Sociological Methodology
2020, Vol. 50(1) 350–385
© American Sociological Association 2019
DOI: 10.1177/0081175019862593
http://sm.sagepub.com

8

HETEROGENEOUS TREATMENT EFFECTS IN THE PRESENCE OF SELF-SELECTION: A PROPENSITY SCORE PERSPECTIVE

Xiang Zhou*
Yu Xie[†]

Abstract

An essential feature common to all empirical social research is variability across units of analysis. Individuals differ not only in background characteristics but also in how they respond to a particular treatment, intervention, or stimulation. Moreover, individuals may self-select into treatment on the basis of anticipated treatment effects. To study heterogeneous treatment effects in the presence of self-selection, Heckman and Vytlacil developed a structural approach that builds on the marginal treatment effect (MTE). In this article, we extend the MTE-based approach through a redefinition of MTE. Specifically, we redefine MTE as the expected treatment effect conditional on the propensity score (rather than all observed covariates) as well as a latent variable representing unobserved resistance to treatment. As with the original MTE, the new MTE also can be used as a building block for evaluating standard causal estimands. However, the weights associated with the new MTE are simpler, more intuitive, and easier to compute. Moreover, the new MTE is a bivariate function and thus is easier to visualize than the

*Harvard University, Cambridge, MA, USA
[†]Princeton University, Princeton, NJ, USA

Corresponding Author:
Xiang Zhou, Department of Government, Harvard University, 1737 Cambridge St., Cambridge, MA 02138, USA.
Email: xiang_zhou@fas.harvard.edu

original MTE. Finally, the redefined MTE immediately reveals treatment-effect heterogeneity among individuals who are at the margin of treatment. As a result, it can be used to evaluate a wide range of policy changes with little analytical twist and design policy interventions that optimize the marginal benefits of treatment. We illustrate the proposed method by estimating heterogeneous economic returns to college with National Longitudinal Study of Youth 1979 data.

Keywords

causal effects, heterogeneity, instrumental variable, marginal treatment effect, propensity score, selection bias

1. INTRODUCTION

An essential feature common to all empirical social research is variability across units of analysis. Individuals differ not only in background characteristics but also in how they respond to a particular treatment, intervention, or stimulation. In the language of causal inference, the second type of variability is called *treatment-effect heterogeneity*. Due to the ubiquity of treatment-effect heterogeneity, all statistical methods designed for drawing causal inferences can identify causal effects only at an aggregate level; they overlook within-group, individual-level heterogeneity (Holland 1986; Xie 2013). Moreover, when treatment effects vary systematically by treatment status, the average difference in outcome between the treated and untreated units is a biased estimate of the average treatment effect in the population (Winship and Morgan 1999).

Depending on data and assumptions about how individuals select into treatment, three major approaches have been proposed to studying heterogeneous treatment effects. First, we simply can include interaction terms between treatment status and a set of effect modifiers in a standard regression model. A drawback of this approach is that the results may be sensitive to the functional form specifying how treatment and covariates jointly influence the outcome of interest. Fortunately, recent developments in nonparametric modeling have allowed the idea to be implemented without strong functional form restrictions (e.g., Hill 2011). Second, recent sociological studies have focused on how treatment effect varies by the propensity score, that is, the probability of treatment given a set of observed covariates (e.g., Brand and Xie 2010; Xie, Brand, and Jann 2012). The methodological rationale for this approach is that under the assumption of ignorability, the interaction between treatment status and the propensity score captures all the treatment-effect heterogeneity that is consequential for selection bias (Rosenbaum and Rubin 1983). Treatment-effect heterogeneity along the propensity score also has profound policy implications. For instance, if the benefits of a job training program are greater among individuals who are more likely to enroll in the program, expanding the size of the program may reduce its average effectiveness.

The aforementioned two approaches for studying heterogeneous treatment effects both rely on the assumption of ignorability, that is, after controlling for a set of observed confounders, treatment status is independent of potential outcomes. This assumption is strong, unverifiable, and unlikely to be true in most observational studies. Two types of unobserved selection may invalidate the ignorability assumption. On the one hand, if treatment status is correlated with some fixed unobserved characteristics such that treated units would have different outcomes from untreated units even without treatment, traditional regression and matching methods would lead to biased estimates of average causal effects. This bias is usually called *pretreatment heterogeneity bias* or *Type I selection bias* (Xie et et al. 2012). As Breen, Choi, and Holm (2015) show, this type of selection easily could contaminate estimates of heterogeneous treatment effects by observed covariates or the propensity score. A variety of statistical and econometric methods, such as instrumental variables (IV), fixed-effects models, and regression discontinuity (RD) designs, have been developed to address pretreatment heterogeneity bias.

The second type of unobserved selection arises when treatment status is correlated with treatment effect in a way that is not captured by observed covariates. This is likely when individuals (or their agents) possess more knowledge than the researcher about their individual-specific gains (or losses) from treatment and act on it (Bjorklund and Moffitt 1987; Heckman and Vytlacil 2005; Roy 1951). The bias associated with this type of selection has been termed *treatment-effect heterogeneity bias* or *Type II selection bias* (Xie et al. 2012). For example, research considering heterogeneous returns to schooling has argued that college education is selective because it disproportionately attracts young persons who would gain more from attending college (e.g., Carneiro, Heckman, and Vytlacil 2011; Moffitt 2008; Willis and Rosen 1979). Similar patterns of self-selection have been observed in a variety of contexts, such as migration (Borjas 1987), secondary-school tracking (Gamoran and Mare 1989), career choice (Sakamoto and Chen 1991), and marriage dissolution (Smock, Manning, and Gupta 1999).

The third approach, developed by Heckman and Vytlacil (1999, 2001a, 2005, 2007b), accommodates both types of unobserved selection through the use of a latent index model for treatment assignment. Under this model, all the treatment-effect heterogeneity relevant for selection bias is captured in the marginal treatment effect (MTE), a function defined as the conditional expectation of treatment effect given observed covariates and a latent variable representing unobserved, individual-specific resistance to treatment. This approach has been called the *MTE-based approach* (Zhou and Xie 2016). As Heckman, Urzua, and Vytlacil (2006) show, a wide range of causal estimands, such as the average treatment effect (ATE) and the treatment effect of the treated (TT), can be expressed as weighted averages of MTE. Moreover, MTE can be used to evaluate average treatment effects among individuals at the margin of indifference to treatment, thus allowing researchers to assess the efficacy of marginal policy changes (Carneiro et al. 2010). For example, using data from the 1979 National Longitudinal Survey of Youth (NLSY79), Carneiro and colleagues (2011) found that if a policy change expanded each individual's probability of attending college by the same proportion, the estimated return to one year of college education among marginal entrants to college would be only 1.5 percent, far lower than the estimated population average of 6.7 percent.

In the MTE framework, the latent index model ensures that all unobserved determinants of treatment status are summarized by a single latent variable and that the variation of treatment effect by this latent variable captures all the treatment-effect heterogeneity that may cause selection bias. Our basic intuition is that under this model, treatment-effect heterogeneity that is consequential for selection bias occurs only along two dimensions: (1) the observed probability of treatment (i.e., the propensity score) and (2) the latent variable for unobserved resistance to treatment. In other words, after unobserved selection is factored in through the latent variable, the propensity score is the only dimension along which treatment effect may be correlated with treatment status. Therefore, to identify population-level and subpopulation-level causal effects such as ATE and TT, it would be sufficient to model treatment effect as a bivariate function of the propensity score and the latent variable. In this article, we show that such a bivariate function is not only analytically sufficient but also essential to the evaluation of policy effects.

Specifically, we redefine MTE as the expected treatment effect conditional on the propensity score (rather than the entire vector of observed covariates) and the latent variable representing unobserved resistance to treatment. This redefinition offers a novel perspective to interpret and analyze MTE that supplements the current approach. First, although projected onto a unidimensional summary of covariates, the redefined MTE is sufficient to capture all the treatment-effect heterogeneity that is consequential for selection bias. Thus, as with the original MTE, it can be used as a building block for constructing standard causal estimands such as ATE and TT. The weights associated with the new MTE, however, are simpler, more intuitive, and easier to compute. Second, by discarding treatment effect variation that is orthogonal to the two-dimensional space spanned by the propensity score and the latent variable, the redefined MTE is a bivariate function, thus easier to visualize than the original MTE. Finally, and perhaps most importantly, the redefined MTE immediately reveals treatment-effect heterogeneity among individuals who are at the margin of treatment. It can thus be used to evaluate a wide range of policy effects with little analytical twist and design policy interventions that optimize the marginal benefits of treatment. To facilitate practice, we also provide an R package, localIV, for estimating the redefined MTE as well as the original MTE via local instrumental variables (Zhou 2019), which is available from the Comprehensive R Archive Network (CRAN).

This article is clearly not the first to characterize the problem of selection bias using the propensity score. Since the seminal work of Rosenbaum and Rubin (1983), propensity score–based methods, such as matching, weighting, and regression adjustment, have been a mainstay strategy for drawing causal inferences in the social sciences. In a series of articles, Heckman and colleagues established the key roles of the propensity score in a variety of econometric methods, including control functions, instrumental variables, and the MTE approach (Heckman 2010; Heckman and Hotz 1989; Heckman and Navarro-Lozano 2004; Heckman and Robb 1986).[1] In the MTE approach, for example, incremental changes in the propensity score serve as “local instrumental variables” that identify the MTE at various values of the unobserved resistance to treatment. Moreover, the weights with which MTE can be aggregated up to standard causal estimands depend solely on the conditional distribution of the propensity score given covariates. We show that the propensity score offers not only a tool for identification but also a perspective from which we can better summarize, interpret, and analyze treatment-effect heterogeneity due to both observed and unobserved characteristics.

The rest of this article is organized as follows. Section 2 reviews the MTE-based approach for studying heterogeneous treatment effects. Specifically, we discuss the generalized Roy model for treatment selection, the definition and properties of MTE, and the estimation of MTE and related weights. Section 3 presents our new approach that builds on the redefinition of MTE. The redefined MTE enables us to directly examine the variation of ATE, TT, and policy-relevant causal effects across individuals with different values of the propensity score. In this framework, designing a policy intervention boils down to weighting individuals with different propensities of treatment. Section 4 illustrates our new approach by estimating heterogeneous economic returns to college with NLSY79 data. Section 5 discusses our conclusions.

2. THE MTE-BASED APPROACH: A REVIEW

2.1. *The Generalized Roy Model*

The MTE approach builds on the generalized Roy model for discrete choices (Heckman and Vytlacil 2007a; Roy 1951). Consider two potential outcomes, Y_1 and Y_0, a binary indicator D for treatment status, and a vector of pretreatment covariates X. Y_1 denotes the potential outcome if the individual were treated ($D=1$), and Y_0 denotes the potential outcome if the individual were not treated ($D=0$). We specify the outcome equations as

$$Y_0 = \mu_0(X) + \epsilon \tag{1}$$

$$Y_1 = \mu_1(X) + \epsilon + \eta, \tag{2}$$

where $\mu_0(X) = \mathbb{E}[Y_0|X]$, $\mu_1(X) = \mathbb{E}[Y_1|X]$, the error term ϵ captures all unobserved factors that affect the baseline outcome (Y_0), and the error term η captures all unobserved factors that affect the treatment effect ($Y_1 - Y_0$). In general, the error terms ϵ and η need not be statistically independent of X, although they have zero conditional means by construction. The observed outcome Y can be linked to the potential outcomes through the switching regression model (Quandt 1958, 1972):

$$\begin{aligned} Y &= (1-D)Y_0 + DY_1 \\ &= \mu_0(X) + (\mu_1(X) - \mu_0(X))D + \epsilon + \eta D. \end{aligned} \tag{3}$$

Treatment assignment is represented by a latent index model. Let I_D be a latent tendency for treatment, which depends on both observed (Z) and unobserved (V) factors:

$$I_D = \mu_D(Z) - V \tag{4}$$

$$D = \mathbb{I}(I_D > 0), \tag{5}$$

where $\mu_D(Z)$ is an unspecified function and V is a latent random variable representing unobserved, individual-specific resistance to treatment, assumed to be continuous with a strictly increasing distribution function. The Z vector includes all the components of X, but it also includes some instrumental variables (IV) that affect only the treatment status D. The key assumptions associated with Equations 1, 2, 4, and 5 are

Assumption 1. (ϵ, η, V) are statistically independent of Z given X (independence).

Assumption 2. $\mu_D(Z)$ is a nontrivial function of Z given X (rank condition).

The latent index model characterized by Equations 4 and 5 combined with Assumptions 1 and 2 is equivalent to the Imbens-Angrist (Imbens and Angrist 1994) assumptions of independence and monotonicity for the interpretation of IV estimands as local average treatment effects (LATE; Vytlacil 2002). Given Assumptions 1 and 2, the latent resistance V is allowed to be correlated with ϵ and η in a general way. For example, research considering heterogeneous returns to schooling has argued that individuals may self-select into college on the basis of their anticipated gains. In this case, V will be negatively correlated with η because individuals with higher values of η tend to have lower levels of unobserved resistance U.[2]

2.2. *Marginal Treatment Effects*

To define the MTE, it is best to rewrite the treatment assignment Equations 4 and 5 as

Table 1. Weights for Constructing ATE(x), TT(x), and TUT(x) from MTE(x, u)

Quantities of Interest	Weight
ATE(x)	$h_{\mathrm{ATE}}(x,u)=1$
TT(x)	$h_{\mathrm{TT}}(x,u)=\frac{\int_u^1 f_{P(Z)\mid X=x}(p)dp}{\mathbb{E}(P(Z)\mid X=x)}$
TUT(x)	$h_{\mathrm{TUT}}(x,u)=\frac{\int_0^u f_{P(Z)\mid X=x}(p)dp}{1-\mathbb{E}(P(Z)\mid X=x)}$

Note: ATE = average treatment effect; TT = treatment effect of the treated; TUT = treatment effect of the untreated; MTE = marginal treatment effect.

$$\begin{aligned} D &= \mathbb{I}(F_{V|X}(\mu_D(Z)) - F_{V|X}(V)>0) \\ &= \mathbb{I}(P(Z) - U>0), \end{aligned} \tag{6}$$

where $F_{V|X}(\cdot)$ is the cumulative distribution function of V given X and $P(Z)=Pr(D=1|Z)=F_{V|X}(\mu_D(Z))$ denotes the propensity score given Z. $U=F_{V|X}(V)$ is the quantile of V given X, which by definition follows a standard uniform distribution. From Equation 6, we can see that Z affects treatment status only through the propensity score $P(Z)$.[3]

The MTE is defined as the expected treatment effect conditional on pretreatment covariates $X=x$ and the normalized latent variable $U=u$:

$$\begin{aligned} \mathrm{MTE}(x,u) &= \mathbb{E}[Y_1 - Y_0|X=x, U=u] \\ &= \mathbb{E}[\mu_1(X) - \mu_0(X)+\eta|X=x, U=u] \\ &= \mu_1(x) - \mu_0(x)+\mathbb{E}[\eta|X=x, U=u]. \end{aligned} \tag{7}$$

Because U is the quantile of V, the variation of MTE(x, u) over values of u reflects how treatment effect varies with different quantiles of the unobserved resistance to treatment. Alternatively, MTE(x, u) can be interpreted as the average treatment effect among individuals who are indifferent between treatment or not with covariates $X=x$ and the propensity score $P(Z)=u$.

A wide range of causal estimands, such as ATE and TT, can be expressed as weighted averages of MTE(x, u) (Heckman et al. 2006). To obtain population-level causal effects, MTE(x, u) needs to be integrated twice, first over u given $X=x$ and then over x. The weights for integrating over u are shown in Table 1. Note that these weights are conditional on $X=x$. To estimate overall ATE, TT, and treatment effect of the untreated (TUT), we need to further integrate estimates of ATE(x), TT(x), and TUT(x) against appropriate marginal distributions of X.

The estimation of MTE(x, u), however, is not straightforward because neither the counterfactual outcome nor the latent variable U is observed. Moreover, the estimation of weights can be practically challenging (except for the ATE case) because it involves estimating the conditional density of $P(Z)$ given X and the latter is usually a high-dimensional vector. We turn to these estimation issues now.

2.3. *Estimation of MTE and Weights in Practice*

Given Assumptions 1 and 2, MTE(x, u) can be nonparametrically identified using the method of local instrumental variables (LIV).[4] To see how it works, let us first write out the expectation of the observed outcome Y given the covariates $X=x$ and the propensity score $P(Z)=p$. According to Equation 3, we have

$$\begin{aligned}\mathbb{E}[Y|X=x,P(Z)=p] &= \mathbb{E}[\mu_0(X)+(\mu_1(X)-\mu_0(X))D+\epsilon+\eta D|X=x,P(Z)=p]\\ &= \mu_0(x)+(\mu_1(x)-\mu_0(x))p+\mathbb{E}[\eta|D=1,X=x,P(Z)=p]p\\ &= \mu_0(x)+(\mu_1(x)-\mu_0(x))p+\int_0^p \mathbb{E}[\eta|X=x,U=u]du.\end{aligned} \tag{8}$$

Taking the partial derivative of Equation 8 with respect to p, we have

$$\begin{aligned}\frac{\partial\mathbb{E}[Y|X=x,P(Z)=p]}{\partial p} &= \mu_1(x)-\mu_0(x)+\mathbb{E}[\eta|X=x,U=p]\\ &= \mathrm{MTE}(x,p).\end{aligned}$$

Because $\mathbb{E}(Y|X=x,P(Z)=p)$ is a function of observed (or estimable) quantities, the previous equation means MTE(x, u) is identified as long as u falls within supp($P(Z)|X$), the conditional support of $P(Z)$ given $X=x$. In other words, MTE(x, u) is nonparametrically identified over supp($X, P(Z)$), the support of the joint distribution of X and $P(Z)$.

In practice, however, it is difficult to condition on X nonparametrically, especially when X is high-dimensional. Therefore, in most empirical work using LIV, it is assumed that (X, Z) is jointly independent of (ϵ, η, V) (e.g., Carneiro et al. 2011; Carneiro and Lee 2009; Maestas, Mullen, and Strand 2013). Under this assumption, the MTE is additively separable in x and u:

$$\begin{aligned}\mathrm{MTE}(x,u) &= \mu_1(x) - \mu_0(x) + \mathbb{E}[\eta|X=x,\,U=u]\\ &= \mu_1(x) - \mu_0(x) + \mathbb{E}[\eta|U=u].\end{aligned} \tag{9}$$

The additive separability not only simplifies estimation, but it allows $\mathrm{MTE}(x,u)$ to be identified over $\mathrm{supp}(X)\times\mathrm{supp}(P(Z))$ (instead of $\mathrm{supp}(X,P(Z))$). The previous equation also suggests a necessary and sufficient condition for the MTE to be additively separable:

Assumption 3. $\mathbb{E}[\eta|X=x,U=u]$ does not depend on x (additive separability).

This assumption is implied by (but does not imply) the full independence between (X,Z) and (ϵ,η,V) (for a similar discussion, see Brinch, Mogstad, and Wiswall 2017).

In most applied work, the conditional means of Y_0 and Y_1 given X are further specified as linear in parameters: $\mu_0(X)=\beta_0^TX$ and $\mu_1(X)=\beta_1^TX$. Given the linear specification and Assumptions 1, 2, and 3, $\mathbb{E}[Y|X=x,P(Z)=p]$ can be written as

$$\mathbb{E}[Y|X=x,P(Z)=p]=\beta_0^Tx+(\beta_1-\beta_0)^Txp+\underbrace{\int_0^p\mathbb{E}[\eta|U=u]du}_{K(p)}, \tag{10}$$

where $K(p)$ is an unknown function that can be estimated either parametrically or nonparametrically.[5]

First, in the special case where the error terms (ϵ,η,V) are assumed to be jointly normal with zero means and an unknown covariance matrix Σ, the generalized Roy model characterized by Equations 1, 2, 4, and 5 is fully parameterized, and the unknown parameters $(\beta_1,\beta_0,\gamma,\Sigma)$ can be jointly estimated via maximum likelihood.[6] This model specification has a long history in econometrics and is usually called the "normal switching regression model" (Heckman 1978; for a review, see Winship and Mare 1992). With the joint normality assumption, Equation 9 reduces to

$$\mathrm{MTE}(x,u)=(\beta_1-\beta_0)^Tx+\frac{\sigma_{\eta V}}{\sigma_V}\Phi^{-1}(u), \tag{11}$$

where $\sigma_{\eta V}$ is the covariance between η and V, σ_V is the standard deviation of V, and $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal distribution function.[7] By plugging in the maximum likelihood estimates

(MLE) of $(\beta_1, \beta_0, \sigma_{\eta V}, \sigma_V)$, we obtain an estimate of $\text{MTE}(x, u)$ for any combination of x and u.

The joint normality of error terms is a strong and restrictive assumption. When errors are not normally distributed, imposition of normality may lead to substantial bias in estimates of the model parameters (Arabmazar and Schmidt 1982). To avoid this problem, Heckman and colleagues (2006) proposed to fit Equation 10 semiparametrically using a double residual procedure (Robinson 1988). In this case, the estimation of $\text{MTE}(x, u)$ can be summarized in four steps:

1. Estimate the propensity scores using a standard logit/probit model, denote them as $\hat{P}$.[8]
2. Fit local linear regressions of Y, X, and $X\hat{P}$ on $\hat{P}$ and extract their residuals e_Y, e_X, and $e_{X\hat{P}}$.
3. Fit a simple linear regression of e_Y on e_X and $e_{X\hat{P}}$ (with no intercepts) to estimate the parametric part of Equation 10, that is, $(\beta_0, \beta_1 - \beta_0)$, and store the remaining variation of Y as $e_Y^* = Y - \hat{\beta}_0^T X - (\hat{\beta}_1 - \hat{\beta}_0)^T X\hat{P}$.
4. Fit a local quadratic regression (Fan and Gijbels 1996) of e_Y^* on $\hat{P}$ to estimate $K(p)$ and its derivative $K'(p)$.

The MTE is then estimated as

$$\widehat{\text{MTE}}(x, u) = (\hat{\beta}_1 - \hat{\beta}_0)^T x + \hat{K}'(u). \tag{12}$$

With estimates of $\text{MTE}(x, u)$, we still need appropriate weights to estimate aggregate causal effects such as ATE and TT. As shown in Table 1, most weights involve the conditional density of $P(Z)$ given X. Because the latter is often a high-dimensional vector, direct estimation of these weights is challenging. In their empirical application, Carneiro and colleagues (2011) conditioned on an index of X, $(\hat{\beta}_1 - \hat{\beta}_0)^T X$, instead of X. In other words, they used $f[\hat{P}|(\hat{\beta}_1 - \hat{\beta}_0)^T X]$ as an approximation to $f[P(Z)|X]$. To estimate the former, we can first estimate the bivariate density $f[\hat{P}, (\hat{\beta}_1 - \hat{\beta}_0)^T X]$ using kernel methods and then divide the estimated bivariate density by the marginal density $f[(\hat{\beta}_1 - \hat{\beta}_0)^T X]$. As we will see, these ad hoc methods for estimating weights are no longer needed with our new approach.

3. A PROPENSITY SCORE PERSPECTIVE

3.1. *A Redefinition of MTE*

Under the generalized Roy model, a single latent variable U not only summarizes all unobserved determinants of treatment status but also captures all the treatment-effect heterogeneity by unobserved characteristics that may cause selection bias. In fact, the latent index structure implies that all the treatment-effect heterogeneity that is consequential for selection bias exists only along two dimensions: (1) the propensity score $P(Z)$ and (2) the latent variable U representing unobserved resistance to treatment. This is directly reflected in Equation 6: a person is treated if and only if his or her propensity score exceeds his or her (realized) latent resistance u. Therefore, given both $P(Z)$ and U, treatment status D is fixed (either 0 or 1) and thus independent of treatment effect:

$$Y_1 - Y_0 \perp D | P(Z), U.$$

This expression resembles the Rosenbaum and Rubin (1983) result on the sufficiency of the propensity score except that we now condition on U in addition to $P(Z)$. Thus, to characterize selection bias, it would be sufficient to model treatment effect as a bivariate function of the propensity score (rather than the entire vector of covariates) and the latent variable U. We redefine MTE as the expected treatment effect given $P(Z)$ and U:

$$\widetilde{\mathrm{MTE}}(p, u) \triangleq \mathbb{E}[Y_1 - Y_0 | P(Z) = p, U = u]. \tag{13}$$

Compared with the original MTE, $\widetilde{\mathrm{MTE}}(p, u)$ has two immediate advantages. First, because it conditions on the propensity score $P(Z)$ rather than the whole vector of X, it captures all the treatment-effect heterogeneity that is relevant for selection bias in a more parsimonious way. Second, by discarding treatment effect variation that is orthogonal to the two-dimensional space spanned by $P(Z)$ and U, $\widetilde{\mathrm{MTE}}(p, u)$ is a bivariate function and thus easier to visualize than $\mathrm{MTE}(x, u)$.

As with $\mathrm{MTE}(x, u)$, $\widetilde{\mathrm{MTE}}(p, u)$ also can be used as a building block for constructing standard causal estimands such as ATE and TT. However, compared with the weights on $\mathrm{MTE}(x, u)$, the weights on $\widetilde{\mathrm{MTE}}(p, u)$ are simpler, more intuitive, and easier to compute. The weights for ATE, TT, and TUT are shown in the first three rows of Table 2. To construct $\mathrm{ATE}(p)$, we simply integrate $\widetilde{\mathrm{MTE}}(p, u)$ against

Table 2. Weights for Constructing ATE, TT, TUT, PRTE, and MPRTE from $\widetilde{\text{MTE}}(p, u)$

Quantities of Interest	Weight
$\text{ATE}(p)$	$h_{\text{ATE}}(p, u) = 1$
$\text{TT}(p)$	$h_{\text{TT}}(p, u) = \frac{1(u < p)}{p}$
$\text{TUT}(p)$	$h_{\text{TUT}}(p, u) = \frac{1(u \geq p)}{1 - p}$
$\text{PRTE}(p, \lambda(p))$	$h_{\text{PRTE}}(p, u) = \frac{1(p \leq u < p + \lambda(p))}{\lambda(p)}$
$\text{MPRTE}(p)$	$h_{\text{MPRTE}}(p, u) = \delta(u - p)$

Note: ATE = average treatment effect; TT = treatment effect of the treated; TUT = treatment effect of the untreated; PRTE = policy-relevant treatment effect; MPRTE = marginal policy-relevant treatment effect. $\delta(\cdot)$ is the Dirac delta function.

the marginal distribution of U—a standard uniform distribution. To construct $\text{TT}(p)$, we integrate $\widetilde{\text{MTE}}(p, u)$ against the truncated distribution of U given $U < p$. Likewise, to construct $\text{TUT}(p)$, we integrate $\widetilde{\text{MTE}}(p, u)$ against the truncated distribution of U given $U \geq p$. To obtain population-level ATE, TT, and TUT, we further integrate $\text{ATE}(p)$, $\text{TT}(p)$, and $\text{TUT}(p)$ against appropriate marginal distributions of $P(Z)$. For example, to construct TT, we integrate $\text{TT}(p)$ against the marginal distribution of the propensity score among treated units.

In practice, $\widetilde{\text{MTE}}(p, u)$ can be estimated as a byproduct of $\text{MTE}(x, u)$. Under Assumptions 1, 2, and 3,[9] $\widetilde{\text{MTE}}(p, u)$ can be written as

$$\widetilde{\text{MTE}}(p, u) = \mathbb{E}[\mu_1(X) - \mu_0(X) | P(Z) = p] + \mathbb{E}[\eta | U = u]. \tag{14}$$

A proof of Equation 14 is given in Appendix A. Comparing Equation 14 with Equation 9, we see that the only difference between the original MTE and $\widetilde{\text{MTE}}(p, u)$ is that the first component of $\widetilde{\text{MTE}}(p, u)$ is now the conditional expectation of $\mu_1(X) - \mu_0(X)$ given the propensity score rather than $\mu_1(X) - \mu_0(X)$. Therefore, to estimate $\widetilde{\text{MTE}}(p, u)$, we need only one more step after implementing the procedures described in Section 2.3: fit a nonparametric curve of $(\hat{\beta}_1 - \hat{\beta}_1)^T X$ with respect to $\hat{P}$ (e.g., using a local linear regression) and combine it with existing estimates of $K'(u)$.

3.2. *Policy-Relevant Causal Effects*

The redefined MTE can be used not only to construct traditional causal estimands but also, in the context of program evaluation, to draw implications for how the program should be revised in the future. To predict the impact of an expansion (or contraction) in program participation, one needs to examine treatment effects for individuals who would be affected by such an expansion (or contraction). To formalize this idea, Heckman and Vytlacil (2001b, 2005) define the policy-relevant treatment effect (PRTE) as the mean effect of moving from a baseline policy to an alternative policy per net person shifted into treatment, that is,

$$PRTE \triangleq \frac{\mathbb{E}(Y|\text{Alternative Policy}) - \mathbb{E}(Y|\text{Baseline Policy})}{\mathbb{E}(D|\text{Alternative Policy}) - \mathbb{E}(D|\text{Baseline Policy})}.$$

They further show that under the generalized Roy model, the PRTE depends on a policy change only through its effects on the distribution of the propensity score $P(Z)$. Specifically, conditional on $X=x$, the PRTE can be written as a weighted average of $\text{MTE}(x, u)$, where the weights depend only on the distribution of $P(Z)$ before and after the policy change. Within this framework, Carneiro and colleagues (2010) further define the marginal policy-relevant treatment effect (MPRTE) as a directional limit of the PRTE as the alternative policy converges to the baseline policy. Denoting by $F(\cdot)$, the cumulative distribution function of $P(Z)$, they consider a set of alternative policies indexed by a scalar α, $\{F_\alpha\colon \alpha \in \mathbb{R}\}$, such that F_0 corresponds to the baseline policy. The MPRTE is defined as

$$\text{MPRTE} = \lim_{\alpha \to 0} \text{PRTE}(F_\alpha).$$

We follow their approach to analyzing policy effects but without conditioning on X. Whereas Carneiro and colleagues (2010) assume that the effects of all policy changes are through shifts in the conditional distribution of $P(Z)$ given X, we focus on policy changes that shift the marginal distribution of $P(Z)$ directly. In other words, we consider policy interventions that incorporate individual-level treatment-effect heterogeneity by values of $P(Z)$, whether their differences in $P(Z)$ are induced by their baseline characteristics X or the instrumental variables $Z\backslash X$. In Section 3.5, we compare these two approaches in more detail and discuss some major advantages of our new approach.

Specifically, let us consider a class of policy changes under which the ith individual's propensity of treatment is boosted by $\lambda(p_i)$ (in a way that does not change his or her treatment effect), where p_i denotes the individual's propensity score $P(z_i)$ and $\lambda(\cdot)$ is a positive, real-valued function such that $p+\lambda(p) \leq 1$ for all $p \in [0,1)$. The policy change thus nudges everyone in the same direction, and two persons with the same baseline probability of treatment share an inducement of the same size. For such a policy change, the PRTE given $P(Z)=p<1$ and $\lambda(p)$ becomes

$$\mathrm{PRTE}(p,\lambda(p))=\mathbb{E}[Y_1-Y_0|p(Z)=p,p\leq U<p+\lambda(p)].$$

As with standard causal estimands, $\mathrm{PRTE}(p,\lambda(p))$ can be expressed as a weighted average of $\widetilde{\mathrm{MTE}}(p,u)$:

$$\mathrm{PRTE}(p,\lambda(p))=\frac{1}{\lambda(p)}\int_p^{p+\lambda(p)}\widetilde{\mathrm{MTE}}(p,u)du.$$

Here, the weight on u is constant (i.e., $1/\lambda(p)$) within the interval of $[p,p+\lambda(p))$ and zero elsewhere.

To examine the effects of marginal policy changes, let us consider a sequence of policy changes indexed by a real-valued scalar α. Given $P(Z)=p$, we define the MPRTE as the limit of $\mathrm{PRTE}(p,\alpha\lambda(p))$ as α approaches zero:

$$\begin{aligned}\mathrm{MPRTE}(p)&=\lim_{\alpha\to 0}\mathrm{PRTE}(p,\alpha\lambda(p))\\&=\mathbb{E}(Y_1-Y_0|p(Z)=p,U=p)\\&=\widetilde{\mathrm{MTE}}(p,p).\end{aligned}\tag{15}$$

Hence, we have established a direct link between $\mathrm{MPRTE}(p)$ and $\widetilde{\mathrm{MTE}}(p,u)$: At each level of the propensity score, the MPRTE is simply the $\widetilde{\mathrm{MTE}}$ at the margin where $u=p$. As shown in the last row of Table 2, $\mathrm{MPRTE}(p)$ can also be expressed as a weighted average of $\widetilde{\mathrm{MTE}}(p,u)$ using the Dirac delta function.

Figure 1 illustrates the relationships between ATE, TT, TUT, and MPRTE. Panel a shows a shaded gray plot of $\widetilde{\mathrm{MTE}}(p,u)$ for heterogeneous treatment effects in a hypothetical setup. In this plot, both the propensity score p and the latent resistance u (both ranging from 0 to 1) are divided into 10 equally spaced strata, yielding 100 grids, and a darker grid indicates a higher treatment effect. The advantage of such a shaded gray plot is that we can use subsets of the 100 grids to represent meaningful subpopulations. For example, we present the grids for treated units in Panel b, untreated units in Panel c, and marginal units in Panel

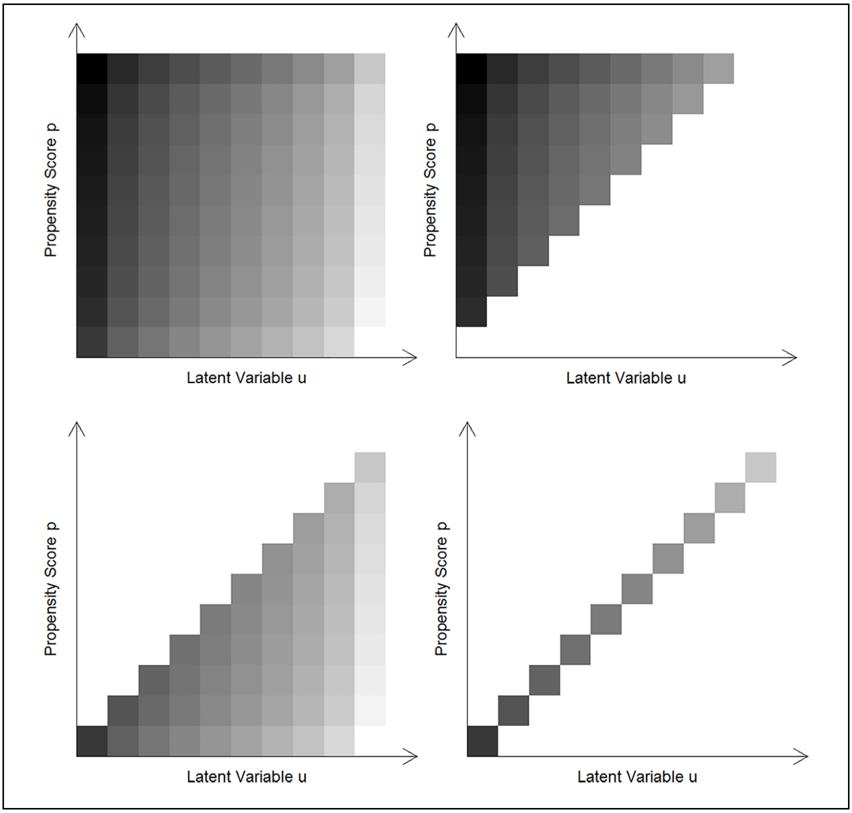

Figure 1. Demonstration of treatment-effect heterogeneity by propensity score $P(Z)$ and latent variable U.
Note: A darker color means a higher treatment effect.

d. Thus, evaluating ATE, TT, TUT, and MPRTE simply means taking weighted averages of $\widetilde{\text{MTE}}(p, u)$ over the corresponding subsets of grids.

3.3. *Treatment-Effect Heterogeneity among Marginal Entrants*

For policymakers, a key question of interest would be how MPRTE(p) varies with the propensity score p. From Equations 14 and 15, we see that MPRTE(p) consists of two components:

$$\text{MPRTE}(p) = E[\mu_1(X) - \mu_0(X)|P(Z) = p] + E(\eta|U = p). \tag{16}$$

The first component reflects how treatment effect varies by the propensity score, and the second component reflects how treatment effect varies by the latent resistance U. Among marginal entrants, $P(Z)$ is equal to U so that these two components fall on the same dimension.

To see how the two components combine to shape MPRTE(p), let us revisit the classic example on economic returns to college. In the labor economics literature, researchers often have found a negative association between η and U, suggesting a pattern of positive selection, that is, individuals who benefit more from college are more motivated than their peers to attend college in the first place (e.g., Blundell, Dearden, and Sianesi 2005; Carneiro et al. 2011; Heckman, Humphries, and Veramendi 2018; Moffitt 2008; Willis and Rosen 1979). In this case, the second component of Equation 16 would be a decreasing function of p. On the other hand, the literature has not paid much attention to the first component, that is, whether individuals who by observed characteristics are more likely to attend college also benefit more from college. A number of observational studies suggest that nontraditional students, such as racial and ethnic minorities or students from less educated families, experience higher returns to college than do traditional students, although interpretation of such findings remains controversial due to potential unobserved selection biases (e.g., Attewell and Lavin 2007; Bowen and Bok 1998; Dale and Krueger 2011; Maurin and McNally 2008; for a review, see Hout 2012).[10] However, if the downward slope in the second component were sufficiently strong, MPRTE(p) would also decline with p. In this case, we would, paradoxically, observe a pattern of negative selection (Brand and Xie 2010): Among students who are at the margin of attending college, those who by observed characteristics are less likely to attend college would actually benefit more from college.

To better understand the paradoxical implication of self-selection, let us revisit Figure 1. From Panel a, we see that in the hypothetical data, treatment effect declines with u at each level of the propensity score, suggesting unobserved self-selection. In other words, individuals may have self-selected into treatment on the basis of their anticipated gains. On the other hand, at each level of the latent variable u, treatment effect increases with the propensity score, indicating that individuals who by observed characteristics are more likely to be treated also benefit more from the treatment. This relationship, however, is reversed among the marginal entrants. As shown in Panel d, among the marginal entrants, individuals who appear less likely to be treated (bottom left grids) have higher treatment effects. This pattern of negative selection at the margin, interestingly, is exactly due to an unobserved positive selection into treatment.

3.4. *Policy as a Weighting Problem*

In Section 3.2, we used $\lambda(p)$ to denote the increment in treatment probability at each level of the propensity score p. Because MPRTE(p) is defined as the pointwise limit of PRTE($p, \alpha\lambda(p)$) as α approaches zero, the mathematical form of $\lambda(p)$ does not affect MPRTE(p). However, in deriving the population-level (i.e., unconditional) MPRTE, we need to use $\lambda(p)$ as the appropriate weight, that is,

$$\text{MPRTE} = C\int_0^1 \text{MPRTE}(p)\lambda(p)dF_P(p). \tag{17}$$

Here $F_P(\cdot)$ is the marginal distribution function of the propensity score, and $C = 1/\int_0^1 \lambda(p)dF_P(p)$ is a normalizing constant (see Appendix B for a derivation). Thus, given the estimates of MPRTE(p), a policymaker could use the previous equation to design a formula for $\lambda(\cdot)$ to boost the population-level MPRTE. This is especially useful if MPRTE(p) varies systematically with the propensity score p. For example, if one found that the marginal return to college declines with the propensity score p, a college expansion program targeted at students with relatively low values of p (e.g., a means-tested financial aid program) would yield higher average marginal returns than would a universal expansion of college enrollment regardless of student characteristics.[11]

In practice, for a given policy $\lambda(p)$, we can evaluate the aforementioned integral directly from sample data using

$$\text{MPRTE} \approx \frac{\sum_i \text{MPRTE}(\hat{p}_i)\lambda(\hat{p}_i)}{\sum_i \lambda(\hat{p}_i)}, \tag{18}$$

where $\hat{p}_i$ is the estimated propensity score for unit i in the sample. When the sample is not representative of the population by itself, sampling weights need to be incorporated into these summations.

3.5. *Comparison with Carneiro and Colleagues (2010)*

In the previous discussion, PRTE and MPRTE are defined for a class of policy changes in which the intensity of policy intervention depends on the individual's propensity score $P(Z)$. In other words, inducements are differentiated between individuals with different values of $P(Z)$, whether their differences in $P(Z)$ are determined by the baseline covariates X or the instrumental variables $Z\backslash X$. This approach to defining MPRTE contrasts sharply with the approach taken by Carneiro and colleagues (2010, 2011), who stipulate that all policy changes have to be "conditioned on X." In their approach, inducements are allowed to vary across individuals with different values of $Z\backslash X$ but not across individuals with different values of X. For convenience, we call Carneiro et al's approach the *conditional approach* and our approach the *unconditional approach.* Compared with the conditional approach, the unconditional approach to studying policy effects has several major advantages.

First, as noted earlier, preferential policies under the conditional approach distinguish individuals with different instrumental variables ($Z\backslash X$) but not individuals with different baseline characteristics (X). To see the limitation of such policies, let us revisit the college education example and consider a simplistic model where the only baseline covariate X is family income and the only instrumental variable $Z\backslash X$ is the presence of four-year colleges in the county of residence. In this case, an "affirmative" policy—a policy that favors students with lower values of $P(Z)$—would be a policy that induces students who happen to live in a county with no four-year colleges, regardless of family income. Given that $P(Z)$ equals U at the margin, this policy benefits students with relatively low Us at all levels of family income. To the extent that there is self-selection into college (i.e., $\mathrm{Cor}(\eta, U)<0$), this policy would yield a larger MPRTE than would a neutral policy. However, if $P(Z)$ is largely determined by family income rather than the local presence of four-year colleges (a plausible scenario), the variation of $P(Z)$ conditional on X would be very limited, as would the gain in MPRTE from a preferential policy. In contrast, the unconditional approach distinguishes individuals with different values of $P(Z)$, most of which may be driven by X rather than $Z\backslash X$. Because $P(Z)$ equals U at the margin, this approach can sort out marginal entrants with different levels of U effectively. Therefore, preferential policies under the unconditional approach are more effective in exploiting unobserved heterogeneity in treatment effects.

Table 3. Weights for Constructing MPRTE(x) from MTE(x, u_D)

Parameters of Interest	Weight
MPRTE(x): $P^* = P + \alpha$	$h_{\text{MPRTE}}(x,u) = f_{P(Z)\|X=x}(u)$
MPRTE(x): $P^* = (1+\alpha)P$	$h_{\text{MPRTE}}(x,u) = \frac{uf_{P(Z)\|X=x}(u)}{\mathbb{E}(P(Z)\|X=x)}$
MPRTE(x): $Z_k^* = Z_k + \alpha$	$h_{\text{MPRTE}}(x,u) = \frac{f_{P(Z)\|X=x}(u)f_V[F_V^{-1}(u)]}{\mathbb{E}[f_V(\gamma' Z)\|X=x]}$

Source: Data from Carneiro, Heckman, and Vytlacil (2011).
Note: MPRTE = marginal policy-relevant treatment effect; MTE = marginal treatment effect.

Second, because treatment effect in general depends on the observed covariates X as well as the latent resistance U, an ideal policy intervention should exploit the variation of treatment effect along both dimensions. The conditional approach, however, differentiates individuals with different Us but not individuals with different observed characteristics (at least in practice). In contrast, by focusing on the propensity score $P(Z)$, the unconditional approach accounts for treatment-effect heterogeneity in both observed and unobserved dimensions. Because $P(Z)$ equals U at the margin, the bivariate function $\widetilde{\text{MTE}}(p,u)$ degenerates into a univariate function of p among marginal entrants (see Equation 16). Thus, by weighting individuals with different values of $P(Z)$, the unconditional approach captures the "collision" of observed and unobserved heterogeneity at the margin. To see why the latter is more effective, consider an extreme scenario where there is no unobserved sorting (i.e., $E(\eta|U)$ is constant) but treatment effect varies considerably by X. In this case, the unconditional approach can partly exploit the variation of treatment effect by X (through the first component of Equation 16), whereas the conditional approach cannot (because it focuses exclusively on the second component of Equation 16).

Finally, the unconditional approach is computationally simpler. $\text{MPRTE}(p) = \widetilde{\text{MTE}}(p,p)$, so no further step is needed to estimate MPRTE(p) once we have estimates of $\widetilde{\text{MTE}}(p,u)$. The conditional approach, by contrast, needs to build MPRTE(x) on MTE(x, u) using policy-specific weights. As shown in Table 3, these policy-specific weights generally involve the conditional density of $P(Z)$ given X. Because X is usually a high-dimensional vector, estimation of these weights is difficult and often tackled with ad hoc methods (see Section 2.3).

4. ILLUSTRATION WITH NLSY DATA

To illustrate the new approach, we reanalyze the data from Carneiro and colleagues' (2011) study on economic returns to college education. We first describe the data, then demonstrate treatment-effect heterogeneity using the newly defined $\widetilde{\mathrm{MTE}}(p, u)$, and finally, evaluate the effects of different marginal policy changes.

4.1. *Data Description*

We reanalyze a sample of white males ($N = 1{,}747$) who were 16 to 22 years old in 1979, drawn from the NLSY 1979. Treatment (D) is college attendance defined by having attained any postsecondary education by 1991. Under this definition, the treated group consists of 865 individuals, and the comparison group consists of 882 individuals. The outcome Y is the natural logarithm of hourly wage in 1991.[12] Following the original study, we include in pretreatment variables (in both X and Z) linear and quadratic terms of mother's years of schooling, number of siblings, the Armed Forces Qualification Test (AFQT) score adjusted by years of schooling, permanent local log earnings at age 17 (county log earnings averaged between 1973 and 2000), and permanent local unemployment rate at age 17 (state unemployment rate averaged between 1973 and 2000) as well as a dummy variable indicating urban residence at age 14 and cohort dummies. Also following Carneiro and colleagues (2011), we use the following instrumental variables ($Z \backslash X$): (1) the presence of a four-year college in the county of residence at age 14, (2) local wage in the county of residence at age 17, (3) local unemployment rate in the state of residence at age 17, and (4) average tuition in public four-year colleges in the county of residence at age 17 as well as their interactions with mother's years of schooling, number of siblings, and the adjusted AFQT score. In addition, four variables are included in X but not in Z: years of experience in 1991, years of experience in 1991 squared, local log earnings in 1991, and local unemployment rate in 1991. More details about the data can be found in Carneiro and colleagues' (2011) online appendix.

4.2. *Heterogeneity in Treatment Effects*

To estimate the bivariate function $\widetilde{\mathrm{MTE}}(p, u)$, we first need estimates of $\mathrm{MTE}(x, u)$. In Section 2, we discussed a parametric and a

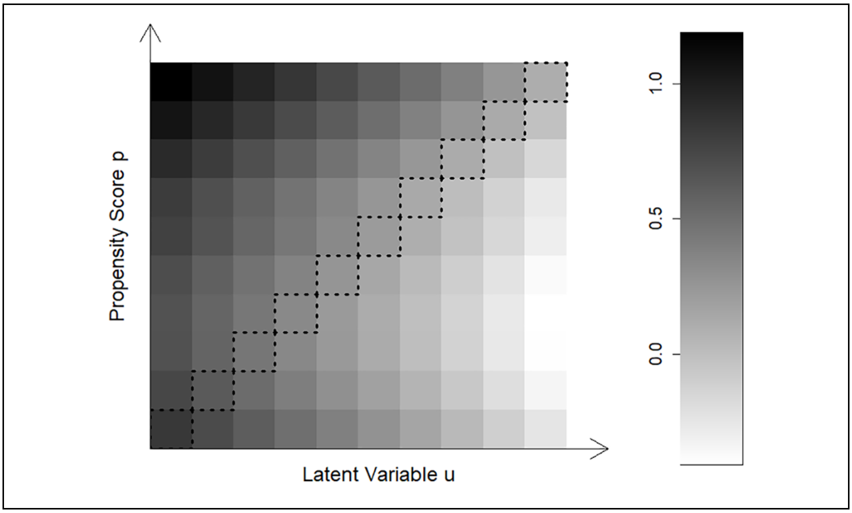

Figure 2. Treatment-effect heterogeneity based on semiparametric estimates of $\widetilde{\text{MTE}}(p, u)$.

Note: MTE = marginal treatment effect.

semiparametric method for estimating $\text{MTE}(x, u)$. Here, we examine treatment-effect heterogeneity with the semiparametric estimates of $\text{MTE}(x, u)$ (Equation 12) and thus $\widetilde{\text{MTE}}(p, u)$.[13] Figure 2 presents our key results for the estimated $\widetilde{\text{MTE}}(p, u)$, with a shaded gray plot in which a darker grid indicates a higher treatment effect. The effect heterogeneity by the two dimensions—the propensity score and the latent resistance to treatment—exhibits an easy-to-interpret but surprising pattern. On the one hand, at each level of the propensity score, a higher level of the latent variable u is associated with a lower treatment effect, indicating the presence of self-selection based on idiosyncratic returns to college. This pattern of "sorting on gain" echoes the classic findings reported in Willis and Rosen (1979) and Carneiro and colleagues (2011). On the other hand, the color gradient along the propensity score suggests that given the latent resistance to attending college, students who by observed characteristics are more likely to go to college also tend to benefit more from attending college.

If we read along the diagonal of Figure 2, however, we find that among students who are at the margin of indifference for attending college, those who appear less likely to attend college would benefit more from a college education, that is, $\text{MPRTE}(p)$ declines with the

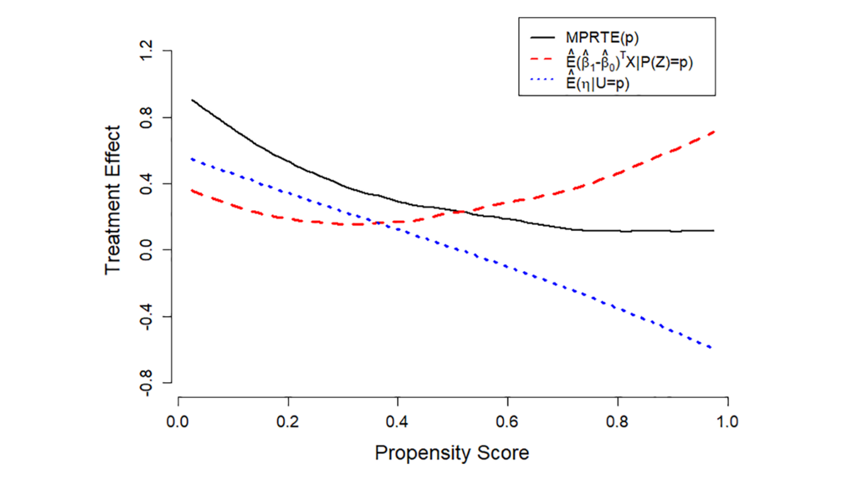

Figure 3. Decomposition of MPRTE(p) based on semiparametric estimates of $\widetilde{\text{MTE}}(p, u)$.

Note: MPRTE = marginal policy-relevant treatment effect; MTE = marginal treatment effect.

propensity score p. Figure 3 shows smoothed estimates of MPRTE(p) as well as its two components (see Equation 16). The negative association between η and the latent resistance U more than offsets the positive association between $(\beta_1 - \beta_0)^T X$ and the propensity score $P(Z)$, resulting in the downward slope of MPRTE(p). Echoing our discussion in Section 3.3, it is unobserved "sorting on gain" that leads to the negative association between the propensity score and returns to college among students at the margin.

We use weights given in Table 2 to estimate ATE, TT, and TUT at each level of the propensity score. Figure 4 shows smoothed estimates of ATE(p), TT(p), TUT(p), and MPRTE(p). Several patterns are worth noting. First, there is a sharp contrast between ATE(p) and MPRTE(p): A higher propensity of attending college is associated with a higher return to college on average (solid line) but a lower return to college among marginal entrants (dot-dash line). Second, TT(p) (dashed line) is always larger than TUT(p) (dotted line), suggesting that at each level of the propensity score, individuals are positively self-selected into college based on their idiosyncratic returns to college. Finally, TT(p) and TUT(p) converge to ATE(p) and MPRTE(p) at the extremes of the propensity score. When p approaches 0, TT(p) converges to MPRTE(p) and

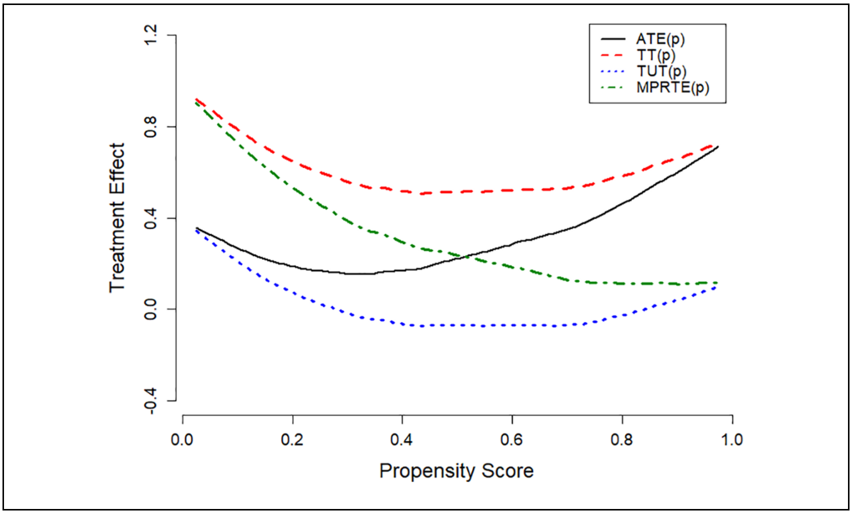

Figure 4. Heterogeneity in ATE, TT, TUT, and MPRTE(p) by propensity score based on semiparametric estimates of $\widetilde{\mathrm{MTE}}(p, u)$.

Note: ATE = average treatment effect; TT = treatment effect of the treated; TUT = treatment effect of the untreated; MPRTE = marginal policy-relevant treatment effect; MTE = marginal treatment effect.

TUT(p) converges to ATE(p). At the other extreme, when p approaches 1, TT(p) converges to ATE(p) and TUT(p) converges to MPRTE(p). Looking back at Figure 1, we see that these relationships simply reflect compositional shifts in the treated and untreated groups as the propensity score changes from 0 to 1.

4.3. *Evaluation of Policy Effects*

Given the estimates of ATE(p), TT(p), and TUT(p), we construct their population averages using appropriate weights across the propensity score. For example, to estimate TT, we simply integrate TT(p) against the marginal distribution of the propensity score among individuals who attended college. The estimates of MPRTE(p) allow us to construct different versions of MPRTE, depending on how the policy change weights students with different propensities of attending college (see Equation 18). Table 4 reports our estimates of ATE, TT, TUT, and MPRTE under different policy changes from the parametric and semiparametric estimates of MTE(x, u). To compare our new approach with Carneiro and colleagues' (2011) original approach, we show estimates built on

Table 4. Estimated Returns to One Year of College

	Parametric (Normal)		Semiparametric	
Building Block	$\mathrm{MTE}(x,u)$	$\widetilde{\mathrm{MTE}}(p,u)$	$\mathrm{MTE}(x,u)$	$\widetilde{\mathrm{MTE}}(p,u)$
ATE	.066 (.038)	.066 (.038)	.082 (.041)	.082 (.041)
TT	.139 (.035)	.142 (.035)	.165 (.048)	.167 (.049)
TUT	−.006 (.067)	−.009 (.067)	.000 (.060)	.000 (.061)
MPRTE				
$\lambda(p)=\alpha$	.066 (.038)	.065 (.039)	.084 (.041)	.083 (.041)
$\lambda(p)=\alpha p$		.061 (.050)		.050 (.048)
$\lambda(p)=\alpha(1-p)$		.068 (.033)		.116 (.042)
$\lambda(p)=\alpha\mathbb{I}(p<0.3)$		.080 (.035)		.155 (.055)

Note: ATE = average treatment effect; TT = treatment effect of the treated; TUT = treatment effect of the untreated; MPRTE = marginal policy-relevant treatment effect; MTE = marginal treatment effect. Numbers in parentheses are bootstrapped standard errors (250 replications).

$\widetilde{\mathrm{MTE}}(p,u)$ and those built on $\mathrm{MTE}(x,u)$. Following Carneiro and colleagues (2011), we annualize the returns to college by dividing all parameter estimates by four, which is the average difference in years of schooling between the treated and untreated groups.

The first three rows of Table 4 indicate that TT $>$ ATE $>$ TUT ≈ 0. That is, returns to college are higher among individuals who actually attended college than among those who attended only high school, for whom the average returns to college are virtually zero. Using either the parametric or semiparametric estimates of $\mathrm{MTE}(x,u)$, our new approach and the original approach yield nearly identical point estimates and bootstrapped standard errors. This consistency affirms our argument that $\widetilde{\mathrm{MTE}}(p,u)$ preserves all of the treatment-effect heterogeneity that is consequential for selection bias. Although the redefined MTE seems to contain less information than the original MTE (as it projects $(\beta_1-\beta_0)^T X$ onto the dimension of $P(Z)$), the discarded information does not contribute to identification of average causal effects.

The last four rows of Table 4 present our estimates of MPRTE under four stylized policy changes: (1) $\lambda(p)=\alpha$, (2) $\lambda(p)=\alpha p$, (3)

$\lambda(p)=\alpha(1-p)$, and (4) $\lambda(p)=\alpha\mathbb{I}(p<0.3)$. Put in words, the first policy change increases everyone's probability of attending college by the same amount, the second policy change favors students who appear more likely to go to college, the third policy change favors students who appear less likely to go to college, and the last policy change only targets students whose observed likelihood of attending college is less than 30 percent. For each policy change, the MPRTE is defined as the limit of the corresponding PRTE as α goes to zero. The first policy change is also the first policy change considered by Carneiro and colleagues (2011:2760), that is, $P_\alpha=P+\alpha$ (see also the first row of Table 3). For this case, we estimated the MPRTE using both the original approach and our new approach. As expected, the two approaches yield the same results. However, the other three policy changes considered here cannot be readily accommodated within the original framework (see Section 3.5). Thus, we evaluate their effects using only our new approach, that is, via Equation 18.

Although the estimates of TUT are close to zero, all four policy changes imply substantial marginal returns to college. For example, under the first policy change, the semiparametric estimate of MPRTE is .083, suggesting that one year of college would translate into an 8.3 percent increase in hourly wages among marginal entrants. However, the exact magnitude of MPRTE depends heavily on the form of the policy change, especially under the semiparametric model. Whereas the marginal return to a year of college is about 5 percent if we expand everyone's probability of attending college proportionally (policy change two), it can be as high as 15.5 percent if we only increase enrollment among students whose observed likelihood of attending college is less than 30 percent (policy change four). Figure 5 shows how different policy changes produce different compositions of marginal college entrants. Because students who benefit the most from college are located at the low end of the propensity score, a college expansion program targeted at those students will yield the highest marginal returns to college. Fortuitously, earlier research that did not account for the presence of unobserved selection reached similar policy implications (Brand and Xie 2010).

5. DISCUSSION AND CONCLUSION

Due to the ubiquity of population heterogeneity in social phenomena, it is impossible to evaluate causal effects at the individual level. All efforts

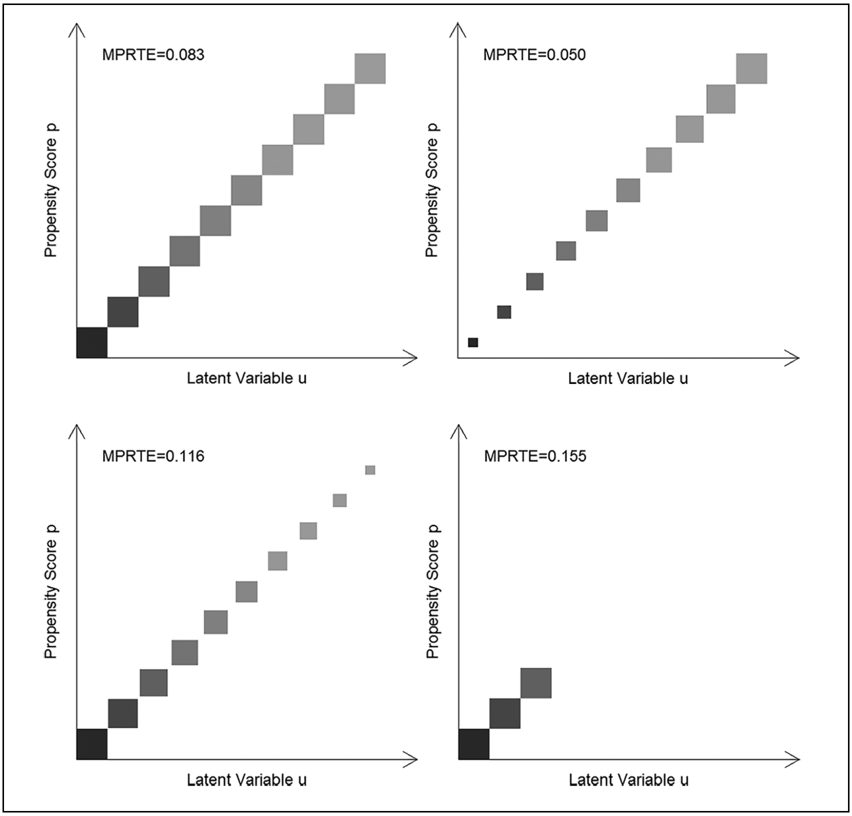

Figure 5. Semiparametric estimates of marginal policy-relevant treatment effect under four policy changes.

Note: MTE = marginal treatment effect.

to draw causal inferences in social science must be at the group level. Yet with observational data, even group-level inference is plagued by two types of selection bias: Individuals in the treated and comparison groups may differ systematically not only in their baseline outcomes but also in their treatment effects. Depending on whether unobserved selection is assumed away, traditional methods for causal inference from observational data can be divided into two classes, as shown in the first row of Table 5. The first class, including regression adjustment, matching, and inverse probability of treatment weighting (Robins, Hernan, and Brumback 2000), rests on the assumption of ignorability: After controlling for a set of observed covariates, treatment status is independent

Table 5. Methods for Identifying and Estimating Causal Effects from Observational Data

		Allowing for Unobserved Selection?	
		No[a]	Yes
Systematically modeling treatment-effect heterogeneity?	No	Regression adjustment, matching, IPW, etc.	IV, RD design, fixed-effects models, etc.
	Yes	$\mathbb{E}(Y_1 - Y_0 \mid X = x)$, $\mathbb{E}(Y_1 - Y_0 \mid P = p)$	$\text{MTE}(x, u)$, $\widetilde{\mathbf{MTE}}(\boldsymbol{p}, \boldsymbol{u})$

Note: IV = instrumental variables; RD = regression discontinuity; IPW = inverse probability weighting; MTE = marginal treatment effect.

[a]When there is unobserved selection by treatment effect but not by the baseline outcome, matching and weighting methods can still be used to consistently estimate treatment effect of the treated (but not average treatment effect).

of both baseline outcomes and treatment effects. The second class of methods, including instrumental variables (IV), regression discontinuity (RD) designs (Hahn, Todd, and Van der Klaauw 2001; Thistlethwaite and Campbell 1960), and fixed-effects models, allows for unobserved selection into treatment but requires exogenous variation in treatment status—either between or within units—to identify causal effects.

Both classes of methods allow treatment effects to vary in the population, but in common practices neither systematically models treatment-effect heterogeneity.[14] When treatment effects are heterogeneous, some of these methods estimate quantities that are not of primary interest to the researcher. For example, when treatment effect varies according to the level of a covariate, main-effects-only regression models cannot recover standard causal estimands such as ATE or TT but instead estimate a conditional-variance-weighted causal effect that has little substantive meaning (Angrist and Pischke 2008; Elwert and Winship 2010). Moreover, when treatment effect is heterogeneous, IV and RD designs can only identify the average causal effect among individuals whose treatment status is influenced by the IV (Imbens and Angrist 1994) or in the case of fuzzy RD designs, by whether the running variable surpasses the "cutoff point" (Hahn, Todd, and Van der Klaauw 2001). Similarly, fixed-effects models can only identify the average causal effect among individuals who change their treatment status over the study period.

The second row of Table 5 summarizes the four approaches that can be used to systematically study treatment-effect heterogeneity, especially treatment-effect heterogeneity by pretreatment characteristics. The first approach, denoted as $\mathbb{E}(Y_1 - Y_0|X)$, includes the long-standing practice of adding interaction terms between treatment status and covariates in conventional regression models as well as recent proposals to fit nonparametric surfaces of potential outcomes and their difference (e.g., Hill 2011). The second approach, denoted as $E(Y_1 - Y_0|P)$, models treatment effect as a univariate function of the propensity score (e.g., Xie et al. 2012; Zhou and Xie 2016). Because the propensity score is the only dimension along which treatment effect may be correlated with treatment status, this approach not only provides a useful solution to data sparseness, but it also facilitates projection of treatment effects beyond particular study settings (Stuart et al. 2011; Xie 2013). However, as noted earlier, these two approaches rely on the assumption of ignorability. When ignorability breaks down, interpretation of the observed heterogeneity in treatment effects becomes ambiguous (Breen et al. 2015).

The latter two approaches, that is, the MTE-based approach and our extension of it, accommodate unobserved selection through use of a latent index model for treatment assignment. In this model, a scalar error term is used to capture all the unobserved factors that may induce or impede treatment. As a result, treatment status is determined by the "competition" between the propensity score $P(Z)$ and the latent variable U representing unobserved resistance to treatment. Therefore, the propensity score $P(Z)$ and the latent variable U are the only two dimensions along which treatment status may be correlated with treatment effects. The MTE, as in Heckman and Vytlacil's (1999, 2001a, 2005, 2007b) original formulation, is asymmetrical with respect to these two dimensions because it conditions on the entire vector of observed covariates X as well as the latent variable U. Because of this asymmetry, the original MTE-based approach has a number of drawbacks, including (1) an exclusive attention (in practice) to unobserved heterogeneity (rather than observed heterogeneity) in treatment effects, (2) difficulty of implementation due to unwieldy weight formulas, and (3) inflexibility in the modeling of policy effects (see Section 3.5).

To overcome these limitations, we presented an extension of the MTE framework through a redefinition of MTE. By conditioning on the propensity score $P(Z)$ and the latent variable U, the redefined MTE not only treats observed and unobserved selection symmetrically, but it more parsimoniously summarizes all the treatment-effect heterogeneity that is consequential for selection bias. As a bivariate function, it can be easily visualized. As with the original MTE, the redefined MTE also can be used as a building block in evaluating aggregate causal effects. Yet the weights associated with the new MTE are simpler, more intuitive, and easier to compute (compare Table 2 with Tables 1 and 3). Finally, the new MTE immediately reveals heterogeneous treatment effects among individuals who are at the margin of treatment, thus allowing us to design more cost-effective policy interventions.

Our extension of the MTE approach is not a panacea. Like the original approach, it hinges on credible estimates of $\text{MTE}(x, u)$. Identification of $\text{MTE}(x, u)$ requires at least a valid IV in the treatment assignment equation. Moreover, under either the parametric or semiparametric model, the statistical efficiency of estimates of $\text{MTE}(x, u)$ depends heavily on the strength of IVs (Zhou and Xie 2016). When the IVs are relatively weak in determining treatment status, MTE-based estimates of aggregate causal effects can be imprecise. Nonetheless, as long as valid instruments are present, more precise estimates can always be achieved with a larger sample size.

APPENDIX A: IDENTIFICATION OF $\widetilde{\textbf{MTE}}(p, u)$ UNDER ASSUMPTIONS 1, 2, AND 3

From Assumption 1, we know $V \perp Z | X$. Because U and $P(Z)$ are functions of V and Z, respectively, $U \perp P(Z) | X$. U follows a standard uniform distribution for each $X = x$, so we also have $U \perp X$. By the rules of conditional independence, we have $U \perp X | P(Z)$. Using this fact and the law of total expectation, we can write $\widetilde{\text{MTE}}(p, u)$ as

$$\begin{aligned}\widetilde{\text{MTE}}(p, u) &= \mathbb{E}_{X|P(Z)=p, U=u}\mathbb{E}[Y_1 - Y_0 | P(Z) = p, U = u, X] \\ &= \mathbb{E}_{X|P(Z)=p}\mathbb{E}[Y_1 - Y_0 | P(Z) = p, U = u, X] \\ &= \mathbb{E}_{X|P(Z)=p}\mathbb{E}[Y_1 - Y_0 | U = u, X] \quad (\text{because} (\eta, U) \perp P(Z) | X) \\ &= \mathbb{E}_{X|P(Z)=p}\text{MTE}(X, u).\end{aligned} \tag{19}$$

Thus $\widetilde{\text{MTE}}(p, u)$ is simply the conditional expectation of $\text{MTE}(X, u)$ given $P(Z) = p$. Given Assumption 3, $\text{MTE}(X, u)$ can be written as Equation 14. Substituting it into Equation 19 yields

$$\widetilde{\text{MTE}}(p, u) = \mathbb{E}[\mu_1(X) - \mu_0(X) | P(Z) = p] + \mathbb{E}[\eta | U = u].$$

APPENDIX B: DERIVATION OF EQUATION 17

To see why Equation 17 holds, consider the overall PRTE for a given α. Given that $P(Z) = p$, the size of inducement $\alpha\lambda(p)$ reflects the share of individuals that are induced into treatment ("compliers"), and the overall PRTE is a weighted average of $\text{PRTE}(p, \alpha\lambda(p))$ with weights $\alpha\lambda(p)$:

$$\text{PRTE}_\alpha = \frac{\int_0^1 \alpha\lambda(p)\text{PRTE}(p, \alpha\lambda(p))dF_P(p)}{\int_0^1 \alpha\lambda(p)dF_P(p)} = \frac{\int_0^1 \lambda(p)\text{PRTE}(p, \alpha\lambda(p))dF_P(p)}{\int_0^1 \lambda(p)dF_P(p)},$$

where $F_P(\cdot)$ denotes the marginal distribution function of the propensity score. We then define the population-level MPRTE as the limit of PRTE_α as α approaches zero. Under some regularity conditions,[15] we can take the limit inside the integral

$$\begin{aligned}\text{MPRTE} &= \lim_{\alpha\to 0} \text{PRTE}_\alpha \\ &= \frac{\int_0^1 \lambda(p) \lim_{\alpha\to 0} \text{PRTE}(p, \alpha\lambda(p))dF_P(p)}{\int_0^1 \lambda(p)dF_P(p)} \\ &= \frac{\int_0^1 \lambda(p)\text{MPRTE}(p)dF_P(p)}{\int_0^1 \lambda(p)dF_P(p)}.\end{aligned}$$

Denoting $C = 1/\int_0^1 \lambda(p)dF_P(p)$, we obtain Equation 17.

Acknowledgments

The authors benefited from communications with Daniel Almirall, Matthew Blackwell, Jennie Brand, James Heckman, Jeffrey Smith, and Edward Vytlacil.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Grant No. R01-HD-074603-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Notes

1. Heckman and Robb (1986) also framed propensity score matching as a special case of control function methods.

[2] In the classic Roy model (Roy 1951), $I_D = Y_1 - Y_0$. In that case, $Z = X$ and $V = -\eta$.

[3] The property that Z affects treatment status only through the propensity score in an additively separable latent index model is called index sufficiency (Heckman and Vytlacil 2005).

[4] An alternative method to identify the MTE nonparametrically is based on separate estimation of $\mathbb{E}[Y|P(Z), X, D=0]$ and $\mathbb{E}[Y|P(Z), X, D=1]$ (see Brinch, Mogstad, and Wiswall 2017; Heckman and Vytlacil 2007b).

[5] In estimating $K(p)$, we need to impose constraints on β_0 and β_1 such that $K(0) = K(1) = 0$. $K(0) = 0$ is from its definition. $K(1) = \int_0^1 \mathbb{E}[\eta|U=u]du = \mathbb{E}_U\mathbb{E}[\eta|U] = \mathbb{E}[\eta] = 0$.

[6] The maximum likelihood estimation can be easily implemented in R using the `sampleSelection` package (see Toomet and Henningsen 2008).

[7] Because the treatment assignment model is now a probit model, σ_V is usually normalized to 1.

[8] More flexible methods, such as generalized additive models and boosted regression trees, also can be used to estimate propensity scores (e.g., McCaffrey, Ridgeway, and Morral 2004).

[9] In a companion paper (Zhou and Xie forthcoming), we discuss the regions over which $\widetilde{\text{MTE}}(p, u)$ can be nonparametrically identified with and without the assumption of additive separability.

[10] Studies that use compulsory schooling laws, differences in the accessibility of schools, or similar features as instrumental variables also find larger economic returns to college than do least squares estimates (Card 2001). However, this comparison does not reveal how returns to college vary by covariates or the propensity score.

[11] Admittedly, the discussion here provides no more than a theoretical guide to practice. In the real world, designing specific policy instruments to produce a target form of $\lambda(p)$ can be a challenging task.

[12] Hourly wage in 1991 is defined as an average of deflated (to 1983 constant dollars) nonmissing hourly wages reported between 1989 and 1993.

[13] Results based on parametric estimates of $\text{MTE}(x, u)$ (Equation 11) are substantively similar.

[14] Although matching and weighting methods are well equipped to estimate ATE, TT, and TUT under the assumption of ignorability, they are seldom used to study treatment-effect heterogeneity by individual characteristics.

[15] A sufficient (but not necessary) condition is that $\widetilde{\text{MTE}}(p, u)$ is bounded over $[0, 1] \times [0, 1]$. By the mean value theorem, $\text{PRTE}(p, \alpha\lambda(p))$ can be written as $\widetilde{\text{MTE}}(p, p^*)$ where $p^* \in [p, p + \alpha\lambda(p)]$. $\text{PRTE}(p, \alpha\lambda(p))$ is thus also bounded. By the dominated convergence theorem, the limit can be taken inside the integral.

References

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Arabmazar, Abbas, and Peter Schmidt. 1982. “An Investigation of the Robustness of the Tobit Estimator to Non-normality.” *Econometrica* 50(4):1055–63.

Attewell, Paul, and David Lavin. 2007. *Passing the Torch: Does Higher Education for the Disadvantaged Pay off across the Generations?* New York: Russell Sage Foundation.

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

Blundell, Richard, Lorraine Dearden, and Barbara Sianesi. 2005. “Evaluating the Effect of Education on Earnings: Models, Methods and Results from the National Child Development Survey.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168(3):473–512.

Borjas, George J. 1987. “Self-Selection and the Earnings of Immigrants.” *The American Economic Review* 77(4):531–53.

Bowen, William G., and Derek Bok. 1998. *The Shape of the River. Long-Term Consequences of Considering Race in College and University Admissions*. Princeton, NJ: Princeton University Press.

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.

Breen, Richard, Seong-soo Choi, and Anders Holm. 2015. “Heterogeneous Causal Effects and Sample Selection Bias.” *Sociological Science* 2:351–69.

Brinch, Christian N., Magne Mogstad, and Matthew Wiswall. 2017. “Beyond LATE with a Discrete Instrument.” *Journal of Political Economy* 125(4):985–1039.

Card, David. 2001. “Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems.” *Econometrica* 69(5):1127–60.

Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil. 2010. “Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin.” *Econometrica* 78(1):377–94.

Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil. 2011. “Estimating Marginal Returns to Education.” *American Economic Review* 101(773):2754–81.

Carneiro, Pedro, and Sokbae Lee. 2009. “Estimating Distributions of Potential Outcomes Using Local Instrumental Variables with an Application to Changes in College Enrollment and Wage Inequality.” *Journal of Econometrics* 149(2):191–208.

Dale, Stacy, and Alan B. Krueger. 2011. “Estimating the Return to College Selectivity over the Career Using Administrative Earnings Data.” Technical report, National Bureau of Economic Research, Cambridge, MA.

Elwert, Felix, and Christopher Winship. 2010. “Effect Heterogeneity and Bias in Main-Effects-Only Regression Models.” Pp. 327–36 *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, edited by R. Dechter, H. Geffner, and J. Y Halpern. London: College Publications.

Fan, Jianqing, and Irene Gijbels. 1996. *Local Polynomial Modelling and Its Applications*. Vol. 66. London: Chapman and Hall.

Gamoran, Adam, and Robert D. Mare. 1989. “Secondary School Tracking and Educational Inequality: Compensation, Reinforcement, or Neutrality?” *American Journal of Sociology* 94(5):1146–83.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69(1):201–209.

Heckman, James J. 1978. "Dummy Endogenous Variables in a Simultaneous Equation System." *Econometrica* 46(4):931–59.

Heckman, James J. 2010. "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy." *Journal of Economic literature* 48(2):356–98.

Heckman, James J., and V. Joseph Hotz. 1989. "Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84(408):862–74.

Heckman, James J., John Eric Humphries, and Gregory Veramendi. 2018. "Returns to Education: The Causal Effects of Education on Earnings, Health and Smoking." *Journal of Political Economy* 126(S1):S197–S246.

Heckman, James J., and Salvador Navarro-Lozano. 2004. "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models." *The Review of Economics and Statistics* 86(1):30–57.

Heckman, James J., and Richard Robb. 1986. "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes." Pp. 63–107 in *Drawing Inferences from Self-selected Samples*, edited by H. Wainer. New York: Springer.

Heckman, James J., Sergio Urzua, and Edward J. Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *The Review of Economics and Statistics* 88(3):389–432.

Heckman, James J., and Edward J. Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proceedings of the National Academy of Sciences of the United States of America* 96(8):4730–34.

Heckman, James J., and Edward J. Vytlacil. 2001a. "Local Instrumental Variables." Pp. 1–46 in *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, edited by C. Hsiao, K. Morimune, and J. L. Powel. New York: Cambridge University Press.

Heckman, James J., and Edward J. Vytlacil. 2001b. "Policy-Relevant Treatment Effects." *American Economic Review* 91(2):107–11.

Heckman, James J., and Edward J. Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73(3):669–738.

Heckman, James J., and Edward J. Vytlacil. 2007a. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." Chapter 71 in *Handbook of Econometrics*. Vol. 6, edited by J. J. Heckman and E. E. Leamer. Elsevier.

Heckman, James J., and Edward J. Vytlacil. 2007b. "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments." Chapter 71 in *Handbook of Econometrics*, Vol. 6, edited by J. J. Heckman and E. E. Leamer. Elsevier.

Hill, Jennifer L. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217–40.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–60.

Hout, Michael. 2012. "Social and Economic Returns to College Education in the United States." *Annual Review of Sociology* 38:379–400.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2):467–75.

Maestas, Nicole, Kathleen J. Mullen, and Alexander Strand. 2013. "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt." *The American Economic Review* 103(5):1797–829.

Maurin, Eric, and Sandra McNally. 2008. "Vive la Révolution! Long-Term Educational Returns of 1968 to the Angry Students." *Journal of Labor Economics* 26(1):1–33.

McCaffrey, Daniel F., Greg Ridgeway, and Andrew R. Morral. 2004. "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies." *Psychological Methods* 9(4):403–25.

Moffitt, Robert. 2008. "Estimating Marginal Treatment Effects in Heterogeneous Populations." *Annales d'Economie et de Statistique* (91/92):239–61.

Quandt, Richard E. 1958. "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes." *Journal of the American Statistical Association* 53(284):873–80.

Quandt, Richard E. 1972. "A New Approach to Estimating Switching Regressions." *Journal of the American Statistical Association* 67(338):306–10.

Robins, James M., Miguel Angel Hernan, and Babette Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11(5):550–60.

Robinson, Peter M. 1988. "Root-N-Consistent Semiparametric Regression." *Econometrica* 56(4):931–54.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41–55.

Roy, Andrew Donald. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3(2):135–46.

Sakamoto, Arthur, and Meichu D. Chen. 1991. "Inequality and Attainment in a Dual Labor Market." *American Sociological Review* 56(3):295–308.

Smock, Pamela J., Wendy D. Manning, and Sanjiv Gupta. 1999. "The Effect of Marriage and Divorce on Women's Economic Well-Being." *American Sociological Review* 64(6):794–812.

Stuart, Elizabeth A., Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf. 2011. "The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(2):369–86.

Thistlethwaite, Donald L., and Donald T. Campbell. 1960. "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment." *Journal of Educational Psychology* 51(6):309–317.

Toomet, Ott, and Arne Henningsen. 2008. "Sample Selection Models in R: Package sampleSelection." *Journal of Statistical Software* 27(7):1–23.

Vytlacil, Edward. 2002. “Independence, Monotonicity, and Latent Index Models: An Equivalence Result.” *Econometrica* 70(1):331–41.

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

Winship, Chris, and Robert D. Mare. 1992. “Models for Sample Selection Bias.” *Annual Review of Sociology* 18:327–50.

Winship, Chris, and Stephen Morgan. 1999. “The Estimation of Causal Effects from Observational Data.” *Annual Review of Sociology* 25:659–706.

Xie, Yu. 2013. “Population Heterogeneity and Causal Inference.” *Proceedings of the National Academy of Sciences* 110(16):6262–68.

Xie, Yu, Jennie Brand, and Ben Jann. 2012. “Estimating Heterogeneous Treatment Effects with Observational Data.” *Sociological Methodology* 42(1):314–47.

Zhou, Xiang. 2019. *localIV: Estimation of Marginal Treatment Effects using Local Instrumental Variables*. R package version 0.2.1, available at the Comprehensive R Archive Network (CRAN).

Zhou, Xiang, and Yu Xie. 2016. “Propensity Score-Based Methods Versus MTE-Based Methods in Causal Inference: Identification, Estimation, and Application.” *Sociological Methods & Research* 45(1):3–40.

Zhou, Xiang, and Yu Xie. Forthcoming. “Marginal Treatment Effects from a Propensity Score Perspective.” *Journal of Political Economy*.

Author Biographies

Xiang Zhou is an assistant professor in the Department of Government at Harvard University. He received a PhD in sociology and statistics from the University of Michigan. His research broadly concerns quantitative methodology, economic inequality and mobility, and contemporary Chinese society. His work has appeared in *American Sociological Review, American Journal of Sociology, Journal of Political Economy*, and *Proceedings of the National Academy of Sciences*, among other peer-reviewed journals.

Yu Xie is Bert G. Kerstetter ’66 University Professor of Sociology and director of Paul and Marcia Wythes Center on Contemporary China, Princeton University. He is also a Visiting Chair Professor of the Center for Social Research, Peking University. His main areas of interest are social stratification, demography, statistical methods, Chinese studies, and sociology of science. His recently published books include *Marriage and Cohabitation* (University of Chicago Press 2007) with Arland Thornton and William Axinn, *Statistical Methods for Categorical Data Analysis* (Emerald 2008, second edition) with Daniel Powers, and *Is American Science in Decline?* (Harvard University Press 2012) with Alexandra Killewald. His methodological work is on categorical data analysis, causal inference, and survey research. Xie is also a former editor of *Sociological Methodology.*