Population Studies Center

Research Report

Yu Xie, Jennie Brand, and Ben Jann

Estimating Heterogeneous Treatment Effects with Observational Data

Report 11-729
February 2011

Population Studies Center
University of Michigan
Institute for Social Research

Estimating Heterogeneous Treatment Effects with Observational Data

Yu Xie
University of Michigan

Jennie E. Brand
University of California – Los Angeles

Ben Jann
University of Bern, Switzerland

Population Studies Center Research Report 11-729

February 2011

Direct all correspondence to Yu Xie, Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48104, USA; email: yuxie@umich.edu. Financial support for this research was provided by the National Institutes of Health, Grant 1 R21 NR010856-01 and by the Population Studies Center at the University of Michigan and the California Center for Population Research at UCLA, which receive core support from the National Institute of Child Health and Human Development, Grants R24HD041028 and R24 HD041022. A version of this paper was presented at the 2010 Stata Users meeting in Germany. The ideas expressed herein are those of the authors.

AUTHORS

Yu Xie is Otis Dudley Duncan Distinguished University Professor of Sociology and Statistics 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, and *Statistical Methods for Categorical Data Analysis* (Second edition, Emerald 2008).

Jennie E. Brand is Associate Professor of Sociology at the University of California – Los Angeles and incoming Associate Director of the California Center for Population Research. Her research focuses on the relationship between social background, educational attainment, job conditions, and socioeconomic attainment and well-being over the life course. This substantive focus accompanies a methodological focus on causal inference and the application and innovation of statistical models for panel data. Current research projects include evaluation of heterogeneity in the effects of education on socioeconomic outcomes and the social consequences of job displacement.

Ben Jann is Professor of Sociology at the University of Bern, Switzerland. His research interests include social-science methodology, statistics, social stratification and labor market sociology. Recent publications include a paper on the Randomized Response Technique in *Sociological Methods & Research* (with Elisabeth Coutts) and a paper on statistical results processing in the *Stata Journal* (with J. Scott Long).

ABSTRACT

Heterogeneous treatment effects are widely recognized but seldom studied empirically in quantitative sociological research. We suspect that lack of accessible statistical methods is one reason why heterogeneous treatment effects are not routinely assessed and reported. In this paper, we discuss a practical approach to studying heterogeneous treatment effects, under the same assumption commonly underlying regression analysis: ignorability. We specifically describe two methods. For the first method (SM-HTE), we begin by estimating propensity scores for the probability of treatment given a set of observed covariates for each unit and construct balanced propensity score strata; we then estimate propensity score stratum-specific average treatment effects and evaluate a trend across the strata-specific treatment effects. For the second method (MS-HTE), we match control units to treated units based on the propensity score and transform the data into treatment-control comparisons at the most elementary level at which such comparisons can be constructed; we then estimate treatment effects as a function of the propensity score by fitting a non-parametric model as a smoothing device. We illustrate the application of the two methods with a concrete empirical example.

1 Introduction

A feature common to all social phenomena is variability across units of analysis (Xie 2007). The development of statistical methods to better understand and accommodate such variability has been a major methodological achievement of modern quantitative social science, especially micro economics (e.g., Heckman 2001). Individuals differ not only in their background characteristics, but also in how they respond to a particular treatment, intervention, or stimulation. We call the latter type of variability "heterogeneous treatment effects." Heterogeneous treatment effects are widely recognized but seldom studied empirically in quantitative sociological research. Researchers using traditional regression modeling assume, often unknowingly, constant effects. We suspect that lack of accessible statistical methods is one reason why heterogeneous treatment effects are not routinely assessed and reported. In this paper, we describe a straightforward approach to exploring and estimating heterogeneous treatment effects with observational data.

Heterogeneity in treatment effects has important implications for social and behavioral research and for social policy (e.g., Heckman, Urzua, and Vytlacil 2006; Manski 2007). On the one hand, if a treatment is costly and difficult to administer and, as a result, is available only to those subjects who are likely to benefit most from it, increasing the pool of subjects receiving the treatment may reduce its average effectiveness. On the other hand, if individuals with current access to a treatment are not the individuals likely to benefit most from the treatment, increasing the availability of the treatment may increase the average effect among the treatment recipients. If policy makers understand patterns of treatment effect heterogeneity, they can more effectively assign different treatments to individuals so as to balance competing objectives, such as reducing cost, maximizing average outcomes, and reducing variance in outcomes within a given population (Manski 2007). The study of treatment effect heterogeneity can also yield important insights about how scarce social resources are distributed in an unequal society.

We propose an approach to studying treatment effect heterogeneity that builds on a common framework for estimating causal effects. We develop two specific methods under the ignorability assumption. The first method has been applied in several recent empirical applications (Brand 2010; Brand and Davis [forthcoming]; Brand and Xie 2010; Tsai and Xie 2008; Xie and Wu 2005). In this paper, we focus on estimating issues of this previously applied method. The second method is a non-parametric counterpart to the first method. Our discussion proceeds as follows. In section 2, we discuss population heterogeneity, selection into treatments, causal inference, and the sociological significance of studying heterogeneous treatment effects. In section 3, we present our approach to studying treatment effect heterogeneity. In section 4, we present an empirical example in which we use the methods. In section 5, we discuss the benefits and limitations of this approach and conclude the paper.

2 Background and Significance

2.1 Causal Inference in the Population Sciences

Sociology is a population science, concerned with understanding characteristics of what Neyman called “populations” – “categories of entities satisfying certain definitions but varying in their individual properties” (quoted by Duncan 1984, p. 96). Population sciences, including economics, demography, epidemiology, psychology, and sociology, all treat individual-level variation as a main object of scientific inquiry, rather than a mere nuisance or measurement error while assuming that all units of analysis are essentially the same (Angrist and Kruger 1999; Ansari and Jedidi 2000; Bauer and Curran 2003; Greenland and Poole 1988; Heckman 2001, 2005; Heckman and Robb 1985; Lubke and Muthen 2005; Manski 2007; Moffitt 1996; Rothman and Greenland 1998; Winship and Morgan 1999; Xie 2007). The idea that scientists can fruitfully study categories of entities that vary from one another was a revolutionary concept that originated with Darwin (1859) (see Mayr 1982, 2001). The recognition of inherent individual-level heterogeneity has important consequences for research designs in the social sciences. Because individuals differ from one another and differ in their response to treatments, results can vary widely depending on population composition.

The large methodological literature on causal inference using statistical methods recognizes the importance of and consequently allows for population heterogeneity (Heckman 2005; Holland 1986; Manski 1995; Rubin 1974; Winship and Morgan 1999). Suppose that a population, *U*, is being studied. Let *Y* denote an outcome variable of interest (with its realized value being *y*) that is a function for each member in *U*. Let us define treatment as an externally induced intervention that can, at least in principle, be given to or withheld from a unit under study. For simplicity, we consider only dichotomous treatments and use *D* to denote the treatment status (with its realized value being *d*), with *D*=1 if a member is treated and *D*=0 if a member is not treated. 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). The framework for counterfactual reasoning in causal inference (Heckman 2005; Holland 1986; Manski 1995; Morgan and Winship 2007; Rubin 1974; Sobel 2000; Winship and Morgan 1999) states that we should conceptualize a treatment effect as the difference in potential outcomes associated with different treatment states for the *same* member in *U*:

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

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."

The scientific solution capitalizes on homogeneity in assuming that all members in *U* are the same, in either the treated state or the control state: $y_i^1 = y_j^1$and$y_i^0 = y_j^0$, where $j \neq i$ in *U*. That is, there is only a difference between the treated state and the untreated state, but there is no variability across units of analysis within the same state, so that all treated units are identical and untreated units are identical. This strong homogeneity assumption would enable a researcher to identify individual-level treatment effects. Indeed, if the strong homogeneity 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 population, as the following would hold true:

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

for any $j \neq i$, where we can drop the subscript of δ because it does not vary across members in the population. While strong homogeneity may hold true in experimental science, sociology, as we discussed above, is a population science so that pervasive heterogeneity across units in *U* is the norm rather than the exception. Thus, in general, the formula under the strong homogeneity assumption [equation (2)] has no practical value in social and behavioral sciences in general.

For a population science, the statistical solution is a necessity. The statistical approach is to compute quantities of interest that reveal treatment effects only at the group level. For example, we may 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. For simplicity, we focus our discussion on identification issues and ignore the problem of statistical inference (Manski 1995). That is, we assume that we have a very large sample so that we are not concerned with statistical inference. The comparison of the two groups randomly selected from the whole population as described above yields a quantity that is called the Average Treatment Effect (ATE):

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

While ATE is defined for the whole population, the researcher may wish to focus and define a treatment effect for a well-defined subpopulation. 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:

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

Although various statistical quantities of interest can easily be defined theoretically with the statistical "solution," there is actually no simple solution, as estimating these quantities in social research can be very difficult, indeed often impossible, in practice. In order to compute ATE, TT, and TUT with observational data, it is necessary to invoke assumptions, a topic to be discussed later.

Because all statistical quantities of interest can be computed only at the group level, the researcher is forced to overlook, within the context of a given study, within-group individual-level variation, which we know must exist (Xie 2007). This is always true despite various efforts to allow for or to recover some degree of heterogeneity through various (but related) statistical approaches, such as regression models with interactions between treatment indicators and contextual or individual level variables, as commonly practiced in multilevel models (Raudenbush and Bryk 2002; Vermunt 2003), Bayesian analysis (Gelman et al. 2004), growth-curve analysis (Muthén and Muthén 2000), meta-analysis (Hedges 1982), the latent class model (D'Unger et al. 1998; Heckman and Singer 1984; Vermunt 2002), and Rasch models (Duncan, Stenbeck, and Brody 1988; Rasch 1966). Most notably, Heckman and his associates have extensively discussed heterogeneous treatment effects, they call "essential heterogeneity," in a class of models relying on instrumental variables (Carneiro, Heckman, and Vytlacil forthcoming; Heckman, Urzua, and Vytlacil 2006). In our view, the real challenge in a research setting is thus not to establish absolute homogeneity across units of analysis, which is impossible, but to realize that any approach to analyzing observational data is essentially marginal: in order to focus on differences across subpopulations to answer questions of research interest, we temporarily overlook heterogeneity within subpopulations defined by observable characteristics by aggregating over heterogeneous units of analysis within subpopulations thus defined. Different analysis specifications essentially boil down to different ways to define such subpopulations.

2.2 Two Types of Heterogeneity

In the preceding section, we established the need to conduct group-level comparisons for causal inference. The rationale is to compare groups that are essentially comparable except for their treatment status. However, due to population heterogeneity, there is no guarantee that the group that actually receives the treatment is comparable, in observed and particularly in unobserved contextual and individual characteristics, to the group that does not receive the treatment. Individuals may self-select into treatment based on their anticipated monetary and nonmonetary benefits and costs of treatment. To see this, let us partition the total population U into the subpopulation of the treated U_1 (for which D=1) and the subpopulation of the untreated U_0 (for which D=0). We can thus decompose the expectation for the two counterfactual outcomes as follows:

$$E(Y^1) = E(Y^1|D=1)P(D=1) + E(Y^1|D=0)P(D=0) \quad (6)$$

and

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

What we observe from data are: $E(Y^1|D=1), E(Y^0|D=0), P(D=1),$ and $P(D=0)$. A concern with selection bias is due to:

$$E(Y^1|D=1) \neq E(Y^1|D=0) \neq E(Y^1) \quad (8)$$

and

$$E(Y^0|D=1) \neq E(Y^0|D=0) \neq E(Y^0). \quad (9)$$

To see how selection into treatment may cause biases on treatment effects, 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:

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

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:

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

Note that the simple comparison estimator from observed data is $E(Y^1_{D=1}) - E(Y^0_{D=0})$. If we use this naive estimator for ATE, we see two sources of bias from equation (10):

(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."

When there is treatment-effect heterogeneity bias, we have the following

$$TT \neq TUT;$$
$$ATE \neq TT;$$
$$ATE \neq TUT.$$

Consider two concrete examples representing the two different sources of selection bias. First, pre-school children from poor families are selected into Head Start programs and thus would compare unfavorably to other children who do not attend Head Start programs without an adequate control for family socioeconomic resources (Xie 2000). Second, economic theory predicts that attainment of college education may be selective because it may attract young persons of greater ability and persons who may gain more than persons who do not attend college (Willis and Rosen 1979). While the first example illustrates the importance of controlling for pre-treatment heterogeneity bias that may be represented by "covariates" or "fixed effects," the second example underscores treatment-effect heterogeneity bias – sorting on the treatment effects – that cannot be "controlled for" by covariates or fixed-effects.

Experimental studies rely on random assignment to eliminate both sources of selection bias. Random assignment ensures that a sample of units in U receives either the treatment or control condition by chance only. Let $\perp\!\!\!\perp$ denote independence. Random assignment, in theory, assures:

$$(Y^1, Y^0) \perp\!\!\!\perp D, \qquad (11)$$

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:

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

Note that the commonly used fixed-effects estimator for observational studies eliminates only (time-invariant) pre-treatment heterogeneity bias (Type I selection bias), but not treatment-effect heterogeneity bias (Type II selection bias), because the fixed effects estimator (Angrist and Krueger 1999) eliminates pre-treatment (i.e., fixed) differences between the treated and untreated groups:

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

In general, with data from observational studies, where subjects have made their own choices regarding the treatment status of interest, the independence condition of equation (11) does not hold true. Most likely, subjects were sorted into treatment or control groups for a number of reasons, some of which may be unknowable to the researcher.

When assignment to treatment is not random, researchers primarily resort to two strategies. First, they may capitalize on a variable (or variables) that affects assignment to treatment exogenously but affects the outcome only indirectly through treatment. For example, draft lottery may be associated with military enlistment but should not affect economic outcomes directly (Angrist 1990). Such variables are called “instrumental variables” (IV) (Angrist and Krueger 1999; Angrist, Imbens, and Rubin 1996; Bound, Jaeger, and Baker 1995; Heckman, Urzua, and Vytlacil 2006; Winship and Morgan 1999). An IV must satisfy the assumption of exclusion restriction in that it affects the outcome of interest (*Y*) only indirectly via affecting the treatment status (*D*). However, the IV approach has three major weaknesses: (a) it is very hard to find a meaningful IV that satisfies the exclusion restriction assumption; (b) a weak IV may give rise to imprecise IV estimates and estimates that are biased in finite samples (Bound, Jaeger, and Baker 1995); and (c) if treatment effects are heterogeneous, the estimates using the IV formula should be only interpreted as local average treatment effects (*LATE*), average effects that pertain only to the units whose treatment assignment statuses are affected by the instrument (Angrist and Krueger 1999; Angrist, Imbens, and Rubin 1996; Heckman, Urzua, and Vytlacil 2006; Imbens and Angrist 1994). We will discuss the IV method in relation to treatment effect heterogeneity in the next subsection.

The second approach is to control for differences between the treatment group and the control group with observed covariates (Angrist and Krueger 1999; Cochran 1972; Cornfield et al. 1959; Heckman, LaLonde, and Smith 1999; Imbens 2004). While it is not possible for a researcher to claim that he or she has controlled for all variables that may affect the outcome, the weaker assumption of having controlled for all relevant pre-treatment variables that simultaneously affect the treatment assignment and the outcome is more plausible. Only covariates that meet the condition of affecting both the treatment assignment and the outcome confound the observed relationship between treatment and outcome (Rubin 1997). It is thus hoped that through control of the relevant covariates the treatment will 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:

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

Comparison of equations (11) and (14) highlight the role of covariates *X*. Because we can never be sure after inclusion of which covariates equation (14) would hold true, the ignorability condition is always held as an assumption, indeed an unverifiable assumption. Substantive knowledge about the subject matter needs to be brought in before a researcher can entertain the ignorability assumption. Measurement of theoretically meaningful confounders makes ignorability tentatively plausible, but not necessarily true. Pearl (2009) provides conditions for including relevant covariates as appropriate controls.

However, the researcher can always 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). Results for causal inference under the ignorability assumption thus should always be interpreted provisionally and cautiously, as we do in our own work (Brand 2010; Brand and Davis [forthcoming]; Brand and Xie 2010; Tsai and Xie 2008; Xie and Wu 2005).

Conditioning on X can be difficult in applied research due to the "curse of dimensionality." However, Rosenbaum and Rubin (1983, 1984) show that, when the ignorability assumption holds true, it is sufficient to condition on the propensity score as a function of *X*. Thus, equation (14) is changed to

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

where $P(D = 1|X)$ is the propensity score, the probability of treatment that summarizes all the relevant information in covariates X. In other words, covariates X may confound the observed relationship between treatment D and outcome Y only through the propensity score $P(D = 1|X)$. The literature on propensity score matching recognizes the utility of the propensity score as a solution to data sparseness in a finite sample (Morgan and Harding 2006). We will discuss this approach and its relation to treatment effect heterogeneity more fully in section 3.

2.3 The IV Approach and Treatment Effect Heterogeneity

In a series of papers, Heckman and his associates (Heckman, Urzua, and Vytlacil 2006; Heckman and Vytlacil 1999, 2001, 2005) have proposed a method of estimating heterogeneous treatment effects using instrumental variables (IV). An IV is a predictor of the treatment probability that satisfies the following assumption of exclusion restriction: affecting the substantive outcome variable (Y) *only* indirectly via the treatment status (D). At the core of this IV approach is the use of a new concept -- the marginal treatment effect (MTE) (Björklund and Moffitt 1987) -- the expected treatment effect conditional on observed covariates as well as the marginal point at which a latent factor determining treatment status is neutral – i.e., does not favor either treatment or control. One important result of this line of work is that once MTE is known, all common treatment parameters of concern, such as ATE, TT, and TUT, can be derived as weighted averages of MTE. For an application of this approach in sociology, see Tsai and Xie (forthcoming).

The estimation of MTE is conditional on covariates $\boldsymbol{X}$. The MTE method attempts to estimate heterogeneous treatment effects over the entire range of the unobserved factor via IVs, given $\boldsymbol{X}$. This IV requirement is very stringent, far more demanding than what is available in actual settings for empirical research. In typical situations employing IVs, the researcher may be in the fortunate situation of affording the assumption that a random event affects the likelihood of treatment status within a small range, but it affects the outcome only indirectly through the treatment. In a typical situation, we may assume that subjects have their otherwise natural propensity of treatment, but some of them were somewhat "induced" into treatment by a randomly occurring event (e.g., IV). Angrist and Pischke (2009) give a good account of such applications in empirical settings. However, in actual applications, this inducement effect of an IV on treatment is usually very small, because social events still occur in their natural course without the interference of IVs. Consider two well-known empirical studies with IVs: (1) the inducement of an IV is only 0.016 for the birth quarter effect on the likelihood of high school graduation; and (2) the inducement of an IV is only 0.060 for the same sex effect of first two children on the likelihood of having a third child (Angrist and

Pischke 2009, p.169). If treatment effects are homogeneous, this low inducement effect of IVs on the treatment likelihood is not a major limitation, as the estimator based on the small proportion of individuals who were induced into treatment can be generalized to the whole population. However, in the presence of heterogeneous treatment effects, we can only limit the interpretation of the resulting estimator to this particular group of individuals being induced, as implied by the term the "local average treatment effect" (LATE) (Angrist, Imbens, and Rubin 1996; Angrist and Pischke 2009). We already know that IVs of this kind that would yield LATE are rare opportunities in empirical research. It stands to reason that IVs that could be utilized to identify MTE for the whole distribution of the latent factor conditional on $\boldsymbol{X}$ are indeed very difficult, if not impossible, to find.[1]

3 Approach for Estimating Heterogeneous Treatment Effects under Ignorability

3.1 The Rationale

Given this practical difficulty of using the IV method to identify heterogeneous treatment effects, we discuss in this paper a simple and straightforward approach under the ignorability assumption. This approach capitalizes on the fact that under the ignorability assumption, any imbalance between the treated group and the untreated group can be adequately captured and characterized by the propensity score, shown in equation (15). We will then detect heterogeneous treatment effects by the propensity score in two ways, the stratification-multilevel method and the matching-smoothing method. Of course, we realize that a narrow focus on estimating heterogeneity in treatment effects by observed covariates is limited, as we by necessity overlook heterogeneity in treatment effects due to unobserved variables. The plausibility of the ignorability assumption is a substantive rather than a methodological issue, depending on the richness of the empirical data. Because we can never be sure which covariates would make equation (14) true, the ignorability condition is always held as an unverifiable assumption. However, the researcher can consider the ignorability assumption and then evaluate its plausibility in a concrete setting through sensitivity or auxiliary analyses (Cornfield et al. 1959; DiPrete and Gangl 2004; Harding 2003; Rosenbaum 2002).

[1] Of course, we are aware of empirical applications of the MTE method (Heckman, Urzua, and Vytlacil 2006; Carneiro, Heckman, and Vytlacil forthcoming; Tsai and Xie forthcoming). However, unlike the examples given by Angrist and Pischke (2009), these studies rely on theory-based assumptions, rather than "natural experiments," to justify the choice of their IVs.

The ignorability assumption allows the researcher to explore empirical patterns of treatment effect heterogeneity as a function of the propensity score, which in turn is a function of observed covariates (Brand 2010; Brand and Davis [forthcoming]; Brand and Xie 2010: Tsai and Xie 2008; Wu 2009; Xie and Wu 2005). Without additional assumptions, analysis under the ignorability assumption is the best that we can do with the information contained in observed data. After we detect heterogeneous treatment effects under the ignorability assumption, as in examples given before, we can then revisit the assumption and ask whether it is indeed plausible given the results. For example, Xie and Wu (2005) interpret a negative selection pattern detected under the ignorability assumption in terms of differential selectivity into treatment status. That is, the ignorability assumption may yield empirical patterns of heterogeneous treatment as a function of covariates, but the empirical results are subject to different interpretations, including those involving selective mechanisms due to unobserved variables.

Why can we eliminate the two types of heterogeneity bias when controlling for the propensity score if the ignorability assumption is true? Recall from equation (15) that, under the ignorability assumption, conditional on the propensity score, treatment status is independent of the two potential outcomes under alternative treatment conditions. In other words, given a level of the propensity score, we assume no bias. Given our earlier discussion that bias can manifest in two types (subsection 2.2), 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

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

Here, we make the propensity score a function of only $\boldsymbol{X}$, not $\boldsymbol{Z}$, because we do not assume the presence of any IVs that satisfy the exclusion restriction.

Hence, the proposed approach places a large emphasis on the propensity score, as it plays a crucial role in both pre-treatment heterogeneity and treatment-effect heterogeneity, as shown in equation (16) and (17). Both types of heterogeneity bias, i.e., systematic differences between the treatment group ($D = 1$) and the control group ($D = 0$) for causal inference, are captured by the propensity score. That is to say, the researcher should pay attention only to the interaction between the treatment indicator and the propensity score, as far as a selection bias is concerned. Given (17), it is easy to show that

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

Equation (18) provides the rationale for estimating heterogeneous treatment effects as a function of the propensity score.

3.2 Estimation of the Propensity Score

In subsection 3.1, we discussed the role of the propensity score as if it is known. In reality, of course, it is unknown and needs to be estimated. For estimating the propensity score we use a probit model.[2] The assumption of exclusion restriction for instrumental variables is unnecessary. All pre-treatment covariates that are unevenly distributed between the treatment group and the control group could potentially contribute to selection biases; they can all be included as predictors of the propensity model. How good the covariates are as predictors of the propensity score is a substantive question. There is a large literature on the estimation and use of the propensity score (see Morgan and Harding 2006; Gangl 2010 for recent reviews). One possible scenario is one of "no common support" – regions of the propensity score in which we do not observe both treated and untreated cases. When there are regions of no common support, the researcher can either impose a structure to essentially impute data for propensity score estimation there or, to be safe, give up inferences for such regions.

We now assume that we have obtained good estimated propensity scores at the individual level, $\hat{p}_i(\boldsymbol{X} = \boldsymbol{x_i})$. These estimated individual-level propensity scores can be used to balance the distribution in covariates between the treatment group and the control group. When the balance is accomplished, we can examine heterogeneous treatment effects as a function of the propensity score. Below, we discuss two different estimation methods based on estimated propensity scores.

[2] Of course, we can also estimate the propensity score with a logit model. Differences between the two models are usually minor (Powers and Xie 2008, p.44).

3.3 The Stratification-Multilevel Method (SM-HTE)

The first method we discuss is what we call the "stratification-multilevel method of estimating heterogeneous treatment effects," or "SM-HTE" for short. It consists of the following concrete steps:

(1) Estimate propensity scores for each unit for the probability of treatment given a set of observed covariates, $P(d=1|X)$, using probit or logit regression models, as discussed in subsection 3.2.

(2) Construct balanced propensity score strata (or ranges of the propensity score) where there are no significant differences in the average values of covariates and the propensity score between the treatment and control groups.[3] The underlying assumption is that we consider all units, treated and untreated, within a stratum as homogeneous for estimating treatment effects. As we discussed earlier, some grouping is necessary when computing statistical quantities representing causal effects. While the assumption of within-stratum homogeneity may not hold true in practice, it is less stringent relative to the original differences between the treatment group and the control group. It is hoped that the stratification of data by the propensity score is an effective way to remove most biases between the treated and untreated groups (Rosenbaum and Rubin 1984).

(3) Estimate propensity score stratum-specific treatment effects within strata. We can do this either by drawing a direct comparison in the outcome variable between the treatment group and the control group within strata, shown in equation (18), or by applying a regression model within strata to further adjust for any remaining covariate imbalance within strata. Results by strata, or level-1 estimates, are obtained from this step of the analysis. Because we do not constrain the comparison of the treatment group and the control group across strata in any way, data analysis at this stage is non-parametric across strata.

(4) Evaluate a trend across the strata using variance-weighted least squares regression of the strata-specific treatment effects, obtained in step (3), on strata rank at level-2. This step departs from the conventional use of propensity scores in constructing strata, where the emphasis is usually on removing biases due to covariate imbalances simply by averaging

[3] The researcher would need to specify a level of significance for testing the differences. As expected, the lower the level, the more stringent the test (i.e., the more likely that some covariates do not satisfy it and remain unbalanced).

the estimated treatment effects across strata (Dehejia and Wahba. 1999; Rosenbaum and Rubin 1984). In contrast, the main research objective emphasized in this paper is to look for a systematic pattern of heterogeneous treatment effects across strata. For simplicity and to preserve statistical power, we mainly suggest modeling the heterogeneity pattern as a linear function across strata ranks. A linearity specification would force the data to tell us whether the treatment effect is either a positive or a negative function of propensity. This strategy has proved useful in empirical research (Brand 2010; Brand and Davis [forthcoming]; Brand and Xie 2010; Tsai and Xie 2008; Xie and Wu 2005). Of course, with more complicated research goals and richer data, the researcher is free to specify different parametric functions at level 2 for the heterogeneity in treatment effects across propensity-score strata, as in ordinary multi-level models (Raudenbush and Bryk 1986, 2002).

The SM-HTE approach has two notable shortcomings. First, the researcher is forced to divide the full range of the propensity score into a limited number of intervals, called strata, within which we assume neither Type I nor Type II selection bias. That is, we impose a form of within-group homogeneity so that both all treated (or untreated) observations are treated as interchangeable within strata. Second, across the strata, we impose a higher-level regression to detect a pattern of treatment heterogeneity. Given the limited number of observations, i.e., strata, for this secondary analysis, a strong functional form, such as the linear form, is often used. To overcome these shortcomings, we introduce a more flexible method below.

3.4 The Matching-Smoothing Method (MS-HTE)

The second method, which we call the "matching-smoothing method of estimating heterogeneous treatment effects," or MS-HTE for short, overcomes the assumption of homogeneity within strata in the SM-HTE method. The researcher using this method can retain individual-level information before making cross-individual comparisons to detect heterogeneous treatment effects. In a way, however, the method can be thought of as the limiting case of SM-HTE with the strata being individual treatment-control comparisons at the most elementary level at which such comparisons can be constructed.

A typical approach to matching is to first define treated units (or untreated) units as the target group to be matched and then select appropriate untreated (or treated) units as matches based on closeness in propensity scores. One convenience of this approach is that the researcher can easily obtain TT (or TUT) by aggregating differences over all matches between treated units and untreated units. For convenience, we focus on the treatment group and find a matched control case for each treated case to illustrate the method. The method consists of the following concrete steps:

(1) Estimate propensity score for each unit, as discussed in subsection 3.2 and step (1) of subsection 3.3.

(2) Match treated units to control units with a matching algorithm. We will discuss matching options below. The basic idea is to find a control unit (or control units) that is a good match for each treated unit based on estimated propensity scores. For convenience, the discussions here in (2)-(4) presume one-to-one matching. With one-to-one matching, the data are paired, i.e., with each pair consisting of a treated unit and an untreated unit with almost the same propensity score. When one-to-multiple matching is used instead, the comparison is made to the group mean of multiple matched controls rather than just a single matched control.

(3) Plot the observed difference in a pair between a treated unit and an untreated unit against a continuous representation of the propensity score. While we cannot treat the difference between only two observations in a pair as a true "estimate," it is the building block for the next step of the analysis. In other words, we transform the data so that the differences in pairs between treated units and their matched untreated units constitute the "observed" data subject to further modeling.

(4) Use a nonparametric model to smooth the variation in matched differences to obtain the pattern of treatment effect heterogeneity as a function of the propensity score. That is, we will obtain, typically in a graphic form, a non-parametric smoothed curve for the trend in matched differences as a function of the propensity score. The researcher can then interpret the curve to answer substantive research questions.

Algorithmic matching estimators differ according to the number of matched control units and how multiple matched control units are weighted if more than one control unit is matched to each treated unit (Abadie and Imbens 2009; Morgan and Harding 2006). In one-to-one matching, we match to the nearest neighbor. One can either use replacement or no replacement of controls to match to treated units. We recommend using replacement. The original motivation of matching is to change the observed distribution of the control cases to that of the treatment cases to estimate treatment effects for the treated. As such, many observed units may not be used in matching procedures.

All matching estimators of TT take the following general form:

$$\widehat{TT} = \frac{1}{n_1}\sum_{i}^{n_1}\left[y_{i,d=1} - \sum_{j(i)}^{J(i)} w_{j(i)}\, y_{j(i),d=0}\right], \tag{19}$$

where n_1 is the number of treatment cases, i is the index over treatment cases, *j(i)* is the index over control cases for treated case i (*j(i)*=1,…*J(i)*), and $w_{j(i)}$ is the scaled weight (with sum of one) that measures the relative importance of each control case. For one-to-one matching, $J(i) = 1$, $w_{j(i)} = 1$, treated unit i is matched by a single case, which can be denoted by $y^M_{i,d=0}$. We thus simplify equation (28) to:

$$\widehat{TT} = \frac{1}{n_1}\sum_{i}^{n_1}\left[y_{i,d=1} - y^M_{i,d=0}\right], \tag{20}$$

Two basic variants of matching include nearest neighbor matching and kernel matching. With nearest neighbor matching, treated units are matched to control units that are closest to the treated units in their estimated propensity scores (Rubin 1973a, 1973b). In one-to-multiple matching, $w_{j(i)}$ is set to equal for a given i, i.e., $1/k_i$ for the matched nearest neighbors, where k_i is the number of matches selected for each treated unit. A caliper can be applied to restrict matched control units to some maximum distance from the target treated unit. With kernel matching, all control units are used and weighted according to the distance from the estimated propensity score of the target treated unit (Heckman, Ichimura, and Todd 1997, 1998). New developments in genetic matching offer some potentially useful weights for achieving balance (Diamond and Sekhon 2005; Sekhon and Grieve 2008). Still, there is no clear consensus as to which matching estimator performs best in each application (Morgan and Harding 2006). We compare nearest neighbor matching with 1 and 5 controls to kernel matching. We also compare local polynomial to Lowess smoothing as we examine the variation in the differences between the treated units and their matched control units along the propensity score axis.

So far, we focused on a matching estimator for *TT*. We could instead match treated units to control units to construct an estimate of *TUT*. These different estimators require different independence assumptions, as described in the large literature on matching (Brand and Halaby 2006; Morgan and Harding 2006). Consider the example of the effects of college attendance on women's fertility. For the TT, we assume that the fertility observed for women who did not attend college represents what the fertility of women who attend college would have been had they instead not attended college. For the TUT, we assume the corollary, that the fertility observed for women who attend college represents what the fertility of women who did not attend college would have been had they instead attended college.

As we discussed around equation (18), the ignorability assumption states that there is no bias resulting from using the naive estimator for estimating the treatment effect conditional on the propensity score. This also means that TT and TUT are the same conditional on the propensity score. As a result, the distinction between choosing treated units or untreated units as the target group is of minor consequence for the MS-HTE method. Note that the choice of the target group is consequential for unconditional TT and TUT because the choice dictates the weights over which conditional treatment effects are aggregated over the range of the propensity score. That is, in theory, the MS-HTE provides the same pattern for heterogeneous treatment effects as a function of the propensity score, no matter whether treated units or untreated units are taken as the target group for matching. In practice, results of the two approaches would often differ somewhat due to data limitation, typically at the tails of the propensity score.

4 Empirical Example

To demonstrate the two methods that we discussed earlier, we draw the example from Brand and Davis's [forthcoming] study in analyzing the effects of college attendance on women's fertility. We replicate one segment of their original analysis with our first estimator (SM-HTE) and then demonstrate our second estimator (MS-HTE). In the subsections that follow, we (1) describe our data, (2) report results for our propensity score model, (3) present results for effects of college attendance on women's fertility under an assumption of treatment effect homogeneity, and then (4) discuss results for heterogeneous college effects using SM-HTE and MS-HTE.

4.1 Data Description

We use panel data from the National Longitudinal Survey of Youth 1979 (NLSY), a nationally representative sample of 12,686 respondents who were 14-22 years old when they were first interviewed in 1979. NLSY respondents were interviewed annually through 1994 and are currently interviewed on a biennial basis. We use data gathered from 1979 through 2006. We restrict our sample to women (n = 6,283) who were 14-17 years old at the baseline survey in 1979 (n = 2,736), who had completed at least the 12th grade when they were 19 years old (n = 2,090), who did not have missing data on college attainment (n = 2,013) or fertility in the 2006 survey (n = 1,512). We set these sample restrictions so that all measures we use are pre-college, particularly ability, and to compare college-educated women with women who completed at least a high school education.[4]

[4] We impute missing values for our set of pre-treatment covariates based on all other covariates. Most variables have 1-2% missing values. Only two variables are missing for more than 5% of the sample: parents' income (355 cases) and high school college-preparatory program (135 cases).

The women we lose due to missing data and attrition tend to be from more disadvantaged family backgrounds and levels of achievement than those women we retain.

Our treated group is comprised of women who completed at least the first year of college by age 19, and our control group is comprised of women who completed high school but did not attend college by age 19. Of those women who attend college by age 19, roughly half complete college by age 23 and two-thirds complete by their early 40s. About 40% of non-college attendees attend college later, although less than 14% complete college. Non-college attendees who attend college at some future point represent a distinct treatment group who are on average more disadvantaged than timely college attendees (Rosenbaum, Deli-Amen, and Person 2006). We do not restrict the control group to women who never attend college; we follow Brand and Xie (2007) in this regard and collapse all future paths when assessing the treatment at a particular time. That is, we focus on whether or not a college education occurs at a particular time and remain agnostic about future educational acquisition, allowing the reference to be a composite of future counterfactual paths.

The likelihood of attending college varies by race and ethnicity, social origins, ability,[5] academic achievement, and pre-college fertility in expected directions. Blacks, Hispanics, teenage mothers, and women with disadvantaged social backgrounds and low levels of academic achievement and ability are less likely to go to college than white women, women who are not teen mothers, and women with advantaged social backgrounds and high levels of academic achievement and ability. See Appendix A for descriptive statistics.

4.2 Propensity Score Estimation

The first step in our analysis is to estimate propensity scores for each woman in the sample for the probability of timely college attendance given a set of observed covariates using a probit regression model, predictors being observed pre-college covariates. Table 1 provides results for the propensity model, which support the literature on the determinants of college attendance.

[5] In 1980, 94% of the NLSY respondents were administered the Armed Services Vocational Aptitude Battery (ASVAB), ten intelligence tests measuring knowledge and skill in areas such as mathematics and language. We residualize separately by race and ethnicity each of the ASVAB tests on age at the time of the test, standardize the residuals to mean zero and variance one, and construct a scale of the standardized residuals ($\alpha = .92$) with a mean of zero, a standard deviation of 0.75, and a range of -3 to 3 (Cawley et al. 1997).

Table 1. Propensity Score Probit Regression Models Predicting College Attendance (N=1,512)

Black	-0.133		(0.116)
Hispanic	0.051		(0.158)
Mother's education	-0.136		(0.080)
Mother's education^2	0.009	**	(0.003)
Father's education	0.038	*	(0.018)
Parents' inc. (1979 $1,000s)	-0.202		(0.448)
Intact family	0.038		(0.112)
Number of siblings	-0.040	†	(0.024)
U.S. born	0.353		(0.258)
Rural residence	-0.157		(0.110)
Southern residence	0.283	**	(0.099)
Catholic	-0.018		(0.108)
Jewish	0.292		(0.458)
Mental ability	0.638	***	(0.079)
College-preparatory	0.380	***	(0.098)
Parents' encouragement	0.455	***	(0.119)
Friends' plans	0.058	*	(0.023)
Child by age 18	-1.230	***	(0.240)
Non-missing on covariates	0.001		(0.098)
Constant	-2.437	***	(0.608)
Wald χ^2	299.43		
P > $\chi 2$	0.000		

Notes: Numbers in parentheses are standard errors. Dependent variable is college attendance by age 19 (1) versus high school completion but no college attendance by age 19 (0).

* p <. 05 ** p < .01 *** p < .001 (two-tailed tests)

4.3 Homogenous Effect Estimates

Before turning to our heterogeneous effect estimates, we estimate the effect of education on women's fertility under a, likely unrealistic, assumption of college effect homogeneity. We evaluate the average effect of college attendance by age 19 on number of children by age 41 using a Poisson regression model controlling for the estimated propensity score.[6]

[6] We use a Poisson rather than a negative binomial model because we did not find evidence of overdispersion (i.e., the variance of the outcome is not greater than the mean of the outcome).

Our estimator takes the following form:

$$log\mu_i = \alpha + \delta d_i + \beta p_i, \tag{21}$$

where μ_i is the conditional expected number of children for the i^{th} observation; d_i indicates whether or not a woman attends college; and p_i represents the propensity for college attendance.
We report the estimated average effects in Table 2. The results from Model 1 suggest a 17% reduction in the number of children for college-educated women relative to less-educated women. However, this is a zero-order relationship. Controlling for the propensity for college attendance in Model 2, or factors that might lead to pre-treatment heterogeneity bias, we find an 11% reduction in the number of children women bear by age 41 associated with college attendance. However, these average effects, whether or not we control for factors that predispose women to attend college, conceal underlying systematic heterogeneity in the effects of college attendance shaped by the population composition of college goers. To this heterogeneity issue we now turn.

Table 2. Homogenous Effects of College Attendance on Fertility, Poisson Regression Models (N=1,512)

	Model 1		Model 2	
College Attendan	-0.171	**	-0.106	†
	(0.049)		(0.057)	
Incidence rate ratio	0.843		0.889	
Propensity Score	---		-0.225	*
			(0.108)	
Incidence rate ratio			0.824	
Constant	0.647	***	0.697	***
	(0.024)		(0.034)	
Wald χ^2	12.10		15.59	
$P > \chi 2$	0.001		0.000	

Notes: Numbers in parentheses are standard errors. Dependent variable is number of children by age 41. Propensity scores were generated by a probit regression model of college attendance by age 19 on the set of pre-college covariates.

† p<.10 * p <.05 *** p < .001 (two-tailed tests)

4.4 Heterogeneous Effect Estimates using the SM-HTE

To estimate heterogeneous treatment effects with the stratification-multilevel method (SM-HTE), we first group respondents into propensity score strata such that average values of the propensity score and each covariate do not significantly differ between college and non-college women (p<.001) (Becker and Ichino 2002). The frequency distributions for college and non-college women run in opposite directions: for college-educated women the frequency count increases with the propensity score whereas for non-college-educated women the count decreases, as shown in Table 3. There is, however, overlap within each stratum (i.e., for each propensity score stratum there are both college and non-college educated women).

Two issues may emerge when studying effect heterogeneity with SM-HTE. First, there may not be a sufficient number of treated and control cases within each stratum to estimate level-1 effects. The number of treated cases in the lowest propensity score stratum and the number of control cases in the highest propensity score stratum, the "against the odds" units, pose the most likely problem. There is a tension between achieving balance in the covariate distribution and stability in the estimated effects. We suggest at least 20 treated and 20 control cases within each stratum. For our empirical example, we did not have a sufficient number of non-college goers in the final stratum (i.e., initially we had 14), and therefore collapsed the final two strata and adjusted for the estimated propensity score in the level-1 stratum 5 regression. Second, some covariates may not balance within some strata. We suggest trying different specifications of the propensity score, such as adding interactions and quadratic terms, to achieve balance. But if there is no reasonable adjustment that renders all strata balanced, the analyst may adjust for the unbalanced covariate(s) in the level-1 models. Our indicator of Hispanic ethnicity was not balanced in stratum 1 for the college attendance model and was thus added as a covariate in our level-1 stratum 1 model.

Table 3 provides covariate means by propensity score strata and college attendance. These statistics demonstrate the characteristics of a typical woman within each stratum. For the kth covariate in the jth stratum, we estimate the standardized mean covariate difference to quantify the bias between the treatment and the control groups (DiPrete and Gangl 2004; Morgan and Winship 2007):

$$B_{k,j} = \frac{|\bar{X}_{k,j,D=1} - \bar{X}_{k,j,D=0}|}{\sqrt{\frac{S^2_{k,j,D=1} + S^2_{k,j,D=0}}{2}}} \tag{22}$$

where $\bar{X}_{k,j,D}$ is the sample mean and $S^2_{k,j,D}$ is the sample variance of the kth covariate in the jth stratum for the treated and control groups as indexed by D=(1,0).

Table 3. Covariate Means by Propensity Score Strata and College Attendance (N=1,512)

	Stratum 1			**Stratum 2**			**Stratum 3**			**Stratum 4**			**Stratum 5**		
	[0.0-0.1)			[0.1-0.2)			[0.2-0.4)			[0.4-0.6)			[0.6-1.0]		
	***E(X)* |**	***E(X)* |**		***E(X)* |**	***E(X)* |**		***E(X)* |**	***E(X)* |**		***E(X)* |**	***E(X)* |**		***E(X)* |**	***E(X)* |**	
Variables	***d* =0**	***d* =1**	***B***	***d* =0**	***d* =1**	***B***	***d* =0**	***d* =1**	***B***	***d* =0**	***d* =1**	***B***	***d* =0**	***d* =1**	***B***
Black	0.18	0.07	0.65	0.15	0.22	0.32	0.16	0.13	0.06	0.07	0.10	0.21	0.10	0.06	0.16
Hispanic	0.07	0.35	1.11	0.05	0.05	0.02	0.04	0.04	0.06	0.05	0.02	0.27	0.06	0.02	0.38
Mother's edu.	10.34	9.97	0.45	11.19	11.71	0.23	11.74	11.85	0.07	12.58	12.55	0.09	14.10	14.45	0.16
Father's edu.	9.92	9.98	0.38	11.18	10.73	0.24	11.94	12.55	0.11	12.80	13.17	0.10	15.43	15.35	0.08
Parents' inc/1000	18.07	16.06	0.18	19.95	16.32	0.22	20.22	0.20	0.03	23.21	0.26	0.20	0.29	0.32	0.16
Intact family	0.71	0.77	0.02	0.72	0.72	0.06	0.75	0.62	0.01	0.80	0.89	0.06	0.80	0.85	0.16
Num. of siblings	3.96	3.68	0.27	3.34	3.48	0.07	2.87	2.90	0.22	2.69	2.43	0.18	2.20	2.33	0.01
U.S. born	0.96	0.74	0.51	0.97	0.96	0.10	0.96	0.98	0.04	0.94	0.98	0.27	0.98	0.97	0.09
Rural res.	0.25	0.02	0.54	0.24	0.35	0.22	0.26	0.20	0.09	0.18	0.21	0.11	0.15	0.16	0.06
Southern res.	0.31	0.29	0.47	0.25	0.38	0.26	0.34	0.37	0.12	0.40	0.37	0.15	0.40	0.35	0.16
Catholic	0.28	0.40	0.63	0.30	0.44	0.06	0.35	0.31	0.01	0.37	0.38	0.01	0.39	0.31	0.17
Jewish	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.01	0.14	0.01	0.00	0.13	0.05	0.05	0.02
Mental ability	-0.34	-0.47	0.14	-0.03	0.04	0.16	0.29	0.28	0.04	0.60	0.67	0.09	0.98	1.04	0.03
College-prep.	0.06	0.03	0.06	0.16	0.14	0.08	0.32	0.32	0.01	0.57	0.61	0.15	0.72	0.77	0.02
Parents' enc.	0.44	0.58	0.19	0.68	0.63	0.03	0.83	0.82	0.03	0.92	0.91	0.05	0.95	0.96	0.05
Friends' plans	12.79	13.75	0.38	13.76	13.83	0.14	14.40	14.67	0.16	15.08	14.91	0.03	16.17	15.88	0.11
Child by age 18	0.23	0.11	0.32	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.12
Propensity score	0.05	0.06	0.45	0.15	0.16	0.41	0.28	0.31	0.38	0.48	0.50	0.29	0.73	0.74	0.07
Sample Size	420	21		223	55		245	98		112	116		72	150	

Notes: E(X) |*d* =0 indicates the mean of *X* for women who did not attend college by age 19 but completed high school, and *E(X)* |*d* =1 indicates the mean of *X* for women who attended college by age 19.

The standardized difference is clearly larger in some strata than in others for some covariates, suggesting that what differentiates those who attend college from those who do not differs between more and less advantaged women. Bias between college and non-college goers' characteristics is largest in stratum 1 for our empirical example; we should take this differential bias into consideration when we interpret our results.

We next report results of estimating heterogeneous treatment effects with the SM-HTE. In level-1, we first present the non-parametric results after stratification only, i.e., propensity score stratum-specific college effects on number of children using Poisson regression models:

$$log\mu_{ij} = \alpha_j + \delta_j d_i \tag{23}$$

where all the terms are defined above. Subjects indexed by *i* are nested in propensity score strata indexed by *j*. Separate Poisson regression models are estimated for each propensity score stratum as indicated by the subscript *j*. Intercepts and slopes are allowed to be estimated freely within propensity score strata.

To detect patterns in treatment effects across propensity-score strata, we now take the estimated stratum-specific slopes as observations in a level-2 model. For simplicity, we summarize the pattern in heterogeneous treatment effects across propensity-score strata with the following linear model:

$$\delta_j = \delta_0 + \phi j + \eta_j, \tag{24}$$

where level-1 slopes (δ_j) are regressed on propensity score rank indexed by *j*, δ_0 represents the level-2 intercept (i.e., the predicted value of the effect of college for the lowest propensity women), and ϕ represents the level-2 slope (i.e., the change in the effect of college on fertility with each one-unit change to a higher propensity score stratum). Normality of $\eta_j$is assumed for inference. We use variance-weighted least squares to estimate equation (33), and thus do not assume homogeneity of variances across the δ's. Variances across the δ's come from two sources: sampling variation (due to different sample sizes by group) and true population variance (heteroskedasticity). When we account for varying precision of level-1 slopes estimated within strata due to sampling variation, the level-2 slope estimate is more efficient. Heteroskedasticity in this case is substantively significant, as it suggests that the uncertainty of treatment effects may vary across groups (Raudenbush and Bryk 2002).

Table 4 and Figure 1 report our multilevel model results for heterogeneous effects of college on fertility. These results are virtually the same as those reported in Brand and Davis [forthcoming], except that we constructed the strata based on a different sample size that led to some moderately different estimates. To facilitate implementation of our method, we use the newly developed Stata module --hte-- (Jann, Brand, and Xie 2010; in Stata "ssc install hte"). The level-2 slope indicates a significant decline in the fertility-decreasing effect of college attendance, a difference of 0.12 for each unit change in propensity score rank. Level-1 estimates range from a 61% decrease in the number of children for women with the lowest propensity to attend college (stratum 1), to a 16% decrease in stratum 2, to a 7% increase in the number of children for women with the highest propensity to attend college (stratum 5). Figure 1 summarizes the results in Table 5. “Dots” in Figure 1 represent point estimates of level-1 slopes, stratum-specific Poisson regression effects of college on number of children by age 41. The linear plot in the figure is the level-2 variance-weighted least squares slope. We reverse the *y*-axis to emphasize the fertility-decreasing effect of college.

Table 4. Heterogeneous Effects of College Attendance on Fertility, SM-HTE (N = 1,512)

Level-1 Slopes		
Poisson Regression		
P-Score Stratum 1	-0.944	**
[0.0-0.1)	(0.349)	
Incidence rate ratio	0.389	
P-Score Stratum 2	-0.18	
[0.1-0.2)	(0.150)	
Incidence rate ratio	0.835	
P-Score Stratum 3	-0.165	
[0.2-0.4)	(0.101)	
Incidence rate ratio	0.848	
P-Score Stratum 4	-0.064	
[0.3-0.6)	(0.115)	
Incidence rate ratio	0.938	
P-Score Stratum 5	0.07	
[0.6-1.0]	(0.128)	
Incidence rate ratio	1.072	
Level-2 Slope	0.129	*
Variance Weighted Least Squares	(0.055)	

Notes: Numbers in parentheses are standard errors.

Dependent variable is number of children by age 41. Propensity scores were generated by a probit regression model of college attendance by age 19 on the set of pre-college covariates. Propensity score strata were balanced such that mean values of covariates did not significantly differ between college and non-college goers.

† p < .10 * p <. 05 ** p<.01 (two-tailed tests)

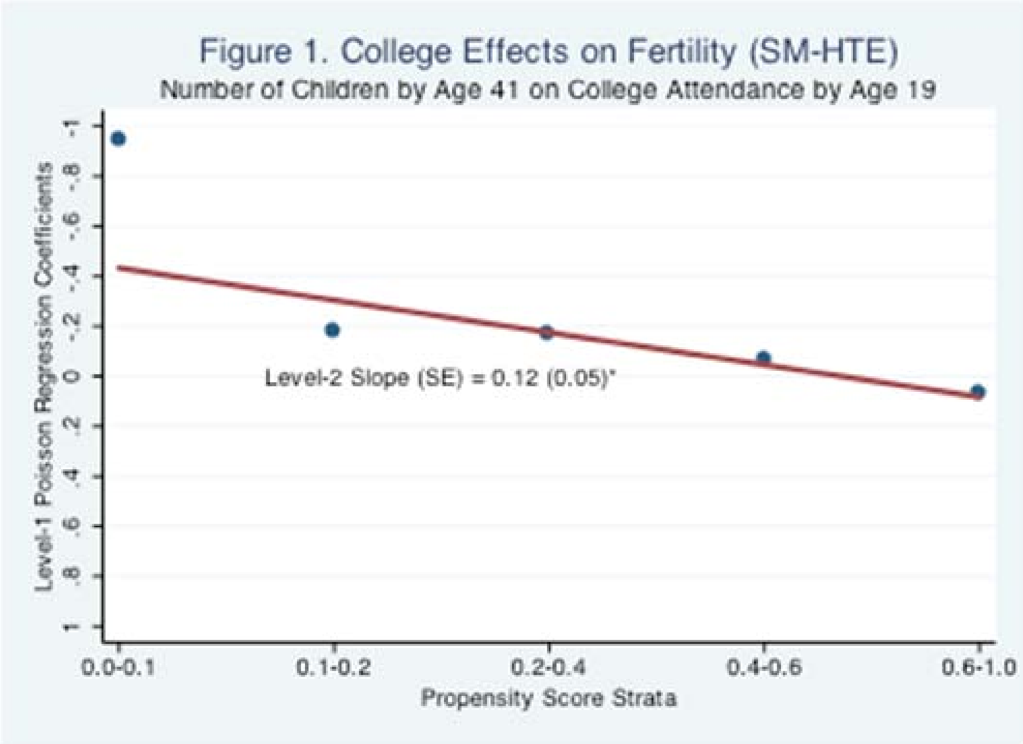


Figure 1. College Effects on Fertility (SM-HTE)

Number of Children by Age 41 on College Attendance by Age 19

A few additional issues about the SM-HTE approach should be noted. First, as shown in Table 4, few of the level-1 estimated effects are significant, despite the significant level-2 slope. Second, there may be differential bias in observed and unobserved factors influencing the treatment and the outcome across propensity score strata. If the bias is greater in stratum 1 than in stratum 5, for example, what can we say about the estimated trend in effects? Perhaps the researcher should resort to sensitivity tests to gauge the susceptibility of the level-2 slope to the presence of stratum-specific omitted variable bias. This issue requires future research.

Finally, Figure 1 depicts the close correspondence between the level-1 college effects on fertility and the level-2 linear slopes. Although this example demonstrates a linear trend in effects of college, linearity is unlikely to hold in most applications (see, for example, Brand and Xie [2010] and Brand 2010). In an actual research setting, linearity should be taken as the first-order approximation of a trend. If an analyst has a larger sample size and more strata, and finds evidence of non-linearity, he or she might fit a quadratic or cubed term in level-2. However, in practice, the researcher would have difficulty identifying higher-order interactions with the SM-HTE approach because typically only a few strata are formed so that only a limited number of δ's are estimated in the stratification step, too limited for the identification of higher-order functions. It is precisely the need for detecting potential non-linearity that motivates our second method, the matching-smoothing (MS-HTE) method, which is non-parametric.

4.5 Heterogeneous Effect Estimates using the MS-HTE

To estimate heterogeneous treatment effects with the matching-smoothing method (MS-HTE), we begin once again by estimating the propensity score for treatment. In our example, we estimate the propensity score of college attendance. The second step is to match treated and control units by the estimated propensity scores and generate differences between treated and control units. As we discussed earlier in subsection 3.4, there are several options for matching treated and control units. We choose three options for illustration and compare the results from these options: (1) nearest neighbor matching with 1 control, (2) nearest neighbor matching with 5 controls, and (3) kernel matching (Leuven and Sianesi 2003). [7] For the main substantive results, we plot the matched differences between treated and control units along a propensity score *x*-axis and fit a smoothed curve. Since the main objective of using the MS-HTE method is to be non-parametric with respect to the pattern of the heterogeneous treatment effects over the range of the propensity score, the researcher would want to use a flexible modeling device to fit the data. In our example, we use both local polynomial smoothing (degree-1, bandwidth 0.2) and Lowess smoothing.

The results from the different smoothing methods yield very similar results. To conserve space, we choose to present results from only one of them. Figure 2 depicts the estimated results for the treatment group with local polynomial smoothing and nearest neighbor matching with 5 controls. The curve for the treatment effect as a function of the propensity score in Figure 2 can be interpreted as a non-parametric regression for the individually matched differences given in Appendix C. In other words, the "raw" data for the second step of smoothing analysis (Figure 2) are differences of matched comparisons in the first step (Appendix C).[8]

Figure 2 differs from Figure 1 in that the *x*-axis is a continuous representation of the propensity score rather than discrete strata. Moreover, we now have a fully non-parametric depiction of treatment effect heterogeneity, rather than the imposition of a functional form on the heterogeneity in effects. In the empirical example, it appears that linearity is a reasonable

[7] Appendix B provides matching estimates of TT and TUT for these various matching algorithms. These estimates suggest heterogeneity in treatment effects, as we observe substantially greater negative effects for the TUT than for the TT, irrespective of which matching algorithm we use, although none of the estimates reflect statistically significant differences. A greater effect for a randomly selected non-college attendee relative to a randomly selected college attendee would support the results from Figure 1, i.e., that the fertility-decreasing effects of college are larger for women with a low propensity than for women with a high propensity for college attendance.

[8] The *y*-axis is wider in Appendix C than in Figure 2 in order to fit all the data points.

functional form. Hence, the substantive conclusion from using either the SM-HTE or the MS-HTE method is the same. That is, we still observe a progressively smaller fertility-decreasing effect of college attendance as women's propensity for college increases. While the data analyzed for this empirical example are well suited to SM-HTE because the trend in effects appears linear, there are surely scenarios in which MS-HTE can be shown to be advantageous over SM-HTE, and vice-versa.

We observe generally the same pattern in Figure 2 using either the local polynomial smoothing or the Lowess smoothing, although we see more curvature at the ends with the latter. In the interest of space, and because results are largely the same, we do not show figures for alternative matching algorithms. In results not shown, we observe a moderately steeper slope using nearest neighbor matching with 5 controls than with kernel matching and we observe a larger effect for women with a low likelihood of college attendance when we apply the method to the untreated rather than the treated. This is due to the increased data mass at the lower end of the propensity score, which makes the smoothing more adaptive in this region.

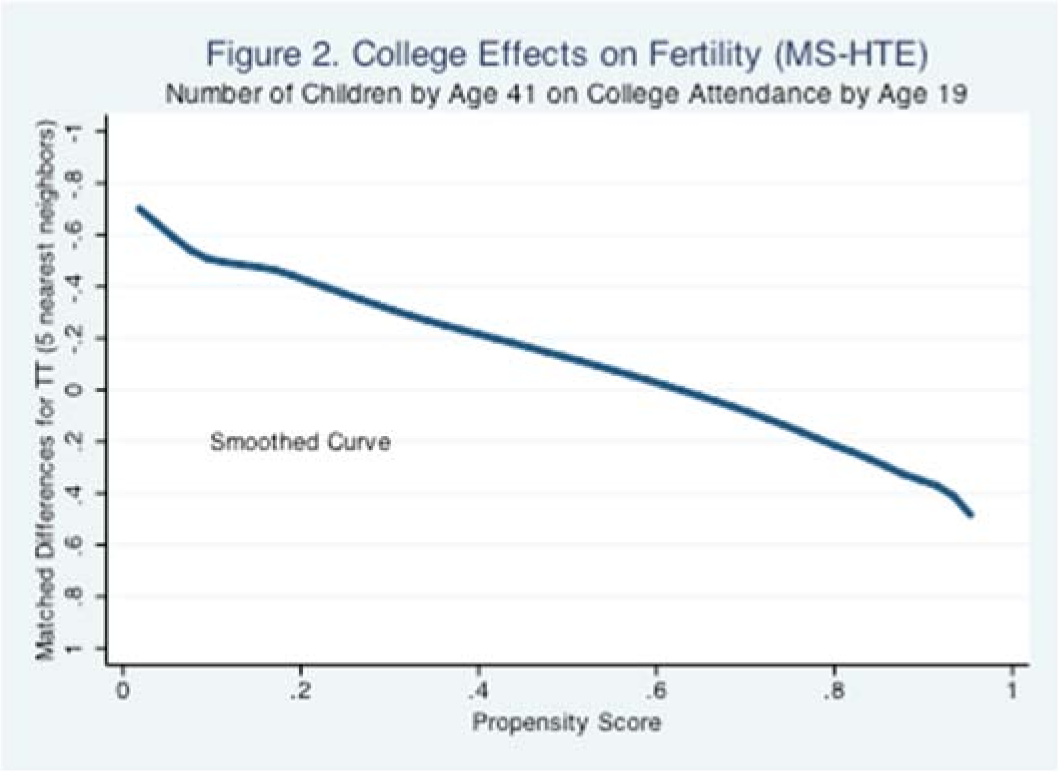


Figure 2. College Effects on Fertility (MS-HTE)
Number of Children by Age 41 on College Attendance by Age 19

5 Discussion and Conclusion

Heterogeneous treatment effects, while widely recognized in sociological research, are seldom studied. If heterogeneity in treatment effects is such that the treatment effect size is correlated with the propensity score, average treatment effects for units at the margin, units being treated, and units not being treated all change when selection criteria for receiving treatment change. This is true even when the proportion receiving treatment simply increases or decreases, as in situations when the pool of treatment expands due to either eligibility criteria becoming lower or incentives becoming higher. By revealing how effects differ among subpopulations, we can contribute to sociological knowledge about the mechanisms through which treatments affect individuals' opportunity structures and enable policy makers to make informed decisions to maximize population benefits.

In this paper, we discuss a practical approach to studying heterogeneous treatment effects, under the same assumption commonly underlying regression analysis: ignorability. We describe two methods within this general approach. For the first method (SM-HTE), we begin by estimating propensity scores for the probability of treatment given a set of observed covariates for each unit and construct balanced propensity score strata; we then estimate propensity score stratum-specific average treatment effects and evaluate a trend across the strata-specific treatment effects across strata. For the second method (MS-HTE), we match control units to treated units based on the propensity score and transform the data into treatment-control comparisons at the crudest level possible; we then estimate treatment effects as a function of the propensity score by fitting a non-parametric model as a smoothing device. There are tradeoffs between the two methods. The first method (SM-HTE) generates stratum-specific estimates that aid the interpretation of treatment effects across strata, which can be compared to regression estimates for the population under an assumption of homogeneity. This method can also provide an estimate of the across-stratum slope, indicating whether or not effect heterogeneity is roughly linear across propensity-score strata. The second method (MS-HTE) does not have these advantages, but overcomes some disadvantages specific to SM-HTE, as it does not assume a global functional form on the heterogeneity in treatment effects. It also allows for heterogeneous treatment effects as a continuous function of the propensity score rather than imposing homogeneity within strata with a sufficient number of observations. We suggest using these methods concurrently to test the corresponding underlying assumptions.

A few comments are in order as to the benefits of the general approach of focusing on observable heterogeneity in treatment effects. First, while the ignorability assumption is unlikely to be true for most sociological applications, its plausibility depends on how rich the covariates are and is thus a substantive issue in actual research rather than a methodological question that can be debated in general. Second, we assume ignorability only so as to see how much we can learn from the data. Without this assumption, strong parametric assumptions are needed about unobservable variables (Heckman 1978; Willis and Rosen 1979). Third, we can always revisit the assumption of ignorability after the analyses are conducted (Harding 2003; Rosenbaum 2002; Xie and Wu 2005). Fourth, ignorability is not incompatible with the common notation of selection bias. Indeed, we hypothesize that much selection bias may reveal unobserved heterogeneity in selection into treatment. Our approach facilitates investigation of heterogeneity bias across the observed likelihood of treatment. Our work on heterogeneous treatment effects therefore complements a large literature that capitalizes on unobserved variables and identification strategies through parametric assumptions and instrumental variables (Heckman 1978; Heckman, LaLonde, and Smith 1999; Heckman, Urzua, and Vytlacil 2006; Willis and Rosen 1979). Finally, while the kind of heterogeneity in treatment effects we discuss is potentially observable in empirical research using regression analyses without any additional assumptions, it does not mean that it is *actually* observed or reported in empirical research. That is, while treatment effect heterogeneity under ignorability has long been recognized and accepted, few researchers actually examine patterns of treatment effect heterogeneity by observed covariates. We suspect that lack of ready-to-use statistical methods is a reason why heterogeneous treatment effects are not routinely checked and reported.

There are alternative methods to those we discussed in this paper. One possibility is stratification by key covariates, allowing the interaction of treatment and certain covariates that are believed to be of primary importance in heterogeneous treatment effects, such as gender and race. Another related method is to test cross-level interactions: estimation of a hierarchical linear model that allows the regression coefficient at the lower level to vary across higher-level units (Raudenbush and Bryk 1986, 2002). However, for the question of whether there are potential selection biases, i.e., systematic differences between the treatment group and the control group, the interaction between the propensity score and the treatment indicator is the only interaction that should concern the researcher. In this paper, we discussed methods that can be used to detect this important interaction pattern, under the *same* assumption that underlies most of the empirical analyses currently practiced in sociology, no matter whether they are interested in homogeneous effects or interaction effects. That is, while we maintain the ignorability assumption, we relax the strict homogeneity assumption.

Of course, a study of heterogeneous effects using methods discussed in this paper does not solve the main methodological challenge facing empirical researchers: selection on the unobservables. Thus, the methods we proposed in this paper are limited, only because they use the same information and presume the same assumption as conventional methods. However, without any additional assumption or additional data, the new methods yield new information of potential importance that is often overlooked in empirical research. Given that no more new data or assumptions are required for the methods being proposed here, continuing the practice of ignoring this kind of information seems unwarranted. Thus, we recommend that researchers use the methods we have proposed here in their empirical work, if not to test theoretically derived hypotheses about heterogeneous treatment effects, then merely as a new way to explore and better understand their empirical data.

REFERENCES

Abadie, Alberto and Guido Imbens. 2009. “Matching on the Estimated Propensity Score.” Unpublished manuscript.

Angrist, J.D. 1990. “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records.” *American Economic Review* 80:313-335.

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): 444-455.

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

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

Ansari, Asim and Jedidi Kamel. 2000. “Bayesian Factor Analysis for Multilevel Binary Observations.” *Psychometrika* 65:475-496.

Bauer, Daniel J. and Patrick J. Curran. 2003. “Distributional Assumptions of Growth Mixture Models: Implications for Overextraction of Latent Trajectory Classes*.” Psychological Methods* 8:338-363.

Becker, Sascha and Andrea Ichino. 2002. “Estimation of Average Treatment Effects Based on Propensity Scores.” *Stata Journal* 2:358-377.

Bjorklund, 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.

Bound, J., D.A. Jaeger, and R.M. Baker. 1995. “Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak.” *Journal of the American Statistical Association* 90(430):443-450.

Brand, Jennie E. 2010. "Civic Returns to Higher Education: A Note on Heterogeneous Effects." *Social Forces* 89(2): 417-433.

Brand, Jennie E. and Dwight Davis. [forthcoming, 2011]. "The Impact of College Education on Fertility: Evidence for Heterogeneous Effects." *Demography.*

Brand, Jennie E. and Charles N. Halaby. 2006. “Regression and Matching Estimates of the Effects of Elite College Attendance on Educational and Career Achievement.” *Social Science Research* 35:749-770.

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

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.

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

Cawley, John, Karen Conneely, James Heckman, and Edward Vytlacil. 1997. “Cognitive Ability, Wages, and Meritocracy.” Pp. 179-192 in *Intelligence, Genes, and Success: Scientists Respond to the Bell Curve,* edited by B. Devlin, S. E. Feinberg, D. Resnick, and K. Roeder. New York: Springer.

Cochran, W.G. 1972. "Observational Studies." Pp. 77-90 in *Statistical Papers in Honor of George Snedecor,* edited by T.A. Bancroft. Ames, IA: Iowa State University Press.

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.

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.

Diamond, Alexis and Jasjeet Sekhon. 2005. “Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies.” Unpublished manuscript.

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.

Duncan, Otis Dudley, Magnus Stenbeck, and Charles Brody. 1988. “Discovering Heterogeneity: Continuous versus Discrete Latent Variables.” *American Journal of Sociology* 93:1305-21.

D'Unger, A. V., Kenneth C. Land, P. L. McCall, and Daniel S. Nagin. 1998. “How Many Latent Classes of Delinquent/criminal Careers? Results from Mixed Poisson Regression Analyses.” *American Journal of Sociology* 103:1593-1630.

Frangakis, Constantine E and Donald B. Rubin. 2002. “Principal Stratification in Causal Inference.” *Biometrics* 58 (1):21–29.

Gangl, Markus, 2010. “Causal Inference in Sociological Research.” *Annual Review of Sociology* 36:21-47.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis,* second edition. Boca Raton, FL: Chapman & Hall/CRC.

Greenland S, and C. Poole. 1988. “Invariants and Noninvariants in the Concept of Interdependent Effects.” *Scandinavian Journal of Work, Environment & Health* 14:125-129.

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. 1978. “Dummy Endogenous Variables in a Simultaneous Equation System.” *Econometrica* 46(4): 931-959.

Heckman, James J. 2001. “Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture.” *Journal of Political Economy* 109: 673-748.

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

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economics and Statistics* 64:605-654.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economics and Statistics* 65:261-294.

Heckman, James J., R. LaLonde, and J. Smith. 1999. “The Economics and Econometrics of Active Labor Market Programs.” Pp.1865-2097 in *Handbook of Labor Economics* vol. 3, edited by A. Ashenfelter and D. Card. New York: Elsevier Science.

Heckman, James J. and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." Pp.156-245 in *Longitudinal Analysis of Labor Market Data,* edited by James Heckman and Burton Singer. Cambridge: Cambridge University Press.

Heckman, James and B. Singer. 1984. "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data." *Econometrica* 52:271-320.

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.

Hedges, Larry V. 1982. "Fitting Categorical Models to Effect Sizes from a Series of Experiments." *Journal of Education Statistics* 7:119-137.

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

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

Imbens, Guido W. and Joshua Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62:467-476.

Jann, Ben, Jennie E. Brand, and Yu Xie. 2010. "hte – Stata module to perform heterogeneous treatment effect analysis by propensity score strata with a multilevel model." in *Stata*: ssc install hte (http://econpapers.repec.org/software/bocbocode/s457129.htm).

Leuven, Edwin and Barbara Sianesi. 2003. "psmatch2 – Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing." in *Stata*: ssc install psmatch2 (http://econpapers.repec.org/software/bocbocode/s432001.htm).

Lubke, Gitta H. and Bengt Muthen. 2005. "Investigating Population Heterogeneity with Factor Mixture Models." *Psychological Methods* 10:21-39.

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

Manski, Charles. 2007. *Identification for Prediction and Decision.* Cambridge: 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.

Moffitt, Robert. 1996. "Identification of Causal Effects Using Instrumental Variables: Comment." *Journal of the American Statistical Association* 91:462-465.

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.

Muthén, B. and L. K. Muthén. 2000. Integrating Person-centered and Variable-centered Analyses: Growth Mixture Modeling with Latent Trajectory Classes. *Alcoholism-Clinical and Experimental Research* 24(6):882-891.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. Second Edition. New York: Cambridge University Press.

Powers, Daniel and Yu Xie. 2008. *Statistical Methods for Categorical Data Analysis, 2nd Edition*. Academic Press.

Rasch, Georg. 1966. "An Individualistic Approach to Item Analysis." Pp. 89-108 in *Readings in Mathematical Social Science,* edited by Paul F. Lazarsfeld and Neil W. Henry. Chicago: Science Research Associates, Inc.

Raudenbush, S. and A. S. Bryk. 1986. "A Hierarchical Model for Studying School Effects." *Sociology of Education* 59(1):1-17.

Raudenbush, Stephen W. and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods* . Newbury Park, CA : Sage.

Rosenbaum, James E., Regina Deli-Amen, and Ann E. Person. 2006. *After Admission: From College Access to College Success.* New York: Russell Sage Foundation.

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.

Rothman, K.J. and S. Greenland, eds. 1998. *Modern Epidemiology*, 2nd Edition. Lippincott-Raven Publishers: Philadelphia, PA.

Rubin, Donald. 1973a. "Matching to Remove Bias in Observational Studies." *Biometrics* 29:159-83.

Rubin, Donald. 1973b. "The Use of Matched Sampling to and Regression Adjustment to Remove Bias in Observational Studies." *Biometrics* 29:185-203.

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.

Sekhon, Jasjeet and Richard Grieve. 2008. "A New Non-Parametric Matching Method for Bias Adjustment with Applications to Economic Evaluations." Unpublished manuscript.

Sobel, Michael E. 2000. "Causal Inference in the Social Science." *Journal of the American Statistical Association* 95: 647-651.

Tsai, Shu-Ling and Yu Xie. 2008. ''Changes in Earnings Returns to Higher Education in Taiwan since the 1990s.'' *Population Review* 47:1–20.

Tsai, Shu-Ling and Yu Xie. Forthcoming. "Heterogeneity in Returns to College Education: Selection Bias in Contemporary Taiwan." *Social Science Research*.

Vermunt, Jeroen K. 2002. "Latent Class Analysis of Complex Sample Survey Data: Application to Dietary Data." *Journal of the American Statistical Association* 97(459):736-737.

Vermunt, Jeroen K. 2003. "Multilevel Latent Class Models." *Sociological Methodology 33*(1):213-239.

Willis, Robert J. and Sherwin Rosen. 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.

Wu, Xiaogang. 2009. “Voluntary and Involuntary Job Mobility and Earnings Inequality in Urban China, 1993-2000.” *Social Science Research* 39: 382-395.

Xie, Yu. 2000. “Assessment of the Long-Term Benefits of Head Start.” Pp.139-167 in *Into Adulthood: A Study of the Effects of Head Start*, edited by Sherri Oden, Lawrence J. Schweinhart, and David P. Weikart. Ypsilanti, MI: High/Scope Press.

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

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

Appendix A. Descriptive Statistics of Pre-College Covariates and Fertility by College Attendance, NLSY Women (N=1,512)

	No College Attendance by Age 19		College Attendance by Age 19	
Variables	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
Race				
Black (0-1)	0.149	0.356	0.100	0.301
Hispanic (0-1)	0.055	0.228	0.035	0.185
Family background				
Mother's education (years)	11.402	2.286	13.050	2.430
Father's education (years)	11.425	3.095	13.636	3.247
Parents' income (1979 dollars)	20418	11636	26143	13139
Intact family age 14 (0-1)	0.737	0.440	0.800	0.401
Number of siblings	3.285	2.195	2.610	1.686
U.S. born (0-1)	0.960	0.197	0.971	0.167
Rural residence, age 14 (0-1)	0.235	0.423	0.197	0.167
Southern residence, age 14 (0-1)	0.321	0.466	0.360	0.476
Catholic (0-1)	0.322	0.467	0.345	0.476
Jewish (0-1)	0.005	0.071	0.025	0.156
Ability and academics				
Mental ability*	0.089	0.616	0.659	0.564
College-prep. (0-1)	0.253	0.423	0.559	0.492
Social-psychological				
Parents' enc. college (0-1)	0.678	0.457	0.881	0.322
Friends' plans (years schooling)	13.905	2.060	15.133	1.900
Fertility history				
Had a child by age 18 (0-1)	0.084	0.276	0.004	0.063
Fertility				
Number of children age 41	1.909	1.301	1.610	1.246
Sample Size	1072		440	
Weighted Sample Prop.	0.68		0.32	

Notes: Ability is measured with a scale of standardized residuals of the ASVAB. All statistics are weighted by a NLSY panel weight.

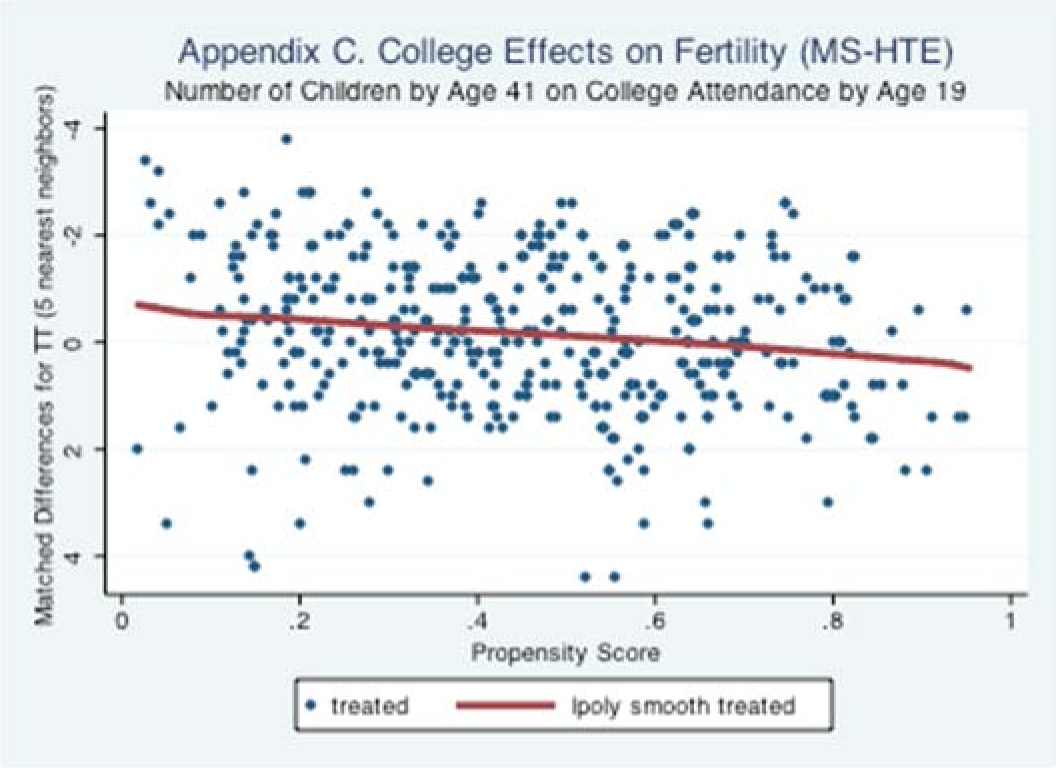
Appendix B. Matching Estimates of Effects of College Attendance on Fertility (N = 1,512)

	TT	TUT
Nearest neighbor matching, 1 contr	-0.123 (0.123)	-0.24
Nearest neighbor matching, 5 contr	-0.157 (0.095)	-0.49
Kernel matching	-0.099 (0.090)	-0.41

Notes: Numbers in parentheses are standard errors.

Dependent variable is number of children by age 41. Treatment is college attendance by age 19. Propensity scores were generated by a probit regression model of college attendance on the set of pre-college covariates.

† p < .10 * p <. 05 ** p<.01 (two-tailed tests)



Appendix C. College Effects on Fertility (MS-HTE)

Number of Children by Age 41 on College Attendance by Age 19

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/