Information Knowledge Systems Management 10 (2011) 279–289
DOI 10.3233/IKS-2012-0197
IOS Press

Chapter 15

Causal inference and heterogeneity bias in social science

Yu Xie[*]

Abstract: Because of population heterogeneity, causal inference with observational data in social science may suffer from two possible sources of bias: (1) bias in unobserved pretreatment factors affecting the outcome even without treatment; and (2)bias due to heterogeneity in treatment effects. Even when we control for observed covariates, these two biases may occur if the classic ignorability assumption is untrue. In cases where the ignorability assumption is true, "composition bias" can occur if treatment propensity is systematically associated with heterogeneous treatment effects.

1. Introduction

Social and behavioral sciences can be considered population sciences in that they try to understand what Neyman called "populations" – "categories of entities satisfying certain definitions but varying in their individual properties" (quoted by Duncan 1984, p. 96). The idea that scientists can fruitfully study categories of entities that are variant from each other, according to Mayr (1982, 2001), originated with Darwin (1859) and was a revolutionary concept. Mayr (1982) called this idea "population thinking," contrasting it with the "typological thinking" that he claimed originated with Plato.

Typological thinking has influenced physical science enormously and is still arguably dominant in determining what can be considered scientific truth. From a typological perspective, the main goal of science should be to discover universally valid, unchanging laws. Thus, scientists should, by eliminating the influences of extraneous, confounding factors, distil their representations of the universe down to abstract but conceptually homogenous relationships. Whether they are constructing thought experiments in developing scientific theories for typical objects or conducting actual experiments to try to verify theories under controlled laboratory conditions, the knowledge that results, according to typological thinking, should be valid anywhere in the universe. Homogeneity is a strong assumption which has worked well in natural science: we need only obtain knowledge about a type of phenomena so as to generalize that knowledge to individual, concrete cases. The typological thinker treats observed variation in the real world as a mere matter of appearances and thus as inconsequential. This ancient philosophical view was supported in the seventeenth century by measurement theorists, who revealed that measurement errors give rise to observed variation and also developed methods of handling these errors (Stigler 1986).

[*]Corresponding author. E-mail: yuxie@umich.edu.

The first person to fundamentally challenge typological thinking was Charles Darwin (1809–1882) (Mayr 1982, 2001). Indeed, the proposition that individual variability is real, not merely apparent, is essential to Darwin's theory of evolution by natural selection.[1] Darwin and his successors saw deviations from population averages not as scientifically trivial but as the very basis of evolution. One of Darwin's successors was Francis Galton (1822–1911), who introduced the principle of variation into social science. Instead of focusing on typical phenomena as typological thinking dictated, Galton focused on "how the quality is distributed" (Galton 1889, pp. 35–36). One modern historian of science described Galton as a scientist to whom "individual differences . . . were almost the only thing of interest" (Hilts 1973, p. 221).

Population thinking, as pioneered by Darwin and Galton, soon gave birth to a new kind of science known as population science. The population scientist does not assume that all concrete units in a population are basically the same, i.e., homogeneous, but instead recognizes that units of analysis in a population are different from one another, i.e., heterogeneous. One might say that most social science disciplines – economics, demography, psychology, sociology, political science, and anthropology – are population sciences, since they cannot afford to ignore individual-level variation. The acceptance of individual-level heterogeneity in social science has important consequences for our research practices. In this paper, I will show how certain biases for causal inference in social science may potentially result from population heterogeneity.

When we begin to take individual-level variability seriously and to treat it as a reality in population sciences rather than imperfection or measurement error, we can no longer rely on a scientific method that has always served physical science well, namely the laboratory experiment. If homogeneity cannot be maintained, how can we know, even at the level of individual units of analysis, if differences in outcomes in units subjected to different experimental conditions are caused by experimental treatment or intrinsic individual-level differences (Holland 1986)? All population scientists can do is to conduct field experiments, in which units of analysis are randomized into experimental conditions (Fisher 1926; Neyman 1923).

Thus, in all population sciences, it is no longer possible to assume that a category of homogeneous entities exists, or that relationships between particular causes and effects are homogeneous. Statistical methods provide the population scientist with an alternative strategy in accounting for this intrinsic and inevitable variability (Xie 2007). Although we cannot understand how a given experimental condition might affect every unit in a population, we can assess its average consequence by means of field experiments and statistical analyses (Fisher 1926; Heckman 2005; Holland 1986; Manski 1995; Rubin 1974).

2. Heterogeneity and possible biases in causal inference

Numerous scholars who study causal inference in social science have previously recognized the importance of population heterogeneity and considered its implications for potential biases (e.g., Heckman 2005; Holland 1986; Manski 1995; Morgan and Winship 2007; Rubin 1974; Tsai and Xie 2011; Winship and Morgan 1999; Xie, Brand, and Jann 2012). In this section, I will consider why population heterogeneity may lead to biases in causal inference.

Let us assume that a population, U, is being studied. Let Y denote an outcome variable of interest that is a real-valued function for each member of U, and let D denote a dichotomous treatment variable (with

[1] Chapters one and two of Darwin's *On the Origins of Species* (1859) are entitled "Variation under Nature" and "Variation under Domestication."

its realized value being *d*) with $D = 1$ if a member is treated and $D = 0$ if a member is not treated. For clarity, let subscript *i* represent the i^{th} member in *U*. We further denote y_i^1 as the i^{th} member's potential outcome if treated (i.e., when $d_i = 1$), and y_i^0 as the i^{th} member's potential outcome if untreated (i.e., when $d_i = 0$). Since population heterogeneity is ever present, let us conceptualize a treatment effect as the difference in potential outcomes associated with different treatment states for the *same* member in *U*:

$$\delta_i = y_i^1 - y_i^0, \tag{1}$$

where δ_i represents the hypothetical treatment effect for the i^{th} member.[2] The fundamental problem of causal inference (Holland 1986) is that, for a given unit *i*, we observe either y_i^1 (if $d_i = 1$) or y_i^0 (if $d_i = 0$), but not both. In light of this fundamental problem, how can we estimate treatment effects? Holland presents two possible solutions: the "scientific solution" and the "statistical solution."

The scientific solution, which is based on typological thinking, assumes that all members in *U* are exactly the same, i.e., homogenous: $y_i^1 = y_j^1$, and $y_i^0 = y_j^0$, where $j \neq i$ is a different member in *U*. This strong assumption would allow the researcher to identify individual-level treatment effects. In fact, if the strong assumption can be maintained and there is no measurement error, one would need only two cases in *U* (say *i* and *j* with different treatment conditions) to reveal treatments effects for all members in the entire population, for the following would hold true:

$$\delta = y_i^1 - y_i^0 = y_j^1 - y_j^0 = y_i^1 - y_j^0, \tag{2}$$

for any $j \neq i$, where we can drop the subscript of δ as it does not vary across members in the population. As previously stated, however, in the social sciences, which are inherently population sciences, heterogeneity is the rule rather than the exception. In general, then, the formula under the strong homogeneity assumption Eq. (2) is of no practical value in social science.

For any population science, the ubiquity of population heterogeneity makes the statistical solution a necessity. One limitation of the statistical approach is that we can compute quantities of interest about causal effects only at the group level. For example, let us compare the average difference between a set of members that were randomly selected for treatment and another set of members that were randomly selected for control. Since this quantity is essentially the average treatment effect over the entire population, it is called the Average Treatment Effect (*ATE*):

$$ATE = E(Y^1 - Y^0). \tag{3}$$

Quantities of interest in the statistical approach can also be defined for other groups (or subpopulations), as long as they are well defined. For example, Treatment Effect of the Treated (*TT*) can be defined as average difference in *Y* between treatment and control among those individuals who are actually treated:

$$TT = E(Y^1 - Y^0 | D = 1). \tag{4}$$

Analogously, Treatment Effect of the Untreated (*TUT*) is the average difference by treatment status among those individuals who are not treated:

$$TUT = E(Y^1 - Y^0 | D = 0). \tag{5}$$

[2] This formulation has limitations, as it presumes that fixed future outcomes are associated with different treatment conditions at the time of treatment. Social outcomes are complex and unpredictible. See Dawid (2000) for an approach based on Bayesian decision analysis. Also see Brand and Xie (2007) for a method of averaging future outcomes.

In order to compute quantities of *ATE*, *TT*, and *TUT*, however, we need to invoke assumptions.

For an elaboration of the above statement, let us partition the total population *U* into the subpopulation of the treated U_1 (for which $D = 1$) and the subpopulation of untreated U_0 (for which $D = 0$). We can thus decompose the expectation for the two counterfactual outcomes as follows:

$$E(Y^1) = E(Y^1|D=1)P(D=1) + E(Y^1|D=0)P(D=0) \tag{6}$$

and

$$E(Y^0) = E(Y^0|D=1)P(D=1) + E(Y^0|D=0)P(D=0). \tag{7}$$

Ignoring issues of statistical inference and focusing only on identification, we can estimate from observed data: $E(Y^1|D=1), E(Y^0|D=0), P(D=1)$, and $P(D=0)$. Selection bias arises if:

$$E(Y^1|D=1) \neq E(Y^1|D=0) \neq E(Y^1) \tag{8}$$

and

$$E(Y^0|D=1) \neq E(Y^0|D=0) \neq E(Y^0). \tag{9}$$

Recall that we can only observe either Y^1 or Y^0 for any unit in *U*. Therefore, we can only make inferences about, but cannot directly estimate, a quantity of interest representing causal effect, such as *ATE*. If we use the naive estimator $E(Y^1|D=1) - E(Y^0|D=0)$ for $E(Y^1) - E(Y^0)$, which is *ATE*, what are potential sources of bias? To answer this question, we can further decompose an overall selection bias in the naive estimator as follows. In doing so, we will use the following abbreviated notations:

p = the proportion treated (i.e., the proportion of cases *D* = 1),

q = the proportion untreated (i.e., the proportion of cases *D* = 0),

$$E(Y^1_{D=1}) = E(Y^1|D=1),$$
$$E(Y^0_{D=1}) = E(Y^0|D=1),$$
$$E(Y^1_{D=0}) = E(Y^1|D=0),$$
$$E(Y^0_{D=0}) = E(Y^0|D=0).$$

Using the iterative expectation rule, we can decompose *ATE* as follows:

$$\begin{aligned} ATE &= E(Y^1 - Y^0) \\ &= E(Y^1_{D=1})p + E(Y^1_{D=0})q - E(Y^0_{D=1})p - E(Y^0_{D=0})q \\ &= E(Y^1_{D=1}) - E(Y^1_{D=1})q + E(Y^1_{D=0})q - E(Y^0_{D=1}) + E(Y^0_{D=1})q - E(Y^0_{D=0})q \\ &= E(Y^1_{D=1}) - E(Y^0_{D=0}) - [E(Y^0_{D=1}) - E(Y^0_{D=0})] - (TT - TUT)q, \end{aligned} \tag{10}$$

where, as previously defined in Eqs (4) and (5), *TT* is the average Treatment Effect of the Treated, and *TUT* is the average Treatment Effect of the Untreated:

$$TT = E(Y^1_{D=1} - Y^0_{D=1}),$$
$$TUT = E(Y^1_{D=0} - Y^0_{D=0}).$$

Thus, we can see from Eq. (10) that if we use the naive estimator from observed data $E(Y^1_{D=1}) - E(Y^0_{D=0})$ for *ATE*, there are two possible sources of bias:

(1) The average difference between the two groups in outcomes if neither group receives the treatment: $E(Y^0_{D=1}) - E(Y^0_{D=0})$, which we will call this the "pre-treatment heterogeneity bias," or "Type I selection bias."

(2) The difference in the average treatment effect between the two groups $(TT - TUT)$, weighted by the proportion untreated q. The weight of q results from our choice to define pre-treatment heterogeneity bias for the untreated state. We call this the "treatment-effect heterogeneity bias," or "Type II selection bias."

These two sources of bias may exist, because subjects may be sorted into treatment or control groups by either their differences in the base-line level (i.e., Type I selection bias) or their differences in the effect of treatment (i.e., Type II selection bias). Let me now reiterate that the treatment-effect heterogeneity bias or Type II selection bias is the situation in which we encounter the following:

$$TT \neq TUT;$$

$$ATE \neq TT;$$

$$ATE \neq TUT.$$

In particular, when $TT - TUT > 0$, there is a sorting gain so that the average treatment effect for the treated is greater than the average treatment effect of the untreated. Conversely, if $TT - TUT < 0$, there is a sorting loss.

3. The ignorability assumption and its implications

In the last section, I established the difficulty, perhaps the impossibility, of drawing causal inference in social science. Given this difficulty, how can social science researchers study causal effects? There are two possible solutions to this problem: the experimental solution and the observational solution.

The experimental solution uses random assignment to get rid of both sources of selection bias that we looked at earlier. Random assignment means that a unit in *U* receives either the treatment or control condition by chance only. Let $\perp\!\!\!\perp$ denote independence. Random assignment ensures:

$$(Y^1, Y^0) \perp\!\!\!\perp D, \tag{11}$$

so that

$$E(Y^1_{D=1}) = E(Y^1_{D=0}) = E(Y^1) \tag{12}$$

and

$$E(Y^0_{D=1}) = E(Y^0_{D=0}) = E(Y^0). \tag{13}$$

Under these conditions, we can easily compute *ATE*, *TT*, and *TUT* as:

$$ATE = TT = TUT = E(Y^1_{D=1}) - E(Y^0_{D=0}).$$

In social science research, experimental studies are uncommon. Even when subjects are randomly assigned to experimental conditions, their compliance may not be random. In such cases, the actual treatment condition may not be truly independent with respect to potential outcomes, as required in

Eq. (11). In other situations, often called "natural experiments," we can assume that some factors that affect treatment conditions may be random and extraneous, even though treatment conditions may not be independent with respect to potential outcomes. In both types of situations, we can take a general approach called "instrumental variable (IV) estimation." For a variable to qualify as instrumental, it must meet the exclusion restriction assumption: it affects the likelihood of treatment condition (*D*) but affects the substantive outcome variable (*Y*) *only* indirectly via the treatment status (*D*). For example, draft lottery may be associated with military enlistment but should affect economic outcomes only indirectly through military enlistment (Angrist 1990).

A large literature has grown out of the application of IVs in causal inference (Angrist, Imbens, and Rubin 1996; Angrist and Pischke 2009; Heckman, Urzua, and Vytlacil 2006). Unfortunately, true and strong IVs are difficult to find in practice. Furthermore, even when true experiments are successfully carried out, or good IVs are found, they are typically based on a particular subpopulation at a particular location or time, say students at a certain college, or applicants to a certain federally funded program. In addition, because of population heterogeneity, it is problematic to generalize findings from such studies based on narrowed-defined subpopulations (Manski and Garfinkel 1992). Thus, despite its methodological appeal and growing popularity, the experimental approach does not, in actuality, provide an adequate solution.

When random assignment is not feasible, and no suitable IV is available, the researcher may turn to the second approach, observational solution. The main idea in this approach is to collect rich data measuring population heterogeneity, called covariates, that pertain to potential systematic differences between the treatment and control groups in either the baseline level or the treatment effect. Since only covariates that affect both the treatment assignment and the outcome have the potential to bias the observed relationship between treatment and outcome (Rubin 1997), the researcher assumes that he/she can adequately control for all covariates that simultaneously affect the treatment assignment and the outcome. Once covariates have been controlled, the hope is that treatment status will then be independent of potential outcomes. This conditional independence assumption is called "ignorability," "unconfoundedness" or "selection on observables." Let *X* denote a vector of observed covariates. The ignorability assumption states:

$$(Y^1, Y^0) \perp\!\!\!\perp D|X. \tag{14}$$

Comparing Eqs (11) and (14) highlights the crucial role of covariates $\boldsymbol{X}$. Note that the ignorability condition is always an unverifiable assumption. Although it is written as a statistical property in Eq. (14), whether the assumption is plausible or not is actually a substantive subject matter, since much depends on what covariates are included. In any case, the researcher can tentatively consider the ignorability assumption and then assess its plausibility in a concrete setting through sensitivity or auxiliary analyses (Cornfield et al. 1959; DiPrete and Gangl 2004; Harding 2003; Rosenbaum 2002).

Conditioning on $\boldsymbol{X}$ can be difficult in applied research due to the "curse of dimensionality." However, Rosenbaum and Rubin's (1983, 1984) important work has shown that, under the ignorability assumption, it is sufficient to condition on the propensity score as a function of $\boldsymbol{X}$. Let $P(D = 1|\boldsymbol{X})$ denote the propensity score of treatment given $\boldsymbol{X}$. Rosenbaum and Rubin essentially changed Eq. (14) to:

$$(Y^1, Y^0) \perp\!\!\!\perp D|P(D = 1|X). \tag{15}$$

That is, it is sufficient to condition on the propensity score $P(D = 1|\boldsymbol{X})$. In actuality, the propensity score is unknown and can be estimated from observed data, for example through a logit model or a probit model. In the current literature on causal inference using observational data, almost all methods are based on the propensity score (e.g., Dehejia and Wahba 1999; Morgan and Harding 2006; Xie, Brand, and Jann 2012).

The main function of the propensity score is to balance out the distribution of observed covariates $\boldsymbol{X}$ between the treatment group and the control group (within a given level of the propensity score). For this purpose, the absolute level of the propensity score does not matter. What matters is the *relative* magnitudes of propensity scores associated with different values of covariates $\boldsymbol{X}$.

The result of Eq. (15) states that, under ignorability, treatment is independent of potential outcomes conditional on the propensity score. That is to say, there is no bias after controlling for the propensity score. In light of our earlier discussion stating that bias can manifest in two types, this amounts to two "no-bias" conditions:

(1) There is no pre-treatment heterogeneity bias, or Type I selection bias, conditional on $p(\boldsymbol{X})$. In reference to Eq. (10), this means

$$E[Y^0_{d=1}|p(\boldsymbol{X})] = E[Y^0_{d=0}|p(\boldsymbol{X})]. \tag{16}$$

(2) There is no treatment-effect heterogeneity bias, or Type II selection bias, conditional on $p\ (\boldsymbol{X})$. In reference to Eq. (10), this means

$$E[Y^1_{d=1} - Y^0_{d=1}|p(\boldsymbol{X})] = E[Y^1_{d=0} - Y^0_{d=0}|p(\boldsymbol{X})]. \tag{17}$$

Given Eqs (16) and (17), the researcher can apply the naive estimator

$$E(Y^1_{d=1}) - E(Y^0_{d=0})$$

conditional on the propensity score, because there is no selection bias conditional on the propensity score. In other words, if the ignorability assumption is true, we can assume away both sources of bias, or systematic differences between treated units and untreated units, at the same level of the propensity score. More precisely, we have

$$\begin{aligned} E[Y^1 - Y^0|p(\boldsymbol{X})] &= E[Y^1_{D=1} - Y^0_{D=1}|p(\boldsymbol{X})] = E[Y^1_{D=0} - Y^0_{D=0}|p(\boldsymbol{X})] \\ &= E[Y^1_{d=1}|p(\boldsymbol{X})] - E[Y^0_{d=0}|p(\boldsymbol{X})]. \end{aligned} \tag{18}$$

Of course, unconditional comparisons of the treatment group and the control group, such as *ATE*, *TT*, and *TUT*, involve aggregation of conditional comparisons over the actual distribution of the propensity score.

4. Composition bias

The methodological literature on causal inference can be divided into two groups, based on whether or not the ignorability assumption is adopted. When it is not, the researcher is concerned with residual selection bias conditional on the propensity score that is attributable to unobservable variables. However, even when the ignorability assumption is true, there can be "composition bias" if treatment propensity is systematically associated with heterogeneous treatment effects (Xie 2011).

"Composition bias" results from a dynamic process of recruitment of units into treatment. Recall that recruitment of units into treatment is always selective. This is acknowledged even by the classic ignorability assumption. A well-known property of a dynamic survival process is that the extent of selectivity changes as the proportion surviving changes (Vaupel and Yashin 1985). As a result, the compositions of both the group that is treated and the group that is untreated change constantly. Imagine that as the proportion of treated units in *U, p,* increases from p_1 to p_2, the resulting changes in the composition of U_1 and U_0 give rise to biases in aggregate measures of treatment effects, such as *TT* and *TUT*.

In a simulation analysis (Xie 2011), I demonstrate how composition bias occurs even when ignorability is satisfied. In a situation in which there is a strong positive association between propensity of treatment and treatment effect, I showed that, as the proportion being treated increases, both *TT* and *TUT* decrease. However, because *TUT* decreases at a faster rate than TT, the amount of the sorting gain bias, *TT–TUT*, increases. The last finding – that the sorting gain bias increases as the proportion being treated increases toward 1 – is surprising.

I should emphasize that the composition bias that I discuss here is different from Type II selection bias, because the former is compatible with ignorability but the latter violates ignorability. To investigate Type II selection bias when ignorability is not true, the researchers may resort to methods based on Marginal Treatment Effect (*MTE*), developed by Heckman and his associates (Björklund and Moffitt 1987; Carneiro, Heckman, and Vytlacil Forthcoming; Heckman, Urzua, and Vytlacil 2006). *MTE* is the expected treatment effect at the marginal point at which a latent factor determining a unit's treatment status is neutral – i.e., does not favor either treatment or control. Zhou and Xie (2011) compares propensity-score based methods to *MTE*-based methods.

MTE is closely related to the IV approach, as it can be conceptualized as the average treatment effect for a small segment of units whose propensity of treatment is altered by an IV. It is this change in propensity that shifts the proportion being treated. The work of Heckman, Urzua, and Vytlacil (2006) shows that it is possible to derive various summary quantities of interest, such as *ATE*, *TT*, and *TUT*, from individual-level *MTE*, using appropriate weights. Thus, it is possible to study the sorting gain (or loss), the difference between *TT* and *TUT*, using *MTE*-based methods.

5. Discussion and conclusion

The ubiquity of population heterogeneity in social phenomena makes it impossible to draw causal inferences at the individual level. Instead, the best that can be achieved in any social science is inference at the group level. However, focusing on group comparison also means inattention to individual heterogeneity, resulting in comparisons essentially assuming relatively homogeneous groups. This is a fundamental dilemma facing all researchers in social science.

Many possible methods of formulating comparison groups can be used in actual research settings. Besides the usual treatment-versus-control group comparison, one useful tool meriting research attention is the propensity score, which summarizes information in a multi-dimensional space from multivariate covariates into a univariate variable. Thus, one potential source of heterogeneity that should receive particular attention in causal inference is the interaction between the treatment effect and the propensity score (Xie, Brand, and Jann 2012). Such interactions can be detected without any new requirement, as this can be done under the assumption of ignorability. When such interactions are found, however, the interpretation of the results may differ. If the researcher believes that ignorability is true, the estimated effect of heterogeneity may be generalized. Alternatively, the researcher may interpret the heterogeneous pattern in the estimated effects as an indication that the selection process into treatment may be selective, driven by unobserved factors (Xie and Wu 2005; Zhou and Xie 2011).

In this paper, I have demonstrated two forms of selection bias when the ignorability assumption is violated. The first is bias in unobserved pretreatment factors affecting the outcome even in the absence of treatment. The second is bias due to heterogeneity in treatment effects. Even when the ignorability assumption is true, there can be "composition bias" if treatment propensity is systematically associated with heterogeneous treatment effects. Composition bias arises when the exposure population, the population at risk for being selected into treatment, is changed dynamically by the selection of units into the treatment group. As the treatment proportion expands, the degree of over-presentation of units with high intrinsic propensities among the newly recruited into treatment declines. This results in a shift in composition among newly recruited increments away from high propensity toward low propensity recruits, thus altering average treatment effects at the group levels.

In conclusion, I wish to warn researchers hoping to draw causal inferences in social science, particularly when using observational data, of several potential sources of bias caused by population heterogeneity. Casual inference is an ideal and sometimes ultimate goal in any science, including social science. However, an essential characteristic of social phenomena – population heterogeneity – makes the task of causal inference in social science extremely difficult, if not insurmountable.

Acknowledgements

Financial support for this research was provided by the National Institute of Health, Grant 1 R21 NR010856-01. I am grateful to my collaborators, Jennie Brand, Ben Jann, Shu-Ling Tsai, and Xiang Zhou, for their contributions to related work, from which this paper is a spin-off. I am grateful to Cindy Glovinsky, Debra Hevenstone, Tony Perez, and Xiang Zhou for their valuable research assistance.

References

Angrist, J.D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80: 313-35.

Angrist, J. D., G.W. Imbens, and D.B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434): 444-55.

Angrist, Joshua D. and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Björklund, A. and R. Moffitt. 1987. "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models." *Review of Economics and Statistics* 69: 42-49.

Brand, Jennie and Yu Xie. 2007. "Identification and Estimation of Causal Effects with Time-Varying Treatments and Time-Varying Outcomes." *Sociological Methodology* 37: 393-434.

Carneiro, Pedro, James J. Heckman, and Edward Vytlacil. Forthcoming. "Estimating Marginal Returns to Education." *American Economic Review*.

Cornfield, J., W. Haenszel, E.C. Hammond, A.M. Lilienfeld, M.B. Shimkin, and E.L. Wynder. 1959. "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions." *Journal of the National Cancer Institute* 22: 173-203.

Dawid, A.P. 2000. "Causal Inference Without Counterfactuals." *Journal of American Statistical Association* 95:407-24.

Darwin, Charles. 1859. *On the Origin of Species by Means of Natural Selection or the Preservation of Favored Races in the Struggle for Life.* London: Murray.

Dehejia, R. H. and S. Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of American Statistical Association* 94: 1053-62.

DiPrete, Thomas and Markus Gangl. 2004. "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments." *Sociological Methodology* 34: 271-310.

Duncan, Otis Dudley. 1984. *Notes on Social Measurement, Historical and Critical*. New York: Russell Sage Foundation.

Fisher, R.A. 1926. The Arrangement of Field Experiments. *Journal of the Ministry of Agriculture* 33: 503-13.

Galton, Francis. 1889. *Natural Inheritance.* London, Macmillan.

Griliches, Zvi. 1977. “Estimating the Returns to Schooling: Some Econometric Problems.” *Econometrica* 45: 1-22.

Harding, David J. 2003. “Counterfactual Models of Neighborhood Effects: The Effect of Neighborhood Poverty on High School Dropout and Teenage Pregnancy.” *American Journal of Sociology* 109(3): 676-719.

Heckman, James J. 2005. “The Scientific Model of Causality.” *Sociological Methodology* 35: 1-98.

Heckman, James, Sergio Urzua, and Edward Vytlacil. 2006. “Understanding Instrumental Variables in Models with Essential Heterogeneity.” *The Review of Economics and Statistics* 88: 389-432.

Hilts, Victor. 1973. “Statistics and Social Science.” pp. 206-233 in *Foundations of Scientific Method, the Nineteenth Century*, edited by R. N. Giere and R. S. Westfall. Bloomington: Indiana University Press.

Holland, Paul W. 1986. “Statistics and Causal Inference” (with discussion). *Journal of American Statistical Association* 81:945-70.

Manski, Charles. 1995. *Identification Problems in the Social Sciences.* Boston, MA: Harvard University Press.

Manski, C.F., and Garfinkel, I. 1992. “Introduction.” pp. 1-21 in Evaluating Welfare and Training Programs, edited by C. F. Manski and I. Garfinkel. Cambridge, MA: Harvard University Press.

Mayr, Ernst. 1982. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. Cambridge, MA: Harvard University Press.

Mayr, Ernst. 2001. “The Philosophical Foundations of Darwinism.” *Proceedings of the American Philosophical Society* 145(4): 488-95.

Morgan, Stephen and David Harding. 2006. “Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice.” *Sociological Methods and Research* 35(1): 3-60.

Morgan, Stephen and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge, UK: Cambridge University Press.

Neyman, J. 1923. “On the Application of Probability Theory to Agricultural Experiments.” Essay on Principles, Section 9 *Statistical Science* 5(4): 465-80.

Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 70: 41-55.

Rosenbaum, Paul R. and Donald B. Rubin. 1984. “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score.” *Journal of the American Statistical Association* 79: 516-24.

Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66: 688-701.

Rubin Donald B. 1997. “Estimating Causal Effects from Large Data Sets Using Propensity Scores.” *Annals of Internal Medicine* 5;127(8 Pt 2): 757-63.

Tsai, Shu-Ling and Yu Xie. 2011. “Heterogeneity in Returns to College Education: Selection Bias in Contemporary Taiwan.” *Social Science Research* 40: 796-810.

Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900.* Cambridge, MA: Harvard University Press.

Vaupel, James, and Anatoli Yashin. 1985. “Heterogeneity’s Ruses: Some Surprising Effects of Selection on Population Dynamics.” *The American Statistician* 39: 176-85.

Winship, Christopher and Stephen L. Morgan. 1999. “The Estimation of Causal Effects From Observational Data.” *Annual Review of Sociology* 25: 659-707.

Xie, Yu. 2007. “Otis Dudley Duncan’s Legacy: the Demographic Approach to Quantitative Reasoning in Social Science.” *Research in Social Stratification and Mobility* 25: 141-56.

Xie, Yu. 2011. “Population Heterogeneity and Causal Inference.” Research Report 11-731, Population Studies Center, University of Michigan. (http://www.psc.isr.umich.edu/pubs/pdf/rr11-731.pdf).

Xie, Yu, Jennie Brand, and Ben Jann. Forthcoming, 2012. “Estimating Heterogeneous Treatment Effects with Observational Data.” *Sociological Methodology*.

Xie, Yu and Xiaogang Wu. 2005. “Market Premium, Social Process, and Statisticism.” *American Sociological Review* 70: 865-70.

Zhou, Xiang, and Yu Xie. 2011. “Propensity-Score-Based Methods versus MTE-Based Methods in Causal Inference.” Unpublished paper. Institute for Social Research, University of Michigan.

Author Bio

Yu Xie is Otis Dudley Duncan Distinguished University Professor of Sociology, Statistics, and Public Policy at the University of Michigan. He is also a Research Professor at the Population Studies Center and Survey Research Center of the Institute for Social Research, and a Faculty Associate at the Center for Chinese Studies. His main areas of interest are social stratification, demography, statistical methods, Chinese studies, and sociology of science. His recently published works include *Women in Science: Career Processes and Outcomes* (Harvard University Press 2003) with Kimberlee Shauman, *Marriage and Cohabitation* (University of Chicago Press 2007) with Arland Thornton and William Axinn, *Statistical Methods for Categorical Data Analysis* (Second edition, Emerald 2008), and *Is American Science in Decline?* (Harvard University Press 2012) with Alexandra Killewald.