College Choice, Abilities and Lifetime Earnings: a Local IV Approach with Swedish Registry Data*

Martin Nybom

Institute for Social Research (SOFI), Stockholm University

DRAFT, Dec 2012

Abstract

A vast amount of research has been devoted to estimating the return to education. Recent work has however clarified that if returns both vary across people and influence their schooling decisions, standard methods typically fail to identify relevant treatment effect parameters. In this paper, I use Swedish registry data to estimate lifetime returns to college and to what extent they vary with respect to observed and unobserved characteristics. The data set includes nearly complete measures of lifetime earnings and high-quality ability measures, which allow me to examine how returns vary with respect to cognitive and noncognitive ability. I implement the local instrumental variables (LIV) procedure (Heckman and Vytlacil, 1999, 2001, 2005) to recover marginal and average treatment effects in the presence of self selection on idiosyncratic gains using the distance to nearest university and local labor market conditions in late adolescence as conditionally exogenous shifters of the selection probability. The large data set allows me to implement the method with more flexibility than in previous studies. My findings lend support to the notion of self selection, but mainly on observed characteristics. In particular, I find that returns vary substantially with respect to both cognitive and noncognitive ability.

*This version corresponds to one of the chapters of my PhD thesis. I thank Anders Björklund and Markus Jäntti for advice and encouragement. I am also grateful to Magne Mogstad, Philipp Eisenhauer, Tuomo Suhonen and seminar participants at SOFI (Stockholm University) and the Sunstrat workshop (Stockholm) for help and comments. Financial support from Swedish Council of Social and Working Life Research (FAS) is gratefully acknowledged.

Introduction

The literature on the returns to education is a classical but still vibrant part of labor economics. The main objective has traditionally been to estimate *the* return to education, either as the return to an additional year of schooling or to a specific level such as college (see Card, 1999, for a review). With the primary focus on solving the standard omitted-variable problem (Griliches, 1977), a large amount of estimates have been produced using standard quasi-experimental methods. However, both the notion of a homogeneous return, and the usefulness of such conventional methods, have been questioned in recent work that relates back to Willis and Rosen's (1979) notion of self selection on heterogeneous gains.

Carneiro et al. (2011) provide evidence that returns vary and that individuals select into college based on their idiosyncratic gain from doing so. In such cases, an additional identification problem arises that is distinct from standard selection bias. Angrist and Imbens (1995) and Heckman and Vytlacil (1999) have clarified that a standard instrumental variable (IV) approach then identifies a local average treatment effect (LATE) of potentially low or unclear external relevance.[1] One remedy that enables one to interpret IV estimates is to impose some structural assumptions; in particular, a clear specification of the choice model is crucial. A recent example of such an approach is Heckman and Vytlacil (1999, 2001, 2005) who use the generalized Roy model and Local IV (LIV) estimation.

Despite growing interest in this approach, existing applications are few and have been limited to survey data. A drawback of the method is its heavy data requirements, especially in its semiparametric versions. This paper sheds new light on the LIV approach by estimating the returns to college using a large registry-based data set of Swedish males. The paper has three objectives: (i) to explore the applicability of the LIV approach; (ii) to provide new estimates of the returns to college in Sweden while taking self selection into account; (iii) to examine the extent of heterogeneity and assess the relative importance of heterogeneity that is observable and unobservable to the researcher. In particular, I make use of high-quality data on cognitive and noncognitive ability to analyze their effects on estimated returns.[2]

Such evidence may shed new light on several important questions. To evaluate the effects of educational policy, it is of central interest to know both

[1] The external relevance of the LATE will vary by context and instruments: in some cases it can be nearly impossible to interpret, in other cases it can be both easily interpretable and of high relevance (e.g., if a certain reform constitutes the instrument and the effect of the reform is of primary interest).

[2] There is a well-known conceptual distinction between "abilities", "skills" and, e.g., "test scores". I will predominantly use the term abilities in this paper, although I recognize that my measures are imperfect measurements of underlying ability. Below when I describe the data I will however argue that these proxies are of unusually high reliability.

who gains from schooling and *how much* they gain. An example is the debate on the optimal size of the college sector, for which it is necessary to concentrate on the distribution of returns in the population, and in particular if returns for those at the margin of attending exceeds marginal costs. Exposing potential heterogeneity may also deepen the understanding of earnings inequality in general, and provide explanations to changes in returns and earnings inequality across time. A related question is if the rise in the college premium (and overall inequality) since the 1980s (see, e.g., Autor et al., 2008) primarily reflects a general shift in the demand for college-educated workers, or rather changes in the degree of "school-skill complementarity", i.e., an increased demand for college-educated workers equipped with certain skills that are produced independent of college (Blackburn and Neumark, 1993; Taber, 2001).[3]

The approach of Heckman and Vytlacil has two cornerstones: (i) a choice-theoretic structure that defines each individual's margin of indifference towards selecting into treatment and (ii) local identification of marginal treatment effects (MTE). The MTE, first introduced by Björklund and Moffitt (1987), implicitly contains information about heterogeneity, and can be used with appropriate weights to compute summary treatment effect parameters such as the population average treatment effect (ATE).

The idiosyncratic return has two parts: heterogeneity with respect to characteristics that are observable and unobservable to the researcher. Whereas previous work, including Carneiro et al. (2011), have been heavily focused on the role of unobservable heterogeneity, I also seek to address the role of heterogeneity with respect to observable characteristics. I devote special attention to two high-quality measures of cognitive and noncognitive ability that are based on high-stake tests from the mandatory enlistment to the Swedish military. I examine if these abilities influence returns and assess their relative importance. This will ultimately cast new light on the potential complementarity of formal education and different abilities.

The study thus relates to the ongoing debate on the role of cognitive and noncognitive ability for educational and labor market outcomes. The notion of individual ability has recently shifted from a onedimensional concept primarily related to IQ, such as in the single-skill signaling model (Arrow, 1973) and the *g* factor (Herrnstein and Murray, 1994), to a multidimensional set of skills that especially recognizes the importance of personality or noncognitive ability.[4] A growing literature concerns the reduced-form earnings returns to various measures of abilities: some focus on the returns to cognitive ability or IQ

[3]More specifically, this school-skill complementarity refers to the cross-derivatives of formal schooling and "informal" skills or abilities in a typical production function.

[4]For example, it has been found that class size affects later performance mainly by increasing noncognitive rather than cognitive ability (Chetty et al., 2011), and that early childhood programs such as Headstart and the Perry Preschool Program are effective in terms of long-run outcomes mainly by increasing noncognitive ability (Heckman, 2005).

(e.g., Murnane et al., 1995; Cawley et al., 2001; Zax and Rees, 2002); others on the returns to noncognitive ability and personality traits (e.g., Nyhus and Pons, 2005; Groves, 2005; Mueller and Plug, 2006); and a third group of papers considers both types of ability jointly (e.g., Heineck and Anger, 2010; Lindqvist and Vestman, 2011).

These studies however rarely address how abilities transmit into earnings and wages, in particular via endogenous schooling choices. For example, one could ask if those with high IQ earn more because of their IQ as such, or because high IQ makes acquiring more education less costly or more beneficial, thus indirectly leading to increased earnings. Some evidence suggests that returns to education vary with respect to different measures of cognitive ability. In two recent Swedish studies, Nordin (2008) and Öckert (2012) estimate heterogeneous returns relying on a selection-on-observables assumption. The former finds that the return to a year of schooling increases at a diminishing rate by level of cognitive ability, the latter that the return to a year of college is steadily increasing with respect to secondary school GPA. Carneiro and Lee (2009) estimate returns to college within the LIV framework using NLSY data and find that cognitive test scores are positively related to the return, and at an increasing rate. These studies do not consider the role of noncognitive ability. In a rare exception, Heckman et al. (2006a) consider multiple abilities jointly. They use NLSY data in a factor structure model and find that both cognitive and noncognitive ability are important in explaining several economic and non-economic outcomes within different schooling groups.

A key difference in the present paper is that I apply a semiparametric LIV estimator, with arguably less restrictive parametric assumptions. The main merit, however, is the data source. Whereas most related studies rely on NLSY data, this paper is based on especially rich registry data that cover a large and representative sample of the Swedish male population. The size of the data set provides more flexibility compared to previous studies. I use measures of close to full lifetime earnings to minimize life-cycle effects in my estimates (Bhuller et al., 2011). In contrast to most previous work, my ability measures are collected at a uniform age and prior to college, thus offsetting much of the concern regarding endogeneity in observed ability measures (Hansen et al., 2004). Moreover, my measure of noncognitive ability is unique in that it is an overall judgement of psychological capability that stems from a semi-structured interview with a certified psychologist. This is in contrast with the NLSY measures that are based on combinations of different self-reported answers about one's personality.

I estimate the average treatment effect of a year of college to be a 4.8 percent increase in lifetime earnings. My findings are consistent with Carneiro et al. (2011) in that agents select into college based on their individual returns. The positive selection is manifested in that the difference between the average treatment effect on the treated and the untreated is consistently positive and

statistically significant across specifications. The evidence on unobserved heterogeneity is however more ambiguous. Overall, the results suggest that observed characteristics do capture a significant part of the total heterogeneity that drives self selection. The heterogeneity in returns with respect to both cognitive and noncognitive ability is substantial, and of comparable magnitude. My findings thus corroborate the idea of substantial school-skill complementarities.

The rest of this paper is structured as follows. Section 1 provides a brief theoretical illustration of the college decision. In Section 2, I show how the structure of the generalized Roy model is used to define the MTE, and briefly discuss identification and estimation. I describe the data in Section 3 and present the results in Section 4. I conclude by discussing some implications of my findings.

1 Theoretical Illustration

The decision rule in the generalized Roy model can be seen as a reduced form of a more elaborate theoretical model. To fix ideas, I will illustrate the college decision with a discrete-choice model of college attendance that builds on those in Keane and Wolpin (2001) and Keane (2002). I expand the model by introducing heterogeneity such that cognitive and noncognitive ability is allowed to affect the costs and benefits of acquiring college education. In addition to the model assumptions listed in Keane (2002), I therefore assume that all agents are endowed with a set of abilities $\mathbf{A}$ that are allowed to impact on both the indirect time costs and the direct utility (or consumption value) of going to college, as well as college and non-college earnings capacity.[5] The wage rate in period 1 is $w_1(\mathbf{A})$. In the following periods, the wage rate is $w_2(\mathbf{A}) + \beta(\mathbf{A})$ if the agent attended school, and $w_2(\mathbf{A})$ otherwise.

For a given ability realization $\mathbf{A} = \mathbf{a}$, the value function conditional on college attendance is

$$\begin{aligned} V_S \mid \mathbf{a} = & \max_{\{h,b\}} u\left[y_1 + b + hw_1(\mathbf{a}) - t, L - h - s(\mathbf{a})\right] + \\ & \varphi(\mathbf{a}) + \rho^{-1} u\left[w_2(\mathbf{a}) + \beta(\mathbf{a}) - rb, 1\right], \end{aligned} \tag{1}$$

and the value function for not attending is

[5] In short, the assumptions listed by Keane imply that agents: are infinitely lived in discrete time; decide whether to attend college in period 1 with a direct cost of attending (e.g., tuition, transaction or moving costs) denoted by t; face a discount factor ρ and interest rate r; can borrow or save b in period 1 with fixed annuity payments rb from period 2 and onwards; devote time to work denoted by h and can work while in college; receive an exogenous transfer payment y_1 from their parents in period 1; receive non-monetary utility from college denoted by $\varphi(\cdot)$; receive utility from consumption c and leisure l through a utility function denoted $u(c,l)$ which is concave in both arguments and $L \geq l \geq 0$; and inelastically supply one unit of labor after period 1 so that utility is $u(c,1)$.

$$V_0 \mid \mathbf{a} = \max_{\{h,b\}} u\left[y_1 + b + w_1(\mathbf{a}), L - h\right] + \rho^{-1} u\left[w_2(\mathbf{a}) - rb, 1\right]. \tag{2}$$

Utility maximization gives

$$u_1(c_1, l_1) = r\rho^{-1} u_1(c_2, l_2), \tag{3}$$

as the first-order condition for *inter*temporal optimality. Moreover, the first-order condition for *intra*temporal optimality is

$$w_1(\mathbf{a}) u_1(c_1, l_1) = u_2(c_1, l_1), \tag{4}$$

given an interior solution. A first-order Taylor expansion of V_S around V_0 at the point of indifference gives the (approximate) decision rule to attend college if and only if

$$\varphi(\mathbf{a}) + \rho^{-1}\beta(\mathbf{a}) u_1(c_2, l_2) - u_2(c_1, l_1) s(\mathbf{a}) - u_1(c_1, l_1) t \geq 0. \tag{5}$$

By using the two first-order conditions, equation (5) can be rewritten as

$$\lambda_1^{-1}\varphi(\mathbf{a}) + r^{-1}\beta(\mathbf{a}) \geq w_1(\mathbf{a}) s(\mathbf{a}) + t, \tag{6}$$

where $\lambda_1 = u_1(c_1, l_1)$. This has several implications. First, parental transfers y_1 affect the decision only through the marginal utility of consumption. If there is no non-monetary utility from schooling, i.e., $\varphi(\mathbf{a}) = 0$, parental transfers do not affect the schooling decision.[6] If $\varphi(\mathbf{a}) > 0$, then larger parental transfers increase attendance rates by decreasing the marginal utility from consumption and thereby increasing $\varphi(\mathbf{a})/\lambda_1$.[7] Second, the higher the interest rate as well as the direct and time costs of getting a degree, the lower is attendance.

The focus of this paper is on the role of individual abilities. For simplicity, consider $\mathbf{a} = a$ as a uni-dimensional ability realization such as a standard notion of cognitive ability. Differentiating the decision rule with respect to a gives:

$$\varphi'(a)/\lambda_1 + \varphi(a)\lambda_1'\lambda_1^{-2} + r^{-1}\beta'(a) \geq w_1'(a) s(a) + w_1(a) s'(a). \tag{7}$$

Ability thus affects the decision rule through several mechanisms: (i) through the non-monetary utility (or consumption value) of college $\varphi'(a)$; (ii) indirectly through the effect of $w_1(a)$ on the marginal utility of consumption; (iii) through the monetary return $\beta(a)$; (iv) through the time cost of acquiring schooling

[6] This is however overturned if individuals are credit constrained (see, e.g., Keane, 2002). Constrained individuals use the parental transfer to pay the direct costs t and to smooth consumption, and higher parental transfers require less borrowing in order to achieve intertemporal optimality. This dimension has been frequently studied in the US perspective, largely motivated by its perceived importance. It is not explicitly addressed in this paper since the absence of tuition fees suggests it is less important (but not necessarily irrelevant) in Sweden.

[7] This is a well-know result and also pointed out in Keane (2002).

$s(a)$; and (v), since $s(a)$ implies foregone earnings, through the monetary opportunity cost $w_1(a)$. For example, assume that ability increases the non-monetary utility of college, as well as both first- and second-period earnings (i.e., absolute advantage), and that it lowers the time cost. In this case the only mechanism that works against attending college is $w_1^{'}(a)s(a)$, i.e., first-period earnings at a given time cost. If we believe that non-college earnings differ relatively little by ability so that $w_1^{'}(a)$ is small or even negative (i.e., comparative advantage), then the effect of ability on the attendance decision is unambiguously positive. Given positive partial ability effects, the first two terms in (7) imply that increased ability both increases $\varphi(a)$, thus directly inducing more consumption of schooling, and lowers λ_1, which indirectly encourages the individual to consume even more schooling through $\varphi(a)$. The term $r^{-1}\beta^{'}(a)$ is the effect of ability on the long-run earnings return from attending college, which I will examine in detail in the empirical part below.

Finally, consider the role of parental transfers as ability varies. Given that the marginal utility of consumption is positive and strictly concave, higher ability increases attendance more for individuals with less wealthy parents (if they give smaller transfers). Moreover, the transfer could be thought of as a composite of parental transfers y_1^p and public tax-based student support y_1^g such that $y_1 = y_1^p + y_1^g$. An important implication is then that higher tax-based student support crowds out the attendance effect of parental transfers. I return in Section 4 to some of the implications of this model to interpret my empirical findings.

2 Econometric Model

Empirical work on the returns to schooling traditionally seeks to estimate variations of the equation

$$Y = \alpha + \beta S + \varepsilon, \tag{8}$$

where Y denotes the post-school wage or earnings, and S can be years of schooling, a vector of schooling levels, or an indicator variable of, e.g., college education. Under familiar assumptions, an OLS regression of Y on S yields an unbiased estimate of β.[8] As discussed above, standard selection bias (S correlated with ε) has typically received most attention, but more recently the issue of heterogeneous "sorting on the gain" (S correlated with β) has been advanced. Heckman et al. (2006b) distinguish between non-essential and essential sources of heterogeneity, where the former implies sorting by

[8]For example, if S is a college dummy, and equation (8) includes control variables $\mathbf{X}$, the OLS estimator gives $E[Y_1 \mid \mathbf{X}, S = 1] - E[Y_0 \mid \mathbf{X}, S = 0]$. The necessary assumptions are: no selection bias, i.e., $Cov(\varepsilon, S \mid \mathbf{X}) = 0$; no self selection based on unobserved heterogeneous gains, i.e., $Cov(\beta, S \mid \mathbf{X}) = 0$; and the parametric assumptions of OLS.

observable and the latter by unobservable characteristics. In what follows, I will use the terms *observed* and *unobserved* in reference to the empirical analyst's perspective. Although a more advanced IV approach may in principle take the former into account, the latter will typically cause IV estimates (i.e., the LATE) to diverge from average and marginal treatment effects.[9]

The econometric method that I apply originates from research by Heckman and Vytlacil (1999, 2001, 2005) in which they use the marginal treatment effect (MTE) to identify and unify different treatment effect parameters under heterogeneous sorting. The MTE, originally introduced by Björklund and Moffitt (1987), is the average treatment effect of those at the margin of indifference for selecting into treatment. This margin of indifference can be identified by imposing the structure of the generalized Roy model on the selection equation.

The Generalized Roy Model

The generalized Roy model offers a discrete-choice framework for policy analysis in which agents self select into treatment based on their expected gains.[10] The decision rule in the binary version of the model can be seen as the reduced form of an economic model of college attendance such as the one outlined in Section 1.

Let S be a binary choice indicator with $S = 1$ if the agent selects into treatment and $S = 0$ if not. Moreover, let the potential outcomes in the two states be

$$Y_j = \mu_j(\mathbf{X}, \mathbf{A}) + U_j, \quad \text{for } j = 0, 1 \tag{9}$$

where $\mathbf{X}$ is a set of observed regressors, $\mathbf{A}$ is a set of observed ability measures, μ_j are unknown functions, and U_j are unobserved random variables that need not be orthogonal to $\mathbf{X}$ and $\mathbf{A}$. The observed outcome can be written in switching regression form:

$$Y = SY_1 + (1 - S)Y_0. \tag{10}$$

Plugging (9) for both states into (10) gives

$$Y = \mu_0(\mathbf{X}, \mathbf{A}) + S\left[\mu_1(\mathbf{X}, \mathbf{A}) - \mu_0(\mathbf{X}, \mathbf{A}) + U_1 - U_0\right] + U_0. \tag{11}$$

The individual benefit of treatment is defined as the difference between potential

[9] Under observable heterogeneity, IV may recover average and marginal treatment effects provided that the functional forms of the regression equations are sufficiently flexible so that the LATE coincides with these parameters. This would, for example, imply non-linearities such as interactions between "sorting variables" and the endogenous treatment variable. The literature appears unsettled on the issue whether two-stage least squares IV is an appropriate estimator in such cases.

[10] The original version of the model is due to Roy (1951). Although different in style and notation, the essence of the model is similar to the one in Willis and Rosen (1979).

outcomes $Y_1 - Y_0 = \mu_1(\mathbf{X}, \mathbf{A}) - \mu_0(\mathbf{X}, \mathbf{A}) + U_1 - U_0$. Thus, the average treatment effect conditional on $\mathbf{X} = \mathbf{x}$ is given by $\text{ATE}(\mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$, and the average ability-specific treatment effect conditional on $\mathbf{X} = \mathbf{x}$ and $\mathbf{A} = \mathbf{a}$ is $\text{ATE}(\mathbf{x}, \mathbf{a}) = \mu_1(\mathbf{x}, \mathbf{a}) - \mu_0(\mathbf{x}, \mathbf{a})$. Moreover, conditioning on $S = 1$ or $S = 0$ defines the average treatment effect of the treated (ATT) and of the untreated (ATU), respectively.[11]

Let I_S denote the (expected) net benefit of selecting into college. An individual's decision rule can then be written as a standard latent variable discrete choice model (see, e.g., Willis and Rosen, 1979) of observed and unobserved variables:

$$\begin{aligned} I_S &= \mu_S(\mathbf{Z}) - V, \\ S &= 1 \text{ iff } I_S \geq 0. \end{aligned}$$

The individual thus selects into college if $I_S \geq 0$, and otherwise not. $\mathbf{Z}$ is an observed vector which may include some or all of the components of $(\mathbf{X}, \mathbf{A})$, but also components $\mathbf{Z} \backslash (\mathbf{X}, \mathbf{A})$ that are excluded from $(\mathbf{X}, \mathbf{A})$. V is unobserved and represents the individual (latent) resistance to select into college. Moreover, assume that V is a continuous variable with a strictly increasing cumulative distribution F_V, and that (U_0, U_1, V) are statistically independent of $\mathbf{Z}$ conditional on $(\mathbf{X}, \mathbf{A})$. $\mathbf{Z} \backslash (\mathbf{X}, \mathbf{A})$ thus work as exogenous cost-shifters that affect the outcome only through the college decision. At this stage, no independence condition is required for the common elements of $\mathbf{Z}$ and $(\mathbf{X}, \mathbf{A})$.

Finally, let the propensity score $P(\mathbf{z}) \equiv \Pr(S = 1 \mid \mathbf{Z} = \mathbf{z}) = F_V[\mu_S(\mathbf{z})]$ denote the probability of college attendance conditional on $\mathbf{Z}$, with the conditioning on all common elements in $(\mathbf{X}, \mathbf{A})$ held implicit. Define $U_S = F_V(V)$ such that U_S corresponds to the quantiles of V and is by construction uniformly distributed. The latent index can be rewritten using $F_V(\mu_S(\mathbf{Z})) = P(\mathbf{Z})$ so that $S = 1$ if $P(\mathbf{Z}) > U_S$. Within this framework, $P(\mathbf{Z})$ and U_S represent the observed and unobserved inducement to college education: the higher is $P(\mathbf{Z})$, the more inducement to attend college from the observables in $\mathbf{Z}$; the higher is U_S, the larger the unobserved resistance to college. For a person of high U_S, it thus takes a high inducement from $\mathbf{Z}$ to attend college. If $P(\mathbf{Z}) = U_S$, the individual is indifferent to attending.

[11] We have $\text{ATT}(\mathbf{x}) = \text{ATE}(\mathbf{x}) + E(U_1 - U_0 \mid S = 1, \mathbf{X} = \mathbf{x})$ and $\text{ATU}(\mathbf{x}) = \text{ATE}(\mathbf{x}) + E(U_1 - U_0 \mid S = 0, \mathbf{X} = \mathbf{x})$, and, for the ability-specific effects, $\text{ATT}(\mathbf{x}, \mathbf{a}) = \text{ATE}(\mathbf{x}, \mathbf{a}) + E(U_1 - U_0 \mid S = 1, \mathbf{X} = \mathbf{x}, \mathbf{A} = \mathbf{a})$ and $\text{ATU}(\mathbf{x}, \mathbf{a}) = \text{ATE}(\mathbf{x}, \mathbf{a}) + E(U_1 - U_0 \mid S = 0, \mathbf{X} = \mathbf{x}, \mathbf{A} = \mathbf{a})$.

Identifying the Marginal Treatment Effect

The marginal treatment effect (MTE) is defined by

$$\begin{aligned} \text{MTE}(\mathbf{x},\mathbf{a},u_S) &\equiv E(Y_1 - Y_0 \mid \mathbf{X}=\mathbf{x},\mathbf{A}=\mathbf{a},U_S=u_S) \\ &= \mu_1(\mathbf{x},\mathbf{a}) - \mu_0(\mathbf{x},\mathbf{a}) + E\left[U_1 - U_0 \mid U_S = u_S\right], \end{aligned} \tag{12}$$

and can be identified across the support of U_S. Since this is conditional on $(\mathbf{X},\mathbf{A})$, it is the local perturbation of the propensity score that is induced by $\mathbf{Z}\backslash(\mathbf{X},\mathbf{A})$ at quantile u_S that provides identification. The return to college can be recovered for persons on the margin of indifference at all quantiles of U_S within the support of $P(\mathbf{Z})$. Persons with high $P(\mathbf{Z})$ identify the return for those with high U_S, and vice versa. Local perturbations at high levels of $P(\mathbf{Z})$ induce persons with high U_S (i.e., high unobserved resistance) to change their treatment status. Those with low U_S are already in treatment for such values of $P(\mathbf{Z})$. Local perturbations at low $P(\mathbf{Z})$ induce those with low U_S to change their treatment status while those with high U_S remain out of treatment for such values of $P(\mathbf{Z})$. If the treatment effect is homogeneous with respect to U_S, then the MTE as a function of U_S would be flat. If the MTE correlates with U_S conditional on $(\mathbf{X},\mathbf{A})$, then there is unobserved heterogeneity.

A key virtue of the MTE approach is that summary parameters – e.g., the ATE, ATT, ATU and conventional IV effects (LATE) – can be recovered using estimates of the MTE and appropriate weights (Heckman et al., 2006b). The LATE is in this framework a discrete form of the MTE, defined on a particular section of U_S. With full support on U_S, the MTE can be estimated at each u_S quantile on the unit interval, with $P(\mathbf{z}) = u_S$ defining the margin of indifference in equation (12). An average MTE at each level of U_S can be obtained by integrating over the joint distribution of $(\mathbf{X},\mathbf{A})$ conditional on $U_S = u_S$. Integrating over the (uniform) distribution of U_S yields the unconditional ATE. The same procedure, conditioning on $S=1$ or $S=0$, gives the unconditional ATT and ATU, respectively.

The ability-specific ATE can be obtained by integrating over $\mathbf{X}$ and U_S while conditioning on $\mathbf{A}=\mathbf{a}$, such that $\text{ATE}(\mathbf{a}) = E_{\mathbf{X},U_S|\mathbf{A}=\mathbf{a}}\left[\text{MTE}(\mathbf{x},\mathbf{a},u_S)\right]$ traces out the ATE at a given value of $\mathbf{A}$. However, the interpretation of $\text{ATE}(\mathbf{a})$ as the contribution of ability to the treatment effect will be confounded if there are heterogeneous treatment effects with respect to variables in $\mathbf{X}$ that correlate with $\mathbf{A}$. An alternative and potentially more robust procedure is instead to impose a linear and separable version of $\mu_1(\mathbf{X},\mathbf{A}) - \mu_0(\mathbf{X},\mathbf{A})$ in equation (11) with $\mu_0(\mathbf{X},\mathbf{A}) = \mathbf{X}\delta_0 + \mathbf{A}\gamma_0$ and $\mu_1(\mathbf{X},\mathbf{A}) = \mathbf{X}\delta_1 + \mathbf{A}\gamma_1$. The ability-specific treatment effect, purged of other heterogeneous treatment effects related to observed covariates, is then given by $\gamma_1 - \gamma_0$.

Estimation using Semiparametric LIV

The MTE can be estimated using local instrumental variables (LIV) as proposed by Heckman and Vytlacil (1999, 2001, 2005). This approach relies on the fact that the expected value of Y depends on the propensity score $P(\mathbf{Z})$, so that $P(\mathbf{Z})$ serves as a local IV. Heckman and Vytlacil show that

$$\Delta_{LIV}(\mathbf{x}, \mathbf{a}, u_S) = \frac{\partial E(Y \mid \mathbf{X} = \mathbf{x}, \mathbf{A} = \mathbf{a}, P(\mathbf{Z}) = p)}{\partial p} \mid_{p=u_S} = MTE(\mathbf{x}, \mathbf{a}, u_S). \quad (13)$$

The computation of the MTE thus involves the estimation of the partial derivative of the conditional expectation of Y with respect to p. For my empirical analysis, in which I consider the linear and separable version of the model, the expected value in (13) can be written as

$$E(Y \mid \mathbf{X} = \mathbf{x}, \mathbf{A} = \mathbf{a}, P(\mathbf{Z}) = p) = \mathbf{x}\delta_0 + \mathbf{a}\gamma_0 + p\left[\mathbf{x}(\delta_1 - \delta_0) + \mathbf{a}(\gamma_1 - \gamma_0)\right] + K(p), \quad (14)$$

where $K(p) = E(U_1 - U_0 \mid S = 1, P(\mathbf{Z}) = p)$. The expression shows that the expected outcome is determined by three components: non-college earnings, the part of the treatment effect that is attributed to observed characteristics, and $K(p)$, which represents the effect that is attributed to unobserved characteristics. Using equations (13) and (14), the estimator becomes

$$\text{MTE}(\mathbf{x}, \mathbf{a}, u_S) = \mathbf{x}'(\delta_1 - \delta_0) + \mathbf{a}'(\gamma_1 - \gamma_0) + \frac{\partial K(p)}{\partial p} \mid_{p=u_S}. \quad (15)$$

In order to compute the MTE, I thus need to estimate $(\delta_1 - \delta_0)$, $(\gamma_1 - \gamma_0)$ and $\partial K(p)/\partial p$. The full estimation procedure that I implement involves several steps.[12] In the first stage, I estimate the college choice equation using a probit model to obtain estimates of $P(\mathbf{Z})$. I then estimate the coefficients in equation (14) using a semiparametric version of the double residual regression procedure. Specifically, I estimate separate local linear regressions of each of the regressors and the outcome variable on the predicted propensity score, and then retrieve their respective residuals.[13] Estimates of δ_0, γ_0, $(\delta_1 - \delta_0)$, and $(\gamma_1 - \gamma_0)$ are then obtained by regressing the residual associated with the outcome on the residuals associated with the variables in $(\mathbf{X}, \mathbf{A})$. Finally, with these estimates at

[12] The implementation follows the guidelines for LIV estimation ("Semiparametric Method 1") presented in Heckman et al. (2006b), and in more detail, at: http://jenni.uchicago.edu/underiv.

[13] Notice that, if $n_X + n_A$ denotes the total number of variables in $(\mathbf{X}, \mathbf{A})$, this step involves the estimation of in total $2 \times (n_X + n_A) + 1$ regressions. This is since equation (14) also contains interaction terms between each of the variables in $(\mathbf{X}, \mathbf{A})$ and the propensity score. The local linear regressions are estimated for the set of values of p that is contained in the support of $P(\mathbf{Z})$ using a kernel function with a bandwidth of 0.4.

hand, $\partial K(p)/\partial p$ can be estimated using standard nonparametric techniques.[14]

The LIV estimate of MTE$(\mathbf{x}, \mathbf{a}, u_S)$ is computed by plugging in the resulting parameter estimates into equation (15). I obtain estimates of the summary treatment effects by applying the respective weights obtained from the data (see Appendix A.1).

An alternative to the semiparametric estimation approach is to impose parametric assumptions on the unobservables and derive the expression for the MTE. This approach, relying on joint estimation of the choice and outcome equations as an endogenous switching regression, is more in line with the work of Willis and Rosen (1979) and Björklund and Moffitt (1987). Similar to above, the parametric MTE estimates can be used together with weights to compute summary treatment effects. As a comparison to the results from the semiparametric LIV, I will also present estimates from a parametric version of the model that assumes joint normality of (U_0, U_1, V). In this case, the MTE can be written as

$$\text{MTE}(\mathbf{x}, \mathbf{a}, u_S) = \mathbf{x}'(\delta_1 - \delta_0) + \mathbf{a}'(\gamma_1 - \gamma_0) - (\sigma_{1V} - \sigma_{2V})\Phi^{-1}(u_S), \tag{16}$$

where $E(U_1 - U_0 \mid U_S = u_S) = -(\sigma_{1V} - \sigma_{2V})\Phi^{-1}(u_S)$ and has a variance that is normalized to one.[15] I estimate the parameters $\delta_1, \delta_0, \gamma_1, \gamma_0, \sigma_{1V}, \sigma_{2V}$ and their standard errors by maximum likelihood and plug them into equation (16) to obtain estimates of MTE$(\mathbf{x}, \mathbf{a}, u_S)$.

3 Data and Sample Restrictions

The data set is based on a representative sample of Swedish men born 1951-1957 and is obtained by merging several registers from Statistics Sweden using unique personal identifiers.[16] These registers include data on earnings, ability test scores, educational attainment, personal and local labor market characteristics, and data on family members. The analysis is restricted to men since the ability data come from military enlistment registers.

[14] This term is also estimated using local linear regression across the common support of $P(\mathbf{Z})$, i.e., the subset of $P(\mathbf{Z})$ for which I obtain positive frequencies in both $S = 1$ and $S = 0$. I use an Epanechnikov kernel function with a bandwidth of 0.1.

[15] Moreover, $\sigma_{1V} = Cov(U_1, V)$, $\sigma_{0V} = Cov(U_0, V)$, and $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cumulative distribution function.

[16] The raw sample is based on a random draw of approximately a third of the Swedish male population. The 1951-1957 cohorts are chosen for two primary reasons: they enable me to use nearly career-long measures to approximate lifetime earnings, and they are the earliest cohorts with data from the military enlistment.

Measure of Lifetime Earnings

To construct a measure of lifetime earnings, I make use of individual information on annual labor earnings from tax-declaration files for the years 1968-2007. These data come with a number of advantages: they are almost entirely free from attrition; pertain to all jobs; are not right-censored; and are believed to suffer relatively little from reporting errors. I approximate each individual's log lifetime earnings by the log of the mean of all non-missing earnings observations over ages 20-50 (i.e., ages for which earnings are observed for all cohorts).[17]

That the data allow for a nearly career-long earnings measure is unusual. In estimation of the returns to education in general, and when based on an explicit decision model in particular, the relevant outcome (or maximand in the model) is the stream of earnings across the lifetime. As a contrast, it has been standard in the literature to use single-year or short-run outcome measures, often from around age 30.[18] If the age-earnings relationship is related to the components in equation (9), such estimates will not in general be unbiased, because heterogeneous earnings profiles cause non-classical measurement error when short-run earnings measures are used as proxies for lifetime earnings.[19] That such "life-cycle bias" can have large quantitative effects on estimates has been shown in recent studies (e.g., Nybom and Stuhler, 2011; Bhuller et al., 2011). Since I use earnings data that span over 31 years, my estimates should be relatively unaffected by such bias.

Measures of Cognitive and Noncognitive Abilities

An attractive feature of the data set is that it includes information from the mandatory military enlistment's tests of cognitive and psychological (noncognitive) ability. The enlistment typically takes place at age 18 and includes two days of physical, intellectual, and psychological tests and evaluations.[20]

[17]The actual measure is "Inkomst av tjänst" in Swedish. The measure includes labor earnings and labor-related benefits such as parental leave benefits. The measure does not include income from self employment. I discount each annual observation to a present value at age 20 using an annual rate of 0.02. Using the slightly different measure "Arbetsinkomst", which includes income from self employment, yields very similar results.

[18]This is particularly true for the numerous studies that rely on NLSY data, including examples such as Heckman et al. (2006a) and Carneiro et al. (2011).

[19]The standard practice of using short-run measures from around age 30 appears to be notably precarious when estimating the returns to college. Individuals with longer schooling may then recently have entered the labor market, which is reflected in relatively low, and possibly also noisy earnings as of higher rates of on-the-job investments and job switching. This is especially a concern in a country like Sweden where higher education is on average both commenced and finished at later ages than in comparable countries.

[20]For the male cohorts born 1951-57, only a tiny fraction were exempted from the enlistment, mainly because of physical or psychological disability. Although the test scores are drawn from a similar age for all, concerns about the joint causality of schooling, latent skills, and test performance (see Hansen et al., 2004) leads me to restrict the sample to individuals with similar

The measure of cognitive ability is based on scores on a test of general intelligence that has been conducted since the 1940s. The test consists of four subtests of logical, verbal, and spatial ability, as well as technical comprehension, each graded on a discrete scale from 1 to 9. The scores on the subtests are transformed to a discrete general variable between 1 and 9 that follows a Stanine scale.[21]

The measure of noncognitive ability is based on standardized interview-based evaluations made by certified psychologists. In the interview, the enlistee's psychological profile and capacity to fulfill the requirements of military duty are evaluated. Central to this is the ability to cope with stress and contribute to group cohesion. Other valued traits include willingness to assume responsibility, independence, emotional stability, outgoing character, persistence, and the ability to take initiatives. Motivation for doing military service is not considered. The interview is semi-structured in the sense that the psychologist follows a manual that states topics to discuss and how to grade answers. Scores are given on four subscales and as an overall assessment that follows a Stanine scale between 1 and 9.

This variable is valuable in two ways: it provides a general "omnibus" measure of noncognitive ability and it is based on a psychologist's experience from a personal encounter with the individual, which is likely to capture more aspects of a personality than what can be deduced from survey questionnaires. Moreover, both the measure of cognitive and noncognitive ability benefit from high comparability and cover large samples.[22]

Data on Education and Background Characteristics

The education data come from a population register that describes the highest level of education, in what year it was achieved, and from what type of study program. I measure educational attainment up until 1990 when the men in the sample were 33-39 years old. The college indicator takes a value one if an individual had a minimum of three years of college studies.[23] To be able to show estimates in annualized values, I impute a measure of years of schooling

pre-academic educational attainment at the age of enlistment by excluding those without a degree from an acedemic high school track. I examine the sensitivity of the results by dropping this restriction in Section 4.

[21] Carlstedt (2000) provides a detailed overview of this test as well as the Swedish military's history of psychometric testing. Carlstedt also provides evidence that the test is a good measure of general intelligence, thus in contrast with tests that tend to measure the more malleable concept of crystallized intelligence (as noted by Lindqvist and Vestman, 2011).

[22] For more information on the military enlistment data, see the excellent data description in Lindqvist and Vestman (2011).

[23] In Section 4, I analyze the sensitivity of the results by using a less strict definition of the college variable (at least one semester).

based on the highest level.[24]

From the registers, I also obtain data on personal characteristics and family background. From censuses, I get data on birth date, country of birth, and geographical residency at different ages. I use a multigenerational register to identify (biological) family members and thus obtain data on birth order, number of siblings, father's and mother's years of schooling, and father's earnings.[25]

Instrumental Variables

As exogenous cost shifters in the choice equation, I use distance from the place of residence to the closest university and short-run fluctuations in unemployment and average earnings in the municipality of residence at the end of high school. Distance to college was first used as an instrument by Card (1995) and has later been frequently utilized, for example in applications of the LIV procedure such as Carneiro et al. (2011).[26] I construct a continuous measure of typical travel distance by car between the central town of the local municipality in which the individual resided in 1965 and the closest university city. This is in contrast with most of the previous literature that has either used dummy variables for whether a college is located in the home county, or, in the case of continuous measures, more crude measures such as "as the crow flies" generated from geographical coordinates.[27]

The unconditional exogeneity of such distance instruments has been questioned in studies based on both US (Cameron and Taber, 2004) and Swedish data (Kjellstrom and Regner, 1999). Given the recommendations in these studies, it is thus crucial that I condition this instrument on measures of ability and family background.

My instruments also include measures of short-run fluctuations in the local labor market at the end of high school, conditioned on permanent local

[24] Around 1970, when most of the men in the sample were about to make their college choice, admission to higher education in Sweden was largely unrestricted. Most higher education was open to anyone with a high school degree and many faculties had no formal application procedure. There were also no tuition fees and the system of student aid was generous. See Erikson and Jonsson (1993) for a detailed account of the history of the Swedish system of higher education.

[25] I compute father's earnings as the average of non-missing annual earnings in the years 1968-1972.

[26] Other papers that have used variations of this instrument include Kling (2001), Currie and Moretti (2003), Cameron and Taber (2004), and Carneiro and Lee (2009).

[27] In 1965 there were 998 local municipalities in my sample, thus ensuring large variation in the distance measure. I consider six university cities: Stockholm, Uppsala, Gothenburg, Lund, Umeå, and Linköping. Up until the late 1970s, nearly all Swedish college students studied in one of these cities. The measure is calculated using the website eniro.se, which is a tool similar to Google Maps. Using a measure of shortest travel time in minutes instead of distance yields similar results.

labor market conditions.[28] The underlying idea is that while both current (or short-run) and permanent local labor market conditions are in the individual's information set at the time of the college decision, the current conditions do not contain any additional information about the future *conditioned on* the permanent component. If this holds true, then such innovations in the local labor markets can be excluded from the outcome equation. As measures of current conditions, I use the unemployment rate and average earnings in the municipality of residence at age 20. As permanent measures, I use municipal averages of unemployment and earnings over the years 1968-1988.[29] Although this type of instrument has been frequently used in the recent literature, one may still be concerned about whether current conditions actually can be excluded from the outcome equation. If individuals would put a higher weight on current measures when forecasting the benefits of the choice alternatives, then these would also enter the outcome equations. To address this concern, I will also provide estimates that do not include these among the instruments.

Model Specification and Sample Statistics

The linear-in-parameter representations of $(\mathbf{X}, \mathbf{A})$ include linear and quadratic terms of cognitive and noncognitive test scores $(\mathbf{A})$, mother's and father's years of schooling, father's log earnings, number of siblings, permanent local earnings and unemployment, as well as region and cohort dummies $(\mathbf{X})$. The exclusion restrictions that enter $\mathbf{Z} \backslash (\mathbf{X}, \mathbf{A})$ are linear and quadratic terms in local short-run earnings and unemployment, and a cubic polynomial of the distance measure. Following Carneiro et al. (2011), I interact the instruments with linear terms in cognitive and noncognitive test scores, mother's years of schooling, and number of siblings. Recognizing that the effect of distance may vary depending on region, I also interact these with the regional dummies.[30] The sample statistics are presented in Table 1.

[28] Previous papers that have done so include Cameron and Heckman (1998), Cameron and Taber (2004), Carneiro and Lee (2009), and Carneiro et al. (2011). As Cameron and Taber (2004) argue, the impact of these variables on schooling choice is theoretically ambiguous. On the one hand, better labor market conditions increase the opportunity cost of schooling. On the other hand, a better labor market also increases the resources of credit constrained households, thus promoting educational attainment.

[29] In effect, I use "non-employment", i.e., one minus the employment rate in the local working-age population, as my measure of unemployment.

[30] The six regions are defined as "university regions" so that all municipalities that share the same "closest university" constitute a region. Moreover, it has been common in applications of the LIV method to include variables in the outcome equation that are not in the selection equation as a means to increase precision. Carneiro et al. (2011), for example, include experience as well as local earnings and unemployment in the municipality of residence at prime age. As these are post-determined, and thus likely endogenous, I avoid including such variables in the baseline analysis. I instead present estimates from such specifications in the sensitivity analysis.

Table 1 Summary Statistics by Treatment Group

	$S=1$		$S=0$	
	Mean	SD	Mean	SD
Log lifetime earnings	12.07	0.50	11.93	0.50
Cognitive test score	0.90	0.78	0.54	0.80
Noncognitive test score	0.46	0.98	0.33	0.92
Mother's years of education	9.99	3.03	8.71	2.38
Number of siblings	2.84	1.17	2.86	1.26
Father's years of education	11.45	3.63	9.66	2.98
Father's log earnings	12.36	1.19	12.05	1.27
Local long-run earnings (SEK/100)	137.94	14.37	136.27	14.82
Local long-run unemployment	0.21	0.04	0.22	0.04
Distance to university (km/100)	0.91	0.99	1.00	1.02
Local short-run earnings (SEK/100)	132.35	19.12	131.06	19.75
Local short-run unemployment	0.26	0.07	0.26	0.07
Non-missing earnings observations	30.72	1.36	30.81	1.15
Years of education	15.90	1.16	12.10	0.98
Number of observations	23186		31840	

Note: Lifetime earnings is computed as the average of all non-missing annual earnings observations for ages 20-50. Test scores are standardized by birth year at the population level (i.e. before any sample restrictions). Father's earnings are computed as the average of annual non-missing earnings for years 1968-1972. Local permanent labor market characteristics are computed as averages across the years 1968-1990 by municipality of residence at age 20. The short-run measures are for age 20. Unemployment is computed as one minus the local working-age employment rate (i.e. a measure of "non-employment"). Distance to university is measured as the closest route by car from the municipality of residence in 1965 (in total 996 entities) to the closest university city (Stockholm, Uppsala, Linköping, Gothenburg, Lund or Umeå). Included in the set of controls are also regional and birth-year dummies (not reported here).

4 Empirical Results

The main objective of this paper is to examine the importance of heterogeneity in the returns to college using the semiparametric LIV estimator. However, it is instructive to first consider more standard approaches, which will then serve as a comparison later in the paper.

Results using Conventional Methods

A natural point of departure is to consider standard OLS estimates. I estimate different versions of eq. (8), with control variables either only entering independently, or also interacted with the college dummy S. Moreover, to illustrate the role of the ability measures, the models are estimated both assuming that these are unobserved (i.e., excluded from the set of controls) and observed (i.e., included as controls). The results are reported in Table 2.

The (discounted) lifetime return to a year of college is estimated to be around 3.7 percent when not controlling for observed abilities (columns 1-2). It falls

Table 2 OLS Estimates of the Return to a Year of College

	OLS Coefficients				
	(1)	(2)	(3)	(4)	(5)
College dummy (*S*)	0.0374	0.0376	0.0319	0.0312	0.0316
	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0012)
*S****A** (Cognitive)	.	.	.	0.0186	0.0175
	.	.	.	(0.0022)	(0.0023)
*S****A** (Noncognitive)	.	.	.	0.0076	0.0067
	.	.	.	(0.0014)	(0.0013)
Ability controls (**A**)	.	.	x	x	x
Interactions *S****A**	.	.	.	x	x
Interactions *S****X**	.	x	.	.	x

Note: This table reports OLS regression coefficients of log lifetime earnings on the college dummy (*S*). The control variables (**X**) include region and cohort dummies, as well as linear and quadratic terms of father's and mother's years of schooling, father's log earnings, number of siblings, local long-run unemployment and earnings in the municipality of residence at age 20. Specifications (3)-(5) also include linear and quadratic terms of the measures of cognitive and noncognitive ability (i.e. **A**). Specifications (2) and (5) include interactions between *S* and all components of **X**, and (4) and (5) include interactions between *S* and all components of **A**. The interaction terms in rows 2 and 3 (*S****A**) are reported as average derivatives (standard errors from 100 bootstrap replications). All coefficients are divided by 3.8 to reflect the difference in years of schooling between those with and without college. Standard errors are in parentheses.

to about 3.1 percent when these are included as controls (columns 3-5). This illustrates in a simple way the potential (positive) ability bias in OLS estimates. Allowing for interactions between the control variables and the college dummy, on the other hand, seems to have little effect on the main estimate. Nevertheless, the observed heterogeneity with respect to the two ability measures is quite substantial (columns 4-5). A one standard deviation increase in cognitive (noncognitive) ability increases the return to a year of college by around 1.7 (0.7) percent. The results thus suggest that there may be considerable variation in individual returns, despite the relative stability of the estimated average effect.

As opposed to OLS, standard IV estimates a causal effect without assuming equal potential outcomes for treated and untreated individuals. I report IV estimates for different sets of instruments in Table 3. Observed heterogeneity is taken into account by including interaction effects in the second stage. In line with several previous studies, my IV estimates are larger than the OLS estimates. There is also variation in the estimated LATEs across different instruments and first-stage models (linear 2SLS or probit). Since different instruments identify the LATE for different subpopulations, such variation is expected in the presence of self selection on heterogeneous returns. Nevertheless, when I use $P(\mathbf{Z})$ with the full set of instruments, the estimate is in the lower range and close to the semiparametric estimate of the ATE (see below). This is not in itself a rejection of the self-selection hypothesis, but is nevertheless noteworthy.

Table 3 IV Estimates of the Return to a Year of College

	IV estimates for different sets of instruments			
	Distance to university	Local earnings	Local unempl.	All
Standard 2SLS	0.0798	0.1150	0.0673	0.0651
	(0.0395)	(0.0461)	(0.0364)	(0.0249)
*P(***Z***)* as instr.	0.0501	0.0847	0.0463	0.0522
	(0.0378)	(0.0376)	(0.0383)	(0.0223)

Note: This table reports IV estimates of the return to college. The respective columns are for different sets of instruments: distance to university at age 20 (cubic) in column 1, local short-run earnings and unemployment (quadratics) in columns 2 and 3, and all these instruments in column 4. Row 1 reports estimates for standard two-stage least squares (2SLS) with a linear first stage, row 2 reports estimates using *P(Z)* as instrument (probit first stage). All specifications include in the second stage interactions between predicted college and all components of **X** and **A**. All coefficients are divided by 3.8 to reflect the difference in years of schooling between those with and without college. Standard errors (in parentheses) are bootstrapped (100 replications).

Results using a Normal Selection Model

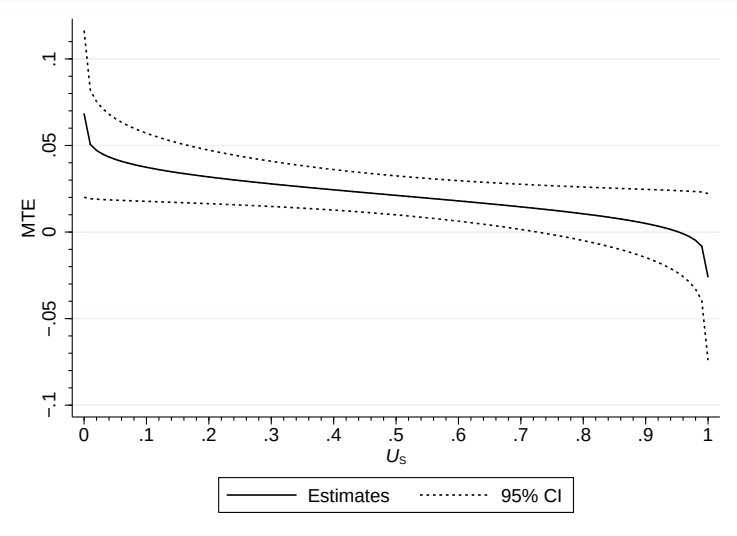
The traditional approach to estimate the model in Section 2 is to specify a parametric joint distribution for the error terms (e.g., Willis and Rosen, 1979). Björklund and Moffitt (1987), for example, estimate the MTE assuming that the error terms are jointly normally distributed. Although my main focus is on the semiparametric method, results based on a normal selection model are useful for purposes of comparison.

Figure 1 shows parametric estimates of the MTE by levels of U_S, conditioned on mean values of $(\mathbf{X}, \mathbf{A})$. The MTE is weakly declining and relatively precisely estimated. A test for selection on unobserved gains is to test whether the slope of the conditional MTE is zero. For the normal selection model this implies testing whether $\sigma_{1V} - \sigma_{2V} = 0$ in eq. (16). I estimate that $\sigma_{1V} - \sigma_{2V} = -0.0481$ with a standard error of 0.0291 (obtained using the delta method). Thus, I cannot reject the hypothesis that the slope of the MTE is zero at the 95 percent confidence level, although it is on the border of rejection at the 90 percent level (z-statistic = 1.6538). As a comparison, Carneiro et al. (2011) estimate that $\sigma_{1V} - \sigma_{2V} = -0.2388$ with a standard error of 0.0982, and thus reject a flat MTE.

If I also account for observed heterogeneity, i.e., the variation in $\mathbf{X}$ and $\mathbf{A}$ and their impact on the MTE through $\mathbf{X}'(\delta_1 - \delta_0) + \mathbf{A}'(\gamma_1 - \gamma_0)$, then the slope becomes much steeper. The magnitude of total heterogeneity is illustrated by the change in the part of the (unconditional) MTE that is attributed to $(\mathbf{X}, \mathbf{A})$ when going from the first to the tenth decile of U_S, which corresponds to about 10 percentage points in terms of the return to a year of college. Across all individuals, the heterogeneity associated with observed characteristics varies between -0.1089 and 0.2328.

Table 4 (column 1) reports estimates of summary treatment parameters based on the parametric MTE estimates and the appropriate weights (reported

Figure 1 MTE by U_S Estimated from a Normal Selection Model



Note: This figure shows point estimates and 95 percent confidence bands of the MTE from the parametric normal selection model in equation 16 estimated by maximum likelihood. All estimates are conditioned on mean values of **X** and **A**.

in Appendix A.2). The estimated ATE implies a return to one year of college of about 2.4 percent. The corresponding estimates for the ATT and ATU are approximately 2.9 and 1.6 percent, respectively. Table 4 also shows tests of equality between ATT and ATE, ATT and ATU, and ATE and ATU, which serve as broad tests for self selection on *total* heterogeneity. All tests reject equality and support the notion that individuals choose schooling based on their own comparative advantage. These results however rest on the potentially restrictive normality assumption, and it is thus not clear how reliable they are.

Results using Semiparametric LIV

A potentially more robust approach for estimating the MTE is to estimate $E(Y \mid \mathrm{X} = \mathbf{x}, \mathbf{A} = \mathbf{a}, P(\mathbf{Z}) = p)$ semiparametrically and then compute its derivative with respect to p, as in eq. (13). This is the essence of the LIV approach. If $(\mathbf{X}, \mathbf{A})$ is not independent of (U_0, U_1, V), a necessary (and very demanding) condition is that P has full support at each value of $(\mathbf{X}, \mathbf{A})$. For each combination of $(\mathbf{X}, \mathbf{A})$, variation in P can only identify the MTE across small intervals of V. To reduce the dimensionality of $(\mathbf{X}, \mathbf{A})$, I therefore use an index of $\mathbf{X}'(\delta_1 - \delta_0) + \mathbf{A}'(\gamma_1 - \gamma_0)$.[31] The support of P for each value of the index is nevertheless small. If I instead follow Carneiro et al. (2011) and invoke the assumption that $(\mathbf{X}, \mathbf{A})$ is independent of (U_0, U_1, V), then each of the intervals from the conditional identification can be put together so that the MTE can

[31] I follow Basu et al. (2007) and condition on demideciles (i.e., 20 uniformly distributed groups) of the scalar index $\mathbf{X}(\delta_1 - \delta_0) + \mathbf{A}(\gamma_1 - \gamma_0)$. The results are robust to conditioning on finer partitions of the index (50 or 100 uniformly distributed groups).

Table 4 Returns to a Year of College

Model	Normal	Semiparametric
ATE	0.0238	0.0484
	(0.0027)	(0.0208)
ATT	0.0321	0.0574
	(0.0029)	(0.0208)
ATU	0.0178	0.0418
	(0.0027)	(0.0210)
ATT - ATU	0.0142	0.0155
	(0.0008)	(0.0032)
ATT - ATE	0.0082	0.0089
	(0.0004)	(0.0018)
ATE - ATU	0.0060	0.0066
	(0.0003)	(0.0013)

Note: This table reports estimates of the average treatment effect (ATE), the average treatment effect on the treated (ATT), and average treatment effect on the untreated (ATU). The estimates in column 1 are based on maximum likelihood estimates of MTEs from the normal switching regression model in equation 16. The estimates in column 2 are based on the semiparametric model, and thus for $\widetilde{ATE}$, $\widetilde{ATT}$, and $\widetilde{ATU}$ (rather than the true ATE, ATT, ATU), with the twiggle indicating that they are sample specific parameters that are conditional on the estimated support of U_S. Rows 4-6 shows the estimated differences between the treatment effect parameters. Standard errors are obtained using the bootstrap (100 replications).

be identified over almost the entire support of V. It is thus only necessary to examine the marginal support of $P(\mathbf{Z})$ as opposed to the support of $P(\mathbf{Z})$ conditional on $(\mathbf{X}, \mathbf{A})$. This assumption also legitimizes the use of interactions between $\mathbf{Z}$ and components of $(\mathbf{X}, \mathbf{A})$ as instruments in the choice equation.

I estimate $P(\mathbf{Z})$ in a probit model and present estimated average marginal derivatives in Table 5. I also report average marginal effects for each of the polynomials of the instruments. The average effect of the distance instrument on college attendance is negative and highly significant. The average effect of local unemployment at age 20 is also negative and significant, whereas local earnings at age 20 is a weak predictor of college attendance. The instruments are jointly strong predictors of college attendance, as are mother's and father's years of schooling, father's earnings, the measure of noncognitive ability, and, in particular, the measure of cognitive ability. In fact, cognitive ability is in terms of average derivatives around eight times stronger than noncognitive ability as a predictor of going to college.

Figure 2 shows the support of the estimated $P(\mathbf{Z})$. There is a lack of support in the lowest tenth of the interval, whereas the support in the upper part of the interval nearly reaches one.[32] Given the estimates of $P(\mathbf{Z})$, the next step is to estimate the components of equation (15) and compute the MTEs.

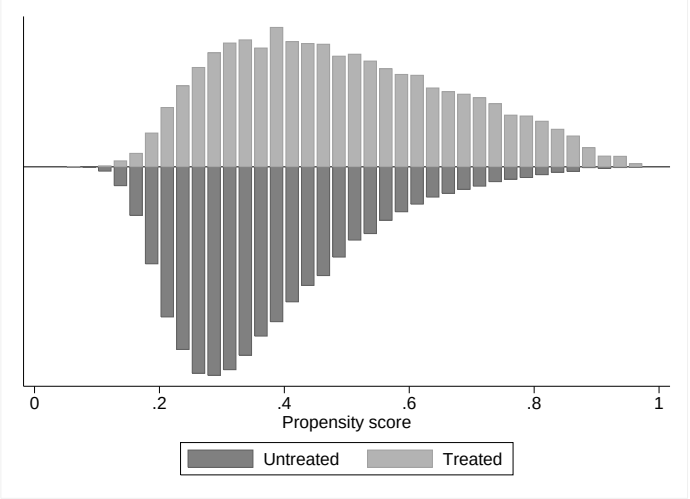
[32]The common support is defined as the intersection of the support of $P(\mathbf{Z} \mid S = 0)$ and the support of $P(\mathbf{Z} \mid S = 1)$. I trim observations for which the estimated $P(\mathbf{Z})$ is either lower than the minimum, or higher than the maximum value of $P(\mathbf{Z})$ for which there is common support.

Table 5 College Decision Model

Controls (**A**, **X**)	Avg. derivative	Avg. marginal effect
Cognitive test score	0.1047	
	(0.0030)	
Non-cognitive test score	0.0130	
	(0.0024)	
Mother's years of schooling	0.0164	
	(0.0011)	
Number of siblings	-0.0024	
	(0.0018)	
Father's years of schooling	0.0173	
	(0.0009)	
Father's log earnings	0.0406	
	(0.0031)	
Local long-run earnings	0.0018	
	(0.0015)	
Local long-run unempl.	0.7832	
	(0.3828)	
Instruments (**Z**)		
Distance to university (km/100)	-0.0586	-0.1114
	(0.0159)	(0.0312)
Distance to univ. (km/100) quadratic		0.2372
		(0.0676)
Distance to univ. (km/100) cubic		-0.1278
		(0.0410)
Local short-run earnings	-0.0018	-0.0003
	(0.0015)	(0.0040)
Local short-run earnings quadratic		-0.0000
		(0.0000)
Local short-run unempl.	-0.7292	-0.5483
	(0.3208)	(0.6411)
Local short-run unempl. quadratic		-0.3892
		(1.0025)
Joint significance test of **Z**: p-value	0.0000	

Note: The table reports average derivatives and marginal effects from a probit regression of a college indicator on the set of variables listed in the table and cohort and region dummies (see Section 3 for exact specification). The average derivatives are obtained by computing for each individual the effect including all polynomial terms of increasing a variable by one unit (keeping all the others constant) on the probability of enrolling in college and then average across all individuals. The average marginal effects (reported for the instruments) are obtained in the same manner but separately for each polynomial term of the respective variable. Standard errors are obtained using the bootstrap (100 replications).

Figure 2 Support of $P(\mathbf{Z})$ for untreated ($S = 0$) and treated ($S = 1$)



Note: This figure shows the support of $P(\mathbf{Z})$ for the treated and the untreated. $P(\mathbf{Z})$ is the probability of going to college estimated in a probit regression of the college choice equation (see Table 5).

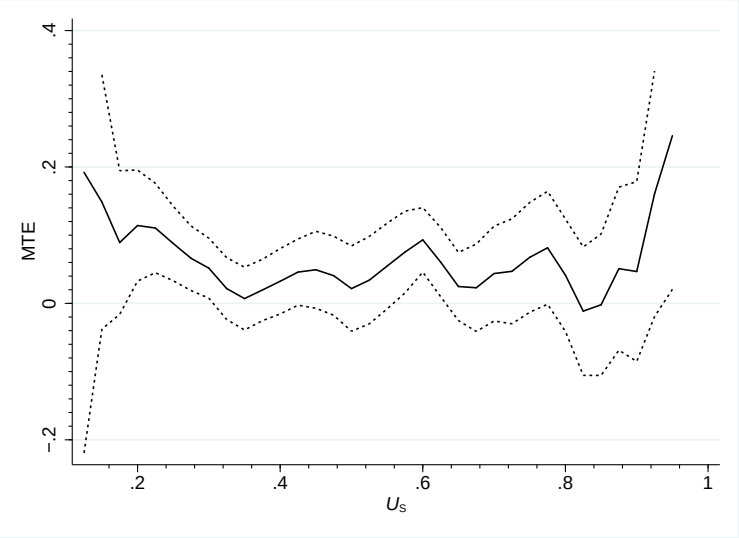
Figure 3 shows how the MTE depends on V across the quantiles of V (i.e., across U_S), with the components of $(\mathbf{X}, \mathbf{A})$ fixed at their sample means. Two main results emerge. First, the variation in unobserved heterogeneity is in terms of point estimates quite substantial. The difference between the sections of U_S with the highest and the lowest MTEs corresponds to about 20 percentage points in the returns to a year of college. For a major part of the U_S interval, however, the results suggest an almost flat MTE and thus little unobserved heterogeneity. Second, the evidence on self selection on unobserved gains is mixed. For values of U_S up until about 0.35, the MTE declines in U_S (i.e., positive selection). For intermediate values of U_S, there is not much of a clear pattern and for the top decile of U_S, the MTE is even increasing (i.e., negative selection). This indicates that the normal selection model provides an incorrect representation of unobserved heterogeneity. The overall picture is thus mixed, although the evidence on self selection on unobserved gains is clearly weaker than what is reported in Carneiro et al. (2011).[33]

A simple test of selection on unobserved gains consists of comparing the average MTE across equally spaced adjacent intervals along the support of U_S, i.e., LATEs defined over different subpopulations (see Heckman et al., 2010).[34]

[33] Although the semiparametric estimates have larger standard errors than the estimates based on the normal model, the precision of my semiparametric estimates are larger than in Carneiro et al. (2011). The main difference is instead that they find evidence of an MTE with an unambiguously steep negative slope.

[34] The test is based on 100 bootstrap replications of the MTE, evaluated at mean values of $\mathbf{X}$ and $\mathbf{A}$. I take the average of the MTE in equally spaced intervals along the support of U_S and compute the statistics $T = \mid \text{LATE}^j - \text{LATE}^{j+1} \mid$ (the absolute value of the difference between two adjacent LATEs j and j+1) and $T_b = \mid (\text{LATE}_b^j - \text{LATE}_b^{j+1}) - (\text{LATE}^j - \text{LATE}^{j+1}) \mid$, where LATE_b^j is the b^{th} bootstrap replication of LATE^j. The corresponding statistics for the joint test

Figure 3 MTE by U_S Estimated by Semiparametric LIV



Note: This figure shows semiparametric point estimates and 95 percent confidence bands of the MTEs from the model in equation 12. The model is estimated by the local linear regression procedure that is further described in Section 2. All estimates are conditioned on mean values of **X** and **A**. Standard errors are bootstrapped (100 replications).

Table 6 reports the outcome of the test. I cannot reject the joint hypothesis that all adjacent LATEs are equal.

Table 6 Test for Equality of LATEs over Different Intervals

Range of LATE^j	(.125;.200)	(.275;.350)	(.425;.500)	(.575;.650)	(.725;.800)
Range of LATE^{j+1}	(.275;.350)	(.425;.500)	(.575;.650)	(.725;.800)	(.875;.950)
LATE^j-LATE^{j+1}	0.0876	0.0138	0.0147	0.0029	0.1079
p-value	0.4600	0.8400	0.6500	0.9300	0.0400
Joint p-value			0.5000		

Note: This table reports a test of essential heterogeneity conducted by testing the equality of LATEs in pairwise adjacent intervals of U_S. I construct intervals of U_S and average the MTE within these intervals by computing $E(Y_1 - Y_0 \mid \mathbf{X} = \bar{\mathbf{x}}, U_S^{L_j} \leq U_S \leq U_S^{U_j})$, where $U_S^{L_j}$ and $U_S^{U_j}$ are the lower and upper bounds of U_S in interval j. This gives the different LATEs and the null of the tests are $H_0: \text{LATE}^j(U_S^{L_j}, U_S^{U_j}) - \text{LATE}^{j+1}(U_S^{L_{j+1}}, U_S^{U_{j+1}}) = 0$. The bottom row reports the outcome of the test that all adjacent LATEs are jointly equal. All tests take the multiple estimation steps into account by using the bootstrap (100 replications).

Lastly, I turn to my estimates of the ATE, ATT, and ATU. Since I do not have full support for P, these parameters cannot be estimated in exact accordance with their definitions. I can, however, compute approximations of these parameters, denoted $\widetilde{\text{ATE}}$, $\widetilde{\text{ATT}}$, and $\widetilde{\text{ATU}}$, for which I rescale the weights (reported in Appendix A.2) to integrate to one over the common support.

Table 4 (column 2) reports the estimates together with a set of simple tests

are $C = \sum_{j=1}^{J-1} (\text{LATE}^j - \text{LATE}^{j+1})^2$ and $C_b = \sum_{j=1}^{J-1} \left[(\text{LATE}_b^j - \text{LATE}_b^{j+1}) - (\text{LATE}^j - \text{LATE}^{j+1}) \right]^2$. The p-value of the tests is the proportion of bootstrap replications for which $T_b > T$ (or $C_b > C$ for the joint test).

for self selection on total heterogeneity. The semiparametric estimate of the $\widetilde{\text{ATE}}$ suggests a return to one year of college of about 4.8 percent. As expected, the estimated $\widetilde{\text{ATT}}$ is larger and $\widetilde{\text{ATU}}$ smaller, although the differences are relatively small. Nevertheless, the differences are all statistically significant, thus indicating sorting into college based on overall heterogeneity. It is those who actually have selected into college that, on average, also have the highest estimated ex-post returns. What is more surprising is the large and positive returns for those who have chosen not to go to college. This is in sharp contrast with Carneiro et al. (2011), who report an $\widetilde{\text{ATU}}$ that is close to zero. Finally, it is worthwhile to compare with the estimates from the normal selection model (column 1). These are substantially lower, but the pattern in terms of the differences across ATE, ATT, and ATU is very similar. As unobserved heterogeneity seems to be of modest importance in my sample, I now turn to examine the role of observed heterogeneity, and in particular ability-specific heterogeneity.

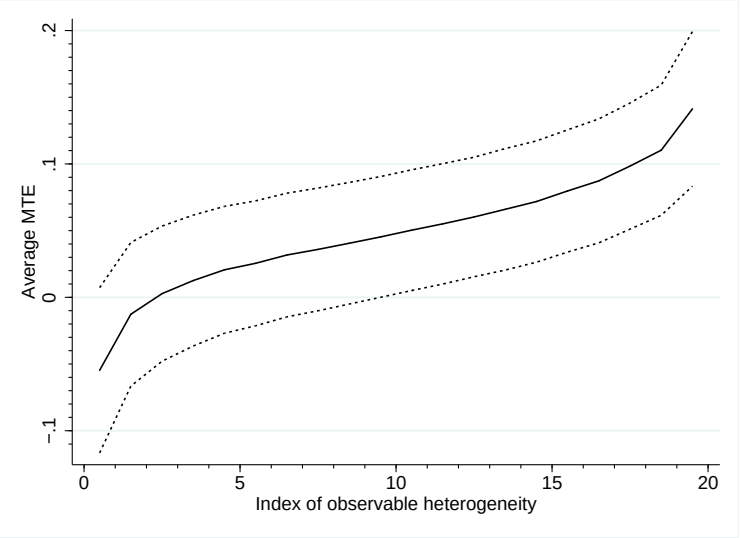
Evidence on Observable Heterogeneity and Ability Heterogeneity

Figure 4 shows the component of the MTE that is attributable to total observed heterogeneity along the scalar index $\mathbf{X}(\delta_1 - \delta_0) + \mathbf{A}(\gamma_1 - \gamma_0)$. First, note that this curve is not comparable to Figure 3, which plotted the MTE across the distribution of U_S. U_S is an unobserved variable, while the scalar index on the x-axis in Figure 4 is itself estimated. Both do however govern expected returns and thereby selection into college. Figure 4 implies that the variation in total observed heterogeneity is substantial and the slope of the curve indicates that observed characteristics impact on returns across the entire distribution (in terms of point estimates). The curve suggests that those with the most favorable characteristics (i.e., that complement formal college education the most) on average have a return that is around 20 percentage points higher than those with the least favorable characteristics. In reality, the heterogeneity is even larger since there is also considerable variation within the groups on the x-axis.

Figures 5a and 5b show the ATE conditional on the measures of cognitive and noncognitive ability, respectively. There is a strong relationship between both measures and the estimated ATE. Moreover, both the pattern and the magnitude of the heterogeneity are roughly similar for the two measures, although the negative effects at the low end are more pronounced for cognitive ability. At the top end, the positive complementarity with college education, as suggested by the point estimates, is even somewhat larger for noncognitive ability, although the difference is marginal. Belonging to the top category in either cognitive or noncognitive ability implies a return to a year of college that is around 10 percentage points higher than the average.

A potential explanation to the high resemblance across the two measures

Figure 4 Average MTE by Total Observed Heterogeneity



Note: This figure plots the average MTE with 95 percent confidence bands across the index of observed heterogeneity. The index is computed by estimating $\mathbf{X}(\delta_1 - \delta_0) + \mathbf{A}(\gamma_1 - \gamma_0)$ for each individual and splitting the sample into demideciles (i.e., 20 uniformly distributed groups), thus following the procedure of Basu et al. (2007) and Carneiro et al. (2011).

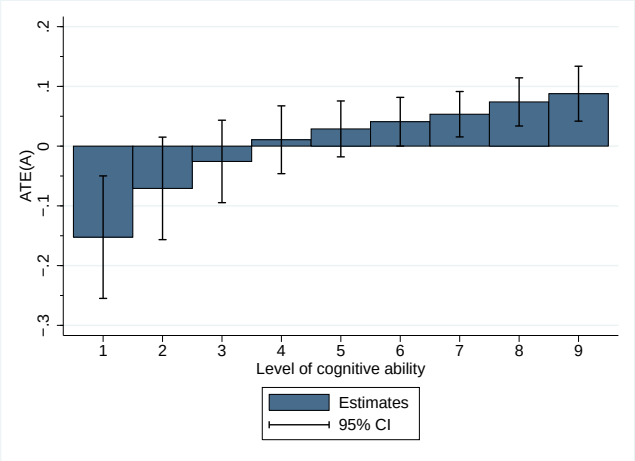
would be that they are highly correlated. This correlation is about 0.19 in my sample, suggesting that this could only be a partial explanation. Moreover, it is possible that the ability measures are correlated with other control variables that impact on observed heterogeneity. In Table 7 (column 2), I therefore report the semiparametric estimates of observed heterogeneity with respect to the (standardized) ability measures from the outcome equation. Despite the fact that these estimates are thus conditional on any potential impacts on observed heterogeneity from other covariates, they imply a pattern similar to the comparison of the conditional ATEs.[35] The estimates, presented as average derivatives, imply that an increase of one standard deviation in the measure of cognitive (noncognitive) ability on average increases the return by about 3.5 (2.5) percentage points. These estimates are larger than the comparable OLS estimates in Table 2, but roughly similar in terms of the relative importance of the two measures. Moreover, the semiparametric estimates in Table 7 (column 1) imply very modest, or even zero, direct effects from the ability measures in the outcome equation. The evidence thus lends support to the notion of comparative advantage in college education, whereas the support for absolute advantage is weak.

The resemblance between the estimates in Table 7 (column 2) and the condi-

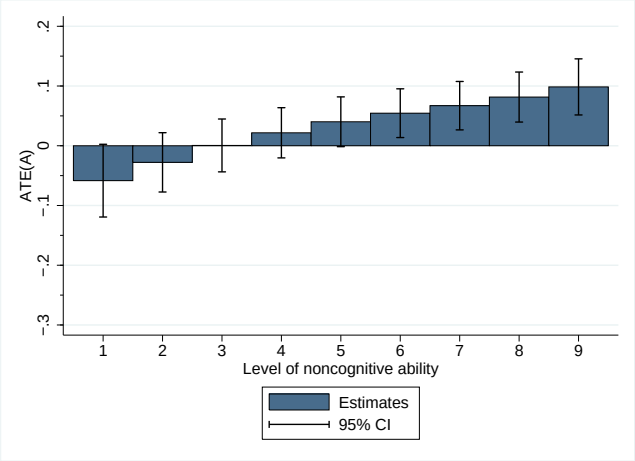
[35]Note however that these estimates are not conditional on unobserved heterogeneity as they rely on the auxiliary assumption that the observed and unobserved heterogeneity components are uncorrelated. The previously discussed estimates of unobserved heterogeneity are derived conditional on estimated observed heterogeneity, I can thus not control for unobserved heterogeneity when I estimate observed heterogeneity.

Figure 5 Observed Ability Heterogeneity in the Return to a Year of College

(a) ATE by Cognitive Ability



(b) ATE by Noncognitive Ability



Note: The figures show semiparametric estimates of average treatment effects (ATE) with 95 percent confidence bands conditional on levels of cognitive and noncognitive abilities. The ability measures are plotted on the x-axes in their original unstandardized form, although standardized measures are used in all estimations. Standard errors are obtained using the bootstrap (100 replications).

Table 7 Average Derivatives for Abilities in the Outcome Equation

	γ_0	$\gamma_1 - \gamma_0$
Cognitive ability	-0.0011	0.0353
	(0.0038)	(0.0082)
Noncognitive ability	0.0069	0.0250
	(0.0017)	(0.0037)

Note: This table reports average derivatives of the (standardized) measures of cognitive and noncognitive ability in the outcome equations for the semiparametric model. The model is estimated by local linear regression. This procedure, and exact specifications of the full set of control variables (not reported here), are further described in Section 2. The average derivatives are obtained by computing for each individual the effect of increasing a variable by one unit (keeping all the others constant) on log lifetime earnings and then average across all individuals. Column 1 reports the main effects, whereas column 2 reports the interaction effects (i.e. observable heterogeneity). Standard errors are bootstrapped (100 replications).

tional ATEs is not surprising. First, the impact on observed heterogeneity from other covariates is very small as compared to the impact from the measures of cognitive and noncognitive ability.[36] Most observed heterogeneity thus seems to be captured by these two variables. Second, the previous evidence did not suggest any dramatic effects from unobserved heterogeneity, although estimates were imprecise. What is maybe more surprising is that the heterogeneity with respect to noncognitive ability is so large, and roughly comparable to the one with respect to cognitive ability. The probit estimates of the choice equation implied that cognitive ability is a much stronger predictor of selection into college than noncognitive ability. If selection were purely driven by expected benefits (i.e., monetary returns), then the two ability types should have a more equal predictive power of college attendance. However, the model of college choice in Section (1) illustrates some potential explanations for why this need not be the case. Cognitive ability might impact on the cost of going to college more than noncognitive ability, either in terms of time costs (yielding more leisure) or psychic costs (less headache) for a given achievement. It could also be due to heterogeneity in the valuation of college as a consumption good; the level of cognitive ability might influence the direct utility derived from going to college more positively than the level of noncognitive ability. Such explanations could each contribute to the result that cognitive ability seems to trigger selection into college much more strongly than noncognitive ability, despite having comparable effects on monetary returns.[37]

Sensitivity to Sample, Specifications and Variable Definitions

A simple way to analyze the robustness of my estimates is to examine how $\widetilde{\text{ATE}}$, $\widetilde{\text{ATT}}$, and $\widetilde{\text{ATU}}$ vary across specifications. In addition, I report a straightforward test of selection on returns: a test of the null that $\widetilde{\text{ATT}} = \widetilde{\text{ATU}}$, i.e., whether the average person attending college has the same return as the average person not attending. Results are reported in panels A, B and C of Table 8.

First, the results in panel A concern choice of sample and specification of the outcome equation. I excluded from my baseline sample everyone without an (academic) high school degree. This is in line with Willis and Rosen (1979), whereas Carneiro et al. (2011) include dropouts. Column 2 in panel A indicates that my estimates are relatively unaffected when including those, although estimated heterogeneity ($\widetilde{ATT} - \widetilde{ATU}$) increases. The model that I use has the limitation that it restricts the college variable to be binary. An intuitive critique of estimates of heterogeneous returns is that different people might choose

[36] The estimates of observed heterogeneity with respect to other control variables are not shown here, but are available from the author upon request.

[37] There are of course other potential explanations that go beyond this simple model, e.g., differences in preferences such as discount factors (as emphasized by Willis and Rosen, 1979) or risk attitudes, and differential forecasting errors.

Table 8 Returns to a Year of College: Sensitivity Analyses

(a) Different Samples and the Specification of the Outcome Equation

	Baseline	Including high school dropouts	Type of college in $\mathbf{X} \setminus \mathbf{Z}$	$\mathbf{X} \setminus \mathbf{Z}$ as Carneiro et al. (2011)	$\mathbf{X} \setminus \mathbf{Z}$ excluding experience
$\widetilde{ATE}$	0.0484	0.0528	0.0137	0.0339	0.0408
	(0.0208)	(0.0206)	(0.0201)	(0.0130)	(0.0210)
$\widetilde{ATT}$	0.0574	0.0786	0.0604	0.0378	0.0508
	(0.0208)	(0.0173)	(0.0212)	(0.0123)	(0.0209)
$\widetilde{ATU}$	0.0418	0.0421	-0.0256	0.0311	0.0335
	(0.0210)	(0.0224)	(0.0198)	(0.0130)	(0.0212)
$\widetilde{ATT} - \widetilde{ATU}$	0.0155	0.0365	0.0860	0.0067	0.0173
	(0.0032)	(0.0071)	(0.0079)	(0.0028)	(0.0032)

(b) Specification of the Choice Equation

	Any college as treatment	Only linear terms in $\mathbf{Z}$	No interactions with $\mathbf{Z}$	Only distance in $\mathbf{Z}$	Sample, $\mathbf{X}$ as Carneiro et al.
$\widetilde{ATE}$	0.0418	0.0279	0.0318	0.0207	0.0501
	(0.0204)	(0.0286)	(0.0344)	(0.0349)	(0.0122)
$\widetilde{ATT}$	0.0503	0.0375	0.0413	0.0299	0.0613
	(0.0206)	(0.0284)	(0.0325)	(0.0328)	(0.0077)
$\widetilde{ATU}$	0.0332	0.0209	0.0248	0.0141	0.0463
	(0.0202)	(0.0288)	(0.0359)	(0.0350)	(0.0121)
$\widetilde{ATT} - \widetilde{ATU}$	0.0171	0.0166	0.0165	0.0158	0.0150
	(0.0033)	(0.0034)	(0.0031)	(0.0031)	(0.0056)

(c) Definitions of the Outcome Variable and analysis of life-cycle effects

	Avg. wage ages 20-50	Average earnings ages 26-30	Average earnings ages 36-40	Average earnings ages 46-50
$\widetilde{ATE}$	0.0432	-0.0403	0.1088	0.1005
	(0.0138)	(0.0312)	(0.0378)	(0.0389)
$\widetilde{ATT}$	0.0479	-0.0288	0.1170	0.1060
	(0.0137)	(0.0312)	(0.0376)	(0.0366)
$\widetilde{ATU}$	0.0398	-0.0487	0.1027	0.0965
	(0.0139)	(0.0313)	(0.0381)	(0.0391)
$\widetilde{ATT} - \widetilde{ATU}$	0.0080	0.0198	0.0143	0.0095
	(0.0022)	(0.0047)	(0.0047)	(0.0053)

Note: This table reports estimates of the return to a year of college for the semiparametric model for various samples and specificiations. The estimates of the average treatment effect (ATE), the average treatment effect on the treated (ATT), and average treatment effect on the untreated (ATU) are computed such that their weights integrate to one over the respective common support. The table also reports a simple test of self selection: if $\widetilde{ATT} - \widetilde{ATU} = 0$. Panel A reports estimates for different samples and specifications of the outcome equation, Panel B for different specifications of the outcome equation, and Panel C for the definition of the outcome variable and life-cycle effects. Standard errors are obtained using the bootstrap (100 replications).

different types of college in terms of quality or field of study. One could, in principle, extend the method used in this paper to multiple schooling types, but that would require distinct instruments for each schooling transition (Heckman et al., 2006b). To explore this limitation, I report estimates with indicators for three broad categories of college (bachelor or less, master, and doctoral degree) included in the outcome equation. The resulting estimates in column 3 surprisingly indicate larger heterogeneity, which might potentially be because these indicators are not exogenous.

An efficient way of reducing the residual variance in the estimation of the outcome equation is to include additional controls in the outcome equation only. Column 4 shows estimates from using total experience (quadratic) and local unemployment and earnings at about age 35, as such additional controls (i.e., similar to Carneiro et al., 2011). This approach is generally questionable, as each of these variables may be endogenous. Since I use lifetime earnings rather than short-run wage as outcome variable, this is most obviously the case with the experience variable, which I exclude from the estimates in column 5. The estimates in columns 4 and 5 are somewhat smaller than the baseline, and when including experience, both estimated heterogeneity and standard errors are much smaller.

Second, the results in panel B concern the specification of the choice equation. Column 1 shows that the estimates are relatively unaffected by using "any college" (minimum one semester) as the college indicator. Columns 2-4 show alternative specifications of the instruments. Using only linear terms in $\mathbf{Z}$ (column 2), or excluding interactions with $\mathbf{Z}$ (column 3), produce somewhat smaller estimates but larger standard errors. One might worry that the local labor market instruments affect selection into college by shifting expected returns, despite the fact that I only use the innovations in these variables in $\mathbf{Z}$. In column 4, I therefore report estimates for which only the distance variable is included in $\mathbf{Z}$. The estimates are now substantially smaller than the baseline but too imprecise to draw any firm conclusions. However, the pattern of estimated heterogeneity is remarkably stable across all the different specifications of the choice equation. Column 5 shows estimates for a setup in which I mimic Carneiro et al. (2011) as closely as possible in terms of sample and specification.[38] Both the estimates and the extent of heterogeneity are similar to my baseline, while the standard errors are considerably lower.

Finally, in panel C I exploit the nearly career-long earnings data to examine how the estimates vary across different definitions of the outcome variable. This sensitivity analysis relates to the interpretation (or external validity) of the estimated effects, rather than the internal validity. I first use an imputed wage measure as the outcome in order to examine the role of labor supply

[38] I thus include dropouts in the sample, and include the same type of variables in $\mathbf{X}$ (i.e., excluding noncognitive ability, and father's schooling and earnings) and in $\mathbf{X} \setminus \mathbf{Z}$ (i.e., local labor market characteristics in prime age and experience) as they do.

(reported in column 1).[39] The estimates are similar in size, but the standard errors and estimated heterogeneity are somewhat lower. This may suggest that labor supply plays a role for the evidence on selection on returns, as measured by annual earnings. Columns 2-4 show evidence on life-cycle effects in the estimates. In column 2, average earnings across ages 26-30 are used as the outcome and the resulting estimates are now negative (but statistically insignificant). This is not surprising, as many at this age might still be in school, or are recent entrants on the labor market. If earnings are instead observed in their late 30s (column 3) or late 40s (column 4), the estimated effects are positive and considerably larger than the baseline. The estimated heterogeneity is consistent with the baseline for earnings observed at the earlier ages, but somewhat smaller for older ages. This highlights that it is relatively easy to under- or overstate point estimates depending on at what age the outcome variable is observed.

5 Conclusions

I applied the LIV approach of Heckman and Vytlacil (1999, 2001, 2005) to a large registry-based data set of Swedish males. My analysis of the returns to college revealed a relatively modest role for heterogeneity in general, and for unobserved heterogeneity in particular, at least in comparison to previous evidence (e.g., Carneiro et al., 2011). Nevertheless, total heterogeneity (mainly via observed characteristics) seems to be an important phenomenon, and this holds across various specifications and sample definitions. However, it is unclear whether the divergence from previous evidence is due to differences in data quality or contextual setting (Sweden vs. the US). A possible explanation for both smaller returns and less heterogeneity could be a lower degree of selection into college in Sweden and most notably a more compressed wage structure. Recent quasi-experimental evidence also lends support to the finding of low returns to college in Sweden (see Öckert, 2010).

Moreover, I provided new evidence on ability heterogeneity using measures of cognitive and noncognitive ability from military enlistment tests. The results implied that both cognitive and noncognitive ability have a large influence on the return, thus indicating that "school-skill complementarities" (i.e., between formal schooling investments and independently produced abilities) are potentially important features of the labor market. Since the effect of noncognitive ability is almost as large as that of cognitive ability, it is puzzling that the former has much less influence on the probability of selecting into college. A potential

[39] I follow the procedure of Antelius and Björklund (2000) who show that left truncating these data, so that low earnings observations and likely part-time workers are excluded, gives similar estimates of the returns to schooling as when using wage measures. I thus use average annual earnings conditional on having annual earnings above 75 000 SEK (about $10000).

explanation might be that cognitive ability also has a more positive impact on either the costs (e.g., time or psychic costs) or the consumption value of going to school. An intriguing avenue for future research is to enable a more causal interpretation of the cost side of such ability heterogeneity, for example by introducing exogenous return shifters in the Roy model.

Some lessons regarding the applicability of the LIV approach can also be learned from my analysis. In general, sample size seems important for the applicability of the LIV approach, but also the existence of good continuous instruments. My large sample clearly produces more precise estimates of treatment effect parameters compared to previous applications that use smaller survey data. The evidence on unobserved heterogeneity is nevertheless somewhat inconclusive, as the part of the MTE that varies with respect to the unobserved characteristic remains quite imprecisely estimated.

The tendency of a U-shaped pattern of the MTE is also notable, as it differs from the monotonic curve reported in Carneiro et al. (2011). A pessimistic explanation would be that part of this is caused by a failure of the independence assumption. On the other hand, it is also possible that the effect of unobserved heterogeneity is more complex than what is commonly assumed. For example, Brinch et al. (2012) find robust evidence that supports a U-shaped MTE curve when applying the LIV approach to the quantity-quality tradeoff of children.

Several sources can potentially generate a non-monotonic shape of the MTE, including heterogeneity in time or risk preferences, assymetric information about the costs and benefits of college, and differences in economic resources or access to credit at the time of the college decision. This would be consistent with a population divided into multiple subpopulations that can be represented as a mixture distribution. As Brinch et al. (2012) demonstrate, a non-monotonic MTE can be derived from such underlying data, for example illustrated by a mixture of multiple normal distributions. A methodological implication is that a typical univariate normal selection model will impose a potentially incorrect representation of treatment effect heterogeneity. A practical implication, given that the highest returns are found in the top and bottom of the distribution, is that a mix of targeted policies that absorb students from opposite ends of the spectrum should be preferred over general expansions or contractions of the college sector.

References

Angrist, J., and G. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity," *Journal of the American Statistical Association*, 90(430), 431–442. 1

Antelius, J., and A. Björklund (2000): "How Reliable are Register Data for Studies of the Return on Schooling? An examination of Swedish data," *Scandinavian Journal of Educational Research*, 44(4), 341–355. 30

Arrow, K. (1973): "Higher Education as a Filter," *Journal of Public Economics*, 2(3), 193–216. 2

Autor, D., L. Katz, and M. Kearney (2008): "Trends in US Wage Inequality: Revising the Revisionists," *The Review of Economics and Statistics*, 90(2), 300–323. 2

Basu, A., J. Heckman, S. Navarro-Lozano, and S. Urzua (2007): "Use of Instrumental Variables in the Presence of Heterogeneity and Self-Selection: an Application to Treatments of Breast Cancer Patients," *Health Economics*, 16(11), 1133–1157. 19, 25, 36

Bhuller, M., M. Mogstad, and K. Salvanes (2011): "Life-Cycle Bias and the Returns to Schooling in Current and Lifetime Earnings," *IZA Discussion Paper No. 5788*. 3, 12

Björklund, A., and R. Moffitt (1987): "The Estimation of Wage Gains and Welfare Gains in Self-Selection," *The Review of Economics and Statistics*, 69(1), 42–49. 2, 7, 11, 18

Blackburn, M., and D. Neumark (1993): "Omitted-Ability Bias and the Increase in the Return to Schooling," *Journal of Labor Economics*, pp. 521–544. 2

Brinch, C., M. Mogstad, and M. Wiswall (2012): "Beyond LATE with a Discrete Instrument," *Statistics Norway Discussion Paper No. 703*, (703). 31

Cameron, S., and J. Heckman (1998): "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males," *Journal of Political Economy*, 106(2), 262–333. 15

Cameron, S., and C. Taber (2004): "Estimation of Educational Borrowing Constraints Using Returns to Schooling," *Journal of Political Economy*, 112(1), 132–182. 14, 15

Card, D. (1995): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, ed. by L. Christofides, K. Grant, and R. Swidinsky, pp. 201–222. University of Toronto Press, Toronto. 14

——— (1999): *The Causal Effect of Education on Earnings*vol. 3A of *Handbook of Labor Economics*, chap. 30, pp. 1801–1863. Elsevier, Amsterdam. 1

Carlstedt, B. (2000): "Cognitive Abilities: Aspects of Structure, Process and Measurement," Ph.D. thesis, University of Gothenburg. 13

Carneiro, P., J. Heckman, and E. Vytlacil (2011): "Estimating Marginal Returns to Education," *The American Economic Review*, 101(6), 2754–2781. 1, 2, 3, 12, 14, 15, 18, 19, 22, 24, 25, 27, 29, 30, 31

Carneiro, P., and S. 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. 3, 14, 15

Cawley, J., J. Heckman, and E. Vytlacil (2001): "Three Observations on Wages and Measured Cognitive Ability," *Labour Economics*, 8(4), 419–442. 3

Chetty, R., J. Friedman, N. Hilger, E. Saez, D. Schanzenbach, and D. Yagan (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *The Quarterly Journal of Economics*, 126(4), 1593–1660. 2

Currie, J., and E. Moretti (2003): "Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings," *The Quarterly Journal of Economics*, 118(4), 1495–1532. 14

Erikson, R., and J. Jonsson (1993): *Ursprung och utbildning*, SOU 1993:85. Utbildningsdepartementet. 14

Griliches, Z. (1977): "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, 45, 1–22. 1

Groves, M. (2005): "How Important is your Personality? Labor Market Returns to Personality for Women in the US and UK," *Journal of Economic Psychology*, 26(6), 827–841. 3

Hansen, K., J. Heckman, and K. Mullen (2004): "The Effect of Schooling and Ability on Achievement Test Scores," *Journal of Econometrics*, 121(1-2), 39–98. 3, 12

Heckman, J. (2005): "Invited Comments," in *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40*, ed. by L. Schweinhart, J. Montie, Z. Xiang, S. Barnett, C. Belfield, and M. Nores, vol. 14 of *Monographs of the High/Scope Educational Research Foundation*, pp. 229–233. High/Scope, Ypsilanti, MI. 2

Heckman, J., D. Schmierer, and S. Urzua (2010): "Testing the Correlated Random Coefficient Model," *Journal of Econometrics*, 158(2), 177–203. 22

Heckman, J., J. Stixrud, and S. Urzua (2006a): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24(3), 411–482.

Heckman, J., S. Urzua, and E. Vytlacil (2006b): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *The Review of Economics and Statistics*, 88(3), 389–432. 6, 9, 10, 29, 36

Heckman, J., and E. Vytlacil (1999): "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96(8). , 1, 7, 10, 30

——— (2001): "Local Instrumental Variables," in *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, ed. by C. Hsiao, K. Morimune, and J. Powell, pp. 1–46, New York, NY. Cambridge University Press. , 1, 7, 10, 30

Heckman, J., and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73(3), 669–738. , 1, 7, 10, 30

Heineck, G., and S. Anger (2010): "The Returns to Cognitive Abilities and Personality Traits in Germany," *Labour Economics*, 17(3), 535–546. 3

Herrnstein, R., and C. Murray (1994): *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press. 2

Keane, M. (2002): "Financial Aid, Borrowing Constraints, and College Attendance: Evidence from Structural Estimates," *The American Economic Review (Papers and Proceedings)*, 92(2), 293–297. 4, 5

Keane, M., and K. Wolpin (2001): "The Effect of Parental Transfers and Borrowing Constraints on Educational Ettainment," *International Economic Review*, 42(4), 1051–1103. 4

Kjellstrom, C., and H. Regner (1999): "The Effects of Geographical Distance on the Decision to Enrol in University Education," *Scandinavian Journal of Educational Research*, 43(4), 335–348. 14

Kling, J. (2001): "Interpreting Instrumental Variables Estimates of the Returns to Schooling," *Journal of Business and Economic Statistics*, 19(3), 358–364. 14

Lindqvist, E., and R. Vestman (2011): "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment," *American Economic Journal: Applied Economics*, 3(1), 101–128. 3, 13

Mueller, G., and E. Plug (2006): "Estimating the Effect of Personality on Male and Female Earnings," *Industrial and Labor Relations Review*, 60(1), 3–22. 3

Murnane, R., J. Willett, and F. Levy (1995): "The Growing Importance of Cognitive Skills in Wage Determination," *The Review of Economics and Statistics*, 77(2), 251–66. 3

Nordin, M. (2008): "Ability and Rates of Return to Schooling - Making Use of the Swedish Enlistment Battery Test," *Journal of Population Economics*, 21(3), 703–717. 3

Nybom, M., and J. Stuhler (2011): "Heterogeneous Income Profiles and Life-Cycle Bias in Intergenerational Mobility Estimation," *IZA Discussion Paper No. 5697*. 12

Nyhus, E., and E. Pons (2005): "The Effects of Personality on Earnings," *Journal of Economic Psychology*, 26(3), 363–384. 3

Öckert, B. (2010): "What's the Value of an Acceptance Letter? Using Admissions Data to Estimate the Return to College," *Economics of Education Review*, 29(4), 504–516. 30

——— (2012): "On the Margin of Success? Effects of Expanding Higher Education for Marginal Students," *Nordic Economic Policy Review*, (1), 111–158. 3

Roy, A. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3(2), 135–146. 7

Taber, C. (2001): "The Rising College Premium in the Eighties: Return to College or Return to Unobserved Ability?," *The Review of Economic Studies*, 68(3), 665. 2

Willis, R., and S. Rosen (1979): "Education and Self-Selection," *Journal of Political Economy*, 87(5), 7–36. 1, 7, 8, 11, 18, 27

Zax, J., and D. Rees (2002): "IQ, Academic Performance, Environment, and Earnings," *Review of Economics and Statistics*, 84(4), 600–616. 3

Appendix

A.1 Definitions of Weights for ATE, ATT, ATU

Under the LIV approach (and the parametric), all treatment parameters of concern can be identified by using weighted averages of MTE. Heckman et al. (2006b) show that

$$\begin{aligned}
\text{ATE}(\mathbf{x},\mathbf{a}) &= E\left[B \mid \mathrm{X}=\mathbf{x},\mathbf{A}=\mathbf{a}\right] = \int_0^1 \text{MTE}(\mathbf{x},\mathbf{a},u_S) w_{ATE}(\mathbf{x},\mathbf{a},u_S) du_S \\
\text{ATT}(\mathbf{x},\mathbf{a}) &= E\left[B \mid \mathrm{X}=\mathbf{x},\mathbf{A}=\mathbf{a},S=1\right] = \int_0^1 \text{MTE}(\mathbf{x},\mathbf{a},u_S) w_{ATT}(\mathbf{x},\mathbf{a},u_S) du_S \\
\text{ATU}(\mathbf{x},\mathbf{a}) &= E\left[B \mid \mathrm{X}=\mathbf{x},\mathbf{A}=\mathbf{a},S=0\right] = \int_0^1 \text{MTE}(\mathbf{x},\mathbf{a},u_S) w_{ATU}(\mathbf{x},\mathbf{a},u_S) du_S,
\end{aligned}$$

where the weights are

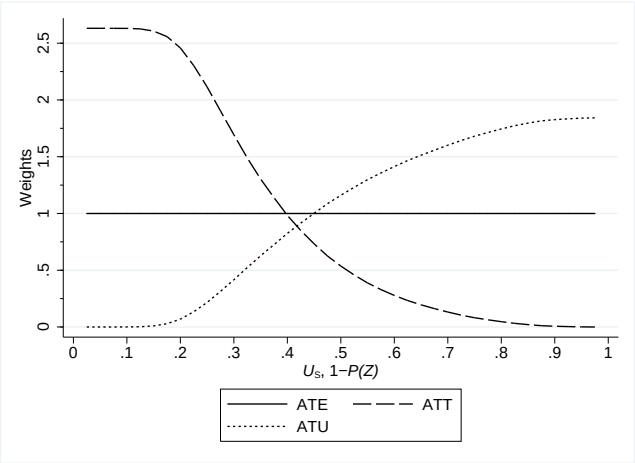
$$\begin{aligned}
w_{ATE}(\mathbf{x},\mathbf{a},u_S) &= 1 \\
w_{ATT}(\mathbf{x},\mathbf{a},u_S) &= \frac{\int_{u_S}^1 f(P(\mathbf{Z})=P(\mathbf{z}) \mid \mathbf{X}=\mathbf{x},\mathbf{A}=\mathbf{a}) dP(\mathbf{z})}{E\left[P(\mathbf{Z}) \mid \mathbf{X}=\mathbf{x},\mathbf{A}=\mathbf{a}\right]} \\
w_{ATU}(\mathbf{x},\mathbf{a},u_S) &= \frac{\int_0^{u_S} f(P(\mathbf{Z})=P(\mathbf{z}) \mid \mathbf{X}=\mathbf{x},\mathbf{A}=\mathbf{a}) dP(\mathbf{z})}{E\left[1-P(\mathbf{Z}) \mid \mathbf{X}=\mathbf{x},\mathbf{A}=\mathbf{a}\right]},
\end{aligned}$$

and f is the density function of $P(\mathbf{Z})$. By integrating the weighted estimates of $\text{MTE}(\mathbf{x},\mathbf{a},u_S)$ over the joint distribution of $(\mathbf{X},\mathbf{A})$ the estimates of $\text{MTE}(u_S)$ are obtained. In practice, however, I do not condition on $(\mathbf{X},\mathbf{A})$ nonparametericallly. Instead, I follow Basu et al. (2007) and others and condition on, and thus also integrate over, demideciles of the (estimated) scalar index $\mathbf{X}(\delta_1-\delta_0)+\mathbf{A}(\gamma_1-\gamma_0)$. Lastly, integrating over $P(\mathbf{z})$ gives the unconditional estimates of ATE, ATT and ATU. The ability-specific ATEs are obtained by evaluating and comparing the treatment parameters at different values of $\mathbf{A}=\mathbf{a}$.

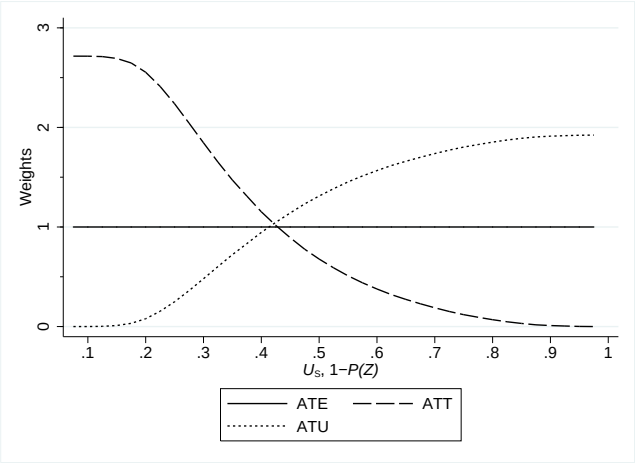
A.2 Computed Weights for ATE, ATT, ATU

Figure A. 1 Sample Weights for Different Treatment Parameters

(a) Normal Selection Model



(b) Semiparametric Model



Note: The figures show the weights for ATE, ATT, and ATU, plotted across U_S, for the normal selection model and the semiparametric model. The weights are defined in Appendix A.1.