Causal Effect Heterogeneity

Jennie E. Brand
Juli Simon-Thomas

PWP-CCPR-2012-005

March 2012

California Center for Population Research
On-Line Working Paper Series

CAUSAL EFFECT HETEROGENEITY

JENNIE E. BRAND
University of California – Los Angeles

JULI SIMON-THOMAS
University of California – Los Angeles

In preparation for *Handbook of Causal Analysis for Social Research*, Springer Series
Edited by Stephen L. Morgan

Version: March 30, 2012
Word count (including abstract and footnotes, excluding references): XXX
References word count: XXX

* Direct all correspondence to Jennie E. Brand, Department of Sociology, University of California – Los Angeles, 264 Haines Hall, Los Angeles, CA 90095-1551, USA; email: brand@soc.ucla.edu; phone: 310.206.1049; FAX: 310.206.9838. This research made use of facilities and resources at the California Center for Population Research, UCLA, which receives core support from the National Institute of Child Health and Human Development, Grant R24HD041022. Simon-Thomas was supported by a UCLA Graduate Summer Research Mentorship program, the Institute for Research on Labor and Employment, and a pre-doctoral advanced quantitative methodology training grant (R305B080016) awarded to UCLA by the Institute of Education Sciences of the U.S. Department of Education. This research was conducted with restricted access to the Bureau of Labor Statistics (BLS) data. The views expressed here do not necessarily reflect the views of the BLS. We thank Yu Xie and Ben Jann for their contributions to related work, and Stephen Morgan and Judea Pearl for comments and suggestions. The ideas expressed herein are those of the authors.

AUTHOR BIOGRAPHY PAGE

JENNIE E. BRAND

Jennie E. Brand is Associate Professor of Sociology at the University of California – Los Angeles and 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 socioeconomic and social-psychological consequences of job displacement.

JULI SIMON-THOMAS

Juli Simon-Thomas is a Ph.D. student at the University of California – Los Angeles, a student affiliate at the California Center for Population Research, and a Fellow of the Institute for Education Sciences' Advanced Quantitative Methods program. Her research centers on the ways in which stratification and the structure of educational systems in the U.S. and Western Europe affect the ways societies – and groups within societies – view and alter education. Methodologically, she focuses on causal inference.

CAUSAL EFFECT HETEROGENEITY

Abstract

Individuals differ not only in background characteristics, often called "pre-treatment heterogeneity," but also in how they respond to a particular treatment, event, or intervention. For causal inference in the social sciences, a principal interaction for understanding selection bias is between the treatment of interest and the propensity of treatment. Although the importance of "treatment effect heterogeneity," so defined, has also been widely recognized in the causal inference literature, empirical quantitative social science research has not fully absorbed these lessons. In this chapter, we describe key estimation strategies for the study of heterogeneous treatment effects; discuss recent research in education that attends to causal effect heterogeneity, and what we gain from such attention; and demonstrate the methods we discuss with an example of the effects of college on civic participation. The primary goal of this chapter is to encourage researchers to routinely examine treatment effect heterogeneity with the same rigor that they devote to pre-treatment heterogeneity.

Keywords: heterogeneity; propensity score matching; instrumental variables; education; civic engagement;

CAUSAL EFFECT HETEROGENEITY

1 Introduction

As attention to questions of causality increasingly occupy social science research, so too has attention to underlying heterogeneity across individuals or other units of analysis. Individuals differ not only in background characteristics, often called "pre-treatment heterogeneity," but also in how they respond to a particular treatment, event, or intervention (Angrist and Krueger 1999; Elwert and Winship 2010; Gangl 2010; Holland 1986; Heckman and Robb 1985; Heckman, Urzua, and Vytlacil 2006; Moffit 2008; Morgan and Winship 2007, 2012; Winship and Morgan 1999; Xie 2011; Xie, Brand, and Jann [forthcoming]). Causal effects should vary across members of a society; it is implausible to assume that different members of a population respond identically to the same treatment condition.

A simple approach to studying variation in causal effects is to examine interactions between the cause or treatment of interest and specific covariates, such as gender or race. For example, we may want to estimate the effect of college on wages, and believe that college effects differ for blacks and whites. Examining effect variation via interaction terms is a straightforward practice in quantitative social science research, although such interactions are perhaps not incorporated as routinely as we might expect (Elwert and Winship 2010; Morgan and Winship 2007, 2012; Xie 2011; Xie, Brand, and Jann [forthcoming]). However, for causal inference and the assessment of selection bias in the social sciences, the subject of this volume, a principal interaction is between the treatment of interest and the propensity of treatment (Heckman, Urzua, and Vytlacil 2006; Xie 2011).

In this chapter, we are concerned with the estimation of the interaction between the treatment and the propensity of treatment. We refer to this – distinguishing it from general covariate interaction – as "treatment effect heterogeneity." Although the importance of treatment effect heterogeneity, so defined, has also been widely recognized in the causal inference literature (Morgan and Winship 2007), empirical quantitative social science research has not fully absorbed these lessons. Yet the study of effect heterogeneity *should* figure prominently in social science research. If there is treatment effect heterogeneity, average treatment effects can vary widely depending on the population composition of the treated and thus, despite common beliefs, simple averages do not have a straightforward interpretation (Angrist 1998; Elwert and Winship 2010; Morgan and Todd 2008; Morgan and Winship 2007, 2012; Xie, Brand, and Jann 2011; Xie 2011).

In addition to attending to matters of selection, effect heterogeneity analyses can yield important insights as to the distribution of scarce social resources in an unequal society and to social policies (Brand 2010; Brand and Davis 2011; Brand and Xie 2010). We can answer such questions as who is most likely to receive a desired social good, and whether they are the optimal beneficiaries under given circumstances. Many policies, such as increasing or decreasing college tuition, free or subsidized immunization for children, subsidized housing, food stamps, Head Start, and increasing or decreasing enrollments at select colleges are issued without regard to group characteristics. In these and many other cases, different marginal persons are "recruited" into treatment as policies, and thus eligibility thresholds, are introduced or revised (Xie 2011). If policymakers understand patterns of treatment effect heterogeneity, they can more effectively assign different treatments to individuals so as to balance competing objectives, including reducing cost and maximizing average outcomes for a given population.

The primary goal of this chapter is to encourage researchers to routinely examine treatment effect heterogeneity with the same rigor they devote to pre-treatment heterogeneity. The chapter has four main sections. The first section describes key estimation strategies for the study of heterogeneous treatment effects. The second section discusses recent research in education that attends to causal effect heterogeneity, what we gain from such attention, and how we reconcile discrepant findings across methods. We focus on education because it is a particularly well-developed example of the study of effect heterogeneity in sociology and economics. The third section offers an empirical demonstration of estimating heterogeneous college effects on civic participation. We summarize and conclude our discussion in the fourth section.

2 Methods for Estimating Heterogeneous Treatment Effects

In this section, we first review pre-treatment heterogeneity. We then discuss treatment effect heterogeneity and a range of analytic approaches for estimating heterogeneous treatment effects under different assumptions: weighted regressions and propensity score matching to recover subpopulation treatment effects; stratification-multilevel, matching-smoothing, and smoothing-differencing for estimating effects across the propensity score distribution; and instrumental variables for estimating local average and marginal treatment effects.

2.1 Pre-Treatment Heterogeneity

We begin by considering a binary treatment, such as receiving a college education or losing one's job, and then 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 $d = 0$). Let Y denote an

outcome variable of interest, and y_i^1 denote the i^{th} member's potential outcome if treated and y_i^0 the i^{th} member's potential outcome if untreated. We 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. We, however, observe only y_i^1 if $d_i = 1$ or y_i^0 if $d_i = 0$ and thus can never compute individual-level treatment effects (Holland 1986).[1] However, due to population heterogeneity, there is no guarantee that the group that receives the treatment is comparable, in observed and unobserved contextual and individual characteristics, to the group that does not receive the treatment. In some cases, individuals may select or be selected into treatment based on anticipated costs and benefits of treatment, or as a result of structural or socioeconomic circumstances (Brand and Xie 2010; Heckman 2001, 2005). For example, children from advantaged families who enroll in college-preparatory classes would be incomparable to more disadvantaged children who do not enroll in high-track classes without an adequate control for family socioeconomic resources and early achievement.

We can decompose the expectation for the two counterfactual outcomes as follows:

$$E(y^1) = E(y^1 \mid d = 1)P(d = 1) + E(y^1 \mid d = 0)P(d = 0) \tag{2}$$

and

$$E(y^0) = E(y^0 \mid d = 1)P(d = 1) + E(y^0 \mid d = 0)P(d = 0). \tag{3}$$

[1] Extending to multi-valued treatments with *j* values, the observed outcome variable contains only 1/*J* of the information in the potential outcome random variable, rather than ½ in the binary treatment set-up. In other words, the proportion of unobservable counterfactual states increases as the number of treatment values *j* increases, such that we have a matrix of potential outcomes with j^2 cells, only the diagonal of which are observed [see Morgan and Winship (2007) for a detailed discussion].

What we observe from the data are: $E(y^1|d=1)$, $E(y^0|d=0)$, $P(d=1)$, and $P(d=0)$. If there is selection bias,

$$E(y^1 \mid d=1) \neq E(y^1 \mid d=0) \neq E(y^1) \tag{4}$$

and/or

$$E(y^0 \mid d=1) \neq E(y^0 \mid d=0) \neq E(y^0). \tag{5}$$

Thus, with observational data, and when selection bias is present, it is clear that the independence condition,

$$d \coprod (y^1, y^0), \tag{6}$$

does not hold because subjects are 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 use two strategies. First, they may control for relevant pre-treatment covariates and assume conditional independence (also called "ignorability," "unconfoundedness" or "selection on observables"):

$$d \coprod (y^1, y^0) \mid X, \tag{7}$$

where $\boldsymbol{X}$ denotes a vector of observed covariates. The ignorability condition is held as an unverifiable assumption. The plausibility hinges on the mechanism governing exposure or assignment to the different values of a given cause. Substantive knowledge about the subject matter needs to be considered before a researcher can entertain the ignorability assumption. Measurement of theoretically meaningful confounders makes ignorability tentatively plausible, but not necessarily true.[2] Pearl (2009) also provides conditions for including covariates as appropriate controls. The researcher can also assess the plausibility of the ignorability

[2] Repeated observations of units of analysis can be used in fixed effects models to control for time-invariant unobserved properties of units, increasing the plausibility of the assumption.

assumption through sensitivity or auxiliary analyses (Harding 2003; Rosenbaum 2002). 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 $\boldsymbol{X}$. Thus, equation (7) is changed to

$$d \coprod \left(y^1, y^0\right) \mid P(d=1 \mid X), \tag{8}$$

where $P(d=1|\boldsymbol{X})$ is the propensity score, the probability of treatment that summarizes all the relevant information in covariates $\boldsymbol{X}$, estimated by a probit or logit regression model. The literature on propensity score methods recognizes the utility of the propensity score as a solution to data sparseness in a finite sample (Morgan and Harding 2006).

Second, researchers can capitalize on an "instrumental" variable (or variables) (IV) to address nonrandom treatment assignment. A valid IV is an exogenous factor that causes at least some of the variation in treatment status and affects the outcome only indirectly through treatment (Angrist and Krueger 1999; Angrist, Imbens, and Rubin 1996; Angrist and Piscke 2009; Bound, Jaeger, and Baker 1995; Heckman, Urzua, and Vytlacil 2006; Morgan and Winship 2007). Identifying a valid IV is a difficult task; a weak IV may give rise to imprecise IV estimates and lead to biased estimates in finite samples (Bound, Jaeger, and Baker 1995).

2.2 Treatment Effect Heterogeneity

An important development of the causal inference literature is the recognition that treatment effects are likely to be heterogeneous (e.g., Angrist and Krueger 1999; Holland 1986; Heckman and Robb 1985; Heckman, Urzua, and Vytlacil 2006; Morgan and Winship 2007, 2012; Winship and Morgan 1999). For example, colleges may select persons who gain more (Willis and Rosen 1979; Carneiro, Heckman, and Vytlacil 2011) or less (Brand and Xie 2010)

than persons who do not attend college. This example underscores the kind of heterogeneity that does not merely reflect group differences at the baseline that can be "controlled for" by covariates or fixed effects. In other words, it reflects treatment effect, rather than just pre-treatment, heterogeneity.[3]

As we note above, researchers are sometimes concerned with stratification by selected covariates, allowing the interaction of treatment and certain covariates that are believed to be of primary importance, such as gender and race. However, the interaction between the propensity score and the treatment indicator is, of course, the key interaction for questions of variation by selection into treatment (Heckman and Robb 1985; Heckman, Urzua, and Vytlacil 2006; Morgan and Winship 2007; Xie 2011). The recognition that treatment effects may vary by the probability of treatment has led to new methods of causal inference and to refined interpretations of effect estimates derived from existing methods (Angrist 1998; Brand and Xie 2010; Elwert and Winship 2011; Morgan and Todd 2008; Morgan and Winship 2007; Xie 2011; Xie, Brand, and Jann [forthcoming]). Despite widespread belief by practitioners, traditional regression estimates do not represent straightforward averages of individual-level causal effects if individual-level variation in the causal effect of interest is not completely random. Instead, they give a peculiar type of average – a conditional variance weighted average of the heterogeneous individual-level effects, where different weights, generated by different population composition,

[3] Some question (e.g., Pearl 2010) whether treatment effect heterogeneity is a "problem" to be solved by the inclusion of all relevant variables. We contend that, rather than a "problem" to be solved, effect heterogeneity is an inherent property of units in a population science. We concede that if we could observe and perfectly predict differential effects of treatment at baseline, we would know treatment effect heterogeneity in advance. However, there are few if any social phenomena in which we can predict who will gain more or less from a treatment until that treatment is experienced. Our task is to uncover differential treatment response resulting from population heterogeneity, an important task for all the reasons identified above.

can produce widely different effect estimates (Angrist 1998; Elwert and Winship 2011; Morgan and Winship 2007; Xie 2011).

Let us define several parameters central to the causal inference literature and to assessing heterogeneity, beginning with the difference between a randomly selected set of individuals in U who were treated to another randomly selected set of individuals who were untreated, i.e. the "Average Treatment Effect" (*ATE*):

$$\bar{\delta}_{ATE} = E(y^1 - y^0) \tag{9}$$

A general definition of causal effect heterogeneity is when:

$$\bar{\delta}_{ATE} \neq \delta_i, \tag{10}$$

i.e. the treatment effect differs across individuals. In this case, conventional regression coefficients have equivocal interpretations. Let us define the average difference among those individuals who were actually treated, the "Treatment Effect of the Treated" (*TT*):

$$\bar{\delta}_{TT} = E(y^1 - y^0 \mid d = 1). \tag{11}$$

And let us define the average difference among those individuals who were not treated, the "Treatment Effect of the Untreated" (*TUT*):

$$\bar{\delta}_{TUT} = E(y^1 - y^0 \mid d = 0). \tag{12}$$

With independence, $\bar{\delta}_{ATE} = \bar{\delta}_{TT} = \bar{\delta}_{TUT} = E(y^1 \mid d = 1) - E(y^0 \mid d = 0)$. Here we define treatment effect heterogeneity more specifically, compared to the general definition given by (10), as:

$$\bar{\delta}_{ATE} \neq \bar{\delta}_{TT} \neq \bar{\delta}_{TUT}. \tag{13}$$

To see how selection into treatment may cause biases in estimates of treatment effects, we use the following abbreviated notations, as described in Xie, Brand, and Jann [forthcoming]:

p = the proportion treated (i.e., the proportion of units $d = 1$),
q = the proportion untreated (i.e., the proportion of units $d = 0$),

$$
\begin{aligned}
E(y^1_{d=1}) &= E(y^1 \mid d=1),\\
E(y^0_{d=1}) &= E(y^0 \mid d=1),\\
E(y^1_{d=0}) &= E(y^1 \mid d=0),\\
E(y^0_{d=0}) &= E(y^0 \mid d=0).
\end{aligned}
$$

Now we decompose $\overline{\delta}_{ATE}$:

$$
\begin{aligned}
\overline{\delta}_{ATE} &= E(y^1 - y^0)\\
&= E(y^1_{d=1})p + E(y^1_{d=0})q - E(y^0_{d=1})p - E(y^0_{d=0})q\\
&= E(y^1_{d=1}) - E(y^1_{d=1})q + E(y^1_{d=0})q - E(y^0_{d=1}) + E(y^0_{d=1})q - E(y^0_{d=0})q\\
&= [E(y^1_{d=1}) - E(y^0_{d=0})] - [E(y^0_{d=1}) - E(y^0_{d=0})] - (\overline{\delta}_{TT} - \overline{\delta}_{TUT})q.
\end{aligned}
\tag{14}
$$

Noting that the simple estimator from observed data is $E(y^1_{d=1}) - E(y^0_{d=0})$, we see two sources of bias for $\overline{\delta}_{ATE}$, both of which are selection biases that may threaten the validity of causal inference with observational data (see also Morgan and Winship (2007), eq. 2.12):

(1) The average difference between the two groups in the absence of treatment, $E(y^0_{d=1} - y^0_{d=0})$, or pre-treatment heterogeneity bias.

(2) The difference in the average treatment effect between the two groups, $\overline{\delta}_{TT} - \overline{\delta}_{TUT}$, weighted by the proportion untreated *q*, or treatment-effect heterogeneity bias.

Therefore, when treatment effects are heterogeneous, an average treatment effect for a population is a weighted average of varying treatment effects, a quantity that depends on population composition (Xie 2011; Xie, Brand, and Jann [forthcoming]). Standard non-experimental evaluation methods, including the fixed effects estimator, eliminate only the first but not the second form of bias.

There have been a few primary approaches to estimating heterogeneous treatment effects in the literature on causal inference. A simple and straightforward approach is to assume ignorability, and to find empirical patterns of treatment effect heterogeneity as a function of observed covariates through the difference between $\overline{\delta}_{TT}$ and $\overline{\delta}_{TUT}$ by way of weighted

regressions (Morgan and Todd 2008) or propensity score matching (Abadie and Imbens 2009; Brand and Halaby 2006; Morgan 2001), or by statistical modeling to explore empirical patterns of effect heterogeneity as a function of the propensity score (Brand 2010; Brand and Davis 2011; Brand, Pfeffer, and Goldrick-Rab 2012; Brand and Simon-Thomas 2011; Brand and Xie 2010; Musick, Brand, and Davis 2012; Tsai and Xie 2008; Xie 2011; Xie, Brand, and Jann [forthcoming]).; Xie and Wu 2005). The plausibility of the (unverifiable) ignorability depends on the richness of the empirical data. The researcher can always evaluate the assumption through sensitivity or auxiliary analyses (DiPrete and Gangl 2004; Harding 2003; Rosenbaum 2002). Indeed, these analyses of treatment effect heterogeneity provide a kind of sensitivity analysis, indicating potential sources of departure from the ignorability assumption.

Just as the presence of effect heterogeneity changes the interpretation of traditional regression estimates, so too effect heterogeneity changes the interpretation of the IV estimator to a local average treatment effect (*LATE*), an effect that pertains only to units whose treatment status is induced by the instrument (Angrist and Krueger 1999; Angrist, Imbens, and Rubin 1996; Angrist and Pischke 2009; Heckman, Urzua, and Vytlacil 2006; Imbens and Angrist 1994).[4] We define the LATE as:

$$\overline{\delta}_{LATE} = E(y^1 - y^0 \mid d_{Z=1} > d_{Z=0}), \tag{15}$$

where Z is the instrumental variable. The IV approach does not rely on the ignorability assumption, but it does rely on its own set of stringent assumptions, to be discussed in more detail below. A limit form of the IV approach (i.e., with a continuous IV variable) is the marginal treatment effect (*MTE*), which focuses on the treatment effect for units at the margin of

[4] Heterogeneity in the effect of a binary endogenous regressor was introduced in the literature on switching regression models (Heckman 1978; Quandt 1972).

treatment assignment (Bjorklund and Moffitt 1987; Heckman, Urzua, and Vytlacil 2006).[5] Heckman, Urzua, and Vytlacil (2006) show that all conventional estimands in causal inference, such as *ATE*, *TT*, and *TUT*, are weighted averages of *MTE* over unobserved variables and $\boldsymbol{X}$. However, IVs that could be utilized to identify *MTE* for the whole distribution of unobserved variables conditional on $\boldsymbol{X}$ are extremely difficult to find.

In each of these approaches, and indeed a defining feature of empirical social science research, is the ceaseless tension between the reality of effect heterogeneity and the practical assumption of effect homogeneity (Xie 2007, 2011). That is, we cannot, nor would we want to, eliminate individual-level response variability, and yet statistical analysis of causal effects for a population science, such as sociology, demography, and economics generally involves a group-level average, and an implicit effect homogeneity assumption. Indeed, each of the methods we describe assumes effect homogeneity for some subpopulation. Different methods essentially differ on how those subpopulations are defined, whether treated or untreated individuals, strata of the propensity distribution, or individuals induced into treatment. Yet the common element for the approaches we describe is that the subpopulations are defined according to their likelihood of selection into treatment. Now let us describe each of these approaches and their estimation methods in more detail.

2.2.1 Heterogeneity Analysis via Differences between $\overline{\delta}_{TT}$ and $\overline{\delta}_{TUT}$

[5] Björklund and Moffitt (1987) introduced the concept of the *MTE*, and showed that the model was observationally equivalent to the switching regression model. See Carneiro, Heckman, and Vytlacil (2011) for a description of related parameters, the policy relevant treatment effect (*PRTE*) and the marginal policy relevant treatment effect (*MPRTE*). See Xie (2011) for a description of the incremental treatment effect (*ITE*), which is the average treatment effect for incremental units when a unit's treatment status changes from d=0 to d=1 when p increases from p_1 to p_2.

Differences between $\overline{\delta}_{TT}$ and $\overline{\delta}_{TUT}$ indicate heterogeneity in treatment effects by selection into treatment. If the $\overline{\delta}_{TT}$ exceeds the $\overline{\delta}_{TUT}$, the effect of treatment is greater for units more likely to select into treatment, sometimes described as "positive selection"; similarly, if the $\overline{\delta}_{TUT}$ exceeds the $\overline{\delta}_{TT}$, the effect of treatment is greater for units less likely to select into treatment, or "negative selection." These different parameters can be estimated in a weighted regression, where the population weights are a function of the predicted probabilities of membership in the treatment group (p_i) (Morgan and Todd 2008):

$$\text{For } d_i = 1,\ w_{i,TT} = 1 \text{ and } w_{i,TUT} = \frac{1-\hat{p}_i}{\hat{p}_i}\text{; for } d_i = 0,\ w_{i,TT} = \frac{\hat{p}_i}{1-\hat{p}_i} \text{ and } w_{i,TUT} = 1.$$

As the goal is to represent the respective population compositions, $w_{i,TT}$ and $w_{i,TUT}$ are used like survey weights. The weight $w_{i,TT}$ makes the control group a representative sample of the treatment group while leaving the treated group unaltered, and the weight $w_{i,TUT}$ works in the opposite direction.

The parameters can also be estimated through matching procedures, where units irrelevant to the estimation of the specified treatment effect are given zero weight (i.e., discarded, in for example, nearest neighbor matching) or weighted (e.g., kernel matching) (Abadie and Imbens 2009; Morgan and Harding 2006; Rubin 1973). The multiple match procedure is generally more efficient but results in greater bias. The motivation of matching, like with weighted regressions, is to change the observed distribution of the control cases to that of the treatment cases to estimate $\overline{\delta}_{TT}$ or to change the observed distribution of the treated cases to that of the control cases to estimate $\overline{\delta}_{TUT}$. Matching estimators of the treatment effects for the treated take the following general form:

$$\bar{\delta}_{TT} = \frac{1}{n_1}\sum_{i}^{n_1}\left\{ y_{i,d=1} - \sum_{i(j)}^{i,j} w_{i(j)} y_{i(j),d=0} \right\}, \tag{16}$$

where n_1 is the number of treatment cases, i is the index over treatment cases, j is the index over control cases, and $w_{i,j}$ represent a set of scaled weights that measure the distance between each treated and control case. The difference in the propensity score is the most commonly used difference measure to construct weights. While in Equation (16) we focus on a matching estimator for the $\bar{\delta}_{TT}$, we could instead match control units to treated units to construct an estimate of the $\bar{\delta}_{TUT}$. These different estimators require different independence assumptions, as described in the large literature on matching [see Morgan and Harding (2006) for a review].

For both the regression and matching routines, there are few examples of systematic tests for whether differences between the respective treatment effects represent statistically significant differences. In one study, Brand and Halaby (2006) take the difference between matching estimates of the $\bar{\delta}_{TT}$ and $\bar{\delta}_{TUT}$ and calculate bootstrap standard errors (generated from 1000 replications).

2.2.2 Heterogeneity Analysis via Statistical Modeling over the Propensity Score Distribution

A few recently developed methods provide statistical tests for differences in effects (i.e., tests for the trend in estimated effects across the propensity score distribution) and an approach to assess possible nonlinearities in subpopulation effects. These methods are described in detail in Xie, Brand, and Jann [forthcoming], as well as applied in studies of college effects in the U.S. (Brand 2010; Brand and Davis 2011; Brand, Pfeffer, and Goldrick-Rab 2012; Brand and Xie 2010; Musick, Brand and Davis 2012) and Taiwan (Tsai and Xie 2008), market processes in

China (Xie and Wu 2005), and effects of maternal job displacement (Brand and Simon-Thomas 2011).

The first method, the stratification-multilevel method (SM) of estimating heterogeneous treatment effects, consists of the following steps: (1) Estimate propensity scores for all units for the probability of treatment given a set of observed covariates, $P(d=1|X)$; (2) Construct balanced propensity score strata where there are no significant differences in the average values of covariates and the propensity score between the treatment and control groups. This practice ignores heterogeneity within a stratum. While within-stratum homogeneity is still implausible, there is greater homogeneity than without stratification. (3) Estimate propensity score stratum-specific treatment effects; and (4) Evaluate a trend across the strata using variance-weighted least squares regression of the strata-specific treatment effects on strata rank at level-2:

$$\delta_s = \delta_0 + \gamma S + \eta_s \tag{17}$$

where level-1 slopes (δ_j) are regressed on propensity score rank indexed by S, δ_0 represents the level-2 intercept (i.e., the predicted value of the treatment effect for the lowest propensity individuals), and γ represents the level-2 slope (i.e., the change in the treatment effect with each one-unit change to a higher propensity score stratum).

The goal of the SM method is to look for a systematic pattern of heterogeneous treatment effects across strata. A linearity specification, typically assumed in order to preserve statistical power, tells us whether the treatment effect is either a positive or a negative function of the propensity of treatment. The SM approach offers useful and easily interpretable estimates of strata-specific treatment effects and the unit change in estimates as we move between strata. However, the SM approach is limited in that the researcher is forced to divide the full range of

propensity scores into a limited number of strata, assume within-strata homogeneity, and use a strong functional form to detect patterns of treatment heterogeneity.

To overcome these shortcomings, Xie, Brand, and Jann [forthcoming] describe two nonparametric methods. First, the matching-smoothing (MS) method of estimating heterogeneous treatment effects consists of the following steps: (1) Estimate the propensity scores for all units; (2) Match treated units to control units with a matching algorithm; (3) Plot the observed difference in a pair between a treated unit and an untreated unit against a continuous representation of the propensity score; and (4) Use a nonparametric model to smooth the variation in matched differences, such as local polynomial or Lowess smoothing, to obtain the pattern of treatment effect heterogeneity as a function of the propensity score. That is, we fit a nonparametric smoothed curve to the trend in matched differences as a function of the propensity score, and thus unlike SM we do not assume homogeneity within strata.[6]

Second, the smoothing-differencing (SD) method of estimating heterogeneous treatment effects is closely related to the MS method as it also uncovers the heterogeneity pattern as a nonparametric function of the propensity score. The steps of the method are: (1) Estimate the propensity scores for all units; (2) For the control group and the treatment group fit separate nonparametric regressions of the dependent variable on the propensity score, such as local polynomial smoothing; and (3) Take the difference in the nonparametric regression line between the treated and the untreated at different levels of the propensity score. The results of MS and SD should be comparable, although both procedures have specific advantages: examination of

[6] The ignorability assumption states that there is no bias resulting from using the naive estimator for estimating the treatment effects conditional on the propensity score; therefore, the *TT* and *TUT* are the same conditional on the propensity score. As a result, in theory, the distinction between choosing treated units or untreated units as the target group is of minor consequence for the MS method.

(raw) observation-level differences between treated and untreated units in the MS method and the simplicity of few modeling decisions in the SD method.

An increase in the treatment effect with an increase in the propensity for treatment using SM, MS, or SD is similar to observing $\overline{\delta}_{TT} > \overline{\delta}_{TUT}$; likewise, a decrease in the treatment effect with an increase in the propensity for treatment is similar to observing $\overline{\delta}_{TUT} > \overline{\delta}_{TT}$. However, we obtain subpopulation treatment effects and a test for the trend in effects using SM, and we may observe situations using MS and SD in which there is a curvilinear pattern of effects across the distribution of the propensity score for which there is no simple analog to the regression and matching estimates of $\overline{\delta}_{TT}$ and $\overline{\delta}_{TUT}$.

2.2.3 Heterogeneity Analysis via Instrumental Variables

As we note above, the presence of effect heterogeneity changes the interpretation of the IV estimator to a local average treatment effect (*LATE*), and a limit form is the marginal treatment effect (*MTE*). Conditioning on ***X***, IV regression is estimated using two-stage least squares (2SLS). In the first step, the instrumental variable(s) *Z* (and other independent variables) are used to predict the instrumented variable *d*. In the second stage, the predicted values of the instrumented variable $\hat{d}$ are used to predict the outcome variable.[7] Use of IVs does not require the strong ignorability assumption, but it relies on its own set of stringent assumptions (Angrist and Pischke 2009). First, we must satisfy an "independence assumption," i.e. that the instrument is as good as randomly assigned. Second, the IV must satisfy the condition that it affects the likelihood of treatment status, even if it does so within a small range, but affects the outcome

[7] With a binary outcome, we cannot use two-stage least squares, but use instead generalized methods of moments (Angrist 2001) or structural mean models or marginal structural models (Robins, Hernán, and Brumback 2000).

only indirectly through the treatment (i.e., it does not affect the outcome independent of treatment selection). This is commonly called the "exclusion restriction." We must assume that subjects have their otherwise natural propensity of treatment, but some of them were "induced" into treatment by some event (i.e., the IV). Third, the IV must satisfy a "monotonicity assumption" (i.e., that although the instrument may have no effect on some people, all those affected are affected in the same direction).

In actual social settings, the inducement effect of an IV on treatment is usually very small. If treatment effects are homogeneous, low inducement effect of IVs on the treatment likelihood is not necessarily 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.[8] However, in the presence of heterogeneous treatment effects, we must limit the interpretation of the resulting estimator to this particular "local" group of units induced, or units on the "margin" of treatment, as implied by the terms "local average treatment effect" (Angrist, Imbens, and Rubin 1996; Angrist and Pischke 2009) and "marginal treatment effect" (Bjorklund and Moffitt 1987; Heckman, Urzua, and Vytlacil 2006). Thus, estimating heterogeneous treatment effects over the entire range of the unobserved factors via IV's given $\boldsymbol{X}$ is far more demanding than what is available in actual settings for empirical research (Morgan and Winship 2007).

Comparisons between estimates of *TT*/*TUT* and *LATE*/*MTE* are complex. We can describe the *TT* as a combination of the effect for individuals induced by the instrument (so-called "compliers") and individuals who are treated regardless of the inducement ("always-takers"); likewise, the *TUT* is a combination of the effect for compliers and individuals untreated

[8] Low inducement is a major limitation if the instrument is so weak as to have very little impact on the treatment of interest. We revisit this issue in the empirical example section.

regardless of the inducement ("never-takers") (Angrist and Pischke 2009).[9] Angrist and Piscke (2009) write: "Because an IV is not directly informative about effects on always-takers and never-takers, instruments do not usually capture the average causal effect on all of the treated or on all of the non-treated" (p. 160). The subpopulation induced into treatment, which cannot actually be identified, can differ on both observed and unobserved characteristics from the treated and the untreated subpopulations. As we cannot know what subpopulation of the treated corresponds to individuals induced into treatment by an instrument, comparisons between strata-specific treatment effect estimates and *LATE*/*MTE* estimates are likewise complex. Moreover, those induced into treatment can differ as the inducement changes, because different instruments will affect treatment status for different segments of the population (Angrist and Pischke 2009; Gangl 2010). Thus, with treatment effect heterogeneity, estimates of treatment effects based on different IVs will differ.

3 Research on Effect Heterogeneity

Although recent research in causal inference recognizes the importance of population heterogeneity and response variation, with notable contributions to the literature from several authors of this volume (Elwert and Winship 2011; Heckman 2005; Holland 1986; Manski 1995; Morgan and Winship 2007, 2012; Winship and Morgan 1999; Xie, Brand, and Jann [forthcoming]), empirical substantive research has been slow to capitalize on these insights. Here we discuss the costs in assuming effect homogeneity and potential benefits in assessing effect heterogeneity using examples from research on education, a substantive area that has more quickly absorbed heterogeneity lessons from the causal literature. We choose examples that

[9] We assume, for simplicity, that there are no "defiers," i.e., those individuals who would always do the opposite of treatment assignment.

demonstrate the use of the different methods for heterogeneity analyses described above, and, where applicable, describe how these different methods complement or challenge one another.

Often a lively unresolved debate surrounds a particular treatment effect of interest, as to whether there is an effect, or whether it is positive or negative. Settling such debates may beg for analyses of variation in effects across the population, as traditional regression coefficients of treatment effects have ambiguous interpretation and depend upon the population composition of the treated in the presence of effect heterogeneity. For example, the ongoing debate over whether there is a positive Catholic school effect, i.e. whether Catholic private schools are more effective than public schools (despite fewer dollars spent per pupil) [see Morgan (2001) for a review], may reflect school effect heterogeneity. Catholic schooling may be more or less beneficial to students who are more or less likely to attend them. Morgan and Todd (2008) using weighted regressions and Morgan (2001) using propensity score matching find that those students who are least likely to attend Catholic schools (in this case, the more disadvantaged students) experience the largest effects ($\overline{\delta}_{TUT} > \overline{\delta}_{TT}$). Understanding the effects of school environments is important to students, parents, and schools making enrollment decisions, and to debates on school choice and vouchers. Disadvantaged students may have poor public school options, such that Catholic schools distribute learning opportunities more evenly and thus more effectively equalize outcomes than do public schools. Or if financially constrained parents select those children they think will be most likely to benefit, there may be greater unobserved selectivity among poor, low propensity students.

Effects of attendance at elite colleges on career achievement provide a similar example underscoring the importance of analyses of treatment effect heterogeneity. An early body of research largely concluded that attendance at highly selective colleges yielded an economic

payoff, while recent studies, which attended more rigorously to issues of selection, yielded mixed results [see Brand and Halaby (2006) for a review]. However, effects may differ for individuals more or less likely to attend elite colleges. Indeed, Brand and Halaby (2006), using propensity score matching, find that the returns to attending an elite college are small by comparison to those that would have been achieved by otherwise equivalent students who attended non-elite colleges ($\overline{\delta}_{TUT} > \overline{\delta}_{TT}$).[10] We can offer similar explanations for this pattern in effects as for those pertaining to the Catholic school effect: an elite college education may be more beneficial for students who have socioeconomically disadvantaged backgrounds and lack social capital, or it may be that low propensity students are the most selective. If unobserved selection is a stronger factor for low propensity students, and the $\overline{\delta}_{TUT}$ is based on a higher proportion of such students, then endogeneity may be a more salient issue for the $\overline{\delta}_{TUT}$ than for the $\overline{\delta}_{TT}$. Understanding this pattern of heterogeneity is once again important to students, parents, and schools making enrollment decisions, and to debates about equal access to college.

School quality is not the only educational effect subject to response variation by selection; heterogeneity in effects may also occur by the years of schooling and credentials an individual receives. College graduates on average earn more money, hold more stable jobs with better working conditions, lead more traditional family lives, are healthier, and participate more in civic life (Hout [forthcoming]). However, each of these average relationships may conceal systematic effect heterogeneity. A recent series of papers finds larger college effects among students with lower propensities for attending and completing college on earnings (Brand and

[10] This finding is further substantiated by a recent study suggesting that economic returns to highly selective college attendance are indistinguishable from zero among the full sample when adjustments for unobserved student characteristics are incorporated, while returns among Black and Hispanic students and students from disadvantaged families remain large (Dale and Krueger 2011). Hout [forthcoming] reviews additional studies with corroborating results.

Xie 2010), on civic engagement (Brand 2010), and on reductions in marriage and fertility (Brand and Davis 2011; Musick, Brand, and Davis 2012). These papers all model heterogeneous effects as a function of the propensity for college attendance using the stratification-multilevel (SM) approach described above.[11]

One interpretation of the results of Brand and Xie (2010) is that a college education may be particularly beneficial among groups targeted by educational expansion efforts — that is, individuals who are otherwise unlikely to attend college based on their observed characteristics. Echoing the theme from this series of papers, Hout [forthcoming] notes: "Young people with the most abilities may learn and ultimately earn the most, but their education augments their success less than it augments less-able people's success" (p. 14). In addition to this probable mechanism, Brand and Xie (2010) note: "… the very pattern of heterogeneous treatment effects of college education by the propensity to complete college suggests an unobserved selection mechanism at work: individuals from disadvantaged social backgrounds, for whom college is not a culturally expected outcome, overcome considerable odds to attend college and may be uniquely driven by the economic rationale" (p. 294). As we suggest earlier, analyses of effect heterogeneity facilitate sensitivity to differential sources of endogeneity. Although we cannot know how strong the observed set of characteristics is relative to the unobserved in influencing selection into treatment, we may hypothesize that we have more unobserved factors influencing selection into treatment among "against the odds" cases, or those individuals with low estimated propensity scores.

[11] Brand and Davis (2011) combine the SM approach with discrete-time event-history models. Xie, Brand, and Jann (2011) also use the MS and SD methods for the effects of college on fertility and find comparable results to those using SM. This is due to a largely linear pattern in effects of college on fertility across the propensity for college.

Another series of papers using instrumental variables, including compulsory schooling laws, secondary and university reforms, and distance to the nearest college or university, have found that IV estimates exceed those of OLS estimates [see Card (2001) and Hout [forthcoming] for reviews]. Recall that such estimates can be interpreted as local average treatment effects (LATE), and LATE estimates that exceed OLS estimates may suggest larger returns for individuals on the margin of school continuation than average individuals. However, an early paper by Willis and Rosen (1979) and more recent work by Heckman and colleagues reached a different conclusion (Carniero, Heckman, and Vytlacil 2011). While the majority of the early studies use a binary instrument, and hence only one portion of the marginal return can be identified, this more recent work uses multiple, multivalued instruments enabling estimates for a wider portion of the return function (Moffit 2008). Carniero, Heckman, and Vytlacil (2011) argue that people select into schooling on the basis of realized returns to schooling; in other words, those who perceive the largest financial benefit from college generally attend, whereas those who do not perceive high financial benefits choose not to attend. Based on these results, they argue that too many people are attending college. As we suggest above, reconciling divergent findings from these methods requires understanding as to how observable and unobservable characteristics influence selection into treatment differently for subpopulations defined by propensity strata and alternative treatment inducements.

We have focused on research from education to illustrate what we gain from attending to treatment effect heterogeneity. However, the usefulness of such analyses is not limited to education research. It extends to a broad array of the effects of access to social resources and social programs, as well as to potentially negative life events and changes in social conditions. For example, Brand and Simon-Thomas (2011) find that, among families with single mothers,

the negative effects of maternal displacement on children's educational attainment and mental health are higher when displacement is an unlikely event. As low propensity mothers are more advantaged, the shock of a displacement event to relatively higher status families may induce larger negative intergenerational effects. Or, larger observed effects among children who have a low propensity for maternal displacement could be the result of greater unobserved selectivity.

4 Empirical Demonstration

In this section, we demonstrate the methods we present above, using the example of civic returns to higher education. Civic returns to education, particularly among disadvantaged members of the population, continue to offer a central justification for public policy promoting equal access to schooling. Education is a key correlate, if not determinant, of civic participation [see Brand (2010) for a review]. Some studies recognize the endogeneity problem associated with assessing the causal effect of education on civic participation, but few recognize potential effect heterogeneity. Brand (2010), an exception to this deficiency, addressed heterogeneous effects of college on civic participation by the propensity of college education using the SM approach. We extend the work of Brand (2010) by comparing a range of methodological approaches to assess treatment effect heterogeneity.

4.1 Data Description

We use panel data from the National Longitudinal Survey of Youth (NLSY) 1979 to assess causal effect heterogeneity of college completion on subsequent civic participation. The NLSY is a nationally representative sample of 12,686 respondents who were 14-22 years old when they were first interviewed in 1979. These individuals were interviewed annually through

1994 and are currently interviewed on a biennial basis. We use information gathered from 1979 through 2006. We restrict the sample to respondents who were 14-17 years old at the baseline survey in 1979 (n = 5,582), who had completed at least the 12th grade by 2006 (n = 4,827), and who did not have missing data on measures of educational attainment or civic participation from the 2006 survey wave (n = 3,452).[12] We set these sample restrictions to ensure all measures we use are pre-college, particularly ability, and to compare college graduates with individuals who completed at least a high school education. The individuals we lose due to attrition and nonresponse tend to be from more disadvantaged family backgrounds and levels of achievement than those individuals we retain.

Appendix A describes measures of pre-college covariates and post-college civic participation. The pre-college measures figure prominently in economic and sociological studies of educational and occupational attainment, and their measurement is straightforward; for details see Brand (2010). The likelihood of college varies by gender, race and ethnicity, family background, academic achievement, friends' plans and parents' encouragement in expected directions. We use two dichotomous indicators of civic participation measured in 2006 asking respondents if they performed any unpaid volunteer work in the past 12 months for: (1) civic, community or youth groups, and (2) charitable organizations or social welfare groups. About 13 percent of college graduates compared to 5 percent of non-college graduates volunteer for civic, community, or youth groups and 9 percent of graduates compared to 4 percent of non-graduates volunteer for charitable organizations or social welfare groups.

[12] 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 and high school college-preparatory program. We control for an imputed value indicator in our models.

4.2 Treatment Effect Analyses

4.2.1 Homogenous Effect Estimates

We first estimate propensity scores for each individual in the sample for the probability of college completion given a set of observed covariates using a logit regression model. Table 1 provides results for the logit model, which support the literature on the determinants of college education. In Table 2, we report average effects of college completion on our two measures of civic participation using logit regression models under an assumption of effect homogeneity. The first model represents the bivariate association; the second model controls for the estimated propensity score.[13] The bivariate models suggest college graduates are about 3.4 times more likely ($e^{1.223}$; predicted probabilities are 0.12 for college graduates and 0.04 for non-college graduates) to volunteer for civic, community, or youth groups than non-college graduates and about 2.4 times more likely ($e^{0.887}$; predicted probabilities are 0.08 for college graduates and 0.04 for non-college graduates) to volunteer for charitable organizations or social welfare groups. Results are highly statistically significant. Controlling for the estimated propensity for college, we find that college graduates are about 2.1 times more likely to volunteer for civic, community, or youth groups than non-college graduates and about 1.4 times more likely to volunteer for charitable organizations or social welfare groups. Propensity for college has a significant positive effect on both forms of volunteering. Point estimates are reduced in the propensity score adjusted models, and the college effect on charitable organizations and social welfare groups no longer reaches statistical significance.[14]

[13] Results controlling for the full set of covariates are very similar. Rosenbaum and Rubin (1983, 1984) demonstrate it is sufficient to condition on the propensity score as a function of $\boldsymbol{X}$ rather than $\boldsymbol{X}$ itself, which we do here for simplicity.

[14] In contrast to Brand (2010), we impute all missing cases; the propensity score model specification also slightly differs from Brand (2010). Thus, our analyses yield marginally different results.

Regression models with homogeneity assumptions such as this one are ubiquitous in empirical social science research. However, in the presence of treatment effect heterogeneity, average effects can vary widely depending on population composition. We next assess whether there is evidence for heterogeneity in the effects of college on civic participation.

4.2.2 Differences between $\overline{\delta}_{TT}$ and $\overline{\delta}_{TUT}$

In Table 3, we report estimates for $\overline{\delta}_{TT}$ and $\overline{\delta}_{TUT}$ using both weighted regression and propensity score nearest neighbor and kernel matching. Weighted regressions estimates of the $\overline{\delta}_{TT}$ suggest that college graduates are about 2 times more likely to volunteer for civic, community, or youth groups than non-college graduates and about 1.3 times more likely to volunteer for charitable organizations or social welfare groups (although the latter effect is not statistically significant). However, estimates are much larger for the $\overline{\delta}_{TUT}$: college graduates are about 2.8 times more likely to volunteer for civic groups than non-college graduates and about 2 times more likely to volunteer for charitable organizations (where both effects are statistically significant).

Matching estimates of the $\overline{\delta}_{TT}$ and $\overline{\delta}_{TUT}$, both kernel and nearest neighbor, are similar in that they too suggest larger college effects on civic volunteering among individuals who went to college but have the characteristics of those who did not. Matching estimates are based on simple differences between predicted probabilities of volunteering between college and non-college graduates, yet we can roughly transform these to odds ratios to compare to our weighted logit regression estimates. For the $\overline{\delta}_{TT}$, college graduates are about 1.8 times more likely to volunteer for civic, community, or youth groups than non-college graduates and about 0.9 times as likely to volunteer for charitable organizations or social welfare groups (the latter effect is not

significant). Effects for the $\overline{\delta}_{TUT}$ once again exceed those for the $\overline{\delta}_{TT}$: college graduates are about 2.8 times more likely to volunteer for civic groups than non-college graduates and about 1.5 times more likely to volunteer for charitable organizations (where for the matching estimates, in contrast to the weighted regressions, the latter effect is not significant).

4.2.3 Statistical Modeling over the Propensity Score Distribution

Differences between the estimates of $\overline{\delta}_{TT}$ and $\overline{\delta}_{TUT}$ suggest subpopulation effect variation by selection into treatment. We next turn to examining effects across the distribution of the propensity score. For the stratification-multilevel model (SM), we first generate propensity score strata such that within each interval of the propensity score the average propensity score and the means of each covariate do not significantly differ between college and non-college graduates (Becker and Ichino 2002). Appendix B provides characteristics of typical individuals within each propensity score stratum, useful statistics we obtain after constructing such strata. Individuals with parents who were high school dropouts, have four siblings, have low ability, were enrolled in a non-college-prep track, and who had friends who had not planned to go to college are characteristic of stratum 1. By contrast, individuals with parents who went to college, have two siblings, have high ability, were enrolled in a college-prep track, who had parents who encouraged college and friends who planned to complete college are characteristic of stratum 7.[15] However, these are averages, albeit more informative than global averages, and not all covariates are positively correlated with one another within propensity score strata.

[15] For the *k*th covariate in the *j*th stratum, we estimate the standardized mean covariate difference to quantify the balance between the treatment and the control groups for each covariate *X* (Morgan and Winship 2007):

In Table 4, we report estimated effects for logit regressions of college on civic participation by propensity score strata (level-1) and the estimated trend in these effects using variance weighted least squares regression (level-2). The level-2 slopes for both indicators of volunteering reveal significant declines in the effect of college completion as the propensity for college increases. For civic, community and youth groups, the level-2 slope indicates a significant 0.26 reduction in the college effect for each unit change in propensity score rank. That is, level-1 estimates range from college graduates being 7.6 times more likely to volunteer for civic groups than non-college graduates in stratum 1 to equally likely to volunteer in stratum 7. Similarly for charitable organizations and social welfare groups, the level-2 slope indicates a statistically significant 0.2 reduction in the college effect for each unit change in propensity score rank. Estimates range from college graduates being 3.1 times more likely to volunteer than non-college graduates in stratum 1 to 0.6 as likely to volunteer in stratum 7. Levels of volunteering by propensity score strata and college completion, reported in Appendix B, provide further evidence as to the pattern in effects. While levels of volunteering by propensity for college are equalized among college graduates, there is a socioeconomic gradient in volunteering among non-college graduates, particularly for civic, community, and youth groups, generating large observed effects of college among disadvantaged individuals who complete college.

Figure 1 graphically depicts the results presented in Table 4. "Points" in Figure 1 represents estimates of level-1 slopes, and the lines in the figure are the level-2 slopes. We plot the mean differences in levels of volunteering rather than the logit regression coefficients for

---

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

where $\bar{X}$ is the sample mean and s^2 is the sample variance of the kth covariate in the jth stratum for the treated and control groups as indexed by D=(1,0). The standardized difference is larger in some strata than in others for some covariates.

comparability to the smoothing-differencing results we present below. The figure depicts the similarity in the decline in the effect of college completion on both forms of civic participation as the propensity for college increases. This figure also suggests potential nonlinearities in effects. To examine this possibility, we now turn to the nonparametric methods.[16]

To estimate heterogeneous treatment effects with the matching-smoothing (MS) method or smoothing-differencing (SD) method, we begin once again by estimating the propensity for treatment. For MS, we then match treated and control units by the estimated propensity scores and calculate differences between outcomes, plot the matched differences between treated and control units along a propensity score *x*-axis, and fit a smoothed curve. For the smoothing-differencing (SD) method, we fit two separate nonparametric regression models for the outcome variable on the propensity score, one for the treatment group and one for the control group. The results from MS and SD yield similar results, and so to conserve space, we choose to present results only from SD. We use local polynomial regression as a smoothing device (degree 1, bandwidth 0.2). The difference between the group-specific regressions gives an estimate of the heterogeneous treatment effects.

Figure 2 displays the resulting curves. Figure 2 differs from Figure 1 in that the *x*-axis is now a continuous representation of the propensity score rather than discrete strata. SD provides a fully nonparametric depiction of treatment effect heterogeneity, rather than the imposition of a functional form on the heterogeneity in effects. For civic groups, we find a flattening in differences at the mid- to high propensity scores, and larger differences for high propensity college goers than we expect given the linear trend reported from the SM method. Linearity was not a reasonable functional form for effects on charitable organizations: we have a relatively flat

[16] We fit alternative SM models in which we include quadratic terms in level-2. These terms were not statistically significant, and we do not present them here.

college effect from the low to mid-propensity scores, and then a significant drop in effects at the upper end of the propensity distribution. Thus, we uncover a potential difference in effects across the propensity for college that we overlook when examining the weighted regressions, propensity score matching, and SM results.

4.2.4 Instrumental Variables

As we describe above, in the presence of treatment effect heterogeneity, instrumental variable (IV) estimates must be interpreted as local average treatment effects (LATE). Following, Carniero, Heckman, and Vytlacil (2011), we aimed to consider two IVs from the private geocode data of the NLSY, one indicating local availability of college at age 14 and one indicating local area unemployment at age 17. Because ours is an analysis of college effects on civic participation, rather than earnings as in Carniero, Heckman, and Vytlacil (2011), we first questioned whether the exclusion restriction held. That is, could we reasonably assume that having a college in a community impacts civic engagement only through its effect on individual educational attainment? A more educated local populace, which we might assume given local availability of college, results in a higher level of civic engagement, which in turn could induce higher civic involvement independent of one's own educational level (Putnam 2000). We have a similar issue using local area unemployment rate as an instrument. Brand and Burgard (2008) show that job displacement has a significant negative impact on civic engagement, suggesting that high levels of unemployment could result in lower levels of civic involvement independent of educational attainment.

However, because we measure college completion in 2006 when respondents were in their early 40s, our concerns with the exclusion restriction may be mitigated, particularly in the

use of local unemployment rate at age 17. We chose to measure college in this way because we were interested in the effect of college on civic involvement, whether college was attended immediately following high school or some time later. But the disjuncture between when some individuals completed college and when they lived in a community for which the instruments were valid led to a potentially even more pressing issue than the exclusion restriction: the instruments were very weak. The *F*-statistic was 3.04 and 1.02 for the first stage regressions of college completion on local unemployment and for college completion on college availability, respectively. The correlations between college and each of the candidate instruments was under |0.05|. As Bound, Jaeger, and Baker (1995) note, candidate instruments are quite commonly only weakly correlated with the endogenous explanatory variable (p. 443). Weak instruments exacerbate the bias due to the other IV assumptions, including the exclusion restriction, as well as the independence and monotonicity assumptions.

At this point, we decided not to pursue the IV analysis. Although we regrettably do not demonstrate the comparison between IV-based and the other heterogeneity effect estimates, there is nevertheless a general, and common, lesson here: although an IV is a potentially useful tool for causal analysis, finding a good one can be very difficult.

5 Conclusion

In this chapter, we have described the importance of studying treatment effect heterogeneity and methods for how to do so. We focused on the interaction between the treatment of interest and the propensity of treatment. We do not contend that this is the only interaction of social significance; indeed, interactions between specific key covariates may be more important for particular studies. However, for questions of treatment effect variation that

relate to matters of causality and selection, the propensity score is consequential. As the propensity score proved beneficial for studies seeking to account for pre-treatment heterogeneity by reducing the problem of dimensionality (Rosenbaum and Rubin 1983, 1984), it is similarly expedient for the study of treatment effect heterogeneity.

Heterogeneity in treatment effects has important implications for understanding how scarce social resources are distributed in an unequal society, for social and behavioral research designs, and for social policy. With a research design that attends to pre-treatment heterogeneity, we assess the internal validity of our effect estimates (i.e., the degree to which we successfully uncover causal effects for the population being studied); but with a design that attends to treatment effect heterogeneity, we also assess the external validity of our effect estimates (the predictive value of the findings in a different context) (Angrist and Pischke 2009). That is, when individuals differ in their response to treatments, treatment effects can vary widely depending on population composition and we must tailor the interpretation of our effect estimates to specific subpopulations. 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. Conversely, if highly resourceful individuals acquire a costly treatment, but not necessarily individuals most likely to benefit, increasing the availability of the treatment may increase the average effect among the treatment recipients. Policy makers who understand patterns of treatment effect heterogeneity can more effectively assign different treatments to individuals to balance competing objectives, such as reducing cost, maximizing average outcomes, and reducing variance in outcomes in a given population.

We discussed and demonstrated a variety of methods used to study treatment effect heterogeneity. In our example of college effects on volunteering, we found larger effects for the treatment effects for the untreated (*TUT*) than for the treatment effects for the treated (*TT*) using weighted regressions and propensity score matching. Point estimates were lower and some effects were not statistically significant using matching; this difference is to be expected, as matching results, which compare units to fewer controls, may achieve less bias and typically at the expense of efficiency. We then examined effects across the distribution of the propensity score using stratification-multilevel (SM) and smoothing-differencing (SD) methods. SM augmented our analysis of differences in *TT* and *TUT*, suggesting larger effects of college on volunteering for individuals least likely to go to college. But by generating estimated effects for seven balanced propensity strata using SM, we exposed more finely graded estimates than our analysis of the *TT* and *TUT*, and we explored a linear trend for the variation in effects. Using SM, we found statistically significant heterogeneity in effects of college on volunteering, suggesting that college effects decrease as the propensity for college increases. This analysis also revealed potential nonlinearities in effects, and our SD analysis confirmed interesting deviations from the linear trend. Finally, we considered instrumental variables to estimate local average treatment effects (LATE), effects that correspond to a particular subpopulation for which the instrument induces a change in the treatment regime. Such analyses, in contrast to the preceding methods, do not rely on the ignorability assumption; however, valid instruments are difficult to come by, and in our demonstration, our instruments were too weak to be considered useful.

A note from Halaby (2004) bears repeating here: "… causal inference cannot be reduced to any one formula applied to data. Because causal inference from observational data is by its

nature precarious, it pays to experiment with the host of basic techniques …" (p. 541). The analytic methods we described for assessing treatment effect heterogeneity have different strengths and weaknesses, and are based on different assumptions. But the methods are also essentially different ways to identify subpopulations with varying probability of selection into treatment, and as such the analysis of the basic techniques yields further insight into effect heterogeneity. As treatment effect heterogeneity is still too infrequently empirically assessed in quantitative social science research, we hope our exposition furthers the absorption of analytic techniques for the study of effect heterogeneity into research practice.

References

Abadie, Alberto, and Guido W. Imbens. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74(1): 235-267.

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact on Voluntary Military Service Using Social Security Date on Military Applicants." *Econometrica* 66:249-288.

Angrist, Joshua D. 2001. "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice." *Journal of Business and Economic Statistics* 19:2-16.

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

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434): 444-455.

Angrist, Joshua D., and Jorn-Steffen Piscke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Becker, Sascha, and Andrea Ichino. 2002. "Estimation of Average Treatment Effects Based on Propensity Scores." *Stata Journal* 2(4):358-77.

Bjorklund, Anders and Robert Moffitt. 1987. "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models." *The Review of Economics and Statistics* 69(1):42-49.

Bound, John, David A. Jaeger and Regina 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 Sarah A. Burgard. 2008. "Job Displacement and Social Participation over the Life Course: Findings for a Cohort of Joiners." *Social Forces* 87(1):211-242.

Brand, Jennie E. and Dwight Davis. 2011. "The Impact of College Education on Fertility: Evidence for Heterogeneous Effects." *Demography* 48(3):863-887.

Brand, Jennie E., Fabian Pfeffer, and Sara Goldrick-Rab. "Interpreting Community College Effects in the Presence of Heterogeneity and Complex Counterfactuals." *California Center for Population Research* Working Paper PWP-2012-004.

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, and Juli Simon-Thomas. 2011. "Job Displacement Among Single Mothers: Effects on Children's Outcomes in Young Adulthood*."* Working paper.

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.

Card, David. 2001. "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems." *Econometrica* 69:1127–60.

Carniero, Pedro, James J. Heckman, and Edward Vytlacil. 2011. "Estimating Marginal Returns to Education." *American Economic Review* 101:2754-2781.

Dale, Stacey and Allan B. Krueger. 2011. "Estimating the Return to College Selectivity over the Career Using Administrative Earning Data." Princeton University: Working Paper.

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.

Elwert, Felix and Christopher Winship. 2010. "Effect Heterogeneity and Bias in Main-Effects-Only Regression Models." Pp. 327-336 in *Heuristics, Probability, and Causality: A Tribute to Judea Pearl*. Rina Dechter, Hector Geffner, and Joseph Y. Halpern (eds.) United Kingdom: College Publications.

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

Halaby, Charles N. 2004. "Panel Models in Sociological Research: Theory into Practice." *Annual Review of Sociology* 30:507-544.

Harding, David J. 2003. "Counterfactual Models of Neighborhood Effects: The Effect of Neighborhood Poverty on Dropping Out 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." Pp. 1–98 in *Sociological Methodology,* Vol. 35, edited by Ross M. Stolzenberg. Boston, MA: Blackwell Publishing.

Heckman, James J. and Richard Robb. Jr. 1985. "Alternative Methods for Evaluating the Impact of Interventions." *Journal of Econometrics* 30: 239-267.

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.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Associations* 81(396):945-960.

Hout, Michael. Forthcoming. "Social and Economic Returns to College Education in the United States." *Annual Review of Sociology*.

Imbens, Guido W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467-475.

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

Moffitt, Robert. 2008. "Estimating Average and Marginal Treatment Effects in Heterogeneous Populations." Working paper.

Morgan, Stephen. 2001. "Counterfactuals, Causal Effect Heterogeneity, and the Catholic School Effect on Learning." *Sociology of Education* 74: 341-374.

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 L. and Jennifer J. Todd. 2008. "A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects With a Demonstration from School Effects Research." *Sociological Methodology* 38:231-81.

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

Morgan, Stephen, and Christopher Winship. 2012. "Brining Context and Variability Back into Causal Analysis." *Oxford Handbook of the Philosophy of the Social Sciences*.

Musick, Kelly, Jennie E. Brand, and Dwight Davis. 2012. "Variation in the Relationship Between Education and Marriage: Marriage Market Mismatch?" *Journal of Marriage and Family* 74:53-69.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2nd Edition. Cambridge: Cambridge University Press.

Pearl, Judea. 2010. "The Foundations of Causal Inference." *Sociological Methodology* 40(1):75-149.

Quandt, Richard E. 1972. "A New Approach to Estimating Switching Regressions." *Journal of the American Statistical Association* 67(338): 306-310.

Robins, James M., M. A. Hernán, and B. Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11:550-560.

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

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

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

Rubin, Donald. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:685-701.

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

Vander Weele, Tyler J. and James Robins. 2007. "Four Types of Effect Modification: A Classification Based upon on Directed Acyclic Graphs." *Epidemiology* 18(5): 561-568.

Willis, Robert J. and Sherwin Rosen. 1979. "Education and Self-Selection." *Journal of Political Economy* 87(5, part 2):S7-S36.

Winship, Christopher and Stephen L. Morgan. 1999. "The Estimation of Causal Effects from Observational Data." *Annual Review of Sociology* 25:659-706.

Xie, Yu. 2007. "Otis Dudley Duncan's Legacy: the Demographic Approach to Quantitative Reasoning in Social Science." *Research in Social Stratification and Mobility* 25:141-156.

Xie, Yu. 2011. "Population Heterogeneity and Causal Inference." *University of Michigan Population Studies Center Research Report* 11-731.

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

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

Table 1. Logit Regression Model Predicting College Completion (N= 3,452)

Male	-0.189	†
	(0.100)	
Black	-0.558	***
	(0.137)	
Hispanic	-0.862	***
	(0.166)	
Mother's education	-0.403	***
	(0.076)	
(Mother's education)^2	0.021	***
	(0.003)	
Father's education	0.077	***
	(0.020)	
Parents' inc. (1979 $10,000s)	0.001	
	(0.001)	
Intact family	0.102	
	(0.120)	
Number of siblings	-0.029	
	(0.024)	
Southern residence	0.232	†
	(0.109)	
Cognitive ability	1.083	***
	(0.168)	
Cog. ability * Par. Income	0.002	*
	(0.001)	
College-preparatory	0.584	***
	(0.107)	
Parents' encouragement	0.511	***
	(0.133)	
Friends' schooling plans	0.830	***
	(0.111)	
Non-missing on Covariates	0.051	
	(0.111)	
Constant	-1.889	***
	(0.500)	
Wald χ^2	1278.55	
P > χ2	0.000	

Notes: Numbers in parentheses are standard errors.

† p < .10 * p<.05 ** p < .01 *** p < .001 (two-tailed tests)

Table 2. Regression Estimates of Homogenous Effects of College Completion on Civic Participation (N= 3,452)

	Civic, Community, or Youth Groups	Charitable Orgs. or Social Welfare Groups
Bivariate Association	0.083 ***	0.047 ***
	(0.009)	(0.008)
Propensity Score Adjusted Logit Regression	0.727 ***	0.339
	(0.187)	(0.211)

Notes: Numbers in parentheses are standard errors. Propensity scores were estimated by a logit regression model of college completion on the set of pre-college covariates as reported in Table 1.

† p<.10 * p <.05 ** p < .01 *** p < .001 (two-tailed tests)

Table 3. Regression and Matching Estimates of Heterogeneous Effects of College Completion on Civic Participation (δ_{TT} and δ_{TUT}; N= 3,452)

	Civic, Community, or Youth Groups	Charitable Orgs. or Social Welfare Groups
Weighted Logit Regression (δ_{TT})	0.670 **	0.284
	(0.211)	(0.240)
Weighted Logit Regression (δ_{TUT})	1.029 ***	0.678 **
	(0.204)	(0.226)
Kernel Matching (δ_{TT})	0.050 **	-0.012
	(0.016)	(0.015)
Kernel Matching (δ_{TUT})	0.067 **	0.027
	(0.022)	(0.018)
Nearest Neighbor Matching (k=5; δ_{TT})	0.051 **	-0.010
	(0.018)	(0.016)
Nearest Neighbor Matching (k=5; δ_{TUT})	0.066 †	0.011
	(0.035)	(0.017)

Notes: Numbers in parentheses are standard errors. Regression estimates are adjusted for propensity scores and matching estimates are matched on propensity scores. Propensity scores were estimated by a logit regression model of college completion on the set of pre-college covariates as reported in Table 1. Standard errors for matching estimates of the δ_{TT} are bootstrapped based on 50 replications.

† p<.10 * p <.05 ** p < .01 *** p < .001 (two-tailed tests)

Table 4. Heterogeneous Effects of College Completion on Civic Participation (N = 3,452)

	Civic, Community, or Youth Groups	Charitable Orgs. or Social Welfare Groups
Level-1 Logit Regressions		
P-Score Stratum 1: [.0-.1)	2.031 ***	1.124 †
n = 1486	(.445)	(.563)
P-Score Stratum 2: [.1-.2)	.473	-.086
n = 553	(.481)	(.637)
P-Score Stratum 3: [.2-.3)	.844 *	.689
n = 345	(.410)	(.463)
P-Score Stratum 4: [.3-.4)	.755 †	.374
n = 234	(.434)	(.573)
P-Score Stratum 5: [.4-.6)	.098	.734 †
n = 343	(.352)	(.410)
P-Score Stratum 6: [.6-.8)	.673	-.301
n = 290	(.515)	(.402)
P-Score Stratum 7: [.8-1.0)	-.020	-.509
n = 201	(.584)	(.677)
Level-2 Variance Weighted Least Squares Regressions	-.259 **	-.196 †
	(.092)	(.104)

Notes: Numbers in parentheses are standard errors. Propensity scores were estimated by a logit regression model of college completion on the set of pre-college covariates as reported in Table 1. Propensity score strata were balanced such that mean values of covariates and the propensity score did not significantly differ between college and non-college graduates.

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

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

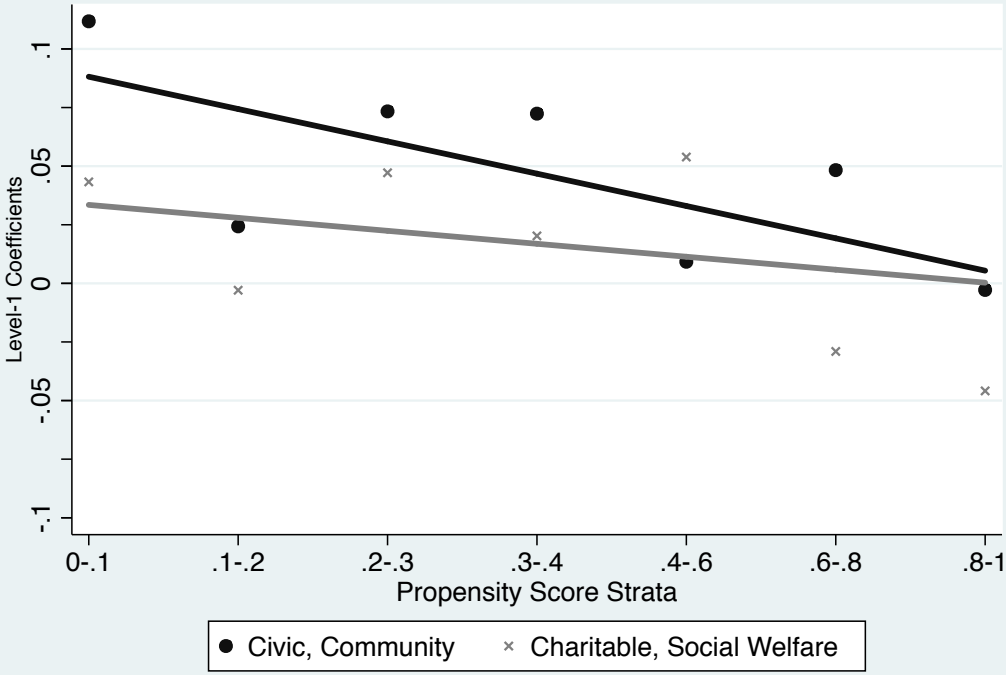
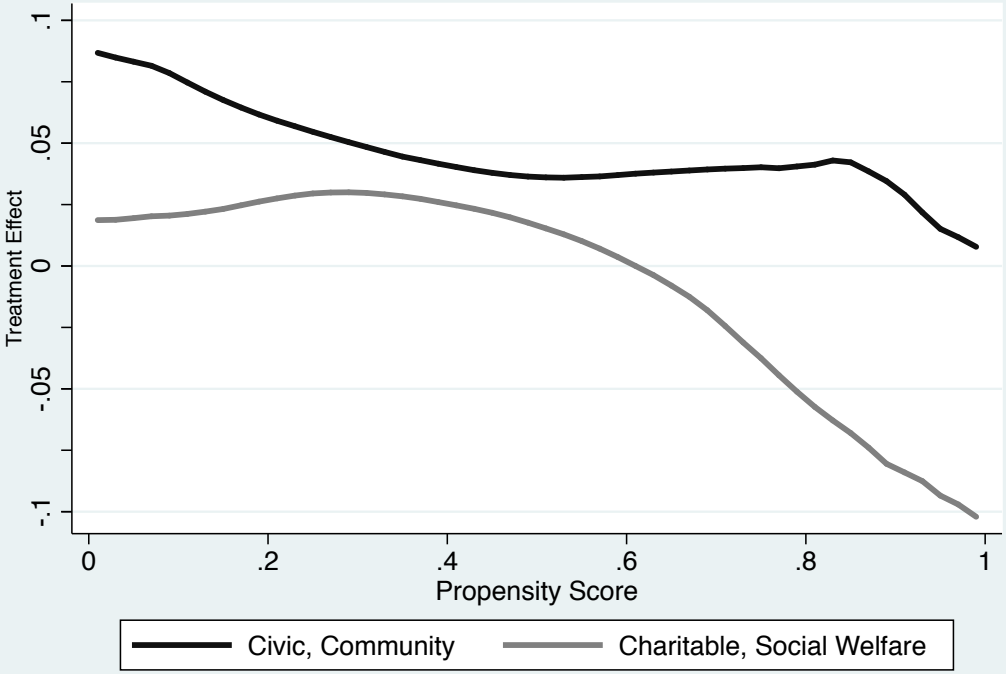


Figure 2. College Effects on Volunteering (SD-HTE)



Appendix A. Descriptive Statistics of Pre-College Covariates and Civic Participation (N=3,452)

	No College Completion		College Completion	
Variables	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
Pre-College Covariates				
Male (0-1)	0.487	0.500	0.484	0.500
Black (0-1)	0.176	0.381	0.083	0.276
Hispanic (0-1)	0.075	0.263	0.032	0.177
Mother's education (years of schooling)	11.130	2.395	13.133	2.383
Father's education (years of schooling)	11.070	3.049	13.933	3.250
Parents' income (1979 dollars)	183.481	107.372	273.763	138.860
Intact family age 14 (0-1)	0.698	0.459	0.826	0.379
Number of siblings age 14	3.372	2.321	2.600	1.686
Southern residence age 14 (0-1)	0.335	0.469	0.296	0.453
Mental ability*	-0.015	0.638	0.718	0.527
College-prep (0-1)	0.223	0.401	0.573	0.485
Parents' encouraged college (0-1)	0.650	0.465	0.882	0.320
Friends' plans (years of schooling)	0.428	0.492	0.803	0.397
Civic Participation				
Civic, Community, Youth Groups (0-1)	0.050	0.219	0.129	0.335
Charitable. Orgs., Social Welfare Groups (0-1)	0.041	0.198	0.085	0.278
Sample Size	2592		860	
Weighted Sample Proportion	0.69		0.31	

Notes: Ability is measured with a scale of standardized residuals of the ASVAB. All statistics are weighted for sample selection and nonresponse.

Appendix B. Covariate and Outcome Means by Propensity Score Strata and College Completion (N=3,452)

	Stratum 1 [.0-.1)			Stratum 2 [.1-.2)			Stratum 3 [.2-.3)			Stratum 4 [.3-.4)			Stratum 5 [.4-.6)			Stratum 6 [.6-.8)			Stratum 7 [.8-1.0)		
Variables	***E(X,Y) | d=0***	***E(X,Y) | d=1***	***B***	***E(X,Y) | d=0***	***E(X,Y) | d=1***	***B***	***E(X,Y) | d=0***	***E(X,Y) | d=1***	***B***	***E(X,Y) | d=0***	***E(X,Y) | d=1***	***B***	***E(X,Y) | d=0***	***E(X,Y) | d=1***	***B***	***E(X,Y) | d=0***	***E(X,Y) | d=1***	***B***	***E(X,Y) | d=0***	***E(X,Y) | d=1***	***B***
Male	0.509	0.426	0.17	0.461	0.472	0.02	0.472	0.412	0.12	0.431	0.477	0.09	0.519	0.518	0.00	0.449	0.507	0.12	0.458	0.435	0.05
Black	0.401	0.361	0.08	0.312	0.427	0.24	0.287	0.250	0.08	0.205	0.273	0.16	0.223	0.189	0.08	0.225	0.129	0.25	0.167	0.101	0.12
Hispanic	0.247	0.311	0.14	0.190	0.191	0.00	0.124	0.162	0.11	0.130	0.114	0.05	0.151	0.085	0.20	0.124	0.065	0.20	0.042	0.068	0.11
Mother's edu.	9.577	9.639	0.02	10.760	10.639	0.05	11.335	11.874	0.25	11.897	11.840	0.02	12.184	12.695	0.22	13.553	12.908	0.27	14.039	15.108	0.34
Father's edu.	9.100	9.289	0.06	10.881	10.447	0.13	11.749	12.070	0.12	11.970	12.207	0.09	12.623	13.087	0.16	14.079	14.229	0.05	15.250	16.028	0.21
Parents' inc./1000	12.474	12.893	0.05	16.184	14.503	0.16	18.177	20.179	0.21	19.645	18.716	0.09	21.147	21.743	0.06	25.522	26.334	0.07	30.128	37.998	0.47
Intact family	0.571	0.525	0.09	0.684	0.539	0.30	0.660	0.750	0.20	0.678	0.648	0.06	0.704	0.823	0.28	0.809	0.866	0.15	0.917	0.881	0.12
Num. of siblings	4.469	4.279	0.07	3.687	3.437	0.11	3.026	3.062	0.02	3.055	3.068	0.01	2.709	2.549	0.09	2.416	2.632	0.13	2.333	2.367	0.02
Southern res.	0.423	0.427	0.01	0.393	0.358	0.07	0.318	0.432	0.24	0.342	0.334	0.02	0.332	0.360	0.06	0.391	0.295	0.20	0.333	0.294	0.08
Mental ability	-0.417	-0.166	0.54	0.162	0.228	0.15	0.410	0.344	0.15	0.532	0.582	0.12	0.751	0.727	0.06	0.922	0.946	0.06	1.176	1.222	0.11
College-prep.	0.087	0.087	0.00	0.262	0.339	0.17	0.362	0.391	0.06	0.439	0.382	0.12	0.571	0.559	0.02	0.690	0.738	0.11	0.792	0.815	0.06
Parents' enc.	0.598	0.586	0.03	0.739	0.779	0.10	0.755	0.809	0.13	0.828	0.838	0.03	0.881	0.871	0.03	0.944	0.940	0.02	1.000	0.959	0.29
Friends' plans	0.200	0.443	0.54	0.558	0.559	0.00	0.692	0.575	0.24	0.783	0.750	0.08	0.838	0.791	0.12	0.865	0.950	0.30	0.917	0.966	0.21
Propensity score	0.042	0.062	0.80	0.143	0.148	0.18	0.247	0.246	0.03	0.345	0.354	0.32	0.482	0.504	0.36	0.680	0.702	0.38	0.857	0.895	0.77
Civic, Community, Youth	0.020	0.131		0.043	0.067		0.065	0.137		0.075	0.148		0.101	0.110		0.056	0.104		0.167	0.164	
Charitable, Social Welfare	0.022	0.066		0.037	0.034		0.053	0.100		0.048	0.068		0.056	0.110		0.124	0.094		0.125	0.079	
Sample Size	1425	61		464	89		265	80		146	88		179	164		89	201		24	177	

Notes: E(X,Y)|d=0 indicates the mean of *X* or *Y* for individuals who did not complete college and *E(X,Y)|d*=1 indicates the mean of *X* or *Y* for individuals who completed college. All statistics are weighted for sample selection and nonresponse. ***B*** is the standardized difference metric between the treated and control groups for ***X***.