate either the direction or magnitude, or both, of a truly heterogeneous effect.

We develop this argument initially using the potential outcomes approach to explain what we mean by selection on baseline differences and by heterogeneous causal effects. Our focus here is on heterogeneity of the causal effects according to the probability of receiving treatment. Then we show how, in both an important special case and in general, selection may confound estimates of heterogeneous causal effects, or, more simply, how selection on baseline differences may easily be mistaken for heterogeneous causal effects. Next we describe a test that allows us to calculate the degree to which estimates of heterogeneous causal effects are robust to different types and kinds of baseline selection bias. We apply this method

Citation: Breen, Richard, Seongsoo Choi and Anders Holm. 2015. "Heterogeneous Causal Effects and Sample Selection Bias." Sociological Science 2: 351-369.

Received: November 4, 2014

Accepted: January 15, 2015

Published: June 8, 2015

Editor(s): Jesper Sørensen, Stephen L. Morgan

DOI: 10.15195/v2.a17

to National Longitudinal Survey of Youth (NLSY79) data with which we replicate the finding of Brand and Xie (2010) that, for the case of earnings, there is negative selection into college in the United States. We show that this result is sensitive to even a small amount of selection bias. We end the paper with a short summary and conclusion.

Potential Outcome Causal Models

Given a binary treatment, D, and observations indexed by i, we use Y_i^0, Y_i^1 to mean the potential outcomes given treatment, $D_i = 1$, and non-treatment, $D_i = 0$, respectively. The individual level causal effect is defined as $Y_i^0 - Y_i^1$, and the average causal effect (or average treatment effect) is

$$E\left(Y_i^1 - Y_i^0\right) \tag{1}$$

For each observation, one or the other of the potential outcomes is counterfactual. The observed outcome is

$$Y_i = Y_i^1 D_i + Y_i^0 \left(1 - D_i\right) \tag{2}$$

The naïve estimator of the causal effect using observed outcomes, $E(Y|D=1) - E(Y|D=0)$, will be biased because of the existence of confounders. Given that some of these have been measured, and denoting them X, we can form an estimator that conditions on X:

$$E\left(Y|D=1,X\right) - E(Y|D=0,X) = E\left(Y^1|D=1,X) - E(Y^0|D=0,X\right) \tag{3}$$

But, considered as an estimator of Equation (1), Equation (3) is still potentially biased because

$$\begin{aligned} &E\left(Y^1|D=1,X\right) - E\left(Y^0|D=0,X\right) \\ =&E\left(Y_i^1 - Y_i^0\right) - c(0,X) + (1-\pi)\left[c(1,X) + c(0,X)\right] \\ =&E\left(Y_i^1 - Y_i^0\right) - \pi c(0,X) + (1-\pi)c(1,X) \end{aligned} \tag{4}$$

In Equation (4), π is the proportion of observations for which $D = 1$ and

$$\begin{aligned} c(0,X) &= E\left(Y^0|D=0,X\right) - E\left(Y^0|D=1,X\right) \\ c(1,X) &= E\left(Y^1|D=1,X\right) - E\left(Y^1|D=0,X\right) \end{aligned} \tag{5}$$

The first term on the righthand side of Equation (4) is the true causal effect; the second term, $c(0,X)$, is usually termed baseline or pre-treatment bias (formerly often termed ‘sample selection bias’); and the third term is the differential treatment effect bias. Baseline bias means that, even conditioning on X, those who receive treatment and those who do not would differ in their average outcome if they had not received treatment. It arises because of the existence of one or more relevant variables, not included in X, that affect both the potential outcomes and the probability of receiving treatment.

Differential treatment effect bias means that the true average treatment effect (the average difference between Y^1 and Y^0) is different for those who received treatment compared with those who did not. We can see this more easily if we rearrange the relevant term of Equation (4):

$$\begin{aligned} & c(1,X) + c(0,X) \\ = & \left[E\left(Y^1|D=1,X\right) - E\left(Y^0|D=1,X\right)\right] \\ & -\left[E\left(Y^1|D=0,X\right) - E\left(Y^0|D=0,X\right)\right] \\ = & \text{ATT} - \text{ATU} \end{aligned}$$

Differential treatment effect bias means that the average treatment effect among the treated (ATT) differs from that for the untreated (ATU). The estimator given by Equation (3) is an unbiased estimate of the average causal effect, Equation (1), only if $c(0,X) = 0$ and $c(1,X) = 0$, so ensuring that the final two terms in Equation (4) are zero. Assuming this to be the case is equivalent to assuming

$$\left(Y^1, Y^0\right) \perp D|X \tag{6}$$

In other words, the joint distribution of the potential outcomes is conditionally (given X) independent of D. This is sometimes called 'the strong ignorability' assumption (Rosenbaum and Rubin 1983).

Heterogeneous treatment effects arise whenever the treatment effect differs according to the values of a third variable. There can be many such variables, but in this article we concentrate on one specific kind of effect heterogeneity. In observational data, unlike in controlled trials where subjects are randomly assigned to treatment, causal effect heterogeneity may arise from the mechanisms of selection into treatment. In some cases, gatekeepers for a particular program (college admissions officers, for example) will select participants on the grounds that they are the ones who will benefit most; in other cases people choose the treatment they think will be most beneficial for them. These considerations point to heterogeneity of causal effects according to the likelihood of receiving (through being allocated to, or choosing) treatment, and recently Brand and Xie (2010), Xie, Brand, and Jann (2012), and Brand and Davis (2011) have analyzed heterogeneity in treatment effects across the values of the estimated propensity score. The propensity score is an estimate, for each observation, of the probability that it received treatment, as a function of observed covariates. Thus, investigating heterogeneity according to propensity score values means examining how average treatment effects differ according the estimated probability of receiving treatment.

Classical Selection Bias

In developing our argument we use linear models because they are both tractable and transparent, but the problems we identify are no less likely to plague analyses that use more flexible functional forms (see Carneiro, Heckman, and Vytlacil 2011; Heckman, Urzua, and Vytlacil 2006).

We begin with a model for an outcome, Y (earnings, for example), that depends on the treatment effect of college, D, and a set of predictors, X, including some, but potentially not all, background characteristics of the students entering college. 'College' in some cases means attendance or enrolment, in others graduation, and sometimes it refers to a four-year college, but sometimes it also includes two-year colleges. While these distinctions are substantively crucial they are not relevant to our general argument. D takes the value 1 if someone received the specific college 'treatment,' 0 if they did not. Commonly the untreated group is made up of people who nevertheless graduated high school, so the causal effect in question is the effect of College relative to having a high school diploma. This may itself lead to selection bias in naïve estimators (see Holm and Jæger 2011).

We write the potential outcomes:

$$\begin{aligned} Y_i^1 &= \beta X_i + \delta + \varepsilon_i \\ Y_i^0 &= \beta X_i + \varepsilon_i \end{aligned} \tag{7}$$

and so the observed Y is equal to

$$Y_i = \beta X_i + \delta D_i + \varepsilon_i \tag{8}$$

where ε is a mean zero error and δ is the treatment effect—that is, the causal effect of college on Y. This effect is homogenous and does not vary over individual observations.

Selection into college is given by an equation that specifies the latent propensity to attend college, D^*, as a function of a set of variables, Z. When the error terms from the equation for the outcome (that is, ε) and from the propensity equation (u, shown in Equation (9) below) are assumed to be correlated, identification requires that X be at most a strict subset of Z (that is, there should be at least one variable in Z that is not also in X).

We can write

$$D_i^* = \gamma Z_i + u_i \tag{9}$$

where u is also a mean zero error. People attend college according to:

$$\begin{aligned} D_i^* &= 1 \text{ if } D_i^* > 0 \\ D_i^* &= 0 \text{ if } D_i^* \leq 0 \end{aligned} \tag{10}$$

A more general linear formulation of the potential outcomes is (omitting the i subscript)

$$Y^1 = \beta_1 X + \varepsilon_1$$
$$Y^0 = \beta_0 X + \varepsilon_0$$

Heterogeneity of causal effects according to measured characteristics is captured by the different β vectors, and heterogeneity according to unmeasured characteristics by differences in the error terms, ε. The constraints imposed by our use of Equation (7) (namely $\beta_0 = \beta_1 = \beta$ and $\varepsilon_0 = \varepsilon_1 = \varepsilon$) define a model of homogenous causal effects in which there is only selection on baseline differences. This baseline or pre-treatment bias arises when $cov(\varepsilon, u) = 0$ as a result of one or more factors that are unmeasured (and thus not part of X) and affect both D and Y. The candidate most frequently mentioned in the literature is ability, which is not fully captured by observed proxies such as an IQ measure or test scores. In this case, $cov(\varepsilon, D|X, Z) \neq 0$, and so the expected value of Y for a given value of D will vary depending not only on δ but also ε:

$$E(Y|D = 1, X) - E(Y|D = 0, X) = \delta + [E(\varepsilon|D = 1, X) - E(\varepsilon|D = 0, X)] \tag{11}$$

The final term in square brackets is the bias that comes from using $E(Y|D = 1, X) - E(Y|D = 0, X)$ as an estimate of the causal effect of D on Y. Invoking the assumption that ε and u have a bivariate normal distribution gives us an analytical expression for this bias (Greene 2003:788):

$$E(\varepsilon|D = 1, X) - E(\varepsilon|D = 0, X) = \rho_{\varepsilon,u}\sigma_\varepsilon \left[\frac{\phi\left(-\frac{\gamma Z}{\sigma_u}\right)}{\Phi\left(-\frac{\gamma Z}{\sigma_u}\right)\left(1 - \Phi\left(-\frac{\gamma Z}{\sigma_u}\right)\right)} \right] \tag{12}$$

In Equation (12), $\rho_{\varepsilon,u}$ is the correlation between the errors of the selection and outcome equations, ϕ is the normal pdf and Φ the normal cdf. But $\Phi\left(-\frac{\gamma Z}{\sigma_u}\right)$ is the probability that an individual receives the college treatment: in other words, it is the estimated propensity score, $p(Z)$, and so we can rewrite the bias as:

$$\rho_{\varepsilon,u}\sigma_\varepsilon \frac{\phi\left(\Phi^{-1}(p(Z))\right)}{p(Z)(1 - p(Z))} \tag{13}$$

That estimates of causal effects are confounded with bias if there is unmeasured selection on baseline differences is hardly news; the important point here is that, as Equation (13) shows, the size of the bias varies according to the value of the propensity score. Thus a truly homogenous causal effect can co-occur with heterogeneous bias across observations; that is, the bias may differ across sub-populations in the sample, depending on their probability of receiving treatment.

In passing we note that, in most empirical applications in sociology, the propensity score is used less as a model of selection into treatment and more as a way of trying to balance the treated and untreated on their distributions of possibly confounding covariates. Thus D^* is a function of the X variables (see, for example, Brand and Xie 2010) and so we can write

$$D_i^* = \gamma X_i + u_i$$

where u is, again, a zero mean error.

The propensity score estimator of the causal effect is given by

$$E(Y|D=1,p(X)) - E(Y|D=0,p(X))$$

In words, we condition on values of the propensity score, $p(X)$, and, if strong ignorability holds (defined in this case as $\left(Y^1, Y^0\right) \perp D|p(X)$, which, as Rosenbaum and Rubin (1983) show, is equivalent to $\left(Y^1, Y^0\right) \perp D|X$ under certain mild conditions), this will give an unbiased estimate of the causal effect. But if it does not hold there are two possible sources of bias: baseline bias and differential treatment effect bias. Supposing, as we have been thus far, that there is baseline bias but not differential treatment effect bias, and that the true causal effect is homogenous (and denoted δ), then we have:

$$\begin{aligned} & E(Y|D=1,p(X)) - E(Y|D=0,p(X)) \\ = \; & E(Y^1 - Y^0) + E(Y^0|D=1,p(X)) - E(Y^0|D=0,p(X)) \\ = \; & \delta + \text{baseline bias} \end{aligned}$$

If we write the potential outcomes as:

$$\begin{aligned} Y^0 &= \mu(X) + \varepsilon \\ Y^1 &= \mu(X) + \delta + \varepsilon \end{aligned}$$

where $\mu(X)$ is an unspecified function of X, we can write the baseline bias in the same form as earlier

$$E(\varepsilon|D=1,p(X)) - E(\varepsilon|D=0,p(X))$$

Assuming that ε and u are bivariate normal allows us to express this bias as Equation (13) (replacing Z with X).

Classical Selection Bias and Heterogeneous Causal Effects

The intuition behind what follows is that both classical selection bias, shown in Equation (13), and heterogeneity of causal effects according to the likelihood of receiving treatment are functions of the propensity score, and so it will be impossible to disentangle one from the other without further assumptions.

Assume that the causal effect is truly homogenous. Unless our data come from a randomized control trial, there is likely to be selection bias. Methods such as propensity score analysis, regression controls, or inverse probability of treatment weighting can only take account of selection on observables; if there is also selection on unobservables the result will be a non-zero bias, as given by Equation (13) in the case we are considering.

If we erroneously believe that the causal effect is heterogeneous according to the propensity to receive or select into treatment we could seek to investigate this in various ways but, following Brand and Xie (2010; see also Xie, Brand, and Jann 2012), one approach is the following. Construct strata from the propensity scores and compute the causal effect of D within each stratum as the average difference in Y between those who did and those who did not receive the college treatment. Brand and Xie (2010) found in their study of the return to college completion that, as they moved from strata at the lower end of the propensity score towards the higher end, the estimated causal effects declined. It appeared that those who were less likely to attend college had a larger positive causal effect of college on earnings than those who were more likely. They term this 'negative selection' into college.

Figure 1 plots the bias in the estimate of a homogenous causal effect against the propensity score. The bias, given by Equation (12) or (13), is due to the non-zero correlation between ε and u, which, in this figure, is assumed to be positive. As can be seen, the bias starts high when the propensity is low, declines, reaches a minimum where $p(Z) = 0.5$, then increases again as the propensity increases. If the correlation between ε and u were negative the curve would be inverted; the bias would be negative, reaching its maximum at $p(Z) = 0.5$. The shape of Figure 1 is determined by the expression in square brackets in Equation (13). This is the ratio of the density function, ϕ, to the product of the propensity score times one minus the propensity score. Both numerator and denominator reach their minimum of zero when $p(Z) = 0$ or $p(Z) = 1$ and their maximum at $p(Z) = 0.5$. Because, for $0 < p(Z) < 1$, $\phi > \Phi(1 - \Phi)$, the bias has the same sign as the correlation between ε and u and, because $\Phi(1 - \Phi)$ is more steeply concave than ϕ, the bias declines to a minimum at $p(Z) = 0.5$ and increases thereafter. Figure 1 shows that, in the presence of classical selection bias, the causal effect of D on Y will appear to vary over the propensity score even if it is really homogeneous.

Common Support and Partial Observability

If there were selection bias and a homogenous causal effect, a semi-parametric approach to estimating the heterogeneous causal effect, such as that suggested by Brand and Xie (2010; see also Xie, Brand, and Jann 2012), would lead us to suppose, erroneously, that the causal effect was heterogeneous. For the case considered in the previous section—a linear model and bivariate normality of the errors—the difference in the average of Y between the treated and untreated in each stratum of the propensity score should follow the trend shown in Figure 1. But it is rare to have observations of treated and untreated individuals equally distributed along the whole range of the propensity score. Typically, at the lower end of the range it is difficult to find cases that received treatment, and at the top it is difficult to find cases that did not. This problem of non-overlapping support sometimes leads investigators to use only part of the range of the propensity score. Furthermore, the widths of the strata often need to be adjusted to ensure that treated and untreated

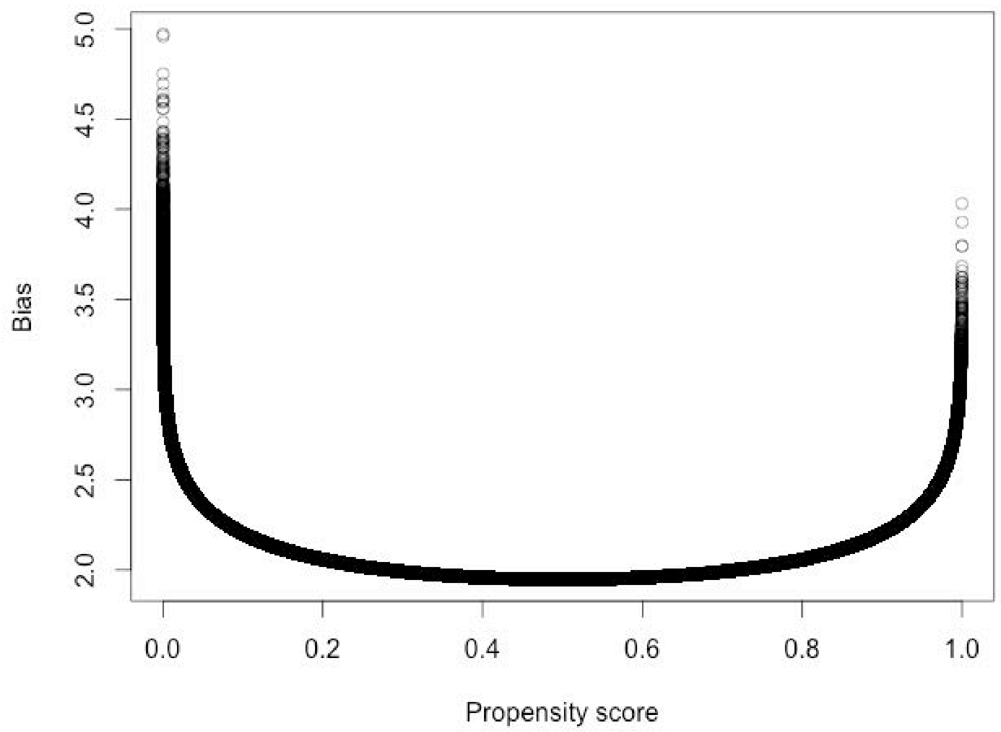

Figure 1: Bias Plotted against Propensity Score

observations in each stratum are balanced (that is, as far as possible, the within-stratum distributions of the X variables should be identical for both groups).

In Brand and Xie's (2010) study, the cut-points for the propensity score strata are 0, 0.05, 0.1, 0.2, 0.4, 0.6, and 1. Thus they have six strata with a finer division of the propensity score at the bottom than at the top (where a single stratum covers the range 0.6 to 1). Figure 2, which follows from Figure 1, shows the average bias (again from Equation [12] or [13]) in each of these strata, plotted at the mid-point of each stratum. If we interpreted this in terms of heterogeneous treatment effects, the estimated causal effects in each stratum would be equal to the bias, as shown in Figure 2, plus a constant (the homogenous treatment effect δ). It can readily be seen that, in Figure 2, there is a declining trend that could easily, but erroneously, be considered evidence of negative selection.

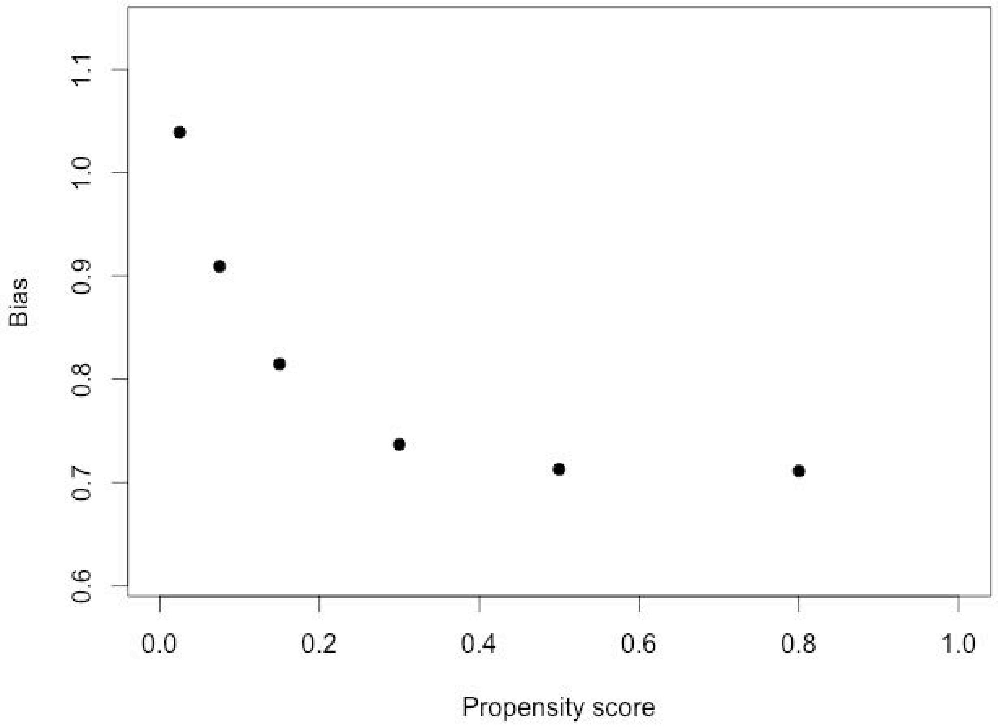

Figure 2: Estimates of Average Within-Stratum Bias

The General Case

If we now relax the assumption that ε and u have a bivariate normal distribution we can write the bias from selection as (Powell 1994):

$$\frac{\int\limits_{-\infty}^{\infty}\int\limits_{-\gamma Z/\sigma_u}^{\infty}\varepsilon f(\varepsilon,u,\theta)\partial u\partial\varepsilon}{\int\limits_{-\infty}^{\infty}\int\limits_{-\gamma Z/\sigma_u}^{\infty}f(\varepsilon,u,\theta)\partial u\partial\varepsilon}-\frac{\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{-\gamma Z/\sigma_u}\varepsilon f(\varepsilon,u,\theta)\partial u\partial\varepsilon}{\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{-\gamma Z/\sigma_u}f(\varepsilon,u,\theta)\partial u\partial\varepsilon} \tag{14}$$

Here θ denotes the parameters of the bivariate distribution of ε and u (the correlation, means, and standard deviations). In the appendix we demonstrate that Equation (14) reduces to

$$\frac{\int\limits_{-\gamma Z/\sigma_u}^{\infty}E(\varepsilon|u,\theta)f(u)du}{1-p(Z)} \tag{15}$$

where $f(u)$ is the distribution of u. We also show in the appendix how, when ε and u have a bivariate normal distribution, Equation (15) can be written as Equation (13).

The derivative of Equation (15) with respect to the propensity score is:

$$\frac{\int\limits_{-\gamma Z/\sigma_u}^{\infty} E(\varepsilon|u,\theta)f(u)du + \left[\frac{\partial\left(\int_{-\gamma Z/\sigma_u}^{\infty} E(\varepsilon|u,\theta)f(u)du\right)}{\partial(p(Z))}\right] \times (1-p(Z))}{(1-p(Z))^2} \tag{16}$$

Without specifying the joint distribution of ε and u it is impossible to make any further progress but, except for degenerate cases, the derivative will always be non-zero and thus the presence of unspecified bias may easily give rise to a mistaken apprehension of heterogeneous causal effects.

Robustness Tests

In Xie, Brand, and Jann's (2012) approach, heterogeneous causal effects are estimated using

$$E(Y|D=1, p(X)=p) - E(Y|D=0, p(X)=p) \tag{17}$$

Here, p denotes a specific value or range of adjacent values of the propensity score, and the causal effect may vary across values of p. The bias in this estimator, for a given p, can be written as

$$c(1, p(X))\Pr(D=0|p(X)=p) - c(0, p(X))\Pr(D=1|p(X)=p) \tag{18}$$

The ignorability assumption invoked by Xie, Brand, and Jann (2012) implies $c(0, p(X)) = 0$ and $c(1, p(X)) = 0$. The former means that controlling for the observed confounders through the propensity score renders the baseline, or non-treatment, potential outcomes mean-independent of the treatment received. The latter means that this also renders the potential treatment outcomes mean-independent of treatment. We leave to one side the question of the effectiveness of the propensity score for this purpose; a more problematic issue is the likely existence of unmeasured factors, U, that help determine both Y^0 and treatment assignment, D, and so invalidate the assumption $c(0, p(X)) = 0$. We considered the consequences of this in the context of a linear model in the first part of the article and showed that, whenever the assumption does not hold, heterogeneous causal effects are confounded by selection bias.

Beyond asking whether the assumption $c(0, p(X)) = 0$ is plausible or not we might try to assess the sensitivity or robustness of estimated heterogeneous causal effects to the assumption. In this part of the article we adopt such a test, drawing on Robins (1999; see also Sharkey and Elwert 2011 and Lawrence and Breen 2014). The underlying idea of Robins's approach is to choose a functional form for $c(d, p(X))$ (where d takes the values 0 or 1) that seeks to capture both the type and magnitude of the bias.

Although the true values of $c(0, p(X))$ and $c(1, p(X))$ are unknown, we can specify functional forms for them and adjust the estimates of the causal effect accordingly in order to examine how robust they are to different types and degrees of bias. The adjustment takes the form of replacing the observed values of Y with Y^*, computed by

$$Y^* = Y - c(D = d, p(X))\mathrm{Pr}(D = 1 - d|p(X) = p) \tag{19}$$

which yields

$$\begin{aligned} Y^* &= Y - c(1, p(X))\mathrm{Pr}(D = 0|p(X) = p) \text{ if } D = 1 \\ Y^* &= Y - c(0, p(X))\mathrm{Pr}(D = 1|p(X) = p) \text{ if } D = 0 \end{aligned} \tag{20}$$

The adjusted causal estimate is then

$$E(Y^*|D = 1, p(X) = p) - E(Y^*|D = 0, p(X) = p) \tag{21}$$

The intuition behind the use of Y^* in Equation (21) is that it subtracts the bias term from the estimator given in Equation (17) (see Brumback et al. 2004:755).

Data and Variables

Following Brand and Xie (2010), we use data from the National Longitudinal Survey of Youth 1979 (NLSY79). The primary purpose of our empirical analysis is to replicate their finding of a negative trend in economic returns to college completion over the propensity score and examine whether it is robust to selection bias. To do this we build our male and female NLSY79 samples following the information provided by Brand and Xie (2010:281–283). All the precollege variables used for estimating propensity scores and all the college and economic outcome variables are constructed as far as possible in the same way as Brand and Xie. We were able to create samples that are very similar to those used in their study in terms of both size and the distributions of the variables. Table 1 shows the summary of our NLSY79 datasets with a comparison to the samples used in the Brand and Xie study.

For pretreatment (precollege) confounders, we include both parents' education, measured in years of schooling; net family income measured in 1978 (in 1979 dollars); growing up in an intact family (living with both parents); number of siblings; living in a rural or metropolitan area; being Jewish; and race. We also include two measures of academic ability: taking college prep courses and the ASVAB test score, which all the sample respondents took in 1980. Like Brand and Xie we use the standardized score, residualized on age separately by race and gender. Whether friends plan to go to college is also included as a proxy for social influence. We define college completion as completing at least four years of college by 1990. Those who had not completed high school by then are excluded from the sample (Brand and Xie 2010:282). For the measure of economic outcome we use logged hourly wages measured in 1994 (aged 29 to 32), in 1998 (aged 33 to 36) and in 2002 (aged 37 to 40). All the respondents who were not employed when wages were measured are excluded from the main analysis (again following Brand and Xie 2010).

Table 1: Summary Statistics of the NLSY79 Samples

	Our Samples				The Brand and Xie Samples			
	Men (N=1274)		Women (N=1211)		Men (N=1265)		Women (N=1209)	
	Non-college Graduates	College Graduates	Non-college Graduates	College Graduates	Non-college Graduates	College Graduates	Non-college Graduates	College Graduates
Race								
Black	0.18	0.09	0.15	0.10	0.18	0.07	0.15	0.07
Hispanic	0.07	0.04	0.08	0.04	0.07	0.03	0.07	0.03
Family Background								
Family income	17987	26136	17972	25684	17870	26538	18174	25991
Mother's education	11.12	13.09	11.06	13.20	11.26	13.32	11.18	13.37
Father's education	11.10	14.05	11.03	13.87	11.23	14.39	11.16	14.14
Intact family	0.76	0.80	0.72	0.84	0.72	0.83	0.67	0.85
Number of siblings	3.36	2.42	3.43	2.51	3.29	2.34	3.40	2.45
Rural residence	0.23	0.22	0.25	0.21	0.25	0.19	0.24	0.21
SMSA residence	0.68	0.72	0.69	0.71	0.77	0.78	0.75	0.80
Jewish	0.01	0.04	0.00	0.03	0.00	0.03	0.00	0.04
Academic ability								
ASVAB score	-0.12	0.63	-0.07	0.58	-0.09	0.69	-0.04	0.64
College prep track	0.20	0.54	0.20	0.47	0.23	0.59	0.23	0.49
Social influence								
Friends' college plan	0.41	0.76	0.45	0.78	0.42	0.79	0.48	0.81
Sample proportion	0.82	0.18	0.83	0.17	0.76	0.24	0.77	0.23

Note: 1. All the mean values of our NLSY79 samples are weighted means adjusted by the 1990 sampling design.
2. College completion indicates whether a respondent completed at least four years of college education.

Table 2: Propensity Score Matching Estimates of Economic Returns to College Education (Homogeneous Effects)

	Economic Return	Standard Error
Men, 1994	$0.212^{\dagger}$	0.059
Men, 1998	$0.323^{\dagger}$	0.066
Men, 2002	$0.472^{\dagger}$	0.085
Women, 1994	$0.289^{\dagger}$	0.056
Women, 1998	$0.193^{\dagger}$	0.059
Women, 2002	$0.391^{\dagger}$	0.129

† $p < .01$

Analysis and Robustness of Negative Selection

We began by estimating the homogenous economic returns to college, using the usual propensity score approach of forming a weighted sum of the differences in outcomes between the treated and untreated across strata of the propensity score. The results are shown in Table 2. Recall that our dependent variable is the logarithm of hourly wages, so the estimates are approximately percentage differences between the earnings received by those who completed college, compared to what they would have earned had they not completed college. The pattern of estimated effects is similar to those reported by Brand and Xie (2010:285, Table 2). Among men, the economic returns to college increase with age, whereas among women there is a curvilinear pattern. But in all cases the returns to college are substantial (under the assumption of strong ignorability).

We estimated the heterogeneous economic returns to college completion that vary over the propensity score under the ignorability assumption using Brand and Xie's method. We examine how the returns vary across the propensity score strata using a hierarchical linear model (see Brand and Xie 2010:280–281 for more details). To perform the analysis we used the Stata program *hte* (Jann, Brand, and Xie 2007).

Figure 3 presents the results from our analysis. We were able to reproduce the negative trend of economic returns to college across propensity score strata that Brand and Xie (2010) found in their study. For the NLSY79 men, the negative slopes are stronger when they are older, but the NLSY79 women show the opposite life course pattern; the tendency for negative selection becomes less pronounced as they get older.[1]

To examine the robustness of the finding of effect heterogeneity to the introduction of selection bias, we apply the method proposed by Robins (Robins 1999; see also Sharkey and Elwert 2011). Since the Brand and Xie analysis relies on the assumption that there is no selection bias but differential treatment effect bias does exist, we model the bias term as

$$\begin{aligned} c(0, p(X)) &= -\alpha \\ c(1, p(X)) &= \alpha \end{aligned}$$

where α is a non-zero positive constant. This corresponds to positive selection bias: the expected potential outcome even without treatment is greater for the treated than for the untreated. Positive selection bias is well established among empirical studies estimating economic returns to education (Card 1999; Manski 2003). In this way we can consider how the claim of negative selection fares against the alternative scenario of a homogeneous treatment effect with positive selection bias.[2]

We adjust the observed value of Y by inserting a value of α into Equation (19) to give the adjusted values:

$$\begin{aligned} Y^* &= Y - \alpha \Pr(D=0|p(X)=p) \text{ if } D=1 \\ Y^* &= Y + \alpha \Pr(D=1|p(X)=p) \text{ if } D=0 \end{aligned} \tag{22}$$

To interpret the results more intuitively, we calibrate α in the same way as Sharkey and Elwert (2011:1975–8). A unit change in α "corresponds to the amount of observed confounding previously eliminated by adjusting for all observed covariates" in the models of propensity score matching with homogeneity. We search for a value of α such that it generates a hypothetical Y (logged wage) which, when used as the outcome variable in a model where observed confounders are not conditioned on, yields an estimate equal to the average treatment effect (for the treated) conditioning on the observed covariates. Given this value, we examine how the results deviate from the original result (where $\alpha = 0$) as the value of α varies (for example, $\alpha/2$, α, 2α, and 3α).[3] With those hypothetical logged wages, we estimate the economic returns to college completion with heterogeneity using the Xie, Brand, and Jann method.

Figure 4 shows graphically, for each age group and for men and women, how the slope of the economic returns to college completion changes with growing positive selection bias. Note that, in our scenario, there is no heterogeneous treatment effect bias (because this is equal to $c(0, p(X)) + c(1, p(X)) = -\alpha + \alpha = 0$). The figure shows that the slopes are very sensitive to the introduction of positive selection bias. In most cases, a weak negative pattern in the slope of the economic returns turns positive even when we allow for only a moderate degree of positive selection bias. When we assume that there is an unobserved factor that affects both college completion and wages as much as the observed covariates do, adjusting for the bias induced by the omitted factor alters all the moderately negative slopes to moderately positive slopes (yellow lines in Figure 4). As the influence of the unobserved confounder increases, the slopes become increasingly positive.[4]

On the other hand, as Table 3 shows, the estimated homogeneous causal effects are, for the most part, very robust to the same test. The sole exception is the estimate for women in 1998. Among men, even if unmeasured selection bias were three times as great as measured bias, the causal effect of college would be substantive and statistically significant. This is also true for women observed in 2002. The 1994 estimates for women are robust up to unmeasured selection bias twice as large as measured bias.

This result from the robustness analysis casts doubt on the degree to which the observed negative slope of economic returns to college across propensity strata is

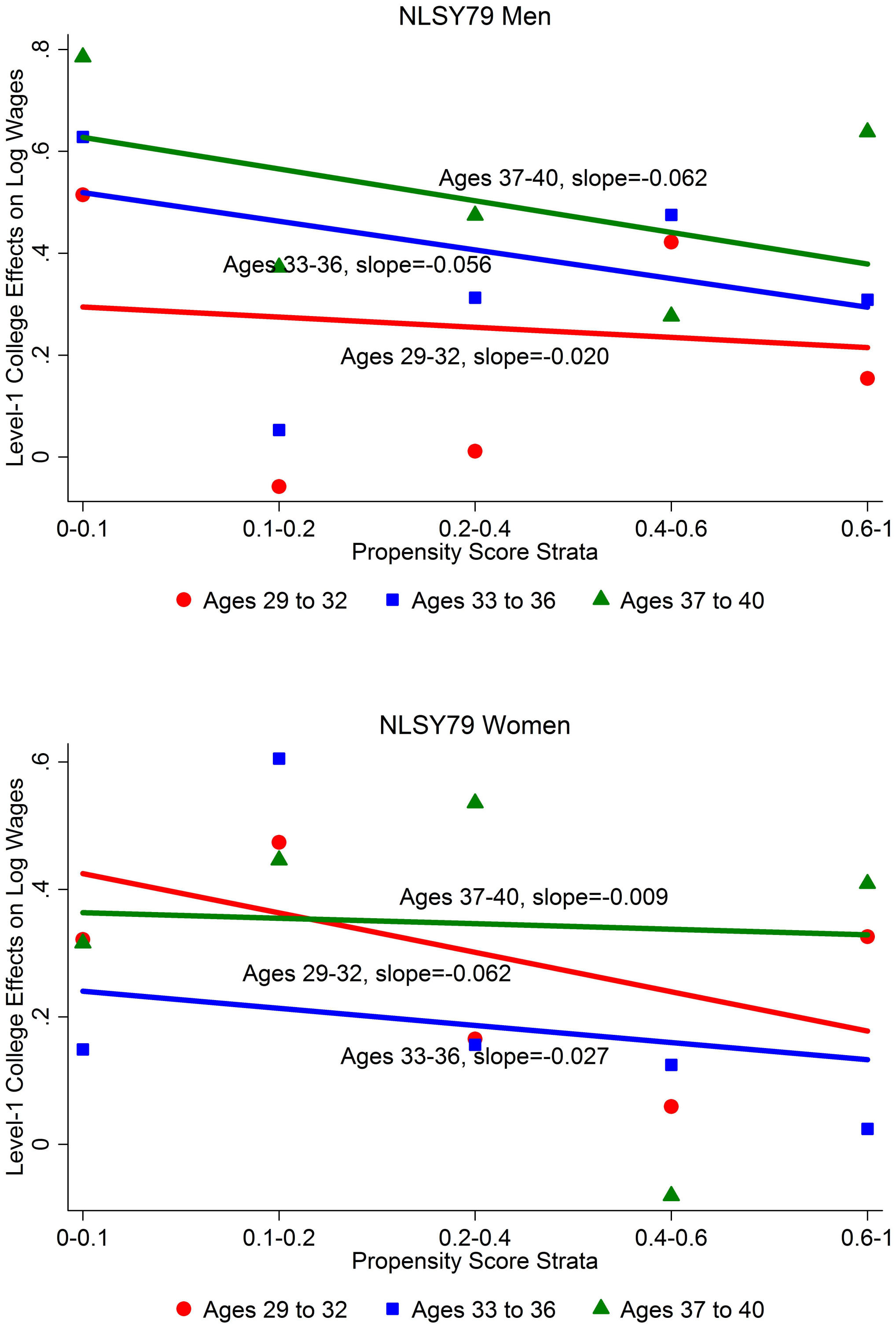

Figure 3: Heterogeneous Economic Returns to College Completion Assuming Ignorability

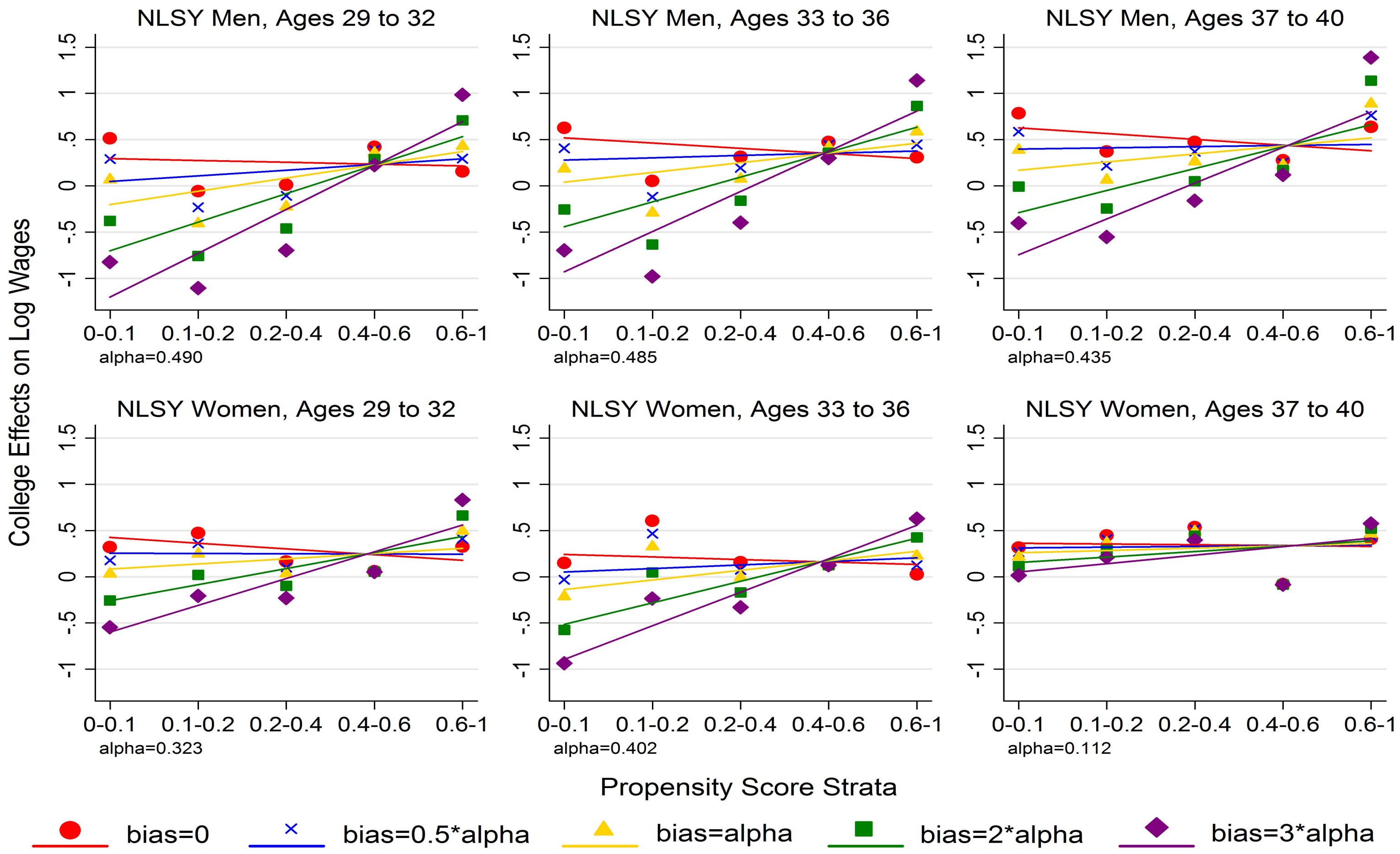

Figure 4: Robustness of Heterogeneous Returns to College Education under the Scenarios of Positive Selection Bias and No Heterogeneous Effect Bias

Table 3: Robustness Analysis of Estimates of Homogenous Causal Effect

	NLSY79 Men			NLSY79 Women		
	1994	1998	2002	1994	1998	2002
	$\alpha = .490$	$\alpha = .485$	$\alpha = .435$	$\alpha = .323$	$\alpha = .402$	$\alpha = .112$
No bias	0.212†	0.323†	0.472†	0.289†	0.193†	0.391†
	(.059)	(0.066)	(0.085)	(0.056)	(0.059)	(0.129)
0.5α	0.202†	0.313†	0.464†	0.257†	0.153†	0.380†
	(0.059)	(0.66)	(0.085)	(0.056)	(0.059)	(0.129)
α	0.192†	0.302†	0.456†	0.226†	0.114	0.368†
	(0.059)	(0.066)	(0.085)	(0.056)	(0.059)	(0.129)
2α	0.173†	0.282†	0.440†	0.162†	0.035	0.346†
	(0.059)	(0.065)	(0.085)	(0.057)	(0.060)	(0.129)
3α	0.154†	0.261†	0.440†	0.98	−0.044	0.323*
	(.059)	(0.065)	(0.085)	(0.057)	(0.061)	(0.129)

$* \ p < .05$; $\dagger \ p < .01$.

driven by the substantive mechanism of negative selection suggested by Brand and Xie (2010:277–280), rather than by selection bias.[5]

Conclusion

We showed that baseline, or selection, bias will vary over values of the propensity score. When using the method that Xie, Brand, and Jann (2012) propose, and Brand and Xie (2010) implement, this bias may very easily be mistaken for causal effects that are heterogeneous according to the probability of receiving treatment. Even in the absence of any true relationship between the propensity to be treated and the magnitude of the treatment effect, we could still find variation in the treatment effect over the propensity score because of uncontrolled selection bias.

There is growing awareness among quantitative sociologists that in the past they may have been too ready to use causal language when it was not justified. But it would be equally unwise for sociologists to adopt a policy of identifying causal effects on the basis of assumptions that are unlikely to hold—ignorability in non-experimental data being a case in point. Assumptions are unavoidable in causal analyses, but as far as possible we should try to determine how robust our causal estimates are to their violation. We have applied such a test in this paper. We believe any causal claim derived from observational data, if it is to be taken seriously, should be shown to be robust using tests like this. In our reanalysis of NLSY79 data we found that our estimates of the average causal effect of college on earnings were very robust. In contrast, the negative selection into college that Brand and Xie (2010) report disappeared as soon as we allowed for even a modest amount of baseline bias.

Notes

[1] Further details of our analysis are available on request. We also ran an analysis using an alternative semi-parametric method proposed by Xie, Brand, and Jann (2012), and the results are consistent with the finding from the linear specification. The results are available from the authors on request.

[2] Modeling the bias as a constant is a very simple approach, but one could equally use more complex functions that would correspond to more subtle kinds of bias. See Brumback et al. 2004 and Lawrence and Breen 2014 for examples.

[3] The value of α for each outcome is shown at the bottom of Figure 4.

[4] It is not difficult to see why this happens, in the light of Equation (22). For $D = 1$, Y^* will be much less than Y when the propensity score is low, whereas for $D = 0$, Y^* will be slightly greater than Y, but not by much (because the probability of receiving treatment will be small). When the propensity is high, however, Y^* will only be slightly smaller than Y for cases where $D = 1$, but for cases where $D = 0$, Y^* will be much greater than Y. Together these adjustments will tend to flatten and then reverse the slope of the estimated treatment effects across the propensity score as it increases.

[5] In further analyses, not reported here but available from the authors on request, we estimated the marginal treatment effect and replicated the finding of Carneiro, Heckman, and Vytlacil (2011) of positive selection into college among the NLSY79 respondents. We used both parametric (control function) and semi-parametric (local instrumental variable) specifications. Most of the estimation results showed that the returns to college are greater for students who are more likely to complete college (the opposite finding to that of Brand and Xie 2010) and, in particular, that those from higher family socioeconomic backgrounds are likely to earn more than those from lower socioeconomic backgrounds. However, with one exception (the results from the parametric model for ages 33 to 36, for both men and women) the positive MTE was not statistically significant.

References

Brand, Jennie E. and Dwight Davis. 2011. "The Impact of College Education on Fertility: Evidence for Heterogeneous Effects." *Demography* 48(3):863–87. http://dx.doi.org/10.1007/s13524-011-0034-3.

Brand, Jennie E. and Yu Xie. 2010. "Who Benefits Most from College? Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education." *American Sociological Review* 75(2):273–302. http://dx.doi.org/10.1177/0003122410363567.

Brumback, Babette A., Miguel A. Hernán, Sebastien J. P. A. Haneuse, and James M. Robins. 2004. "Sensitivity Analyses for Unmeasured Confounding Assuming a Marginal Structural Model for Repeated Measures." *Statistics in Medicine* 23(5):749–67. http://dx.doi.org/10.1002/sim.1657.

Card, David. 1999. "The Causal Effect of Education on Earnings." Pp. 1801–63 in *Handbook of Labor Economics*, vol. 3, part A, edited by Orley C. Ashenfelter and David Card. Amsterdam: Elsevier. http://dx.doi.org/10.1016/s1573-4463(99)03011-4.

Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil. 2011. "Estimating Marginal Returns to Education." *The American Economic Review* 101(6):2754–81. http://dx.doi.org/10.1257/aer.101.6.2754.

Greene, William H. 2003. *Econometric Analysis*. 5th ed. Upper Saddle River, NJ: Prentice Hall.

Heckman, James J., Sergio Urzua, and Edward J. Vytlacil. 2006. “Understanding Instrumental Variables in Models with Essential Heterogeneity.” *Review of Economics and Statistics* 88(3):389–432. http://dx.doi.org/10.1257/aer.101.6.2754.

Heckman, James and Salvador Navarro-Lozano. 2004. “Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models.” *Review of Economics and Statistics* 86(1):30–57. http://dx.doi.org/10.1162/003465304323023660.

Holm, Anders, and Mads Jæger. 2011. “Dealing with Selection Bias in Educational Transition Models: The bivariate Probit Model." *Research in Social Stratification and Mobility* 29(3):239–250. http://dx.doi.org/10.1016/j.rssm.2011.02.002.

Jann, Ben, Jennie E. Brand, and Yu Xie. 2007. HTE: Stata Module to Perform Heterogeneous Treatment Effect Analysis. http://ideas.repec.org/c/boc/bocode/s457129.html.

Lawrence, Matthew and Richard Breen. 2014. “And Their Children After Them? The Effect of College on Educational Reproduction.” Unpublished paper.

Manski, Charles F. 2003. *Partial Identification of Probability Distributions*. New York: Springer.

Powell, James L. 1994. "Estimation of Semi-parametric Models." Pp. 2443–2521 in Engle, R. F. and D. L. McFadden (eds.) *Handbook of Econometrics IV*. Elsevier Science.

Rosenbaum, Paul R. and Donald R. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41–53. http://dx.doi.org/10.1093/biomet/70.1.41.

Robins, James M. 1999. “Association, Causation, and Marginal Structural Models.” *Synthese* 121(1–2):151–79. http://dx.doi.org/10.1023/A:1005285815569.

Sharkey, Patrick and Felix Elwert. 2011. “The Legacy of Disadvantage: Multigenerational Neighborhood Effects on Cognitive Ability.” *American Journal of Sociology* 116(6):1934–81. http://dx.doi.org/10.1086/660009.

Xie, Yu, Jennie E. Brand, and Ben Jann. 2012. “Estimating Heterogeneous Treatment Effects with Observational Data.” *Sociological Methodology* 42(1):314–47. http://dx.doi.org/10.1177/0081175012452652.

Richard Breen: Department of Sociology, Yale University. E-mail: richard.breen@yale.edu.

Seongsoo Choi: Department of Sociology, Yale University. E-mail: seongsoo.choi@yale.edu.

Anders Holm: Department of Sociology, University of Copenhagen. E-mail: ah@soc.ku.dk.