Population Studies Center

Research Report

Yu Xie

Population Heterogeneity and Causal Inference

Report 11-731
March 2011

Population Studies Center
University of Michigan
Institute for Social Research

Population Heterogeneity and Causal Inference

Yu Xie

University of Michigan

Population Studies Center Research Report 11-731

March 2011

Paper prepared for a workshop on Understanding and Influencing the Causality of Change in Complex Socio-technical Systems in Gold Coast, Australia (February 1-18, 2011).

Please direct all correspondence to Yu Xie (e-mail: yuxie@umich.edu), Population Studies Center, Institute for Social Research, 426 Thompson Street, University of Michigan, Ann Arbor, MI 48106.

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 and Ben Jann for their contributions to related work, from which this paper is a spin-off. I am grateful to Debra Hevenstone, Tony Perez, and Xiang Zhou for their valuable research assistance.

ABSTRACT

Population heterogeneity is ubiquitous in social science research. The very objective of social science is not to discover abstract and universal laws but to understand population heterogeneity. Due to population heterogeneity, causal inference with observational data in social science is impossible without strong assumptions. There are two potential sources of bias. 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. In this paper, I show how "composition bias" due to population heterogeneity arises when treatment propensity is systematically associated with heterogeneous treatment effects. Of particular interest is the way in which composition bias, a form of selection bias, arises even under the classic assumption of ignorability, as I demonstrate with a simple simulation example.

INTRODUCTION

Two philosophical views have dominated the practice of science. In the classic view, firmly established by Plato and still well represented in physical science today, scientific discoveries consist of abstract knowledge about observed phenomena that essentially share the same properties.[1] An alternative view, first provided by Darwin and now well represented in social science, holds that members of a population are inherently different from each other and should be studied as such. Mayr (1982, 2001) called the first view "typological thinking" and the later view "population thinking."

Typological thinking has had enormous influence on physical science and remains arguably the dominant view of what constitutes scientific truths. According to typological thinking, science should focus on the discovery of universally valid and unchanging laws. Toward this end, scientists should extract abstract but conceptually homogenous relationships in the universe by eliminating the influences of extraneous, confounding factors. They may construct thought experiments in developing scientific theories for typical objects and conduct actual experiments in controlled laboratory conditions while attempting to verify the theories, but the objective is always to obtain knowledge that would be universally valid anywhere in the universe. A strong assumption, which has worked well in natural science, is homogeneity: once we obtain knowledge about a type of phenomena, we can generalize the knowledge to individual, concrete cases. Observed variation in the real world is treated as apparent and thus insignificant. This view was further reinforced in the seventeenth century by measurement theory, which revealed that measurement errors could give rise to such variations and also developed methods of handling measurement errors (Stigler 1986). In social science, Adolphe Quetelet (1796-1874) applied measurement theory to his "social physics," which naively essentialized population averages, in the form of the "average man," as the main objective of social science (Quetelet 1842).

It was Charles Darwin (1809-1882) who first challenged typological thinking in a fundamental way (Mayr 1982, 2001). In fact, the proposition that individual variability is real

---

[1] Plato separated the "world of being" (or the world of Forms) from the "world of becoming" (or the world of things). The "world of being" is where true knowledge resides. The "world of becoming" is what we actually observe in real life.

rather than apparent is essential to Darwin's theory of evolution by natural selection.[2] Deviations from the average in a population were no longer considered scientifically trivial, as they were when using typological thinking, but were seen as the very basis of evolution. The importance of variation was later introduced to social science by Francis Galton (1822-1911). Instead of focusing on typical phenomena as dictated by the paradigm in typological thinking, Galton was concerned with "how the quality is distributed" (Galton 1889, pp.35-36). A historian of science characterizes Galton as someone to whom "Individual differences . . . were almost the only thing of interest" (Hilts 1973, p.221). Departing from Quetelet, Galton was interested in variations in statistical distributions as objectives meriting the attention of science.

Population thinking pioneered by Darwin and Galton led to the emergence of a new kind of science: population science. Here, I wish to borrow Neyman's definition of populations (Duncan 1984):

> Beginning with the nineteenth century, and increasing in the twentieth, science brought about "pluralistic" subjects of study, categories of entities satisfying certain definitions but varying in their individual properties. Technically such categories are called "populations."(p.96)

Note that in a population science, the scientist no longer assumes that all concrete units in a population are essentially the same – or homogeneous. Rather, it is explicitly recognized that units of analysis in a population are different from one another – or heterogeneous. In my view, most social science disciplines, including economics, demography, psychology, sociology, political science, and anthropology, are population sciences in that they cannot afford to discard individual-level variation as a mere nuisance or measurement error by assuming that all units of analysis are essentially the same. The recognition of inherent individual-level heterogeneity has important consequences for research practices. For example, a social scientist always needs to first define the population being studied before conducting a study. Because units of analysis in a population all differ from one another, scientific (or random) sampling is important to ensure replicability across different studies. In this paper, I will illustrate some implications of this heterogeneity for causal inference.

---

[2] The first two chapters of Darwin's book *On the Origins of Species* (1859) are entitled "Variation under Nature" and "Variation under Domestication."

CAUSAL INFERENCE UNDER POPULATION THINKING

The recent literature on causal inference in social science already recognizes the importance of population heterogeneity (e.g., Heckman 2005; Holland 1986; Manski 1995; Rubin 1974; Winship and Morgan 1999). Let me use some notations to illustrate the problem.

Suppose that a population, *U*, is being studied. Let *Y* denote an outcome variable of interest that is a real-valued function for each member in *U*, and let *D* denote a dichotomous treatment variable (with 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). Due to the ever-presence of population heterogeneity, we should conceptualize a treatment effect as the difference in potential outcomes associated with different treatment states for the *same* member in *U*:

$$\delta_i = y_i^1 - y_i^0, \qquad (1)$$

where δ_i represents the hypothetical treatment effect for the *i*th member. 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. Given this fundamental problem, how can we estimate treatment effects? Holland describes two possible solutions: the "scientific solution" and the "statistical solution."

Based on typological thinking, the scientific solution capitalizes on homogeneity in assuming that all members in *U* are exactly the same: $y_i^T = y_j^T$, and $y_i^C = y_j^C$, where $j \neq i$ is a different member in *U*. This strong assumption would allow the researcher to identify individual-level treatment effects. Indeed, if the strong assumption can be maintained and there is no measurement error, one would need no more than 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:

$$\delta = y_i^1 - y_i^0 = y_j^1 - y_j^0 = y_i^1 - y_j^0, \qquad (2)$$

for any $j \neq i$, where we can drop the subscript of δ because it does not vary across members in the population. However, as I discussed earlier, in social sciences, which are population sciences,

heterogeneity is the rule rather than the exception. Thus, in general, the formula under the strong homogeneity assumption (equation 2) is of no practical value in social science.

For a population science, the ubiquity of population heterogeneity dictates the statistical solution as a necessity. One limitation of the statistical approach is that we can compute quantities of interest about causal effects only at the group level. One example is to compare the average difference between a randomly selected set of members in *U* that were treated to another randomly selected set of members that were untreated. Because this quantity is essentially the average treatment effect over the entire population, it is called the Average Treatment Effect (ATE):

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

Quantities of interest in the statistical approach can also be defined for other groups (or sub-populations), as long as they are well defined. For example, Treatment Effect of the Treated (*TT*) refers to the average difference by treatment status among those individuals who are actually treated:

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

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

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

However, in order to compute quantities of *ATE*, *TT*, and *TUT*, it is necessary to invoke assumptions so that population heterogeneity would not cause selection biases to our estimates of such quantities from observational data.

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:

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

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

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:

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

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

Recall that we can only observed either Y^1 or Y^0 for any unit in U. Thus, 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. Let us now 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:

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

where, as previously defined in equations (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 observe from equation (10) that if we use this naive estimator from observed data $E(Y^1_{D=1}) - E(Y^0_{D=0})$ for ATE, there are two potential sources of bias:

(1) The average difference between the two groups in outcomes if neither group receives the treatment: $E(Y^0_{D=1}) - E(Y^0_{D=0})$. We 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 either by their differences in the base-line level (i.e., Type I selection bias) or by 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.

RANDOM ASSIGNMENT, IGNORABILITY, AND PROPENSITY SCORE

In the previous section, I have established that the difficulty of drawing causal inferences in social sciences is rooted in two fundamental causes: (1) units of analysis are all heterogeneous, and (2) for any given unit of analysis, we observe the outcome associated with only one, actually realized, treatment condition. Given the combination of these two unavoidable difficulties, how can social science researchers study causal effects? There are two solutions: the experimental solution and the observational solution.

The experimental solution relies on random assignment to eliminate both sources of selection bias that we discussed 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 rare. Even when assignment into experimental conditions is random, subjects' compliance with assignments may not be random. In such cases, the true treatment condition may not be truly independent with respect to potential outcomes, as required in equation (11). In other situations, often called "natural experiments," it may be assumed that factors that affect treatment conditions may be random and extraneous, although treatment conditions may not be independent with respect to potential outcomes. In both types of situations, we have a general approach called "instrumental variable (IV) estimation." For a variable to qualify as an IV, 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 not affect economic outcomes directly (Angrist 1990).

A large literature has been developed in 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. In addition, even when true experiments are successfully carried out, or good IVs are found, they are typically based on a specific subpopulation at a specific location or time, say students at a particular college, or applicants to a particular federally funded program. Population heterogeneity also makes it problematic to generalize findings from such studies based on narrowed-defined subpopulations to the general population at large (Manski and Garfinkel 1992). For these reasons, despite its methodological appeal and growing popularity, the experimental approach does not provide a satisfactory solution in practice.

When random assignment is infeasible, and a suitable IV is unavailable, the researcher may resort to the second approach: observational solution. The basic idea 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. Because only covariates that affect both the treatment assignment and the outcome can cause biases to the observed relationship between treatment and outcome (Rubin 1997), the researcher hopes that he/she can adequately control for all covariates that simultaneously affect the treatment assignment and the outcome. After the control of the covariates, treatment status is

independent of potential outcomes. This conditional independence assumption is called "ignorability," "unconfoundedness" or "selection on observables." Let $\boldsymbol{X}$ denote a vector of observed covariates. The ignorability assumption states:

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

Comparison of equations (11) and (14) highlight the crucial role of covariates $\boldsymbol{X}$. Note that the ignorability condition is always an unverifiable assumption. While it is written as a statistical property in equation (14), whether the assumption is plausible or not is a substantive subject matter, as much depends on what covariates are included. In any event, 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, the important work of Rosenbaum and Rubin (1983, 1984) reveals that, under the ignorability assumption, it is sufficient to condition on the propensity score as a function of $\boldsymbol{X}$. Let $P(D = 1|X)$ denote the propensity score of treatment given $\boldsymbol{X}$. Rosenbaum and Rubin essentially changed equation (14) to:

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

That is, it is sufficient to condition on the propensity score $P(D = 1|\boldsymbol{X})$. In practice, the propensity score is unknown and can be estimated from observed data, say 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 2011).

It is important to realize that 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 $\boldsymbol{X}$. This is the justification for the common practice of using response-

based sampling strategy, i.e., combining a dataset for treated units with another dataset for untreated units, in constructing propensity scores in the literature.[3]

The result of equation (15) states that, under ignorability, treatment condition is independent of potential outcomes. In other words, given a level of the propensity score, there is no bias. Given our earlier discussion stating that bias can manifest in two types, this is tantamount 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 equation (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 equation (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 equation (17), the researcher can apply the naïve 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. That is, 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 propensity score. More precisely, we have

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

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

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

---

[3] Response-based sampling is also called "case-control studies." In logit regressions, this means that the intercept can be ignored (Breslow 1996; Xie and Manski 1989).

UNOBSERVABLE BIAS, TREATMENT-EFFECT HETEROGENEITY BIAS, AND COMPOSITION BIAS

It is important to realize that ignorability in the form of equation (14) is merely an assumption. Even in situations where the assumption seems plausible, as long as the researcher deals with observational data, it is not a verifiable assumption. Thus, social research that relies on the ignorability assumption can always be challenged by the likely possibility that the assumption may not hold true in reality.

How does the violation of ignorability threaten research findings yielded by methods assuming ignorability? Again, we can answer this question from the two sources of selection bias. One possibility is that, even at a given level of propensity score, treated units and untreated units may still differ systematically in baseline so that their potential outcomes would be different in the absence of treatment. Another possibility is that even at a given level of propensity score, treated and untreated units may still differ systematically in the effects of treatment so that average treatment effects would differ between the two groups.

Economists have resorted to using unobservable variables to represent the two sources of selection bias that remain after the control of propensity score. Thus, let me call the two sources of residual bias respectively "pre-treatment unobservable heterogeneity bias" or "Type I unobservable selection bias" and "unobservable treatment-effect heterogeneity bias" or "Type II unobservable selection bias." It is interesting to note that the fixed effects method widely used in social science (e.g., Angrist and Krueger 1999) is defined only as handling Type I unobservable selection bias but as powerless regarding Type II unobservable selection bias. When the treatment effect is heterogeneous, the traditional IV approach would only identify the average treatment effect of a segment of the population that is induced by the instrument, called Local Average Treatment Effect (LATE) (Angrist and Krueger 1999; Angrist, Imbens, and Rubin 1996; Heckman, Urzua, and Vytlacil 2006).

The use of an unobservable variable in consideration of potential biases, either Type I or Type II, has strong support in economic theories. Let us use research on economic returns to college education as an example. It is plausible that persons who complete college educations may have higher innate mental ability than their peers who do not, everything else being equal. Because

mental ability should have a positive effect on earnings, omission of this person-specific factor causes what is called "ability bias," a Type I unobservable selection bias in the direction of over-estimation of the economic return to college education (Griliches 1977). Another possibility is that the likelihood of completing education is partly driven by heterogeneous returns to college education so that persons who actually complete college have on average higher economic returns than persons who do not. This is the comparative advantage model of Willis and Rosen (1979) that predicts the "sorting gain bias," a Type II unobservable selection bias studied heavily by Heckman and his associates through the IV approach (Carneiro, Hansen, and Heckman 2003; Carneiro, Heckman, and Vytlacil Forthcoming; Heckman, Urzua, and Vytlacil 2006).

Hence, the literature on causal inference is divided 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. In this paper, I show how another type of selection bias -- "composition bias" -- arises through dynamic processes when treatment propensity is systematically associated with heterogeneous treatment effects. Fundamentally, composition bias results from aggregation of units across heterogeneous subpopulations. What is interesting about composition bias is that it can arise even when the ignorability assumption is satisfied. Furthermore, there is no need to resort to unobservable variables to explain its existence and its consequences.

To understand what I mean by "composition bias," it is useful to conceptualize selection into treatment as a dynamic process, akin to survival analysis. A well-known property of a dynamic survival process is selective attribution/selection so that the composition of the remaining population at risk for selection changes constantly. That is, as the proportion of treated units in *U, p,* increases from p_1 to p_2, the subpopulation of the treated (U_1) changes from $U_1(p_1)$ to $U_1(p_2)$, and the subpopulation of the untreated (U_0) changes from $U_0(p_1)$ to $U_0(p_2)$. For simplicity, we assume a strictly nested structure so that for $p_1 < p_2$, $U_1(p_1)$ is strictly contained in $U_1(p_2)$. Of course, this also means that $U_0(p_2)$ is strictly contained in $U_0(p_1)$. 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*. I call such biases "composition biases."

For convenience in my discussion, I now define a new quantity of interest: "Increment Treatment Effect" (*ITE*). For $p_1 < p_2$,

$$ITE(p_1, p_2) = E[Y^1 - Y^0 | i \in U_1(p_2) \backslash U_1(p_1)], \quad (19)$$

where i is the ith unit in U, and $U_1(p_2) \backslash U_1(p_1)$ is the complement of $U_1(p_1)$ relative to $U_1(p_2)$, additional units recruited into treatment when the proportion of treatment increases from p_1 to p_2. In other words, *ITE* is the average treatment effect for these incremental units when $p = p_1$ changes to $p = p_2$. Like *TT* and *TUT*, *ITE* is the average effect of a group. Differing from *TT* and *TUT*, which are defined by a unit's observed status of treatment, *ITE* is defined by a latent attribute: a unit's treatment status changes from D=0 to D=1 when p increases from p_1 to p_2. Thus, we can alternatively define *ITE* as:

$$ITE(p_1, p_2) = E[Y^1 - Y^0 | D_{p_1} = 0, D_{p_2} = 1], \quad (20)$$

where D_{p_1} and D_{p_2} are the treatment statuses under, respectively, $p = p_1$ and $p = p_2$. Equations (19) and (20) are discrete forms of *ITE* when p increases from p_1 to p_2. The limit form of *ITE* can be defined as:

$$ITE(p) = \lim_{\delta \to 0} E[Y^1 - Y^0 | i \in U_1(p + \delta) \backslash U_1(p)]$$

$$= \lim_{\delta \to 0} E[Y^1 - Y^0 | D_p = 0, D_{p+\delta} = 1]. \quad (21)$$

It is important to note that *ITE* is different from a related quantity of interest -- Marginal Treatment Effect (*MTE*) -- which plays a key role in the IV approach to estimating heterogeneous treatment effect 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. That is, *MTE* is the average treatment effect of a subpopulation defined by units' intrinsic characteristics, i.e., relatively homogeneous propensities of treatment. In contrast, *ITE* is the average treatment effect of a relatively heterogeneous subpopulation defined by treatment regimes.

Another way to consider the differences between *MTE* and *ITE* is to see how a system in an equilibrium state may be changed by an external shock. For *MTE*, the latent propensity of treatment of a certain group of units (or even the entire population) is altered by an IV. It is this change in propensity that shifts the proportion being treated. For *ITE*, the proportion being treated is the exogenous change. It is the exogenous change in the proportion being treated that affects which units are actually included as increments in the treatment group, even though the relative magnitudes in the intrinsic propensity of treatment remain unchanged. In a sense, we can conceptualize changes in p (say from p_1 to p_2) as an IV, as it affects all units' likelihoods of being treated but not the outcome directly. However, we assume that this inducement effect is at the macro regime level; how the macro-level change may result in individual-level changes in treatment status still depends on intrinsic heterogeneous propensities of treatment across units in U.

Why may *ITE* be a useful quantity of interest, distinct from *MTE*? Because *ITE* and *MTE* provide different perspectives when an inducement occurs so that more units are treated than in an equilibrium state. *MTE* views the change of treatment status from the supply perspective. As the propensities of treatment for certain units are changed due to an IV, we ask which units may have changed their treatment status from not being treated to being treated. In contrast, *ITE* views a change from the demand perspective. As the proportion treated is changed at the regime level, we ask who is being newly recruited into the treatment group.

The usefulness of *ITE* is best illustrated in its sensitivity to the proportion treated, even when we keep all individual-level heterogeneity intact. This occurs because selection into treatment is a dynamic process (akin to survival analysis), so that net "composition" changes with the proportion of the subpopulation being treated (p) (Vaupel and Yashin 1985). When p is small, an increment in p is likely to recruit units with high propensities of treatment; *ITE* is then an average of treatment effects weighted heavily by high-propensity units. When p is high, high-propensity units are already in the treatment group; an increment in p is likely to recruit units with relatively lower propensities of treatment, because the representation of high-propensity units in the untreated subpopulation decreases with p. Consequently, *ITE* is weighted towards low-propensity units as p increases. Because the change of *ITE* as a function of p is purely a result of the compositional changes in the treated and untreated subpopulations, I call the bias resulting from this dynamic process the "composition bias."

As is true for *MTE*, we can also aggregate *ITE* over p to obtain *TT*, *TUT*, and *ATE*. Note that in our setup, *TT* and *TUT* depend on proportion treated and thus functions of p. A surprising result is that very simple expressions link *ITE* to *TT* and *TUT*, shown as follows.

$$TT(p) = \frac{1}{p}\int_0^p ITE(u)du. \quad (22)$$

$$TUT(p) = \frac{1}{1-p}\int_p^1 ITE(u)du. \quad (23)$$

$$ATE = \int_0^1 ITE(u)du. \quad (24)$$

A comparison of equations (22) and (23) reveals a selection bias, as in general,

$$TT(p) \neq TUT(p)$$

Note that this bias arises even when the ignorability assumption is satisfied. The source of this bias is a compositional change in either the U_1 or U_0 subpopulation, when the treatment proportion changes.

A TOY EXAMPLE FOR ILLUSTRATION

We now illustrate how the composition bias comes out in a dynamic process with a simple toy example. We conduct a simulation with a closed population of 1,000 units that are divided into ten evenly sized (n=100) strata (denoted by j, j = 1 . . .10). We specify that all 100 units in each stratum have the same intrinsic propensity potential (P_j^*) and the same treatment effect (δ_j). That is, we allow for heterogeneity in both intrinsic propensity of treatment and treatment effect across the ten strata, but for simplicity we assume homogeneity across the 100 units within each stratum. We specify P_j^* to vary linearly from 0.05 to 0.95. Likewise, we specify δ_j to increase linearly from 50 to 950, producing a correlation of 1 between the two parameters across the 10 strata. In this artificial example, ATE = 500. The detailed setup for the toy example is given in Table 1.

For the convenience of illustration, I also make increments discrete, developing in ten steps: [0.0, 0.1], [0.1, 0.2], [0.2, 0.3], [0.3, 0.4], [0.4, 0.5], [0.5, 0.6], [0.7, 0.8], [0.8, 0.9], and [0.9, 1.0]. For the first round of increment [0.0, 0.1], 100 units are moved from the untreated subpopulation (U_0) to the treated subpopulation (U_1). However, the distribution of these 100 units across the ten strata is not even. Again, for simplicity, I use expected numbers rather than realized numbers when employing simple random sampling. This strategy is tantamount to ignoring the influence of the sample size, which is arbitrarily set at 1,000. The first round of increments from $p = 0.0$ to $p = 0.1$ results in 100 new units being treated. The detailed results for the first round of increments are given in the first panel of Table 2.

Table 1: Setup of a Toy Example. A Hypothetical Population (N = 1,000) with Ten Strata

Strata (j)	Propensity Potential (P_j^*)	Treatment Effect (δ_j)	Number of Units (n_j)
1	0.05	50	100
2	0.15	150	100
3	0.25	250	100
4	0.35	350	100
5	0.45	450	100
6	0.55	550	100
7	0.65	650	100
8	0.75	750	100
9	0.85	850	100
10	0.95	950	100

Note: total Population is set to be 1,000. ATE = 500.

Table 2: Dynamic Recruitment of Treated Units at the First Two Rounds ($p = 0.0$ to $p = 0.1$; $p = 0.1$ to $p = 0.2$)

Strata (j)	First Round ($p = 0.0$ to $p = 0.1$)			Second Round ($p = 0.1$ to $p = 0.2$)		
	$\Delta U_{1,j}$	$U_{1,j}$	$U_{0,j}$	$\Delta U_{1,j}$	$U_{1,j}$	$U_{0,j}$
1	1	1	99	1	2	98
2	3	3	97	3	6	94
3	5	5	95	5	10	90
4	7	7	93	8	15	85
5	9	9	91	9	18	82
6	11	11	89	11	22	78
7	13	13	87	13	26	74
8	15	15	85	15	30	70
9	17	17	83	16	33	67
10	19	19	81	18	37	63
Total	*100*	*100*	*900*	*100*	*200*	*800*
Effect Measure	***ITE*** **665**	***TT*** **665**	***TUT*** **482**	***ITE*** **652**	***TT*** **659**	***TUT*** **460**

Note: U1,j and U0,j respectively denote the treated and untreated groups in the jth stratum. For each round of increments, $\square U1,j$ to denote the newly recruited units from the jth stratum that change the treatment status from D=0 to D=1, i.e., increments to U1,j.

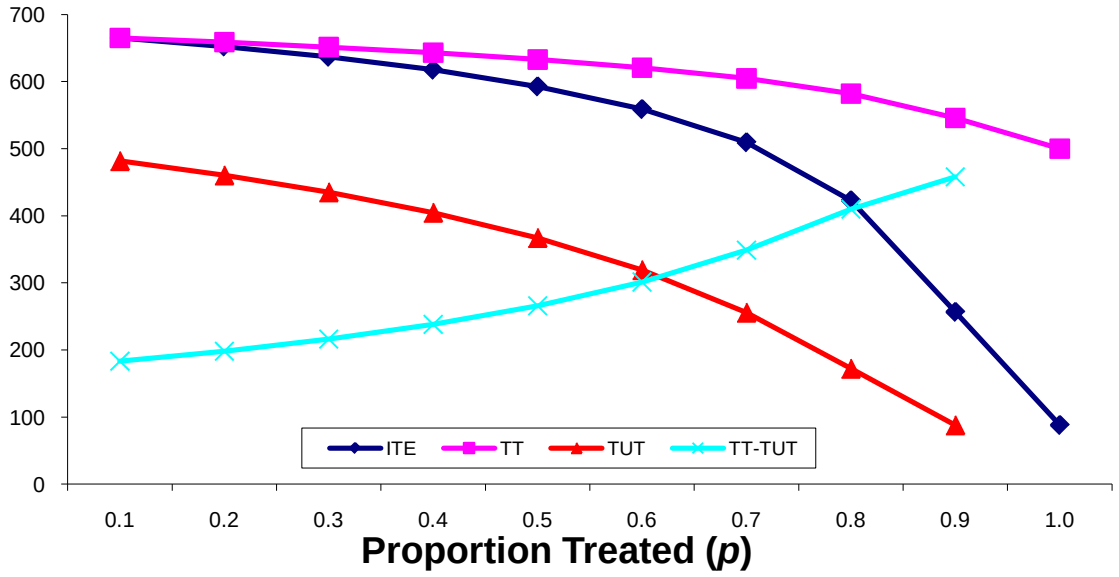
In Table 2, $U_{1,j}$ and $U_{0,j}$ respectively denote the treated and untreated groups in the *j*th stratum. $\Delta U_{1,j}$ denotes the newly recruited units from the *j*th stratum that changed the treatment status from $D = 0$ to $D = 1$, or increments to $U_{1,j}$. Because this is the first round of increments from $p = 0.0$ to $p = 0.1$, $\Delta U_{1,j}$ (the increment to the treated) , given in column 2, is identical to $U_{1,j}$ itself (column 3). The untreated subpopulation, $U_{0,j,}$ is simply the complement of $U_{1,j}$, given in column 4. It is apparent that the 100 newly treated cases (the second column, labeled $\Delta U_{1,j}$) are not evenly distributed across the ten strata, although we started with equal-sized strata in the population. Because the propensity potential P_j^* in a higher-numbered stratum is greater than that in a lower-numbered stratum, the number of units to be recruited into treatment in a higher-numbered stratum is also higher than that in a lower-numbered stratum. In fact, for the first round of increments, the ratio in the number being treated between two strata is directly proportional to the ratio in propensity potential (P_j^*). For example, the ratio in treated cases between stratum 10 and stratum 1 is 19, reflecting their ratio in P_j^*: $0.95/0.05$. Besides $\Delta U_{1,j}$, $U_{1,j}$ and $U_{0,j}$ are also unequally distributed across strata. The uneven distributions constitute different weights in the calculation of respective treatment effects, given in the last row. For this round, $ITE = TT$ at 665, much higher than TUT at 482. None of them is equal to ATE at 500.

We now conduct the second round of increments, from $p = 0.1$ to $p = 0.2$. We use the same recruiting mechanism and keep the intrinsic properties of all units intact. A key difference between the second round and the first round of increments is a compositional change in the exposure population from which increments are drawn. For the first round, the exposure population is the original population with an equal distribution across strata, shown in the last column in Table 1 (labeled n_j). For the second round, the exposure population is now changed to the untreated subpopulation in the first round, shown in the fourth column in Table 2 (labeled $U_{0,j}$). Due to this difference in exposure composition, the resulting increments in the second round, shown in the fifth column (labeled $\Delta U_{1,j}$) have a different across-strata distribution than its counterparts in the first round (second column, also labeled $\Delta U_{1,j}$). Comparing strata 10 and 1 again, for example, we see the ratio in $\Delta U_{1,j}$ between stratum 10 and stratum 1 to be reduced to 18, from 19 in the first round.

The reason for the decline in the representation of high-numbered strata in $\Delta U_{1,j}$ in the second round compared to the first round is simple. Because high-numbered strata have higher intrinsic propensity potentials (P_j^*), they are over-represented in $\Delta U_{1,j}$ in the first round and thus in $U_{1,j}$. As a result, higher-numbered strata are now under-represented in $U_{0,j}$, which serves as the exposure population for the next round of increments. Given the fixed propensity potential (P_j^*), a lower representation in the exposure population results in a lower representation in the newly recruited units, i.e., $\Delta U_{1,j}$ in the second round.

In fact, this dynamic process can continue and further compound the compositional process. In general, units with higher intrinsic propensity potentials are likely to be recruited into treatment when p is low, while units with lower intrinsic propensity potentials are likely to be recruited into treatment only when p is high. When intrinsic propensity and treatment effects are positively correlated, as is the case in this toy example, a positive selection bias arises due to sorting so that TT>TUT. In Figure 1, I present the full results when I carried out the toy example to its end, all the way to $p = 1.0$ with 0.1 as the increment interval. In the figure, I present four lines as functions of the proportion treated (p).

Figure 1: Treatment Effects by Increment Rounds



As we discussed earlier, *ITE* begins at a high level at 665 in the first round. It coincides with *TT* in round one and then diverges from *TT* by moving downward at a faster speed than that of *TT*. In the eighth round ($p = 0.7$ to $p = 0.8$), *ITE* is actually below *ATE* (which is 500). This shows that *ITE* is highly sensitive to changes in the composition of $U_{0,j}$ in the previous rounds. In contrast, *TT* is cumulative, as the average of *ITE* in earlier rounds (equation 22), and it declines more slowly. Note that $TT(p) > ATE$ for all p, due to our setup for a positive selection. However, the gap between *TT* and *ATE* diminishes gradually over p, especially after the eighth round ($p = 0.7$ to $p = 0.8$). Similarly, *TUT* is also cumulative, but reversely from $p = 1.0$ backwards. We normalize that $TUT(p = 1.0)$ is undefined. Because of the way we define *ITE* in discrete intervals, $TUT(p = 0.9) = ITE(p = 0{,}9{,}1.0)$ in our example. We also observe that $TUT(p) < ATE$ for all p. Furthermore, as in the case of *TT*, *TUT* also trends downward with p. This last result is sensible in light of the relationship between *TUT* and *ITE*, but I did not know about it until I obtained the results from the toy experiment.

One way to evaluate the Type II selection bias is to measure the sorting gain (or loss), the difference between *TT* and *TUT*. Hence, in Figure 1, I present $TT - TUT$ as a function of the treatment proportion p. A counterintuitive finding from this exercise is that the amount of bias as measured by the sorting gain actually increases, rather than decreases, as the proportion treated increases. This is due to the fact that the downward trend of *TUT* is steeper than that of *TT*. This pattern results from the shape of *ITE*, as the decline of *ITE* is slower when p is small but accelerates when p is close to 1. The increasing trend in the amount of bias depicted in Figure 1 is surprising because one may think that, as treatment extends to a larger and larger portion of a population, treated units should become less and less selective (which we show is true), and thus selection bias should decline (which is not true). Of course, our conclusion is based on using *TUT* instead of *ATE* as the reference for measuring the amount of bias. In our example, as p increases, units being treated become less positively selective, but units not being treated become more negatively selective. Although the two trends are in the same direction, the decline in selectivity among the treated is slower than the change in selectivity among the untreated. In other words, by the time almost every unit in a population is treated, only those with extremely low intrinsic propensities of treatment remain in the untreated subpopulation – i.e., severe selectivity.

DISCUSSION AND CONCLUSION

Due to the ubiquity of heterogeneity in social phenomena, it is impossible to draw causal inferences at the individual level. All efforts to draw causal inferences in social science must take place at the group level. However, comparison of groups is not possible without some way of combining intrinsic heterogeneous individuals into relatively homogeneous groups. This is a fundamental dilemma facing all researchers in social science.

It is a truism that any group-level comparison can be further decomposed into comparisons of sub-groups. For causal inference, it is now well known that a useful dimension for decomposition is the propensity score, which summarizes information in a multi-dimensional space from multivariate covariates into a univariate variable. Therefore, 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 2011). 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 the ignorability is true, the estimated effect heterogeneity may be subject to generalizations, as discussed in this paper. However, the researcher may alternatively 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).

All quantities of interest at the group level, such as *TT* and *TUT*, are essentially weighted averages of treatment effects across subgroups. Therefore, composition is important in causal inference. In this paper, I have shown the presence of "composition bias," which is a form of selection bias. This composition bias is generated by a dynamic process when the treatment proportion changes. Interestingly, this form of bias selection can be generated even under the ignorability assumption. All that is required is the combination of three things: (1) intrinsic heterogeneity in treatment propensity, (2) intrinsic heterogeneity in treatment effects, and (3) a correlation between heterogeneity in treatment propensity and heterogeneity in treatment effects. Under these simple conditions, all permissible under the ignorability assumption, a classic scenario for selection bias may arise, units more responsive to treatment being more likely to receive treatment than units less responsive (Roy 1951; Willis and Rosen 1979).

A composition bias is essentially driven by the fact that units with higher intrinsic propensity of treatment are likely to be over represented when the treatment proportion is small. Their presence in the exposure population however is reduced by their entry 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 shift in composition among newly recruited increments away from high propensity toward low propensity results in changes in average treatment effects, regardless of how those effects are calculated. In short, I have demonstrated in this paper that treatment effects calculated over heterogeneous populations are highly susceptible to compositional shifts. Researchers should always be mindful of the population or sub-population of interest when deriving and interpreting average causal estimates from potentially heterogeneous subgroups.

A substantive example would be the administration of a medical treatment or social intervention on a graduated schedule. Let us assume that participation is need-based, the poorest persons being most eligible and thus chosen first, and, further, that the poorest persons would also stand to benefit most from treatment. Under these conditions, individuals selected at later stages (i.e., becoming eligible only after the eligibility cut-point is moved up) would exhibit lower average treatment effects simply by virtue of coming from a less responsive subpopulation.

Our results should also serve as a warning regarding efforts to extend research results from a small-scale study, be it observed or experiment, beyond the setting in which the study was conducted. Population heterogeneity means not only that treated units may be incomparable to untreated units in the study – an issue of internal validity – but also that external validity can be difficult to establish. As the researcher generalizes results from a small- scale study to the general population, we cannot know whether the subjects in the study are comparable to those in the population. The potential systematic differences between the subjects in the study and the general population, called "compositional differences" in this paper, may dramatically alter the average treatment effect.

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, J. D. and A. Krueger. 1999. "Empirical Strategies in Labor Economics." Pp. 1277-366 in *Handbook of Labor Economics*, vol. 3A, edited by O. Ashenfelter and David Card. Amsterdam: Elsevier.

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.

Breslow, Norman. 1996. "Statistics in Epidemiology: The Case-Control Study." *Journal of the American Statistical Association* 91(433): 14-28.

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.

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* 94:1053-1062.

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.

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.206-33 in *Foundations of Scientific Method, the Nineteenth Century*, edited by Ronald N. Giere and Richard 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 Manski, Charles F. and Irwin 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):488-495.

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.

Plato. 1997. *Complete Works*, edited by John M. Cooper. Indianapolis, IN: Hackett.

Quételet, Adolphe. 1842. *A Treatise on Man and the Development of his Faculties.* A facsimile reproduction of the English translation of 1842, with an introduction by Solomon Diamond. Gainesville, FL: Scholars' Facsimiles. 1969.

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* 79:516-524.

Roy, A. D. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers*, New Series, 3: 135-146.

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.

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* 39:176-185.

Willis, R. and Rosen, S. 1979. "Education and Self-Selection." *Journal of Political Economy* 87:S7-36.

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

Xie, Yu, Jennie Brand, and Ben Jann. 2011. "Estimating Heterogeneous Treatment Effects with Observational Data." University of Michigan Population Studies Center Research Report.

Xie, Yu and Charles F. Manski. 1989. "The Logit Model and Response-Based Samples." *Sociological Methods and Research* 17:283-302.

Xie, Yu and Xiaogang Wu. 2005. "Market Premium, Social Process, and Statisticism." *American Sociological Review* 70:865-870.

Population Studies Center Research Reports

The **Population Studies Center** (PSC) at the University of Michigan is one of the oldest population centers in the United States. Established in 1961 with a grant from the Ford Foundation, the Center has a rich history as the main workplace for an interdisciplinary community of scholars in the field of population studies.

Currently PSC is one of five centers within the University of Michigan's Institute for Social Research. The Center receives core funding from both the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R24) and the National Institute on Aging (P30).

PSC Research Reports are **prepublication working papers** that report on current demographic research conducted by PSC-affiliated researchers. These papers are written for timely dissemination and are often later submitted for publication in scholarly journals.

The **PSC Research Report Series** was initiated in 1981.

Population Studies Center
University of Michigan
Institute for Social Research
PO Box 1248, Ann Arbor, MI 48106-1248 USA
http://www.psc.isr.umich.edu/pubs/