Estimating Marginal and Average Returns to Education

(Author list removed for review process)

November 3, 2006

Abstract

This paper estimates marginal and average returns to college when returns vary in the population and people sort into college with at least partial knowledge of their returns. Different instruments identify different parameters which do not, in general, answer well-posed economic questions or identify traditional treatment effects. Recent developments in the instrumental variables literature enable analysts to identify returns at the margin for an unidentified margin. We apply recent extensions of the instrumental variables literature to estimate marginal and average returns at clearly identified margins and to construct policy relevant parameters. We find that marginal entrants earn substantially less than average college students, that comparative advantage is a central feature of modern labor markets and that ability bias is an empirically important phenomenon.

JEL Code: J31

Key words: Returns to Schooling; Marginal Return; Average Return; Marginal Treatment Effect

1 Introduction

Returns at the margin are central to economic analysis. So is the contrast between average returns and marginal returns which determine economic rents and profitability. It is thus surprising that so few empirical papers distinguish marginal and average returns. The marginal returns that are reported in the recent instrumental variables literature are at unidentified margins, making it difficult to use the estimates in policy or welfare analysis, or to compare marginal returns across different studies.

This paper applies a framework developed by Heckman and Vytlacil (1999, 2001a, 2005, 2007) and Heckman, Urzua and Vytlacil (2006) to identify both marginal and average returns at well defined margins of choice. We estimate the returns to college for persons at the margin of attending college, as well as the average return of those who go to college, and what the return would be for those who do not go to college.

Consider a simple model of schooling and earnings. In the standard regression model, log earnings, $\ln Y$, are written as a function of schooling S,

$$\ln Y = \alpha + \beta S + U, \tag{1}$$

where α, β are parameters and S is correlated with mean zero error U, the least squares estimators of β are biased and inconsistent. S can stand for any treatment and $\ln Y$ for any outcome. We assume that S is binary valued ($S = 0$ or 1). If an instrument Z can be found so that (a) Z is correlated with S but (b) it is not correlated with U, β can be identified, at least in large samples. This is the most commonly used method of estimating β. Valid social experiments and valid natural experiments can be interpreted as generating instrumental variables.

The standard model makes very strong assumptions. In particular, it assumes that the (causal) effect of S on $\ln Y$ is the same for everyone, so the marginal return is the average return. However, if β varies in the population and people sort into economic sectors on the basis of at least partial knowledge of β, then the marginal β will usually be different from the average β. In this case, there is no single "effect"of S on $\ln Y$. Different policies affect different sections of the distribution of β, and their evaluation requires estimating different parameters. Furthermore, different estimators

produce different scalar summary measures of the distribution of β.

An important paper by Imbens and Angrist (1994) gives conditions under which instrumental variables identify returns to S for persons induced to change their schooling status by the instrument. Card (1999, 2001) interprets their analysis as identifying marginal returns. As noted by Heckman (1996), the actual margin of choice is not identified by the instrument and it is unclear as to which segment of the population the estimated return applies.

This paper shows how to identify marginal returns at explicitly identified margins when β varies in the population and is correlated with S because schooling choices depend on β. Our analysis is based on the Marginal Treatment Effect (MTE), introduced in Björklund and Moffitt (1987) and extended in Heckman and Vytlacil (1999, 2001a, 2005, 2007). The MTE is the mean return to schooling for persons at the margin of indifference between taking treatment or not. We use this parameter to describe heterogeneity in returns; to construct estimates of clearly defined marginal and average returns; to construct policy relevant parameters; to characterize what parameters different instruments estimate; and to identify marginal returns for the entire population. We apply the method of local instrumental variables introduced in Heckman and Vytlacil (1999) to estimate the Björklund-Moffitt parameter.

We contribute to the literature in the following ways:

1. We overcome a problem that plagues the recent literature that estimates marginal returns for persons at unidentified margins. We show how to unify diverse instrumental variables estimates and to determine what margins they identify. Instead of reporting a marginal return for unidentified persons, we report marginal returns for all persons identified by a latent variable that arises from a well defined choice model and is related to the propensity of persons to attend college.

2. We use our framework to interpret the margins of choice identified by various instruments and to place disparate instruments on a common interpretive footing.

3. We document the empirical importance of heterogeneity in the returns to college in the US. Our analysis relaxes the normality assumptions of Willis and Rosen (1979) and estimates the marginal and average returns for their schooling choice model. We show that comparative

advantage and self-selection are empirically important features of schooling choice: marginal college attendees have lower returns than average attendees and the fall off in their returns is sharp.

4. We estimate economically interpretable measures of the return to schooling. In particular, we estimate the return to college for those individuals induced to enroll in college by a specific policy (college construction), which we call the Policy Relevant Treatment Effect (PRTE) (Heckman and Vytlacil, 2001b). We also estimate the Average Marginal Treatment Effect (AMTE), the average return for the set of individuals at the margin of enrolling in college. We distinguish this from the "marginal return"estimated from IV, and from standard parameters such as the Average Treatment Effect or Treatment on the Treated.

5. We characterize the parameter estimated by the instrumental variables method using an interpretable economic model. For our data, OLS and conventional IV estimators substantially underestimate the policy relevant return to schooling. We clarify the meaning of the OLS-IV comparison, which is widely used in the literature. We establish that this is generally an economically meaningless comparison because neither OLS nor IV estimate well defined average or marginal returns. We use recently developed tools to establish that the average marginal individual has a lower return to college than the average individual who attends school, while the IV estimate of this return exceeds the OLS estimate because of selection bias.

The plan of the paper is as follows. The next section presents a short review of the empirical literature on the returns to schooling, and highlights its main findings. Then we present the empirical framework which we use for the rest of the paper. We then discuss the estimation method and present our empirical estimates. We start with standard linear instrumental variables estimates of the return to college, and estimate semiparametric and normal selection models. Using these models we construct marginal returns and policy relevant returns. The last section concludes.

2 Instrumental Variables Estimates of the Returns to Schooling

Instrumental variables (IV) estimates of the return to a year of schooling vary widely across studies. Card (2001) reports values that range from 3.6% to 16.4%. Most of these estimates come from the following specification of the earnings function:

$$\ln Y = \alpha + \beta S + X\gamma + U, \tag{2}$$

where $\ln Y$ is log hourly wage, S is years of schooling, and X is a vector of other controls, which includes polynomials in age or experience, and sometimes also includes test scores, family background, cohort dummies, or regional controls.

Some of the estimates reported in Card (2001) differ because they are estimated from different time periods or economic environments and returns to schooling have increased over time. However, even when we restrict our attention to estimates based on recent data, there exists widespread variation among them. For example, using the 1980 Census, Angrist and Krueger (1991) and Staiger and Stock (1997) produce IV estimates that range from 6% to 10% (the OLS range is 5–6%). Using NLS data from the 1970s, Card's (1995) IV estimates are between 9.5% and 13.2% (OLS is 7.3%), while the estimate in Kane and Rouse (1995) (based on the NLS Class of 1972) is 9.4% (OLS is 6.3%). Using more recent data from the NLSY79, Cameron and Taber's (2004) IV estimates range from 5.7% to 22.8% across different specifications (OLS is about 6%).[1]

Why might this be? One explanation is that returns are heterogeneous.[2] Take the model in (1) and assume that β is a variable coefficient, so that $\beta = \bar{\beta} + \varepsilon$, where $\bar{\beta}$ is the mean of β. In this case (keeping X implicit):

$$\ln Y = \alpha + \bar{\beta} S + \varepsilon S + U.$$

Unlike the case where $\varepsilon = 0$, or ε is distributed independently of S, finding an instrument Z

[1] Kling's (2001) estimate from the same dataset is 46%, although it is very imprecisely estimated.

[2] See Heckman and Robb (1985, 1986), Imbens and Angrist (1994), Card (1999, 2001), Heckman and Vytlacil (2001a, 2005), and Heckman, Urzua and Vytlacil (2006).

correlated with S but not U or ε is not enough to identify $\bar{\beta}$. Simple algebra shows that

$$\operatorname{plim} \hat{\beta}_{\mathrm{IV}}^{Z} = \frac{\operatorname{Cov}(Z, \ln Y)}{\operatorname{Cov}(Z, S)} = \bar{\beta} + \frac{\operatorname{Cov}(Z, U)}{\operatorname{Cov}(Z, S)} + \frac{\operatorname{Cov}(Z, S\varepsilon)}{\operatorname{Cov}(Z, S)}.$$

If $\operatorname{Cov}(Z, U) = 0$ (the standard IV condition), the second term in the final expression vanishes. In general the third term does not vanish, unless $\varepsilon \equiv 0$ (a common coefficient model), or if ε is independent of S and Z.[3] But in general ε is dependent on S and the term does not vanish.

Notice that if we use another instrument $W \neq Z$ (with $\operatorname{Cov}(W, U) = 0$ and $\operatorname{Cov}(W, \varepsilon) = 0$),

$$\operatorname{plim} \hat{\beta}_{\mathrm{IV}}^{W} = \frac{\operatorname{Cov}(W, \ln Y)}{\operatorname{Cov}(W, S)} = \bar{\beta} + \frac{\operatorname{Cov}(W, S\varepsilon)}{\operatorname{Cov}(W, S)}.$$

Only by coincidence will $\operatorname{plim} \hat{\beta}_{\mathrm{IV}}^{Z} = \operatorname{plim} \hat{\beta}_{\mathrm{IV}}^{W}$. Therefore, IV estimates of β may vary across studies just because the instrumental variables used are different in each study: Angrist and Krueger (1991) and Staiger and Stock (1997) use as IV quarter of birth interacted with state and year of birth; Card (1995) uses the availability of a college in the SMSA of residence in 1966; Kane and Rouse (1995) use tuition at 2 and 4 year state colleges and distance to the nearest college; Cameron and Taber (2004) use an indicator for the presence of a college in the SMSA of residence at age 17, and local earnings in the SMSA of residence at age 17. The estimates in these papers do not identify the same parameter so there is no reason why they should be equal to each other.[4]

Furthermore, as emphasized by Card (1999, 2001), OLS estimates of the return to schooling are generally below IV estimates of the same parameter. This is inconsistent with a fixed coefficient model ($\varepsilon \equiv 0$) with positive ability bias ($\operatorname{Cov}(S, U) > 0$), but can be rationalized with a model of heterogeneous returns. There is no obvious ordering between OLS and IV.[5]

In this paper we use NLSY data to show that different instruments produce different estimates,

[3] If $U_1 - U_0$ is independent of Z and if $U_1 - U_0$ does not determine S conditional on Z, then $U_1 - U_0$ will be independent of (S, Z).

[4] In fact, the argument that instrumental variables estimates based on the presence of a college in the SMSA of residence at age 17 or local earnings in the SMSA of residence at age 17 lead to estimates of different parameters is a central argument in Cameron and Taber's (2004) study. These authors claim that, in the presence of credit constraints these instruments affect different groups of the population, and the first estimate should be higher than the second.

[5] See Carneiro and Heckman (2002). Another possible reason why IV estimates may exceed OLS estimates is measurement error. However, as argued in Card (1999), schooling is relatively well measured in the US, and the large discrepancies between OLS and IV estimates of the returns to schooling are unlikely to be explained by measurement error.

which are generally above OLS estimates. We show empirically that the returns to schooling vary across individuals, implying that the data underlying the instrumental variables estimates we present comes from a model of heterogeneous returns to schooling. We contrast instrumental variables estimates with estimates of the average and marginal returns to schooling, and explain why OLS is below IV. We use a simple economic model to show how to place all instruments on a common footing, identifying returns at clearly specified margins of choice.

3 Model with Heterogeneous Returns to Schooling

Using the framework employed in Heckman and Vytlacil (2001a, 2005, 2007), let $\ln Y_1$ be the potential log wage of an individual as a college attendee, and $\ln Y_0$ be the potential log wage of an individual as a high school graduate. Then we can write:

$$\ln Y_1 = \mu_1(X) + U_1 \text{ and } \ln Y_0 = \mu_0(X) + U_0, \tag{3}$$

where $\mu_1(X) \equiv E(Y_1 \mid X)$ and $\mu_0(X) \equiv E(Y_0 \mid X)$. The return to schooling is $\ln Y_1 - \ln Y_0 = \beta = \mu_1(X) - \mu_0(X) + U_1 - U_0$, so that the average treatment effect conditional on $X = x$ is given by $\bar{\beta}(x) = E(\beta \mid X = x) = \mu_1(x) - \mu_0(x)$ and the effect of treatment on the treated conditional on $X = x$ is given by $E(\beta \mid X = x, S = 1) = \bar{\beta}(x) + E(U_1 - U_0 \mid S = 1, X = x)$.[6]

A standard latent variable model determines enrollment in school:

$$\begin{aligned} S^* &= \mu_S(Z) - V, \\ S &= 1 \text{ if } S^* \geq 0. \end{aligned} \tag{4}$$

A person goes to school ($S = 1$) if $S^* \geq 0$. Otherwise $S = 0$. In this notation, (Z, X) are observed and (U_1, U_0, V) are unobserved. V is assumed to be a continuous random variable with a strictly increasing distribution function F_V. V may depend on U_1 and U_0 in a general way. The Z vector may include some or all of the components of X. We assume that (U_0, U_1, V) is independent of Z conditional on X. Appendix A presents an economic model of schooling where individuals choose

[6] Heckman and Vytlacil (1999, 2001a, 2005, 2007) develop their results for a general nonseparable model: $\ln Y_1 = \mu(X, U_1)$ and $\ln Y_0 = \mu(X, U_0)$. They do not assume that $X \perp\!\!\!\perp (U_0, U_1)$ so X may be correlated with the unobservables in potential outcomes.

the level of schooling which maximizes their present value of earnings net of costs, and which is consistent with this representation of the choice model. This model captures the framework of Willis and Rosen (1979).

Let $P(z)$ denote the probability of receiving treatment $S = 1$ conditional on $Z = z$, $P(z) \equiv \Pr(S = 1|Z = z) = F_V(\mu_S(z))$, where we keep the conditioning on X implicit. Define $U_S = F_V(V)$.[7] We can rewrite (4) using $F_V(\mu_S(Z)) = P(Z)$ so that $S = 1$ if $P(Z) \geq U_S$. $P(Z)$ is the mean scale utility function in discrete choice theory (McFadden, 1974).

The marginal treatment effect (MTE), defined by

$$\Delta^{\text{MTE}}(x, u_S) \equiv E(\beta \mid X = x, U_S = u_S)$$

is central to our analysis. This parameter was introduced into the literature by Björklund and Moffitt (1987) and extended in Heckman and Vytlacil (1999, 2001a, 2005, 2007). It is the mean gain to schooling for individuals with characteristics $X = x$ and $U_S = u_S$. Equivalently, it is the mean return to schooling for persons indifferent between going to college or not who have mean scale utility value u_S.

The MTE has two advantages. First, by showing us how the return to college varies with X and U_S, the MTE is a natural way to characterize heterogeneity in returns and the marginal returns to school for persons at the margin at all values of U_S instead of just an unknown range of U_S selected by one instrument as in LATE. By estimating MTE for all values of U_S, we can identify the returns at all margins of choice. Using this parameter, not only can we examine how wide is the dispersion in returns, but we can also relate it to observed and unobserved variables that determine college enrollment. This allows us to understand how individuals sort into different levels of schooling.

Second, Heckman and Vytlacil (1999, 2001a, 2005, 2007) establish that all of the conventional treatment parameters are different weighted averages of the MTE where the weights integrate to one. See Table 1A for the treatment parameters expressed in terms of MTE and Table 1B for the weights. If β is a constant conditional on X or more generally if $E(\beta \mid X = x, U_S = u_S) = E(\beta \mid X = x)$, ($\beta$ mean independent of U_S and conditional on X), then all of these mean treatment parameters conditional on X are the same. Different instruments weight MTE differently. We can characterize

[7] This is a uniform random variable.

those weights and thus can compare the instruments. The MTE unifies all the parameters in the treatment effect literature.

One parameter of particular interest is the Policy Relevant Treatment Effect (PRTE), introduced in the literature by Heckman and Vytlacil (2001b). For example, if a policy consists in the construction of colleges in all counties, then this parameter corresponds to the average return to schooling for individuals induced to enroll in college by college construction. Only by accident do the traditional evaluation parameters such as the average treatment effect or the mean effect of treatment on the treated answers this question. In general, the instrumental variables estimate of the return to schooling does not answer this question. As shown in Heckman and Vytlacil (2001b) (and Appendix B), this question is better answered by finding the corresponding weights and using them to construct the appropriate weighted average of the MTE, where the weights are given in tables 1A and 1B.

A related parameter is the Average Marginal Treatment Effect (AMTE), which can be defined in different ways as shown in Appendix B. In our application, the AMTE corresponds to the average return to schooling for individuals induced to enroll in college by marginal changes in a policy variable, so that we define the AMTE as a particular limit version of the PRTE. The AMTE is the return to schooling for the average marginal person, a concept of central importance in our paper and in economics.

Tables 1A and 1B also show how the OLS and IV estimates of the return to schooling can be expressed as weighted averages of the MTE. We present IV weights for the general case where we use $J(Z)$ as the instrument, where $J(.)$ is a function of Z (see Heckman and Vytlacil, 2005, and Heckman, Urzua and Vytlacil, 2006).

4 Using Local Instrumental Variables to Estimate the MTE

There are several ways to construct the MTE. For example, if we impose parametric assumptions on the joint distribution of (U_1, U_0, V) we can derive the implied expression for the MTE (see Heckman, Tobias and Vytlacil, 2001). However, it is also possible to nonparametrically estimate the MTE using the method of local instrumental variables, developed in Heckman and Vytlacil (1999, 2000, 2001a).

Take the model in equation (3) and keep X implicit. Then we can write:

$$\ln Y_0 = \alpha + U_0, \tag{5a}$$

$$\ln Y_1 = \alpha + \bar{\beta} + U_1, \tag{5b}$$

where $E(U_0) = 0$ and $E(U_1) = 0$ so $E(\ln Y_0) = \alpha$, $E(\ln Y_1) = \alpha + \bar{\beta}$, and $\beta = \bar{\beta} + U_1 - U_0$. Observed earnings are

$$\ln Y = S \ln Y_1 + (1 - S) \ln Y_0 = \alpha + \beta S + U_0 = \alpha + \bar{\beta} S + \{U_0 + S(U_1 - U_0)\}. \tag{6}$$

Using equation (6), the conditional expectation of $\ln Y$ given $P(Z) = p$ is

$$E(\ln Y \mid P(Z) = p) = E(\ln Y_0 \mid P(Z) = p) + E(\ln Y_1 - \ln Y_0 \mid S = 1, P(Z) = p)\, p,$$

where we keep the conditioning on X implicit. Heckman and Vytlacil (2001a, 2005, 2007) show one representation of $E(\ln Y \mid P(Z) = p)$ that reveals the underlying index structure:

$$E(\ln Y \mid P(Z) = p) = \alpha + \bar{\beta} p + \int_{-\infty}^{\infty} \int_{0}^{p} (U_1 - U_0) f(U_1 - U_0 \mid U_S = u_S)\, du_S\, d(U_1 - U_0),$$

where for ease of exposition we assume that $(U_1 - U_0, U_S)$ has a density.[8] Differentiating with respect to p, we obtain MTE:

$$\begin{aligned} \frac{\partial E(\ln Y \mid P(Z) = p)}{\partial p} &= \bar{\beta} + \int_{-\infty}^{\infty} (U_1 - U_0) f(U_1 - U_0 \mid U_S = p)\, d(U_1 - U_0) \\ &= \Delta^{\text{MTE}}(p). \end{aligned}$$

Thus we can recover the return to S for persons indifferent at all margins of U_S within the empirical support of $P(Z)$. Notice that persons with a high mean scale utility function $P(Z)$ identify the return for those with a high value of U_S, i.e., a value of U_S that makes persons *less* likely to participate in schooling. The high $P(Z)$ is required to offset the high U_S and induce people to

[8] More precisely, a density with respect to Lebesgue measure. The result holds more generally. See Heckman and Vytlacil (2001a, 2005, 2007) and Heckman, Urzua and Vytlacil (2006).

attend school.

IV estimates $\bar{\beta}$ if $\Delta^{\text{MTE}}(u_S)$ does not vary with u_S. Under this condition, $E(\ln Y \mid P(Z) = p)$ is a linear function of p. Under our assumptions, a test of the linearity of the conditional expectation of $\ln Y$ in p is a test of the validity of linear IV for $\bar{\beta}$, or a test of selection on returns. This test is simple to execute and interpret and we apply it below.

LIV is an instrumental variables method where we use $P(Z)$ as the instrument and we allow it to affect the outcome in a nonparametric way. We focus on $E(\ln Y|P(Z) = p)$ and differentiate this conditional expectation to obtain MTE. We could also have considered $E(\ln Y|Z)$ or $E(\ln Y|Z_k)$ where Z_k is the kth component of Z. However, conditioning on $P(Z)$ instead Z has several advantages. By examining derivatives of $E(\ln Y|P(Z) = p)$, we are able to identify the MTE function for a broader range of values than would be possible by examining derivatives of $E(\ln Y|Z = z)$ while removing the ambiguity of which element of Z to vary. Also, by connecting the MTE to $E(\ln Y|P(Z) = p)$, we are able to exploit the structure on $P(Z)$ when making out of sample forecasts. If Z_1 is a component of Z that is associated with a policy, but has limited support, we can simulate the effect of a new policy that extends the support of Z_1 beyond historically recorded levels by varying the other elements of Z.[9] See Heckman (2001) and Heckman and Vytlacil (2001b, 2005, 2007).

It is straightforward to estimate the levels and derivatives of $E(\ln Y \mid P(Z) = p)$ and standard errors using the methods developed in Heckman, Ichimura, Smith and Todd (1998a). Software for doing so is presented at the website for Heckman, Urzua and Vytlacil (2006).The derivative estimator of MTE is the LIV estimator of Heckman and Vytlacil (1999, 2001a).

5 Estimating the MTE and Comparing Treatment Parameters, Policy Relevant Parameters and IV Estimands

This section reports estimates of the MTE using a sample of white males from the National Longitudinal Survey of Youth. The data are described in Appendix C. Our estimates are based on data from the National Longitudinal Survey of Youth of 1979 (NLSY). We measure wages, years

[9] Thus if $\mu(Z) = Z\gamma$, we can use the variation in the other components of Z to substitute for the missing variation in Z_1 given identification of the γ up to a common scale. See Heckman and Vytlacil (2005)

of experience and college participation in 1994. Our instruments for schooling include the presence of a four year public college in the SMSA of residence at age 14, log average earnings in the SMSA of residence at age 17, and the average unemployment rate in the state of residence at age 17 (as used, for example, in Card, 1995; Currie and Moretti, 2003; Kane and Rouse, 1995; Kling, 2001, and Cameron and Taber, 2004). The set of controls we use consists of a measure of cognitive ability (AFQT), maternal education, years of experience in 1994, cohort dummies, log average earnings in the SMSA of residence in 1994, and the average unemployment rate in the state of residence in 1994. We present a test for the validity of our exclusion restrictions. In our data set there are 711 high school graduates who never attended college and 903 individuals who attended any type of college.[10] Table 2 documents that individuals who attend college have on average a 34% higher wage than those who do not attend college. They also have two and a half a years less of work experience since they spend more time in school. The scores on a measure of cognitive ability, the Armed Forces Qualifying Test (AFQT), are much higher for individuals who attend college than for those who do not.[11] Those who only attend high school have less educated mothers than individuals who attend college. They also spent their adolescence in counties less likely to have a college. Local labor market variables at 17 are not much different between these two groups of individuals. The wage equations include, as X variables, experience, schooling-adjusted AFQT, mother's education, cohort dummies, log average earnings in the SMSA of residence in 1994 and local unemployment rate in the state of residence in 1994. Our exclusion restrictions (variables in Z not in X) are distance to college, local earnings in the SMSA of residence at 17 and the local unemployment rate in the state of residence at age 17.[12] We include all X variables in Z except work experience, local wages and unemployment in the year the wage outcome we use is measured. These variables are realized after the schooling decision is made.

[10] These are white males, in 1994, with either a high school degree or above and with a valid wage observation, as described in Appendix C. We use as a measure of wage the average of all nonmissing wages reported in 1992, 1993, 1994 and 1996.

[11] We use a measure of this score corrected for the effect of schooling attained by the participant at the date of the test, since at the date the test was taken, in 1981, different individuals have different amounts of schooling and the effect of schooling on AFQT scores is important. We use a version of the nonparametric method developed in Hansen, Heckman and Mullen (2004). We perform this correction for all demographic groups in the population and then standardize the AFQT to have mean 0 and variance 1. See Table A1.

[12] We have constructed both SMSA and state measures of unemployment, but our state measure has better predictive power for schooling (perhaps because of less measurement error), and therefore we choose to use it instead of SMSA unemployment.

The instrumental variables we use for identification of the model (exclusion restrictions) are intended to measure different costs of attending college and are based on the geographic location of individuals in their late adolescence. If the decision to go to college and the (prior) location decision are correlated, then our instruments may not be valid if unobserved determinants of location are correlated with wages. Individuals who are more likely to enroll in college may choose to locate in areas where colleges are abundant. These locations may have higher wages. In our wage equations, we control for measured ability, mother's years of schooling, and current local labor market conditions. Our identifying assumption is that the instruments are valid conditional on measured ability, mother's education, and current local labor market conditions, which are also correlated with location choice at age 17.

Distance to college was first used as an instrument for schooling by Card (1995) and was subsequently used by Kane and Rouse (1995), Kling (2001), Currie and Moretti (2003) and Cameron and Taber (2004). Cameron and Taber (2004) and Carneiro and Heckman (2002) show that distance to college in the NLSY79 is correlated with a measure of ability (AFQT), but in this paper we include this measure of ability in the outcome equation.

Local labor market variables have also been used by Cameron and Heckman (1998, 2001) and Cameron and Taber (2004). If local unemployment and local earnings of unskilled workers at age 17 are correlated with the unobservable in the earnings equation, our measures of local labor market conditions would be invalid instruments. To mitigate this concern, in our outcome equations we include the SMSA of residence average log earnings, and the state of residence unemployment rate in the year in which wages are measured. As argued in Cameron and Taber (2004), local labor market conditions can influence schooling through two possible channels. On the one hand, better labor market conditions for the unskilled increase the opportunity costs of schooling, and reduce educational attainment. On the other hand, better labor market conditions can lead to an increase in the resources of credit constrained households, and therefore to an increase in educational attainment. Therefore, the sign of the total impact of these variables on schooling is theoretically ambiguous.

We use a logit model for schooling choice to construct $P(Z)$. The Z include AFQT and its square, mother's education and its square, an interaction between mother's education and AFQT, cohort

dummies, the presence of a college at age 14, local unskilled earnings and local unemployment at age 17, and interactions of these last three variables with AFQT and its square, mother's education and its square and an interaction between AFQT and mother's education. The specification is quite flexible, and alternative functional form specifications for the choice model produce similar results to the ones reported in this paper. Under standard conditions, the distribution of U_S can be estimated nonparametrically up to scale so our results do not (in principle) depend on arbitrary functional form assumptions about unobservables. Table 3 gives estimates of the average marginal derivatives of each variable in the choice model. The instruments are strong predictors of schooling, as are mother's education and AFQT (we present a test at the bottom of Table 3).

Our instruments predict college attendance and are assumed to be uncorrelated with the unobservables in the wage equation. We report an indirect test of the validity of our instruments in Appendix Table A2. Using the high school transcript data available in the NLSY, we regress the percentage of high school subjects in which each student achieved a grade of A, and the percentage of high school subjects in which each student achieved a grade of B or above, on college attendance and the AFQT (adjusted for schooling at time of test). We also include, as additional controls, mother's education, cohort dummies, local earnings and local unemployment in 1994 (the exact specification is at the base of the table). The OLS estimates show a strong relationship between college participation and both measures of high school grades (see columns 1 and 3). Since college follows high school, the only mechanism for producing this effect is some unobserved motivational or ability variable not captured by the AFQT. This unobserved variable may also appear in the wage equation.

Using $P(Z)$ as an instrument for college attendance eliminates the spurious college-attendance raising-high-school-grades relationship (see columns 2 and 4), while the relationship between AFQT and high school grades becomes stronger. Thus, to the extent that the unobservable in this relationship is in the error of our log wage equation, we can feel confident that our IV has eliminated this source of bias. This gives us further confidence in our instrument but of course does not prove that it is uncorrelated with the errors in the wage equation.

The support of the estimated $P(Z)$ is shown in Figure 1 and it is almost the full unit interval, which may be surprising. Formally, for nonparametric analysis, we need to determine the support

of $P(Z)$ conditional on X. However, if we are willing to assume separability and independence between X and the unobservables of the model, then we do not need to condition on X but we can include all of X, previously described, as components of Z in generating the support of $P(Z)$. We discuss this point further in section 5.2 below.

5.1 Standard IV estimates of the Return to College

Before we proceed to estimate the MTE and several other parameters, we start by presenting standard least squares and linear instrumental variables estimates of the return to college. We start by estimating the following model:

$$\ln Y = \alpha + \beta S + \mu(X) + U, \tag{7}$$

where $\mu(X)$ includes years of experience and its square, AFQT and its square, mother's education and its square, an interaction between mother's education and AFQT, cohort dummies, local earnings in 1994 and local unemployment in 1994. The first column of Table 4 presents the OLS estimate of β, and columns 2 through 6 display linear IV estimates of β using different instruments (distance to college, local earnings at 17, local unemployment at 17, all of them simultaneously, and $P(z)$).[13] The OLS estimate is below standard estimates reported in the literature, perhaps because of the larger than usual set of controls we are able to include in the model, while the IV estimate is above standard IV estimates reported in the literature. As in most of the literature, the IV estimate exceeds the OLS estimate. Card (1999) reports IV estimates as high as 16% from the literature, and in Card (1995), using data from the late 1970s, IV estimates range from 9% to 13%. Our estimates are larger than these, but we conjecture that one possible reason is an increase in the return to schooling during the 1980s and 1990s. Other reasons may be that we are using different instruments, or that we are considering only the returns to college (whereas most of the literature considers a continuous measure of schooling).

The instrumental variables estimates vary considerably depending on which instrument is used,

[13]Except for $P(Z)$, we allow the effect of the instruments to vary with X, by interacting the instruments with AFQT and its square, maternal education and its square, and the interaction of these two variables. The resulting estimates are presented in columns 2-5 of table 4. If we do not include these interactions the standard errors of the IV estimates increases and their point estimates become respectively (from column 2 to 5): 0.1585, 0.2105, 1.3577, and 0.2102

suggesting that β is heterogeneous and that β is correlated with S (below we present a formal test of this hypothesis). In the last two columns of Table 4 we allow the return to college to vary with X by adding an interaction between $\mu(X)$ and S in equation (7). We evaluate our estimates at the average value of X in our sample. The IV estimate exceeds the OLS estimate of the return to college. We also report the average effect of AFQT on the return to college. In the IV specification, this effect is large and statistically strong, which is consistent with previous results in the literature (see Blackburn and Neumark (1993), Bishop (1991) or Grogger and Eide (1995)).

The results in this section suggest that β varies across individuals, that β is correlated with S, and that β varies with X (in particular, AFQT). We now proceed to characterize how β varies with unobservables (and observables), and we identify the marginal return to schooling at well defined margins of choice.

5.2 Estimating the MTE using LIV

Constructing the MTE requires estimating $E(Y|X, P(Z) = p)$ and then computing its derivative with respect to $P(Z)$. However, fully nonparametric estimation of the derivatives of $E(\ln Y|X, P(Z))$ is not feasible due to the curse of dimensionality that plagues nonparametric statistics. We impose additional structure on the model that produces a feasible semiparametric estimation problem. In particular, we assume separability between X and U_1 and U_0 in the outcome equations. We specify

$$\beta = \mu_1(X) - \mu_0(X) + U_1 - U_0,$$

where $\mu_1(X)$ and $\mu_0(X)$ are functions of X with parameters β_1 and β_0 respectively (e.g., $\mu_1(X) = X\beta_1$ and $\mu_0(X) = X\beta_0$). The exact functional form of $\mu_1(X)$ includes linear and quadratic terms in years of experience, AFQT and mother's education, an interaction between AFQT and mother's education, cohort dummies, log average earnings in the SMSA of residence and the average unemployment rate in the state of residence. The same functional form is used for $\mu_0(X)$.

The outcome equation can be written as

$$\ln Y = \mu_0(X) + S[\mu_1(X) - \mu_0(X)] + U_0 + S(U_1 - U_0), \tag{8}$$

with (U_0, U_1, U_S) independent of (X, Z).[14] Combining the model for S with the model for Y implies a partially linear model for the conditional expectation of Y:

$$E(\ln Y \mid X, P(Z)) = \mu_0(X) + P(Z)\left[\mu_1(X) - \mu_0(X)\right] + K(P(Z)), \tag{9}$$

where

$$K(P(Z)) = E(U_1 - U_0 \mid P(Z), S = 1)P(Z) = E\left(U_1 - U_0 \mid U_S \leq P(Z)\right)P(Z).$$

No parametric assumption is imposed on the distribution of (U_0, U_1). Thus $K(\cdot)$ is an unknown function that must be estimated nonparametrically. In general, unless $P(Z)$ has full support in the unit interval, it is not possible to separately identify the intercepts of $\mu_0(X)$, $\mu_1(X)$ and the intercept of the function $K(P)$. However the MTE can still be identified at U_S evaluation points within the support of $P(Z)$ since

$$\begin{aligned}\Delta^{\text{MTE}}(x,p) &= \left.\frac{\partial E\left\{\ln Y \mid X, P(Z)\right\}}{\partial P(Z)}\right|_{P(Z)=p} \\ &= \mu_1(X) - \mu_0(X) + E\left(U_1 - U_0 | U_S = p\right).\end{aligned}$$

Equation (9) suggests that $\mu_1(X)$ and $\mu_0(X)$ can be estimated by a partially linear regression of $\ln Y$ on X and $P(Z)$. Since $P(Z)$ is unobserved, we proceed in two steps. The first step is construction of the estimated $P(Z)$ and the second step is estimation of β_1 and β_0 using the estimated $P(Z)$. The first step is carried out using a logit regression of S on Z. In the second step we use the Robinson (1988) method for estimating partially linear models as extended in Heckman, Ichimura and Todd (1997).[15]

Table A3 in the appendix shows estimates of the average marginal derivatives for each variable in X. Standard errors are computed using the bootstrap.[16] Although the estimates are imprecisely

[14] In theory we only require that (U_0, U_1, U_S) independent of Z given X, so we do not estimate the most general possible model within our framework.

[15] We use a bandwidth of 0.1 for this procedure.

[16] Heckman, Ichimura and Todd (1997) show that the bootstrap provides a better approximation to the true standard errors than asymptotic standard errors for the estimation of β_1, β_0, and $K(P)$ in a model similar to the one we present here. We use 250 bootstrap replications. In each iteration of the bootstrap we reestimate $P(Z)$ so all standard errors account for the fact that $P(Z)$ is itself an estimated object.

determined, cognitive ability (as measured by AFQT) is an important determinant of the returns to schooling.

Local polynomial estimation is used here to estimate $K(P(Z))$ and its partial derivative with respect to $P(Z)$. This is because local polynomial estimation not only provides a unified framework for estimating both a function and its derivative but also has a variety of desirable properties in comparison with other available nonparametric methods.[17]

We can test for selection on the individual returns to attending college by checking whether $E(\ln Y|X, P(Z))$ is a linear or a nonlinear function of $P(Z)$. Nonlinearity in $P(Z)$ means that there is heterogeneity in the returns to college attendance and that individuals select into college based at least in part on their own idiosyncratic return (conditional on X). There are alternative ways to implement this test, which we discuss in Appendix D. As discussed there, we can reject linearity in $P(Z)$ using a normal selection model and a nonparametric test. Figure 2 plots the estimated function for $E(\ln Y \mid P(Z) = p)$ as a function of $P(Z)$ (along with a model which imposes linearity of this expectation in $P(Z)$), evaluated at the mean X in the sample. Visually, it shows a substantial departure from linearity which is supported by our formal statistical tests.

We can partition the MTE into two components, one depending on X and the other on u_S:

$$\begin{aligned} \text{MTE}(x, u_S) &= E(\ln Y_1 - \ln Y_0 \mid X = x, U_S = u_S) \\ &= \mu_1(X) - \mu_0(X) + E(U_1 - U_0 \mid U_S = u_S). \end{aligned}$$

Figure 3 plots the component of the MTE that depends on U_S but not on X.[18] We fix the components of X at their mean values in the sample. Returns are annualized to reflect the fact that college goers on average attend 3.5 years of college.

$E(U_1 - U_0 \mid U_S = u_S)$ is declining in u_S. The most college worthy persons in the sense of having

[17] Fan and Gijbels (1996) provide a detailed discussion of the properties of local polynomial estimators. In general, use of higher order polynomials may reduce the bias but increase the variance by introducing more parameters. Fan and Gijbels (1996) suggest that the order π of the polynomial be equal to $\pi = \mu + 1$, where μ is the order of the derivative of the function of interest that we want to fit. That is, Fan and Gijbels (1996) recommend a local linear estimator for fitting a function and a local quadratic estimator for fitting a first-order derivative. Therefore, $\partial K(p)/\partial p$ is estimated by a local quadratic estimator. We choose the bandwidth that minimize the integrated residual sum of squares, which gives us $h_n = 0.257$. Our results are robust to the choice of bandwidths between 0.1 and 0.4.

[18] For better visualization of the pointwise estimates of the MTE, appendix Figure A1 plots the same curve as in Figure 3 without the confidence interval bands. Figure A2 graphs the two dimensional function MTE(AFQT, U_S), where all Xs are kept at their average value in the sample except for AFQT, which is allowed to vary.

high gross returns are more likely to go to college (they have low U_S). Individuals choose the schooling sector in which they have comparative advantage. The magnitude of the heterogeneity in returns on which agents select on is substantial: returns can vary from 7% (for high U_S persons) to 40% per year of college (for low U_S persons). The magnitude of total heterogeneity is likely to be even higher. Unless the density of returns is degenerate for each U_S ($f(U_1 - U_0|U_S)$ is degenerate), there will be a distribution of returns centered at each value of the MTE. Furthermore, as shown in figure A2, when we add the $\mu_1(X) - \mu_0(X)$ component we observe returns as low as -0.2 and as high as 0.7. Returns are increasing in AFQT (the only varying element of X in this figure), indicating again that those individuals with the highest gross returns are the most likely to enroll in college.

Confidence interval bands are based on bootstrapped standard errors. Even though they are apparently wide, Appendix D establishes that we can reject the hypothesis that the MTE is flat, so marginal and average returns are not equal. Below we present empirical estimates derived from a parametric model, where the joint distribution of the unobservables in the outcome and selection equations is assumed to be normal (as in Willis and Rosen, 1979). The MTE is more precisely estimated but it has a very similar shape.

We estimate the MTE using the same instrumental variables used in standard analyses of the returns to schooling, but we are able to extract more information from the same data. In particular, whereas standard IV provides us with a single return to schooling at an unidentified margin, local IV allows us to estimate returns across a continuum of different margins which can be identified in terms of the position on the U_S scale. The margins we can identify depend on the support of our data. We are able to obtain full support for $P(Z)$ (or close to it) by: (i) using different instruments simultaneously; and (ii) exploring the assumptions that the MTE is additively separable in X and U_S, and that X is independent of the unobservables of the model.

Figure 4 shows the density of $P(Z)$ when we fix the variables in X at their mean values and vary the instruments one at a time. Therefore, all variation in $P(Z)$ is due to variation in the instruments. This experiment informs us about what margin each instrument identifies. To generate the graph labeled "Distance", we not only fix X at its mean, but we also fix all the other instruments at their mean values. Because this variable only takes two values, the density of $P(Z)$ in this case only has

two mass points. The graph labeled "Wage"corresponds to the density of $P(Z)$ we obtain when all variables except local wage at 17 are kept at their mean values, and the line labeled "Unemp"is generated by varying only local unemployment at 17. Finally, the line labeled "All"is the density of $P(Z)$ when all the instruments are allowed to vary and the variables in X are fixed at their mean values. The MTE as a function of U_S (for fixed X) is also plotted, but rescaled to fit the picture.

There are two important aspects of this figure. First, each instrument has different support, and therefore if we were to use each instrument in isolation at mean X we would only be able to identify a small section of the MTE. Instrumental variables, instrument by instrument, identify an average of the MTE taken over the section of the MTE covered by the support of the instrument with weights specified by Heckman and Vytlacil (2005) and Heckman, Urzua and Vytlacil (2006). When we use all instruments simultaneously the support of $P(Z)$ is greatly expanded.

Second, it is striking that even when we use all the instruments simultaneously at mean X the support of $P(Z)$ is very limited. We are able to get full support of $P(Z)$ because X varies across individuals.[19] Figure 5 shows the density of $P(Z)$ when all instruments are allowed to vary and the variables in X are fixed at different values. For simplicity, we group all Xs in an index, and consider a low and a high value of the index.[20] We allow Z to vary within the groups of observations with low and high X generating two densities of $P(Z)$. Since the MTE also varies with X there are two MTEs at two different levels (although both of them are rescaled to fit the picture).

This figure demonstrates that the only reason we have full support of $P(Z)$ is because we allow X to vary as well as the instruments. Notice that X is not used as an exclusion restriction, since all the Xs are included in the outcome equations. Variation in Z at low values of X identifies how the MTE changes with U_S at low values of U_S. Similarly, variation in Z at high values of X identifies how the MTE changes with U_S at high values of U_S. Notice that this procedure is only possible because the MTE is additively separable in X and the unobservables, and because X is

[19] Recall that we exclude X values realized after schooling choices are made.

[20] In particular, the schooling equation takes the following form:

$$S = 1 \text{ if } X\gamma_1 + Z\gamma_2 + ZX\gamma_3 + \varepsilon > 0$$

where Z is the vector of instruments. We pick $X\gamma_1$ as the index of X and we compute the percentiles of its distribution. Then we get all observations for which $X\gamma_1$ is between the 20^{th} and 30^{th} percentiles of its distribution, we compute the average $X\gamma_1$ in this group and call it Low X in the figure, and we allow Z to vary within this set of observations. This generates the density of $P(Z)$ for Low X. For high X we proceed analogously, but we take observations for which $X\gamma_1$ is between the 70^{th} and 80^{th} percentiles of its distribution.

independent of the unobservables. This implies that the two MTEs plotted in Figure 5 are parallel to each other. If that were not the case we would have to do our empirical analysis conditioning nonparametrically on X. It is clear from Figures 4 and 5 that in such a case $P(Z)$ would not have full support for any value of X, and it would not possible to estimate the MTE over the full interval. However, both the separability and independence assumptions are standard in empirical analyses of this type of model, and we do not see them as overly restrictive.[21]

Heckman, Urzua and Vytlacil (2006) examine in detail the behavior of the IV estimator in a model of heterogeneous returns. They study the case of IV for vector Z when the analyst uses $J(Z)$, a scalar function of Z. They show that even though the IV weights always integrate up to 1, they can be negative. In such a case, the IV estimate of the return to college can be negative even if the MTE is positive everywhere, creating an interpretative problem. There are, however, certain instruments or functions of instruments for which the weights are always positive. $P(Z)$ has a special status as an instrument because it always produces non-negative weights, and it also allows us to estimate the MTE, which we can then use to interpret any standard IV estimate.

Figure 6 plots the weight on the MTE for different instruments used one at a time. "All" corresponds to using $P(Z)$ as an instrument. As predicted by the analysis of Heckman and Vytlacil (1999), use of $P(Z)$ as an instrument always produces positive weights on MTE. However, using only one component of Z as an instrument, as is common in "sensitivity analyses" in the empirical IV literature, is not guaranteed to produce positive weights for all points in the support of U_S, although the weights integrate to 1. The weight on "Unemp" is negative for some intervals of U_S while the weights on the other instruments are positive everywhere. Different instruments weight the MTE differently. Negative weights can cause IV to be negative even if IV is everywhere positive (Heckman, Urzua and Vytlacil, 2006). These weights can be estimated and different IV can be compared on a common scale.

[21] A possible concern in this case is that all of the variation in $P(Z)$ is driven by X and not by the instruments, but that is not the case in our data. Even at the extremes of the density of $P(Z)$, there is considerable variation in the instruments. For example, for values of $P(Z)$ as low as 0.05 or as high as 0.96, there exist sets of individuals living near a college and sets of individuals living away from a college at age 14.

5.3 Average and Marginal Returns to College

Table 5 presents estimates of different summary measures of returns to one year of college. The ATE, TT, TUT, AMTE and the return for individuals induced to go to college by construction of colleges in counties where these do not exist (PRTE) are obtained in the following way.[22] First we construct different weighted averages of the MTE by applying the weights of Table 1A. Recall, however, that these weights are defined conditional on X and they define parameters conditional on X. Therefore, after computing each of these parameters for each value of $X = x$, we need to integrate them against the appropriate distribution of X, as shown in Appendix E. We calculate the AMTE weights by shifting SMSA average log earnings at 17 marginally for all individuals and then generate the density of X and U_S for the individuals who change their schooling status.[23]

The limited support of $P(Z)$ near the boundary values of $P(Z) = 0$ and $P(Z) = 1$ creates a practical problem for the computation of the treatment parameters such as ATE, TT, and TUT, since we cannot evaluate MTE for values of U_S outside the support of $P(Z)$. Furthermore, the sparseness of the data in the extremes does not allow us to accurately estimate the MTE at evaluation points close to 0 or 1. Therefore, the numbers presented in Table 5 are constructed by rescaling the weights to integrate to one over the region $[0.05, 0.96]$. These can be interpreted as the parameters defined in the empirical (trimmed) support of $P(Z)$, which is close to the full unit interval. Our evidence of near full support for $P(Z)$ in this paper is in marked contrast to the limited support found in Heckman, Ichimura, Smith and Todd (1998a), where lack of full support of $P(Z)$ and failure to account for it was demonstrated to be an empirically important source of bias for conventional evaluation estimators.

The sensitivity of estimates to lack of support in the tails ($P(Z) = 0$ or $P(Z) = 1$) is important for parameters that put substantial weight on the tails of the MTE distribution, such as ATE or TT. Even with support over most of the interval $[0, 1]$, such parameters cannot be identified unless 0 (for both ATE and TT) and 1 (for ATE) are contained in the support of the distribution of $P(Z)$. Estimates of these parameters are sensitive to imprecise estimation or extrapolation error

[22] This is one policy we consider in this paper, although many other alternatives are possible. In forming the PRTE, we force the variable that denotes presence of a college at 14 to take value 1 for every individual in the sample.

[23] College construction is a dichotomous variable and therefore not amenable to limit operations. Opportunity costs are continuous and hence amenable to limit operations. Thus AMTE is defined for a different policy change than PRTE.

for $E(Y|X, P(Z) = p)$ for values of p close to 0 or 1. Even though empirical economists often seek to identify ATE and TT, usually they are not easily estimated nor are they always the economically interesting parameters. In contrast (depending on the policies being analyzed) PRTE and AMTE parameters typically place little weight on the tails of the MTE distribution, and as a result are often relatively robust to imprecise estimation or extrapolation error in the tails. AMTE and PRTE are thus much easier to estimate, and are likely to be much less sensitive to alternative methods for estimating the MTE than are ATE, TT and TUT.

Integrating over $P(Z)$ in the interval $[0.05, 0.96]$, Table 5 reports estimates of the average annual return to college for a randomly selected person in the population (ATE) of 18.32%, which is between the annual return for the average individual who attends college (TT), 21.65%, and the average return for high school graduates who never attend college (TUT), 16.72%. The average marginal individual (AMTE) has an annual return of 17.93% which is below the annual return for the average person (TT). None of these numbers corresponds to the average annual return to college for those individuals who are induced to enroll in college by an increase in the number of colleges (PRTE), which is 20.13%. This is the relevant return for evaluating this specific policy using a Benthamite welfare criterion. It is below TT, which means that the average entrant induced to go to college by this specific policy has an annual return below that of the average college attendee (but well above the IV estimate of 17.51%).[24]

It is informative to visualize the weights behind each treatment parameter, which show us why some parameters are higher or lower than others. Figure 7 graphs the weights for $E(Y_1 - Y_0|X, U_S = u_S)$ for ATE, TT and PRTE (evaluated at the average X). ATE gives a uniform weight to all U_S, while TT overweights individuals with low levels of U_S (who are individuals with high returns, and also very likely to have enrolled in college), and PRTE puts more weight on individuals in the middle ranges of U_S. Figure 8 presents these weights for $E(Y_1 - Y_0|X)$ (we fix U_S at 0.5). In practice, because X is multidimensional, we only allow the MTE to vary with AFQT, and therefore we fix all the components of X at their mean values except for AFQT. The PRTE places more weight at the center of the distribution of X than does TT or ATE. Individuals attracted into college by

[24] In Appendix F we present estimates of different treatment parameters when we extrapolate the MTE to take values over the full unit interval. We also present bounds for ATE, developed in Heckman and Vytlacil (2001a). Given that we have close to full support, our results do not change substantially.

college construction differ from the average individual who attends college both in terms of U_S and in terms of X.

We next compare all of the estimated summary measures of returns with the OLS and IV estimates of the annual return to college, where the instrument is $\hat{P}(Z)$, the estimated probability of attending college for individuals with characteristics Z. Our OLS estimate is based on equation (8). The IV estimate is derived from the same equation. Since the returns estimated by OLS and by IV both depend on X (in this case, AFQT), we evaluate the OLS and IV returns at the average value of X. The OLS estimate of the return to a year of college is 5.02% while the IV estimate is 17.51%. Figure 9 plots the weight for $E(U_1 - U_0|U_S = u_S)$ for IV and for PRTE. Compared to the IV estimator, PRTE places greater weight at the extremes of MTE. Only by accident does IV identify policy relevant treatment effects when the MTE is not constant in U_S and the instrument is not the policy.

5.4 Comparing OLS and IV Estimates of the Returns to Schooling

A recurrent finding of the recent literature on the returns to schooling is that OLS estimates are below IV estimates of returns to schooling (see Card, 1999, 2001). Figure 10 plots the MTE weight for IV and the MTE weight for OLS on a comparable scale.[25] Because of the large negative components of the OLS weight, it is not surprising that the OLS estimate is lower than the IV estimate. One common interpretation of this finding is that returns are heterogeneous and IV estimates the return for the marginal person, while OLS estimates the return for the average person or is an upward biased estimate of the average return. Therefore the fact that IV estimates are larger than OLS estimates suggests that the return for the marginal person is above the return for the average person (see Card, 1999, 2001). However, we showed that the marginal person has a return substantially below the return for the average person, and still $\hat{\beta}_{\text{IV}} > \hat{\beta}_{\text{OLS}}$. The main reason why $\hat{\beta}_{\text{IV}} > \hat{\beta}_{\text{OLS}}$ is not that the marginal is above the average return, but that OLS places a large negative weight on the MTE due to selection bias.

[25] In order to place the weights on a comparable scale, we rescale the OLS weight. Estimation of the OLS weight requires the estimation of both $E(Y_1|X, U_S)$ and $E(Y_0|X, U_S)$. It is easy to show that $E(Y_1|X, U_S = p) = \frac{\partial E(SY|X,P(Z))}{\partial P(Z)}\Big|_{P(Z)=p}$ and $E(Y_0|X, U_S = p) = -\frac{\partial E[(1-S)Y|X,P(Z)]}{\partial P(Z)}\Big|_{P(Z)=p}$. These derivatives are estimated using the same procedure we described for the estimation of $E(Y_1 - Y_0|X, U_S = p) = \frac{\partial E(Y|X,P(Z))}{\partial P(Z)}\Big|_{P(Z)=p}$.

The comparison between $\hat{\beta}_{\text{IV}}$ and $\hat{\beta}_{\text{OLS}}$ is misleading because neither corresponds to an economically interpretable parameter. The least squares estimator does not identify the return to the average person attending college $E(\beta \mid S=1) = E(\ln Y_1 - \ln Y_0 \mid S=1)$. Rather it identifies treatment on the treated plus a selection bias term (keeping the conditioning on X implicit):

$$E(\ln Y \mid S=1) - E(\ln Y \mid S=0) = E(\beta \mid S=1) + [E(U_0 \mid S=1) - E(U_0 \mid S=0)].$$

In a model without variability in the returns to schooling, $E(\beta \mid S=1) = E(\beta) = \bar{\beta}$ is the same constant for everyone, so it is plausible that if U_0 is ability, the last term in brackets in the final expression will be positive (more able people attend school). This is the model of ability bias that motivated Griliches (1977) and those that preceded him. The ability bias argument suggests that OLS may provide an upward biased estimate of the average return to schooling. However, as noted by Willis and Rosen (1979), if there is comparative advantage, the term in brackets may be negative. Cunha, Heckman and Navarro (2005) demonstrate that this theoretical possibility actually characterizes U.S. data. People who go to college are below average in the no college Y_0 distribution, i.e., $E(U_0 \mid S=1) - E(U_0 \mid S=0) < 0$ even though they are above average in the Y_1 distribution. This could offset a positive sorting effect ($E(U_1 - U_0 \mid S=1) > 0$) and make the OLS estimate below that of the IV estimate, even if the IV estimate is below the return for the average person ($E(\beta|S=1)$). Thus the evidence reported in the recent literature comparing OLS and IV is not informative on the comparison of average and marginal returns.

The OLS weight can be decomposed into the TT ($E(Y_1 - Y_0|S=1)$) weight and the selection bias ($E(Y_0|S=1) - E(Y_0|S=0)$) weight. The TT weight is nonnegative and integrates to 1. The selection bias weight can be negative and may not integrate to 1. Under the assumptions that justify IV, the application of IV eliminates the second term. In Figure 11, we decompose the OLS weight. In our data the selection bias weight is a much more important component of the OLS weight than is the TT weight, which indicates that OLS estimates are not economically interpretable. Finding that IV estimates exceed OLS estimates is a consequence of selection bias.

5.5 Estimates of the MTE from a Normal Selection Model

The empirical results just presented are robust to specification changes, such as the introduction of additional parental background variables as controls, or to the use of alternative years of data. However, they are relatively imprecisely estimated. There are two main sources of imprecision in our model. The first relates to the fact that we are using IV. Most of the literature reports instrumental variable estimates of the returns to schooling which are generally more imprecisely estimated than are least squares estimates of the same parameter (e.g. Card, 1999). The second is our use of semiparametric methods, which impose relatively little structure but are more demanding in terms of data requirements.

In this section we examine what happens to the magnitude and precision of our estimates when we assume that the joint distribution of the unobservables (U_1, U_0 and V) is normal. We expect this additional structure to produce more precise estimates. First, we estimate the selection equation using a probit ($\mu_S(Z)$ is specified exactly as above). Then we compute the selection correction terms and include them in the wage equations for high school and college. This is a version of the estimation procedure pursued by Björklund and Moffitt (1987). As above, standard errors are bootstrapped (250 replications).

The estimates of $\mu_S(Z)$, $\mu_1(X)$ and $\mu_0(X)$ are similar to the ones reported above using semiparametric methods, although (as expected) they are more precisely estimated (in particular, AFQT is a quantitatively important and statistically strong determinant of the returns to college in this specification; see Tables A5 and A6 in the appendix). More interesting is the estimate of how the MTE varies with U_S, as shown in Figure 12. The shape and slope of the normal version of the MTE is very similar to the shape and slope of the semiparametric estimate of the MTE, although the level is slightly lower. Standard errors are much smaller, especially in the tails of the MTE, where the data are more scarce. As above, we can test and reject that the slope of the MTE is equal to zero (see Appendix D).

Table 6 reports the treatment parameters corresponding to the normal model. They are smaller than the ones shown in Table 5, and standard errors are about two thirds as large as those reported in Table 5. The IV estimates based on a probit rather than a logit are lower. Still, the main patterns are similar across the two tables. In particular, the average student in college has a much higher

return than the marginal entrant into college.

6 Summary and Conclusions

This paper estimates marginal and average returns to college when returns differ among individuals and persons select into economic activities based in part on their idiosyncratic return. Different conventional average return parameters and IV estimators are weighted averages of the marginal treatment effect (MTE). Unless the instruments are the policies being studied, these parameters answer well-posed economic questions only by accident.

We show how to identify and estimate the MTE using a robust nonparametric selection model. Our method allows us to combine diverse instruments into a scalar instrument motivated by economic theory. This combined instrument expands the support of any one instrument, and allows the analyst to estimate average returns at a continuum of different identified and interpretable margins, and to perform out-of-sample policy forecasts and to determine at what margin estimates of LATE are being identified. Focusing on a policy relevant question, we construct estimators based on the MTE to answer it, rather than hoping that a particular instrumental variable estimator happens to answer a question of economic interest. We estimate the Policy Relevant Treatment Effect and the Average Marginal Treatment Effect.

We estimate the returns to college using a sample of white males extracted from the National Longitudinal Survey of Youth (NLSY). We propose and implement a test for the importance of comparative advantage and self-selection in the labor market. The data suggest that comparative advantage is an empirically important phenomenon governing schooling choices, consistent with the analysis of Willis and Rosen (1979). Individuals sort into schooling on the basis of gains which are both observed and unobserved by the economist. Marginal expansions of schooling programs produce marginal gains that are well below average returns but well above OLS estimated returns. There is substantial evidence of selection bias that causes OLS to produce downward biased estimators.

Instrumental variables are not guaranteed to estimate policy relevant treatment parameters or conventional treatment parameters. Different instruments define different parameters. In our empirical analysis, IV understates the policy relevant return by 2.6 log points, and the average

marginal return by 1.2 log points. Comparing estimates from our semiparametric IV approach with estimates from a normal selection model of the sort used by Willis and Rosen (1979) and Björklund and Moffitt (1987), the semiparametric estimates produce an MTE with the same general shape consistent with diminishing returns, but with a lower level for the MTE.

A A Model of Schooling Choice

Consider a standard model of schooling choice. Let $Y_1(t)$ be the earnings of the schooled at experience level t while $Y_0(t)$ is the earnings of the unschooled at experience level t. Assuming that schooling takes one period in which earnings are foregone, a person takes schooling if

$$\frac{1}{(1+r)}\sum_{t=0}^{\infty}\frac{Y_1(t)}{(1+r)^t} - \sum_{t=0}^{\infty}\frac{Y_0(t)}{(1+r)^t} - C^* \geq 0,$$

where C^* is direct cost which may include psychic cost components, r is the discount rate, and lifetimes are assumed to be infinite to simplify the algebra. This is the prototypical discrete choice model applied to human capital investments.[26] We follow Mincer (1974) and assume that earnings profiles in logs are parallel in experience across schooling levels. Thus $Y_1(t) = Y_1 e(t)$ and $Y_0(t) = Y_0 e(t)$, where $e(t)$ is a post-school experience growth factor. Think of "1" as college and "0"as high school.

The agent attends school if

$$\left(\frac{1}{(1+r)}Y_1 - Y_0\right)\sum_{t=0}^{\infty}\frac{e(t)}{(1+r)^t} \geq C^*.$$

Let $K = \sum_{t=0}^{\infty}\frac{e(t)}{(1+r)^t}$ and absorb K into C^* so $\widetilde{C} = \frac{C^*}{K}$. Define discount factor $\gamma = \frac{1}{(1+r)}$. Using growth rate g to relate potential earnings in the two schooling choices we may write $Y_1 = (1+g)Y_0$ (in our empirical model, $\beta = \ln(1+g)$). Thus the decision to attend school ($S = 1$) is made if

$$Y_0[\gamma(1+g) - 1] \geq \widetilde{C}.$$

[26] This formulation makes it clear that we are analyzing *ex post* returns, as is conventional in the schooling literature. For an analysis of *ex ante* and *ex post* returns, see Carneiro, Hansen and Heckman (2003) and Cunha, Heckman and Navarro (2005).

This is equivalent to

$$\beta \geq \ln(1+\frac{\widetilde{C}}{Y_0})+\ln(1+r).$$

For $r \approx 0$ and $\frac{\widetilde{C}}{Y_0} \approx 0$, we may write the decision rule as $S=1$ if

$$\beta \geq r+\frac{\widetilde{C}}{Y_0}. \tag{A.1}$$

Ceteris paribus, a higher r or $\widetilde{C}$ lowers the likelihood that $S=1$. As long as $g>r$ (so $\gamma\,(1+g)-1>0$), a higher Y_0 implies a higher absolute return to college and leads people to attend college. Assuming that direct costs are zero ($\widetilde{C}=0$) and that the only cost is the opportunity cost of foregone income, the marginal return for those indifferent between going to school and facing interest rate r is $E\,(\beta \mid \beta=r)$. In the empirical analysis of this paper, we introduce variables Z that shift costs and discount factors ($\widetilde{C}=\widetilde{C}(Z)$, $r=r(Z)$). We use $C\,(Z)=r\,(Z)+\frac{\widetilde{C}(Z)}{Y_0}$ in the text.

B Weights for the PRTE and AMTE

In a model of heterogeneous returns it is possible to define alternative treatment effects, such as the average treatment effect ($E\,(\beta)$, or ATE), treatment on the treated ($E\,(\beta|S=1)$, or TT), or treatment on the untreated ($E\,(\beta|S=0)$, or TUT), among others. Which, if any, of these effects should be designated as "the"causal effect? This question is best answered by stating an economic question and finding the answer to it. In this paper, we adopt a standard welfare framework. Aggregate per capita outcomes under one policy are compared with aggregate per capita outcomes under another. One of the policies may be no policy at all. For utility criterion $V(Y)$, a standard welfare analysis compares an alternative policy with a baseline policy:

$$E(V(Y) \mid \text{Alternative Policy}) - E(V(Y) \mid \text{Baseline Policy}).$$

Adopting the common coefficient model, a log utility specification ($V(Y)=\ln Y$) and ignoring general equilibrium effects, when β is the same constant for everyone, $\beta=\tilde{\beta}$, the mean change in welfare is

$$E(\ln Y \mid \text{Alternative Policy}) - E(\ln Y \mid \text{ Baseline Policy}) = \tilde{\beta}(\Delta P), \tag{B.1}$$

where (ΔP) is the change in the proportion of people induced to attend school by the policy. This can be defined conditional on $X = x$ or overall for the population. In terms of gains per capita to recipients, the effect is $\tilde{\beta}$. In the general case, where β varies across individuals, agents partially anticipate β, and comparative advantage dictates schooling choices, none of the traditional treatment parameters plays the role of $\tilde{\beta}$ in (B.1) or answers the stated economic question.

We consider a class of policy interventions that affect $P(Z)$ but not $(\ln Y_1, \ln Y_0)$. This is the standard assumption in the partial equilibrium treatment effect literature.[27] We suppress the notation for conditioning on X in this section, leaving implicit that all of our analysis is conditional on X. Let $P(Z)$ be the baseline probability of $S = 1$ with cumulative distribution function $F_{P(Z)}$. Define $P^*(Z)$ as the probability produced under an alternative policy regime with cumulative distribution function $F_{P^*(Z)}$. Then we can write

$$E(V(Y) \mid \text{Alternative Policy}^*) - E(V(Y) \mid \text{Baseline Policy}) = \int_0^1 \text{MTE}(u)\omega(u)du,$$

where $\omega(u) = F_{P(Z)}(u) - F_{P^*(Z)}(u)$ where $F_{P(Z)}$ and $F_{P^*(Z)}$ denote the cdf of $P(Z)$ and $P^*(Z)$, respectively.[28]

To define a parameter comparable to $\tilde{\beta}$ in section (3), we normalize the weights by $\Delta P(Z) = E(P^*(Z)) - E(P(Z))$, the change in the proportion of people induced into the program. Thus if we use the weights

$$\tilde{\omega}(u) \equiv \frac{\omega(u)}{\Delta P(Z)} = \frac{F_{P(Z)}(u) - F_{P^*(Z)}(u)}{E(P^*(Z)) - E(P(Z))},$$

we produce the gain in the outcome for the people induced to change into (or out of) schooling by the policy change in the case where the policy change shifts individuals' college choice decision in one direction. The weights are well-defined if $E\left(P^*(Z)\right) \neq E\left(P(Z)\right)$.[29] These weights, developed by Heckman and Vytlacil (2001b), define the Policy Relevant Treatment Effect (PRTE),

$$\text{PRTE} = \int_0^1 \text{MTE}(u)\omega(u)du.$$

[27] For evidence against this in the case of large-scale social programs, see Heckman, Lochner and Taber (1998b, 1999). In the context of schooling, tuition can affect the choice of S and hence $P(Z)$ and also $(\ln Y_1, \ln Y_0)$ if changes in aggregate schooling participation affect skill prices.

[28] For a proof see Heckman and Vytlacil (2001b, 2007). Other criteria produce different weights.

[29] In this paper we do not consider the case where $E\left(P^*\right) = E\left(P\right)$.

Observe that these weights differ from the weights for the conventional treatment parameters (see Table 1B).

The following special case of the PRTE parameter will be particularly important for our analysis. Consider a policy that shifts Z_k (the kth element of Z) to $Z_k + \varepsilon$. For example, Z_k might be the tuition faced by the agent and the policy change might be to provide an incremental tuition subsidy of ε dollars. Suppose that $\mu_S(Z) = Z\gamma$, and that γ_k (the kth element of γ) is nonzero. We define $S = \mathbf{1}[Z\gamma > V]$ and denote by F_V the distribution of V with density f_V. Let PRTE_ε denote the resulting PRTE parameter and let $\tilde{\omega}_\varepsilon$ denote the resulting weights for the policy change of shifting of Z_k to $Z_k + \varepsilon$. We define

$$\text{PRTE}_\varepsilon = \int_0^1 \text{MTE}(u)\tilde{\omega}_\varepsilon(u)du,$$

with

$$\begin{aligned}\tilde{\omega}_\varepsilon(u) &= \frac{\Pr\left[Z\gamma \leq F_V^{-1}(u)\right] - \Pr\left[Z\gamma \leq F_V^{-1}(u) - \varepsilon\gamma_k\right]}{\Pr\left[V \leq Z\gamma + \varepsilon\gamma_k\right] - \Pr\left[V \leq Z\gamma\right]} \\ &= \frac{F_{Z\gamma}\left[F_V^{-1}(u)\right] - F_{Z\gamma}\left[F_V^{-1}(u) - \varepsilon\gamma_k\right]}{E_{Z\gamma}\left[F_V(Z\gamma + \varepsilon\gamma_k) - F_V(Z\gamma)\right]}.\end{aligned}$$

Consider the limit of this expression as ε goes to zero, i.e., the PRTE parameter corresponding to an infinitesimal change in Z_k. Assuming that $Z\gamma$ has a continuous density, then $\lim_{\varepsilon\to 0}\text{PRTE}_\varepsilon$ exists and is given by

$$\lim_{\varepsilon\to 0}\text{PRTE}_\varepsilon = \int_0^1 \text{MTE}(u)\tilde{w}(u)du,$$

where

$$\tilde{w}(u) = \lim_{\varepsilon\to 0}\tilde{\omega}_\varepsilon(u) = \frac{f_{Z\gamma}[F_V^{-1}(u)]}{E_{Z\gamma}(f_V(Z\gamma))}.^{30} \tag{B.2}$$

For example, if Z_k is college subsidy, $\lim_{\varepsilon\to 0}\text{PRTE}_\varepsilon$ corresponds to the effect of making an infinitesimal increment in the subsidy.

Suppose that $\varepsilon\gamma_k > 0$. We obtain

$$\text{PRTE}_\varepsilon = E(\beta \mid Z\gamma \leq V \leq Z\gamma + \varepsilon\gamma_k),$$

[29] We assume that $Z\gamma$ is absolutely continuous with respect to Lebesgue measure.

i.e., PRTE_ε is the average return among individuals who are induced into college by the incremental subsidy. Thus,

$$\lim_{\varepsilon \to 0} \text{PRTE}_\varepsilon = \lim_{\varepsilon \to 0} E(\beta \mid Z\gamma \leq V \leq Z\gamma + \varepsilon\gamma_k)$$

can be seen as the average return among those individuals who would be induced into college based on an infinitesimal change in Z_k. We define this parameter to be the Average Marginal Treatment Effect (AMTE), and we write $E(\beta \mid Z\gamma = V)$ to denote the AMTE parameter.[31] The average marginal treatment effect is thus defined by a sequence of policy alternatives that shift one element of Z by a shrinking additive shift, or equivalently by the average return among those individuals whose decision whether to attend college would be changed by an infinitesimal additive shift in one element of Z. Even though they are widely applied in the estimation of the returns to schooling, instrumental variables methods generally do not estimate PRTE or AMTE.

C Description of the Data

We restrict the NLSY sample to white males with a high school degree or above. We define high school graduates as individuals having a high school degree, or having completed 12 grades and never reporting college attendance.[32] We define participation in college as having attended some college or having completed more than 12 grades in school. GED recipients who do not have a high school degree, who have less than 12 years of schooling completed or who never reported college attendance are excluded from the sample. The wage variable that is used is an average of deflated (to 1983) non-missing hourly wages reported in 1992, 1993, 1994 and 1996. We delete all wage observations that are below 1 or above 100. Experience is actual work experience in weeks (we divide it by 52 to express it as a fraction of a year) accumulated from 1979 to 1993 (annual weeks

[31] Formally, $E(\beta | Z\gamma = V)$ is not uniquely defined since the conditioning set is a set of measure zero. We define it by $\lim_{\varepsilon \to 0} E(\beta \mid Z\gamma \leq V \leq Z\gamma + \varepsilon\gamma_k)$, which is equivalent under our assumptions to defining it by $\lim_{\varepsilon \to 0} E(\beta | -\varepsilon \leq Z\gamma - V \leq \varepsilon)$ and is also equivalent to defining it by $E(\beta \mid Z\gamma - V = t)$ evaluated at $t = 0$. However, it would be equally valid to define AMTE by, e.g., $\lim_{\varepsilon \to 0} E(\beta | -\varepsilon \leq P(Z) - U_S \leq \varepsilon)$ which is equivalent to $E(\beta \mid P(Z) - U_S = t)$ evaluated at $t = 0$; or alternatively defining AMTE by $\lim_{\varepsilon \to 0} E(\beta | -\varepsilon \leq [P(Z)/U_S] - 1 \leq \varepsilon)$ which is equivalent to $E(\beta \mid [P(Z)/U_S] = t)$ evaluated at $t = 1$. Each of these alternative expressions leads to a different value of AMTE with different weights on MTE, and correspond to the limits of alternative sequences of policy changes. For example, the PRTE parameter corresponding to shifting tuition downward proportionally by an infinitesimal amount corresponds to an alternative AMTE defined by $\lim_{\varepsilon \to 0} E(\beta \mid -\varepsilon \leq [Z\gamma/V] - 1 \leq \varepsilon)$ or equivalently $E(\beta \mid [Z\gamma/V] = t)$ evaluated at $t = 1$, and this alternative AMTE parameter will place different weights on the MTE parameter from the one defined by $\lim_{\varepsilon \to 0} E(\beta | -\varepsilon \leq Z\gamma - V \leq \varepsilon)$.

[32] For a description of the NLSY 1979, see Bureau of Labor Statistics (2001).

worked are imputed to be zero if they are missing in any given year). The remaining variables that we include in the X and Z vectors are mother's years of schooling, schooling corrected AFQT, dummies indicating the year of birth, the presence of a four-year college in the SMSA of residence at age 14 (from Kling, 2001),[33] local average earnings in the SMSA of residence at 17 and local unemployment rate in state of residence at age 17, and in 1994. SMSA earnings correspond to the average wage per job in the SMSA constructed using data from the Bureau of Economic Analysis, deflated to 2000. The state unemployment rate data come from the BLS website. However, from the BLS website it is not possible to get state unemployment data for all states for all the 1970s. Data are available for all states from 1976 on, and for 29 states for 1973, 1974 and 1975. Therefore for some of the individuals we have to assign them the unemployment rate in the state of residence in 1976 (which will correspond to age 19 for those born in 1957 and age 18 for those born in 1958). SMSA and state of residence at 17 are not available for everyone in the NLSY, but only for the cohorts born in 1962, 1963 and 1964 (age 17 in 1979, 1980 and 1981). However, SMSA and state of residence at age 14 is available for most respondents. Therefore, we impute location at 17 to be equal to location at 14 for cohorts born between 1957 and 1962 unless location at 14 is missing, in which case we use location in 1979 for the imputation. The NLSY79 has an oversample of poor whites which we exclude from this analysis. We also exclude the military sample. Many individuals report having obtained a bachelors degree or more and, at the same time, having attended only 15 years of schooling (or less). We recode years of schooling for these individuals to be 16. This variable is only used to annualize the returns to schooling. If we did not perform this recoding, when computing returns to one year of college we would divide the returns to schooling by 3.2 instead of dividing by 3.5. This corresponds to multiplying all of the estimated returns in the paper by $3.5/3.2 = 1.09$. To remove the effect of schooling on AFQT we implement the procedure of Hansen, Heckman and Mullen (2004). See the estimates reported in Table A1.

[33] The distance variable we use is the one used in Kling (2001), available at the *Journal of Business and Economics Statistics* website.

D Tests of Selection on Returns

This appendix discusses and applies three tests of the hypothesis that $E(\ln Y \mid X, P(Z))$ is linear in $P(Z)$. Recall that, under our assumptions, we can write

$$\ln Y = \mu_0(X) + S[\mu_1(X) - \mu_0(X)] + U_0 + S(U_1 - U_0),$$

which implies that:

$$E(\ln Y \mid X = x, P(Z) = p) = \mu_0(x) + p[\mu_1(x) - \mu_0(x)] + K(p).$$

The MTE is equal to $E(\ln Y_1 - \ln Y_0 \mid X = x, U_S = p) = \mu_1(x) - \mu_0(x) + K'(p)$, which is the derivative of the previous expression.

We first consider a direct test of whether $K(P(Z))$ is linear in $P(Z)$, due to Yatchew (2003). We estimate $\mu_1(x)$ and $\mu_0(x)$ using partial linear regression (as described in the main text) and compute the residual $R = Y - \{\mu_0(x) + p[\mu_1(x) - \mu_0(x)]\}$. $K(P(Z)) = E(R|P(Z))$, so that $R = K(P(Z)) + \varepsilon$. We obtain N pairs $(R, P(Z))$, where N is the number of observations, which we sort by $P(Z)$, with $p_1 < \cdots < p_N$. We can form a consistent estimator of σ_ε^2: $s_{\text{diff}}^2 = \frac{1}{2N}\sum_{i=2}^{N}(R_i - R_{i-1})^2$ (see Yatchew, 2003). If we impose the null hypothesis that $K(P(Z))$ is linear in $P(Z)$, we can run a regression of R on $P(Z)$. If the linear model is correct then the following is a consistent estimator of the variance of the regression residuals: $s_{\text{res}}^2 = \frac{1}{N}\sum_{i=1}^{N}(R_i - \hat{\gamma}_0 - \hat{\gamma}_1 P(Z))^2$, where $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are the estimated regression coefficients. Yatchew (2003) shows that $V = \frac{N^{1/2}\left(s_{\text{res}}^2 - s_{\text{diff}}^2\right)}{s_{\text{diff}}^2}$ is asymptotically $N(0,1)$ under the null. In our data $V = 1.9749$, which corresponds to a p-value of 0.0482, indicating rejection of the null hypothesis that $K(P(Z))$ is linear in $P(Z)$.

A second test is done under the assumption of normality in the error terms of the selection model we estimate, as in the last part of our empirical section. There we compute $\text{Cov}(U_1 - U_0, V)$ directly, which for our data is -0.3717, with a standard error of 0.1833, indicating that there is selection on returns using a 5% significance level. The estimated negative covariance is consistent with the estimates from our semiparametric model. Those with higher unobserved costs (V) have lower returns (see Figure 11).

A third test is a joint LATE test. Specifically, we divide U_S into 10 intervals (0-0.1, 0.1-0.2, ...,

0.9-1) and compute LATE within each interval ($E(\beta \mid u < U_S \leq u')$, where u and u' are the lower and upper bound of each interval). If there is no selection on returns, all LATEs should be equal to each other. Therefore, we perform a test of joint equality of all pairwise combination of LATEs, using the bootstrapped variance covariance matrix of the LATEs. The chi-square statistic for this test is above 80, indicating a clear rejection of the null of equality of LATEs across intervals.

E Accounting for X in the Weights for Treatment Parameters

In our exposition most of the weights were defined conditional on X and they define parameters conditional on X. Therefore, after computing each of these parameters for each value of $X = x$, we need to integrate them against the appropriate distribution of X, which depends on the parameter we want to compute:

$$\begin{aligned}
\Delta^{\text{ATE}} &= \int \Delta^{\text{ATE}}(x) f_X(x)\, dx, \\
\Delta^{\text{TT}} &= \int \Delta^{\text{TT}}(x) f_X(x \mid S=1)\, dx, \\
\Delta^{\text{TUT}} &= \int \Delta^{\text{TUT}}(x) f_X(x \mid S=0)\, dx, \\
\Delta^{\text{AMTE}} &= \int \Delta^{\text{AMTE}}(x) f_X(x \mid \text{Marginal})\, dx, \\
\Delta^{\text{PRT}} &= \int \Delta^{\text{PRT}}(x) f_X(x \mid \text{PRT})\, dx,
\end{aligned}$$

where $f_X(x \mid \text{PRT})$ is the density of X for individuals induced to go to college by the policy. The schooling choice equation is: $S = \mathbf{1}[P(Z) - U_S \geq 0]$, so

$$\begin{aligned}
f_X(x \mid S=1) &= f_X(x \mid P(Z) - U_S \geq 0), \\
f_X(x \mid S=0) &= f_X(x \mid P(Z) - U_S < 0), \\
f_X(x \mid \text{PRT}) &= f_X(x \mid P(Z) - U_S < 0, P(Z') - U_S \geq 0),
\end{aligned}$$

where Z and Z' are the values of the instruments under the baseline regime and under the new policy regime, respectively. These densities are also weights, but instead of weighting functions of U_S they weight functions of X.[34]

F Extrapolating the MTE and Computing Bounds for Treatment Parameters

Table A4 contrasts the estimates of treatment parameters we obtain by constraining the support of the MTE to be between 0.05 and 0.96 with the estimates we would obtain if we extrapolated the MTE (by evaluating the estimated function over the whole unit interval instead of just using points between 0.05 and 0.96) to exist in the full unit interval. Since the support of our data is close to the unit interval, our estimates do not change significantly.

Another approach to deal with the lack of full support is to bound the treatment parameters. We compute the bounds for ATE developed in Heckman and Vytlacil (2001a), and discussed further in Heckman and Vytlacil (2007), using our data, and found them to be relatively wide, although less wide than the standard bounds reported in the literature, as shown in Table A4. To compute the bounds we assume wages may take a minimum value of \$1 and a maximum value of \$100.

References

Angrist, Joshua D. and Alan B. Krueger, "Does Compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics*, November 1991, *106* (4), 979–1014.

Bishop, John H., "Achievement, Test Scores, and Relative Wages," in Marvin H. Kosters, ed., *Workers and their wages: Changing patterns in the United States*, number 520. In 'AEI Studies.', Washington, D.C.: AEI Press, 1991, pp. 146–186.

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

Blackburn, McKinley L. and David Neumark, "Omitted-Ability Bias and the Increase in the Return to Schooling," *Journal of Labor Economics*, July 1993, *11* (3), 521–544.

[34] To compute these weights we need to estimate $f(P(Z)|X)$. For simplicity, we group X into an index $(\mu_{SX}(X)$, corresponding to the sum of the components of $\mu_S(Z)$ involving only the variables in $X)$. Z is characterized by index $P(Z)$. We estimate $F(P(Z)|\mu_{SX})$ using a local linear regression of $\mathbf{1}[P(Z) \leq p]$ on $\mu_{SX}(X)$.

Bureau of Labor Statistics, *NLS Handbook 2001: The National Longitudinal Surveys*, Washington, DC: U.S. Department of Labor, 2001.

Cameron, Stephen V. and Christopher Taber, "Estimation of Educational Borrowing Constraints Using Returns to Schooling," *Journal of Political Economy*, February 2004, *112* (1), 132–182.

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

_ **and** _ , "The Dynamics of Educational Attainment for Black, Hispanic, and White Males," *Journal of Political Economy*, June 2001, *109* (3), 455–99.

Card, David, "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky, eds., *Aspects of Labour Market Behaviour: Essays in Honor of John Vanderkamp*, Toronto: University of Toronto Press, 1995, pp. 201–222.

_ , "The Causal Effect of Education on Earnings," in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*, Vol. 5, New York: North-Holland, 1999, pp. 1801–1863.

_ , "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, September 2001, *69* (5), 1127–1160.

Carneiro, Pedro and James J. Heckman, "The Evidence on Credit Constraints in Post-Secondary Schooling," *Economic Journal*, October 2002, *112* (482), 705–734.

_ **, Karsten Hansen, and James J. Heckman**, "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review*, May 2003, *44* (2), 361–422. 2001 Lawrence R. Klein Lecture.

Cunha, Flavio, James J. Heckman, and Salvador Navarro, "Separating Uncertainty from Heterogeneity in Life Cycle Earnings, The 2004 Hicks Lecture," *Oxford Economic Papers*, April 2005, *57* (2), 191–261.

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

Fan, Jianqing and Irene Gijbels, *Local Polynomial Modelling and its Applications*, New York: Chapman and Hall, 1996.

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

Grogger, Jeff and Eric Eide, "Changes in College Skills and the Rise in the College Wage Premium," *Journal of Human Resources*, Spring 1995, *30* (2), 280–310.

Hansen, Karsten T., James J. Heckman, and Kathleen J. Mullen, "The Effect of Schooling and Ability on Achievement Test Scores," *Journal of Econometrics*, July–August 2004, *121* (1-2), 39–98.

Heckman, James J., "Identification of Causal Effects Using Instrumental Variables: Comment," *Journal of the American Statistical Association*, June 1996, *91* (434), 459–462.

_ , "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy*, August 2001, *109* (4), 673–748.

_ **and Edward J. Vytlacil**, "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, April 1999, *96*, 4730–4734.

_ **and** _ , "The Relationship Between Treatment Parameters Within a Latent Variable Framework," *Economics Letters*, January 2000, *66* (1), 33–39.

_ **and** _ , "Local Instrumental Variables," in Cheng Hsiao, Kimio Morimune, and James L. Powell, eds., *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, New York: Cambridge University Press, 2001, pp. 1–46.

_ **and** _ , "Policy-Relevant Treatment Effects," *American Economic Review*, May 2001, *91* (2), 107–111.

_ **and** _ , "Structural Equations, Treatment Effects and Econometric Policy Evaluation," *Econometrica*, May 2005, *73* (3), 669–738.

_ **and** _ , "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Economic Estimators to Evaluate Social Programs and to Forecast Their Effects in New Environments," in J. Heckman and E. Leamer, eds., *Handbook of Econometrics, Volume 6*, Amsterdam: Elsevier, 2007. Forthcoming.

_ **and Richard Robb**, "Alternative Methods for Evaluating the Impact of Interventions," in J.J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*, Vol. 10, New York: Cambridge University Press, 1985, pp. 156–245.

_ **and** _ , "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," in H. Wainer, ed., *Drawing Inferences from Self-Selected Samples*, New York: Springer-Verlag, 1986, pp. 63–107. Reprinted in 2000, Mahwah, NJ: Lawrence Erlbaum Associates.

_ **, Hidehiko Ichimura, and Petra E. Todd**, "How Details Make a Difference: Semiparametric Estimation of the Partially Linear Regression Model," 1997. Unpublished manuscript, University of Chicago, Department of Economics.

_ **,** _ **, Jeffrey Smith, and Petra E. Todd**, "Characterizing Selection Bias Using Experimental Data," *Econometrica*, September 1998, *66* (5), 1017–1098.

_ **, Justin L. Tobias, and Edward J. Vytlacil**, "Four Parameters of Interest in the Evaluation of Social Programs," *Southern Economic Journal*, October 2001, *68* (2), 210–223.

_ **, Lance J. Lochner, and Christopher Taber**, "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents," *Review of Economic Dynamics*, January 1998, *1* (1), 1–58.

_ , _ , and _ , "General-Equilibrium Cost-Benefit Analysis of Education and Tax Policies," in G. Ranis and L.K. Raut, eds., *Trade, Growth and Development: Essays in Honor of T.N. Srinivasan*, Elsevier Science B.V., 1999, chapter 14, pp. 291–349.

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

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

Kane, Thomas J. and Cecilia E. Rouse, "Labor-Market Returns to Two- and Four-Year College," *American Economic Review*, June 1995, *85* (3), 600–614.

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

McFadden, Daniel, "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, ed., *Frontiers in Econometrics*, New York: Academic Press, 1974.

Mincer, Jacob, *Schooling, Experience and Earnings*, New York: Columbia University Press for National Bureau of Economic Research, 1974.

Robinson, Chris, "The Joint Determination of Union Status and Union Wage Effects: Some Tests of Alternative Models," *Journal of Political Economy*, June 1989, *97* (3), 639–667.

Robinson, Peter M., "Root-N-Consistent Semiparametric Regression," *Econometrica*, July 1988, *56* (4), 931–954.

Staiger, Douglas and James H. Stock, "Instrumental Variables Regression with Weak Instruments," *Econometrica*, May 1997, *65* (3), 557–586.

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

Yatchew, Adonis, *Semiparametric Regression for the Applied Econometrician*, New York: Cambridge University Press, 2003.

Table 1A
Treatment Effects and Estimands as Weighted Averages of the Marginal Treatment Effect

$\text{ATE}(x) = E\left(Y_1 - Y_0 \mid X = x\right) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, du_D$

$\text{TT}(x) = E\left(Y_1 - Y_0 \mid X = x, D = 1\right) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{TT}}(x, u_D)\, du_D$

$\text{TUT}\,(x) = E\left(Y_1 - Y_0 \mid X = x, D = 0\right) = \int_0^1 \Delta^{\text{MTE}}\,(x, u_D)\ \omega_{\text{TUT}}\,(x, u_D)\ du_D$

Policy Relevant Treatment Effect $(x) = E\left(Y_{a'} \mid X = x\right) - E\left(Y_a \mid X = x\right) = \int_0^1 \Delta^{\text{MTE}}\,(x, u_D)\ \omega_{\text{PRTE}}\,(x, u_D)\ du_D$

for two policies a and a' that affect the Z but not the X

$\text{IV}_J(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{IV}_J}(x, u_D)\, du_D$, given instrument J

$\text{OLS}(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{OLS}}(x, u_D)\, du_D$

Source: Heckman and Vytlacil (2005)

Table 1B
Weights

$\omega_{\text{ATE}}(x, u_D) = 1$

$\omega_{\text{TT}}(x, u_D) = \left[\int_{u_D}^1 f(p \mid X = x)dp\right] \frac{1}{E(P|X=x)}$

$\omega_{\text{TUT}}\,(x, u_D) = \left[\int_0^{u_D} f\,(p|X = x)\, dp\right] \frac{1}{E((1-P)|X=x)}$

$\omega_{\text{PRTE}}(x, u_D) = \left[\frac{F_{P_{a'},X}(u_D) - F_{P_a,X}(u_D)}{\Delta \bar{P}}\right]$

$\omega_{\text{IV}_J}(x, u_D) = \left[\int_{u_D}^1 (J(Z) - E(J(Z) \mid X = x)) \int f_{J,P|X}\,(j, t \mid X = x)\ dt\, dj\right] \frac{1}{\text{Cov}(J(Z),D|X=x)}$

$\omega_{\text{OLS}}(x, u_D) = 1 + \frac{E(U_1|X=x,U_D=u_D)\,\omega_1(x,u_D) - E(U_0|X=x,U_D=u_D)\,\omega_0(x,u_D)}{\Delta^{\text{MTE}}(x,u_D)}$

$\omega_1(x, u_D) = \left[\int_{u_D}^1 f(p \mid X = x)\, dp\right] \left[\frac{1}{E(P|X=x)}\right]$

$\omega_0(x, u_D) = \left[\int_0^{u_D} f(p \mid X = x)\, dp\right] \frac{1}{E((1-P)|X=x)}$

Source: Heckman and Vytlacil (2005)

Table 2
Sample Statistics

	$S = 0$ $(N = 717)$	$S = 1$ $(N = 903)$
Log Hourly Wage	2.4029	2.7406
	(0.5568)	(0.5493)
Years of Experience	10.1838	7.5162
	(4.2233)	(3.9804)
Corrected AFQT	-0.3580	0.5563
	(0.8806)	(0.7650)
Mother's Years of Schooling	11.4895	12.8992
	(2.0288)	(2.2115)
SMSA Log Earnings in 1994	10.2707	10.3277
	(0.1618)	(0.1738)
State Unemployment in 1994	5.7793	5.9292
(in %)	(1.2431)	(1.2851)
Presence of a College at 14	0.4616	0.5825
	(0.4988)	(0.4934)
SMSA Log Earnings at 17	10.2793	10.2760
	(0.1625)	(0.1692)
State Unemployment Rate at 17	7.0945	7.0847
(in %)	(1.8361)	(1.8746)

Note: Corrected AFQT corresponds to a standardized measure of the Armed Forces Qualifying Test score corrected for the fact that different individuals have different amounts of schooling at the time they take the test (see Hansen, Heckman and Mullen, 2004; see also Data Appendix B). This variable is standardized within the NLSY sample to have mean zero and variance 1. High School dropouts are excluded from this sample. We use only white males from the NLSY79, excluding the oversample of poor whites and the military sample. Standard deviations are in parentheses.

Table 3
Average Derivatives for College Decision Model

Corrected AFQT	0.2238
	(0.0279)
Mother's Years of Schooling	0.0422
	(0.0119)
Presence of a College at 14	0.0933
	(0.0231)
SMSA Log Earnings at 17	-0.1543
	(0.0761)
State Unemployment Rate at 17	0.0082
(in %)	(0.0090)
Chi-Squared test for joint significance of instruments	36.03
p-value	0.0070

Note: This table reports the average marginal derivatives from a logit regression of college attendance (a dummy variable that is equal to 1 if an individual has ever attended college and equal to 0 if he has never attended college but has graduated from high school) on polynomials in the set of variables listed in the table and on cohort dummies. For each individual we compute the effect of increasing each variable by one unit (keeping all the others constant) on the probability of enrolling in college and then we average across all individuals. Bootstrapped standard errors (in parentheses) are presented below the corresponding coefficients (250 replications).

Table 4
OLS and IV Estimates of the Return to One Year of College

	Return does not vary with X						Return varies with X	
	OLS	IV					OLS	IV
		Distance	Earnings	Unemployment	All	P		P
β	0.0389	0.1896	0.2431	0.0787	0.1865	0.1379	0.0502	0.1751
	(0.0087)	(0.0960)	(0.1230)	(0.1301)	(0.0573)	(0.0470)	(0.0119)	(0.0661)
$\partial\beta/\partial$AFQT							0.0249	0.0855
							(0.0148)	(0.0385)
F - Statistic (first stage)		2.79	1.90	1.31	2.63	2.23		2.23
p-value		0.01	0.07	0.25	0.00	0.00		0.00

Note: This table reports OLS and IV alternative estimates of the returns to schooling. The model estimated in the first six columns is $\ln Y = \alpha + \beta S + X\gamma + \varepsilon$ where $\ln Y$ is log hourly wage in 1994, S is college attendance, and X is vector of controls (years of experience, AFQT, mother's education, cohort dummies, state unemployment rate in 1994, and SMSA log wage in 1994). The estimate presented in the table corresponds to $\beta/3.5$, since 3.5 is the average difference in the years of schooling of individuals with and without any college attendance in our sample. In columns 2 through 6 we instrument S with different instruments: the presence of a college in the SMSA of residence at 17, SMSA log earnings at 17, and state unemployment at 17. The column labeled ALL corresponds to the use of all the instruments simultaneously, and in the column labeled P we instrument S with P, the predicted probability of going to college (a function of X and all the instruments). In columns 7 and 8 we estimate the following alternative model: $\ln Y = \alpha + \beta S + X\gamma + \theta SX + \varepsilon$, where SX is a vector of interactions between S and X. The estimate presented in the first line of the table corresponds to $[\beta + \theta E(X)]/3.5$, where $E(X)$ is the average value of X in the sample. In the second line we report the average marginal effect of AFQT on the return to a year of college, computed from the interactions between S and X. In the last column of the table we instrument S with P, and SX with PX. The F-statistics and the p-values in the last two rows of the table correspond to a test of whether the instrumental variables belong in a regression of college attendance on the instruments and X.

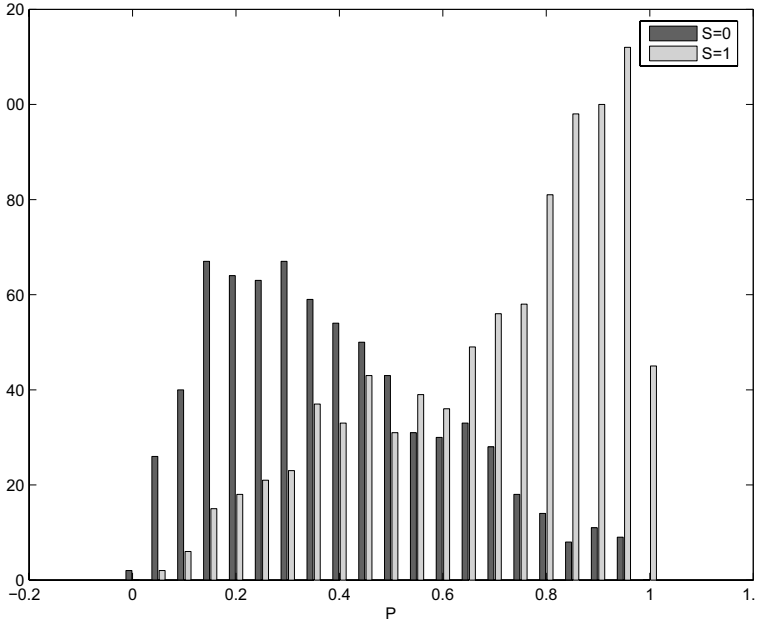
Table 5
Estimates of Various Returns to One Year of College
(Semi-Parametric Model)

	$0.0541 < P < 0.9662$
Average Treatment Effect	0.1832
	(0.0855)
Treatment on the Treated	0.2165
	(0.0978)
Treatment on the Untreated	0.1672
	(0.0875)
Average Marginal Treatment Effect	0.1793
	(0.1114)
Policy Relevant Treatment Effect	0.2013
(Construction of Colleges)	(0.1079)
Ordinary Least Squares	0.0502
	(0.0119)
Instrumental Variables	0.1751
	(0.0661)

Table 6
Estimates of Various Returns to One Year of College
(Normal Model)

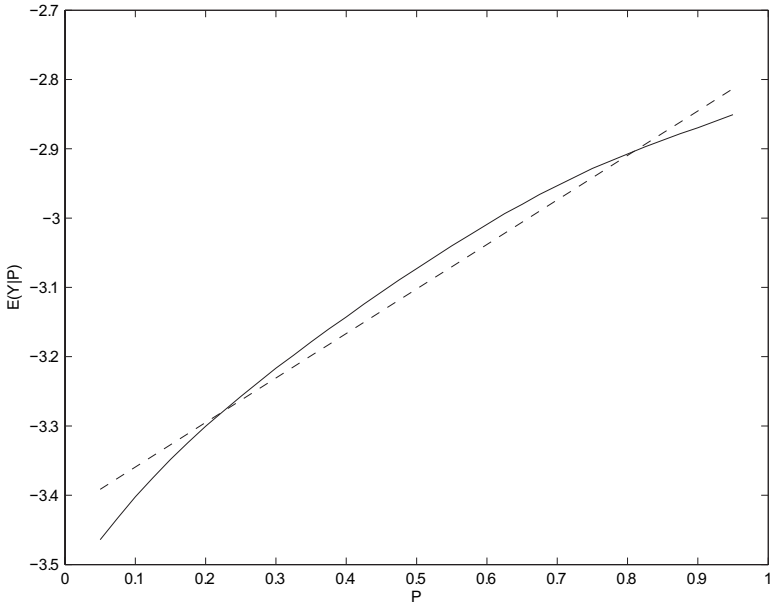
	$0.0541 < P < 0.9662$
Average Treatment Effect	0.1505
	(0.0553)
Treatment on the Treated	0.1854
	(0.0682)
Treatment on the Untreated	0.1187
	(0.0541)
Average Marginal Treatment Effect	0.1563
	(0.0624)
Policy Relevant Treatment Effect	0.1717
(Construction of Colleges)	(0.0659)
Ordinary Least Squares	0.0502
	(0.0119)
Instrumental Variables	0.1834
	(0.0670)

Figure 1: Density of P Given $S = 0$ and $S = 1$
(Estimated Probability of Enrolling in College)



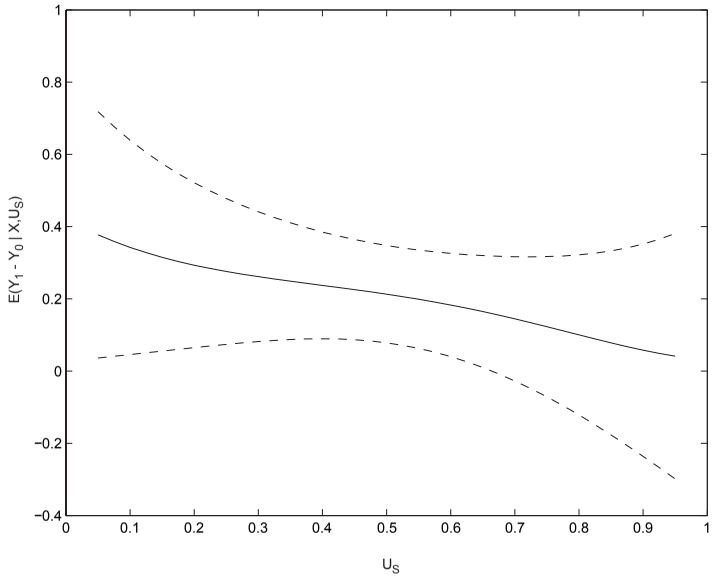
Note: P is the estimated probability of going to college. It is estimated from a logit regression of college attendance on corrected AFQT, mother's education, cohort dummies, a dummy variable indicating the presence of a college in the county of residence at age 14, average unemployment in the state of residence at age 17 and average log earnings in the SMSA of residence at age 17 (see Table 3).

Figure 2: $E(Y|X, P)$ as a Function of P for Average X



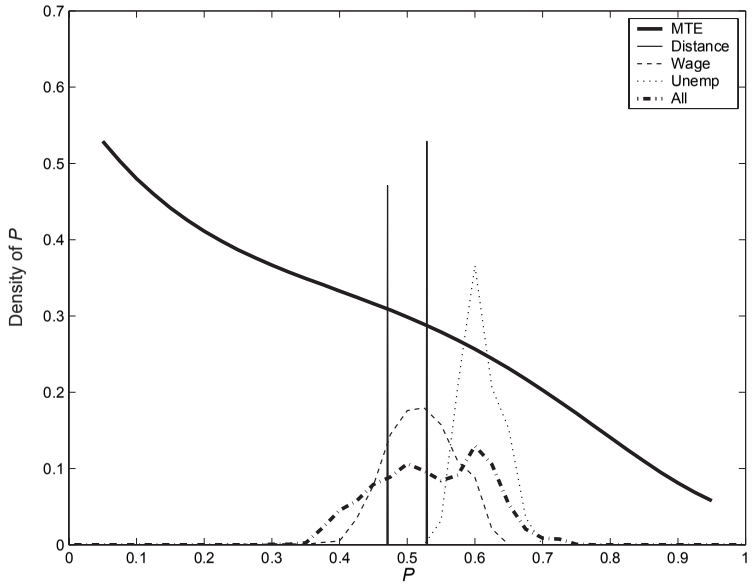
Note: To estimate the nonlinear function in this figure we use a partially linear regression of log wages on polynomials in X, interactions of polynomials in X and P, and $K(P)$, a locally quadratic function of P (where P is the predicted probability of attending college), with a bandwidth of 0.25. X includes years of experience, corrected AFQT, mother's education, cohort dummies, average unemployment in the state of residence and average log earnings in the SMSA of residence, measured in 1994. The straight line is generated by imposing that $K(P)$ is a linear function of P.

Figure 3: $E(Y_1 - Y_0|X, U_S)$ Estimated Using Locally Quadratic Regression (Averaged Over X)



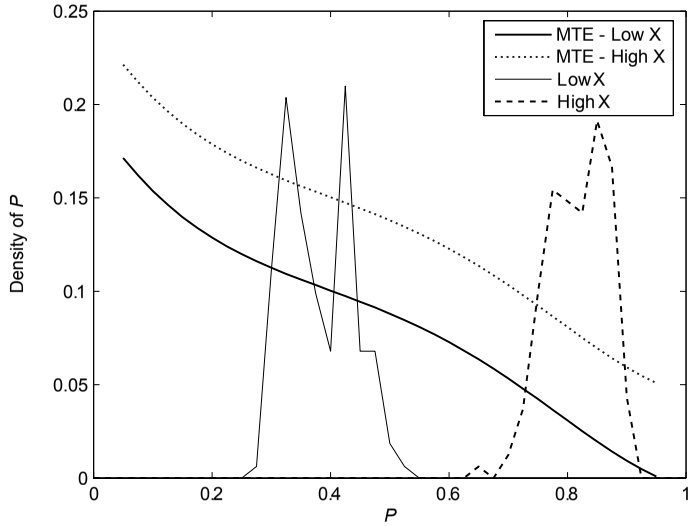
Note: To estimate the function plotted here we first use a partially linear regression of log wages on polynomials in X, interactions of polynomials in X and P, and $K(P)$, a locally quadratic function of P (where P is the predicted probability of attending college), with a bandwidth of 0.25. X includes years of experience, corrected AFQT, mother's education, cohort dummies, average unemployment in the state of residence and average log earnings in the SMSA of residence, measured in 1994. Then the figure is generated by taking the coefficient on the linear term in P from $K(P)$. Standard error bands are obtained using the bootstrap (250 replications).

Figure 4: Support of P for Different Instruments at Mean X



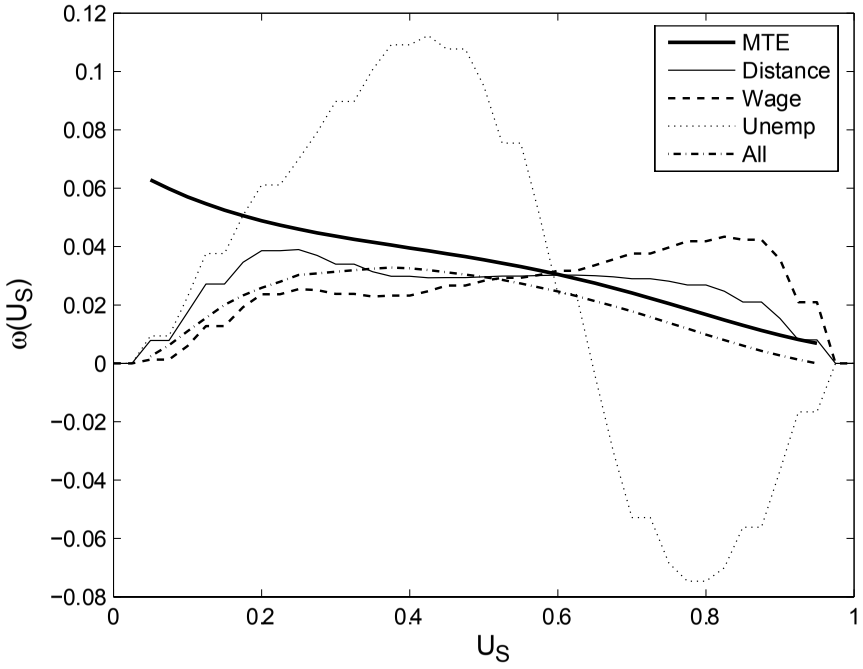
Note: This figure shows the density of P when we fix the variables in X at their mean values. In order to draw the line labeled Distance we not only fix X at its mean, but we also fix all the instruments at their mean values, except for the presence of a college at 14. The line labeled Wage corresponds to the density of P we obtain when all variables except local wage at 17 are kept at their mean values, and the line labeled Unemp is generated by varying only local unemployment at 17. Finally, the line labeled All is the density of P when all the instruments are allowed to vary and the variables in X are fixed at their mean values. The MTE as a function of U_S (for fixed X) is also plotted, but rescaled to fit the picture.

Figure 5: Support of P for Low and High X (Using All Instruments)



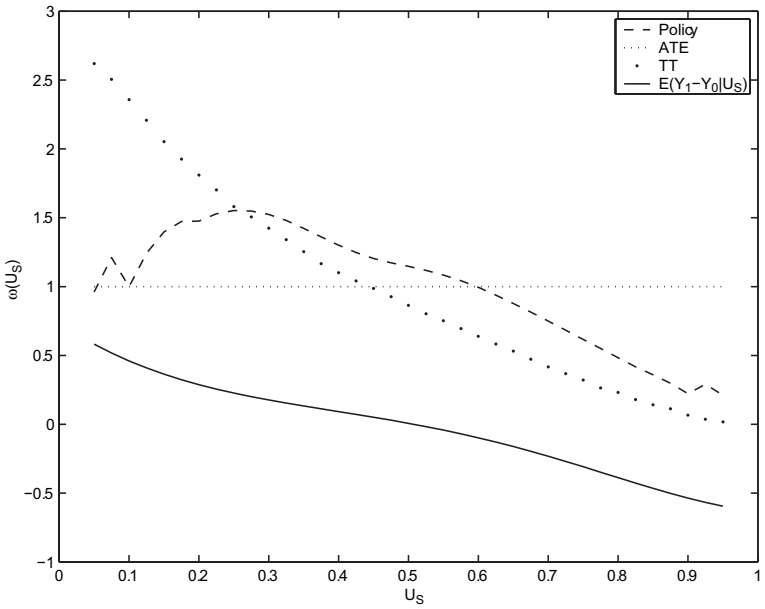
This figure shows the density of P when all instruments are allowed to vary and the variables in X are fixed at different values. We group all Xs in an index, and consider a low and a high value of the index. In particular, the schooling equation takes the following form: $S = 1$ if $X\gamma_1 + Z\gamma_2 + ZX\gamma_3 + \varepsilon > 0$ where Z is the vector of instruments. We pick $X\gamma_1$ as the index of X and we compute the percentiles of its distribution. Then we get all observations for which $X\gamma_1$ is between the 20^{th} and 30^{th} percentiles of its distribution, we compute the average $X\gamma_1$ in this group and call it Low X in the figure, and we allow Z to vary within this set of observations. This generates the density of P for Low X. For High X we proceed analogously, but we take observations for which $X\gamma_1$ is between the 70^{th} and 80^{th} percentiles of its distribution. We allow Z to vary within the groups of observations with low and high X generating two densities of P. Since the MTE also varies with X there are two MTEs at two different levels (although both of them are rescaled to fit the picture).

Figure 6: IV Weights for Different Instruments



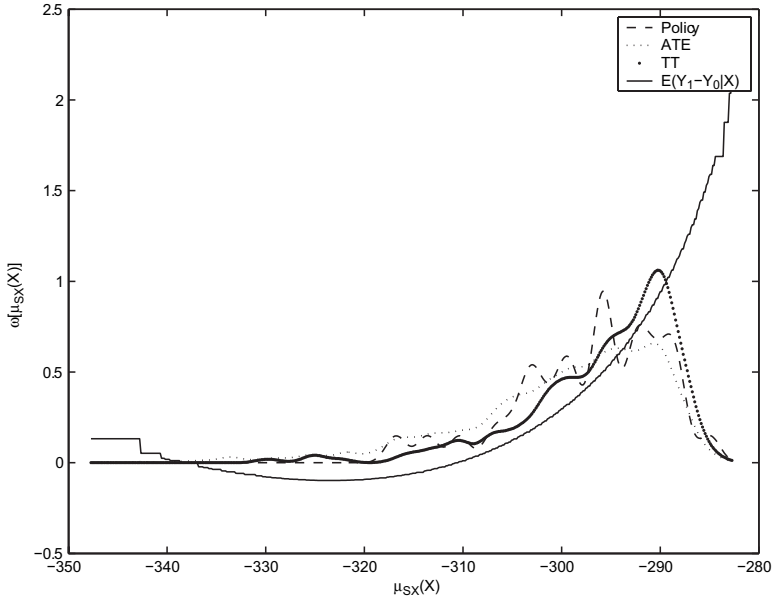
Note: We denote weight by $\omega(\cdot)$. The scale of the y-axis is the scale of the parameter weights, not the scale of the MTE. MTE is scaled to fit the picture. Weights are calculated using the formulae in Heckman, Urzua and Vytlacil (2006)

Figure 7: ATE, TT and Policy Weights for $E(Y_1 - Y_0|X, U_S)$ (Averaged Over X)



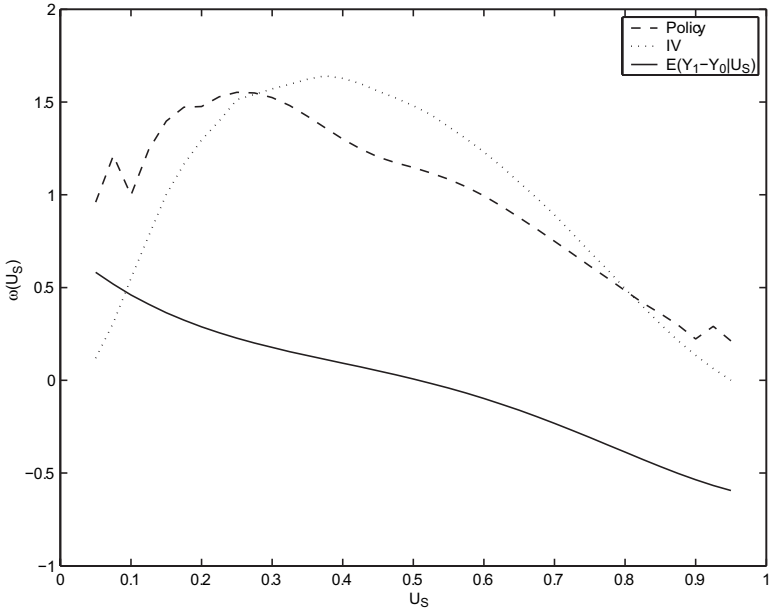
Note: We denote weight by $\omega(\cdot)$. The scale of the y-axis is the scale of the parameter weights, not the scale of the MTE. MTE is scaled to fit the picture.

Figure 8: ATE, TT and Policy Weights for $E(Y_1 - Y_0|X, U_S)$ (Averaged Over U_S)



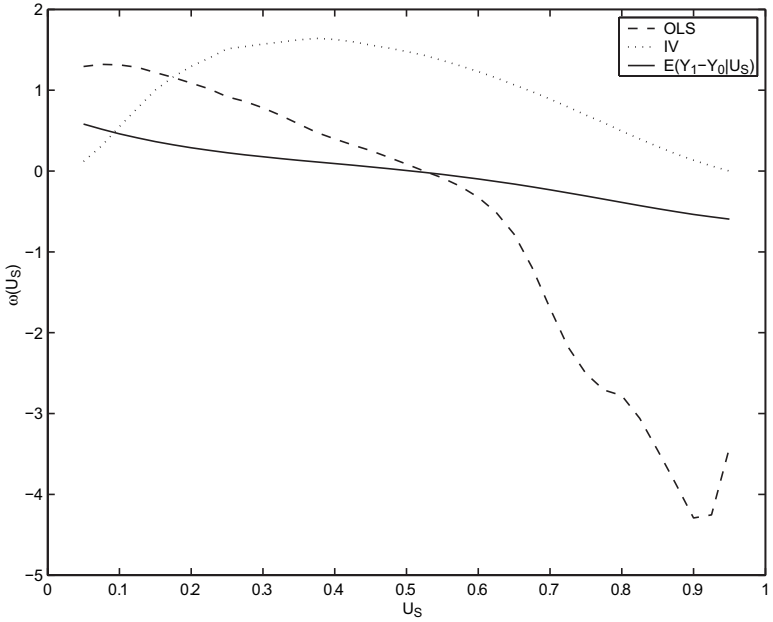
Note: We denote weight by $\omega(\cdot)$. The scale of the y-axis is the scale of the parameter weights, not the scale of the MTE. MTE is scaled to fit the picture.

Figure 9: Policy and IV Weights for $E(Y_1 - Y_0|X, U_S)$
(Averaged Over X)



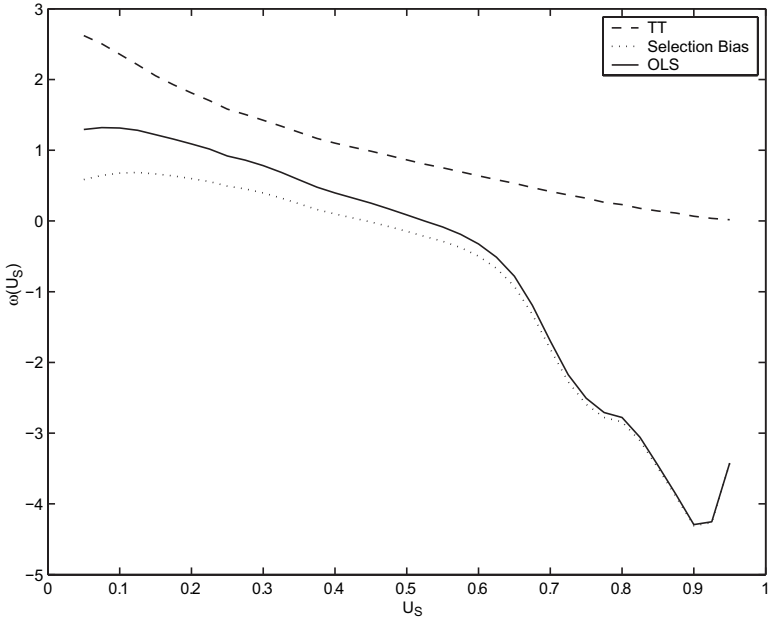
Note: We denote weight by $\omega(\cdot)$. The scale of the y-axis is the scale of the parameter weights, not the scale of the MTE. MTE is scaled to fit the picture.

Figure 10: OLS and IV Weights for $E(Y_1 - Y_0|X, U_S)$
(Averaged Over X)



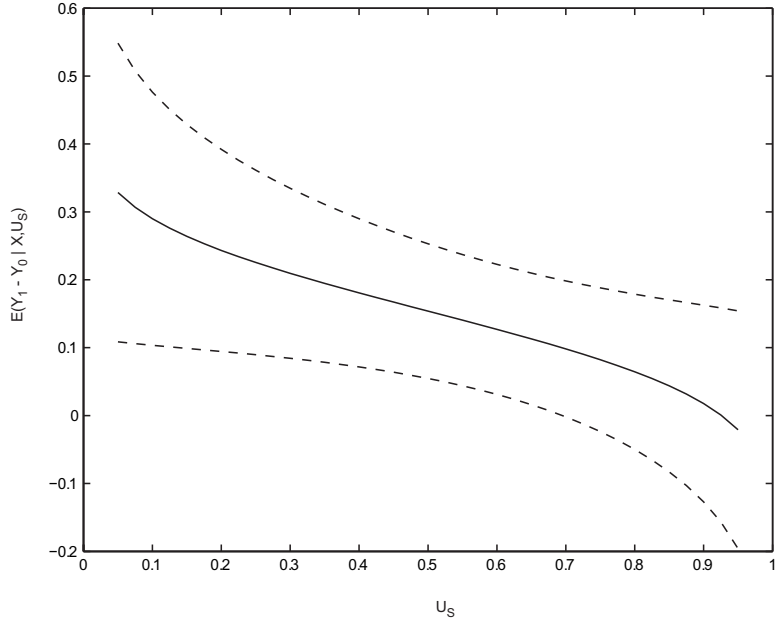
Note: We denote weight by $\omega(\cdot)$. The scale of the y-axis is the scale of the parameter weights, not the scale of the MTE. MTE is scaled to fit the picture. The OLS weight is divided by 100 in order to fit this figure.

Figure 11: Decomposition of OLS Weights for $E(Y_1 - Y_0|X, U_S)$ (Averaged Over X)



Note: We denote weight by $\omega(\cdot)$. The OLS and the Selection Bias weights are scaled by 100 in order to fit this figure.

Figure 12: $E(Y_1 - Y_0|X, U_S)$ Estimated Using a Normal Selection Model (Averaged Over X)



To estimate the function plotted here we use a regression of log wages on polynomials in X, interactions of polynomials in X and P, and $K(P)$, a function of P (where P is the predicted probability of attending college) which is derived from a normal selection model. X includes years of experience, corrected AFQT, mother's education, cohort dummies, average unemployment in the state of residence and average log earnings in the SMSA of residence, measured in 1994. Then the figure is generated by computing $K'(P)$. Standard error bands are obtained using the bootstrap (250 replications).

Table A1
Regression of AFQT on Schooling at Test Date
and Completed Schooling

Schooling at Test Date	Coefficient
9	12.6802
	(1.5105)
10	16.9406
	(1.5158)
11	22.0232
	(1.5354)
12	23.1203
	(1.4901)
13 to 15	26.6032
	(1.7298)
16 or greater	29.0213
	(2.1278)

Note: These are coefficients of the AFQT score on schooling at test date and complete schooling:

$$\text{AFQT} = \delta_0 + \Sigma_{ST} D_{ST} \delta_{ST} + \Sigma_{SC} D_{SC} \delta_{SC} + \eta$$

D_{ST} are dummy variables, one for each level of schooling at test date and δ_{ST} are the coefficients on these variables. D_{SC} are dummy variables, one for each level of completed schooling and δ_{SC} are the coefficients on these variables. The omitted category in the table is "less or equal to eight years of schooling."

Table A2
OLS and IV Estimates of the "Effect"
of College Participation on High School Grades

	% A or Above		% B or Above	
	OLS	IV using P	OLS	IV using P
College	0.0493	0.0029	0.0823	-0.0554
	(0.0112)	(0.0594)	(0.0140)	(0.0792)
AFQT	0.0707	0.1323	0.1059	0.1851
	(0.0367)	(0.0399)	(0.0457)	(0.0532)

Note: In this table we present estimates of OLS and IV regressions of High School Grades (% of subjects where the grade was A or above; % of subjects where the grade was B or above) on College Participation, AFQT and its square, Mother's Education and its square, and interaction between Mother's Education and AFQT, Cohort Dummies, Local Earnings in 1994 and Local Unemployment in 1994. We use P (the predicted probability of going to college) as the instrument for college participation.

Table A3
Average Derivatives of the Wage Equation
(Semi-Parametric Model)

	$\frac{\partial \mu_0(X)}{\partial X_j}$	$\frac{\partial [\mu_1(X)-\mu_0(X)]}{\partial X_j}$
Years of Experience	0.0110	-0.0134
	(0.0090)	(0.0154)
SMSA Log Earnings in 1994	0.6570	0.0010
	(0.1049)	(0.1298)
State Unemployment Rate in 1994	0.0310	-0.0410
(in %)	(0.0304)	(0.0486)
Corrected AFQT	-0.4534	0.5136
	(0.3644)	(0.5373)
Mother's Education	-0.0221	0.0371
	(0.0445)	(0.0579)

Note: The estimates reported in this table come from a regression of log wages on polynomials in experience, corrected AFQT, mother's education, cohort dummies, local earnings and local unemployment in 1994, and interactions of these polynomials with P (where P is the predicted probability of attending college), and $K(P)$, a nonparametric function of P. We use Robinson's (1988) method for estimating a partially linear model. We report the average derivatives of each variable in $\mu_0(X)$ and $\mu_1(X) - \mu_0(X)$ (the average marginal effect of each variable on high school wages and on the returns to college). Bootstrapped standard errors (in parentheses) are presented below the corresponding coefficients (250 replications).

Table A4
Estimates of Various Returns to One Year of College

	$0.0541 < P < 0.9662$	Extrapolate MTE
Average Treatment Effect	0.1832	0.1800
Treatment on the Treated	0.2165	0.2249
Treatment on the Untreated	0.1672	0.1541
Average Marginal Treatment Effect	0.1793	0.1769
Policy Relevant Treatment Effect	0.2013	0.1973
Bounds for ATE	$0.0541 < P < 0.9662$	(0.1037;0.2194)
	$0.1 < P < 0.9$	(0.0402;0.3034)

The numbers on the first column of this table are exactly the same ones reported in Table 5. In the second column of the table, we report estimates of the treatment parameters when we extrapolate the MTE so that it exists for the whole support of U_S. In particular, we extend the MTE at the right and left tails of the function by evaluating the function over the whole support, even for points where there is very little available data. The bounds presented at the bottom of the table are computed using expressions derived in Heckman and Vytlacil (2000).

Table A5
Average Derivatives of the College Decision Model
(Normal Model)

Corrected AFQT	0.2310
	(0.0116)
Mother's Years of Schooling	0.0407
	(0.0057)
Presence of a College at 14	0.0958
	(0.0226)
SMSA Log Earnings at 17	-0.1512
	(0.0707)
State Unemployment Rate at 17	0.0088
(in %)	(0.0063)
Chi-Squared test for joint significance of instruments	42.63
p-value	0.0009

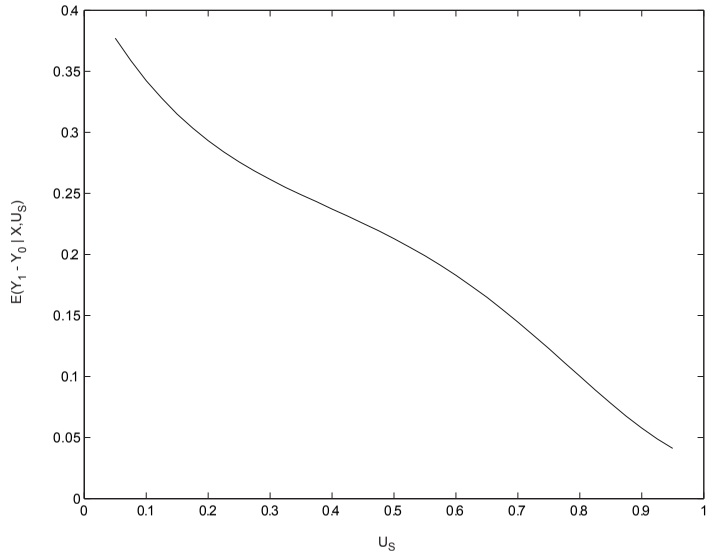
Note: This table reports the average marginal derivatives from a probit regression of college attendance (a dummy variable that is equal to 1 if an individual has ever attended college and equal to 0 if he has never attended college but has graduated from high school) on polynomials in the set of variables listed in the table and on cohort dummies. For each individual we compute the effect of increasing each variable by one unit (keeping all the others constant) on the probability of enrolling in college and then we average across all individuals. Bootstrapped standard errors (in parentheses) are presented below the corresponding coefficients (250 replications).

Table A6
Average Derivatives of the Wage Equation
(Normal Model)

	$\frac{\partial \mu_0(X)}{\partial X_j}$	$\frac{\partial [\mu_1(X)-\mu_0(X)]}{\partial X_j}$
Years of Experience	-0.0015	-0.0041
	(0.0040)	(0.0068)
SMSA Log Earnings in 1994	0.6142	0.0201
	(0.1207)	(0.1532)
State Unemployment Rate in 1994	-0.0075	0.0300
(in %)	(0.0199)	(0.0239)
Corrected AFQT	-0.1320	0.2762
	(0.0887)	(0.1022)
Mother's Education	-0.0244	0.0452
	(0.0198)	(0.0235)

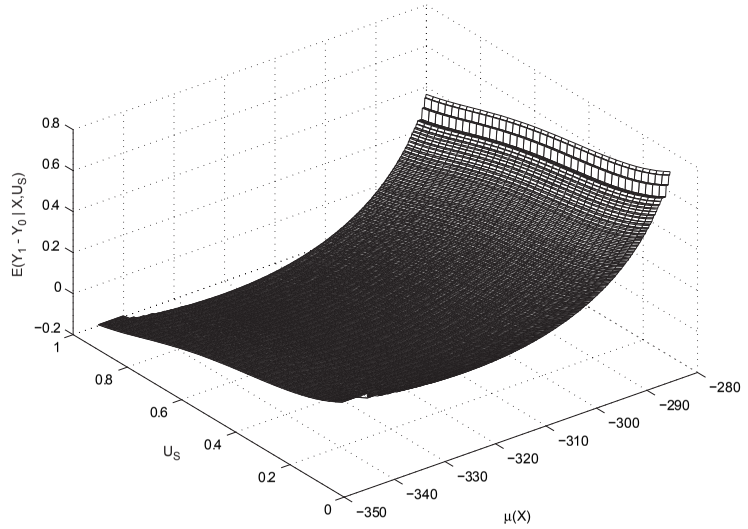
Note: The estimates reported in this table come from a regression of log wages on polynomials in experience, corrected AFQT, mother's education, cohort dummies, local earnings and local unemployment in 1994, and interactions of these polynomials with P. (where P is the predicted probability of attending college), and $K(P)$, a function of P which we derive assuming that the unobservables of the model are jointly normally distributed. We report the average derivatives of each variable in $\mu_0(X)$ and $\mu_1(X) - \mu_0(X)$ (the average marginal effect of each variable on high school wages and on the returns to college). Bootstrapped standard errors (in parentheses) are presented below the corresponding coefficients (250 replications).

Figure A1: $E(Y_1 - Y_0|X, U_S)$ Estimated Using Locally Quadratic Regression (Averaged Over X)



Note: To estimate the function plotted here we first use a partially linear regression of log wages on polynomials in X, interactions of polynomials in X and P, and $K(P)$, a locally quadratic function of P (where P is the predicted probability of attending college), with a bandwidth of 0.25. X includes years of experience, corrected AFQT, mother's education, cohort dummies, average unemployment in the state of residence and average log earnings in the SMSA of residence, measured in 1994. Then the figure is generated by taking the coefficient on the linear term in P from $K(P)$.

Figure A2: $E(Y_1 - Y_0|X, U_S)$ Estimated by Locally Quadratic Regression



Note: To estimate the function plotted here we first use a partially linear regression of log wages on polynomials in X, interactions of polynomials in X and P, and $K(P)$, a locally quadratic function of P (where P is the predicted probability of attending college), with a bandwidth of 0.25. X includes years of experience, corrected AFQT, mother's education, cohort dummies, average unemployment in the state of residence and average log earnings in the SMSA of residence, measured in 1994.