Understanding What Instrumental Variables Estimate: Estimating Marginal and Average Returns to Education

Pedro Carneiro
University of Chicago

James J. Heckman*
University of Chicago and
The American Bar Foundation

Edward Vytlacil
Stanford University

October, 2000,
Revised, April, 2001,
Revised, May, 2001 and
July 19, 2003.

*This research was supported by NSF 97-09-873, NSF-SES-0099195 and NICHD-40-4043-000-85-261. Carneiro benefited from support from Fundaçao Ciência e Tecnologia and Fundaçao Calouste Gulbenkian. We have benefitted from comments received at the Applied Price Theory Workshop, from comments received from Jaap Abbring, Flavio Cunha, Sebastian Gay, Michael Greenstone, Larry Katz, Steve Levitt, Robert Moffitt, Kevin Murphy, Derek Neal, Sergio Urzua and from participants at the Royal Economic Society, Durham, England, April 10, 2001; especially those of Richard Blundell and Costas Meghir. Jingjing Hsee, Maria Isabel Larenas and Maria Victoria Rodriguez provided excellent research assistance.

Abstract

This paper develops and applies new methods for estimating marginal and average returns to economic activities when returns vary in the population and people sort into these activities with at least partial knowledge of their returns. Different valid instruments identify different parameters which do not, in general, answer well-posed economic questions or identify traditional treatment effects. We start with a well-posed economic question and develop methods for answering it. We extend the standard instrumental variables literature to estimate marginal returns and to construct policy relevant parameters. Applying our methods to an analysis of the economic returns to college education, 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

Pedro Carneiro
Department of Economics
University of Chicago
1126 E. 59th Street
Chicago IL 60637
Phone: (773) 256-6268
Fax: (773) 256-6313
Email: pmcarnei@uchicago.edu

James J. Heckman
Department of Economics
University of Chicago
1126 E. 59th Street
Chicago IL 60637
Phone: (773) 702-0634
Fax: (773) 702-8490
Email: jjh@uchicago.edu

Edward Vytlacil
Department of Economics
Stanford University
231 Landau Economics Building
579 Serra Mall
Stanford, CA 94305
Phone: (650) 725-7836
Fax: (650) 725-5702
Email: vytlacil@stanford.edu

Economics is all about returns at the margin. Yet most empirical work on returns in economics estimates average returns. This paper develops methods for estimating both marginal and average returns to economic activities. We apply our methods to estimate the return to education for persons at the margin of attending college. We contrast the higher return earned by all college goers with the lower return earned by marginal entrants to college.

This paper contributes to an emerging literature that documents that people respond differently to the same policy, intervention, or economic choice.[1] There is no single "effect" of a choice but rather a distribution of effects. There are many ways to summarize this distribution. A major contribution of this paper is to define summary measures that answer policy relevant questions and to contrast these measures with those produced from conventional instrumental variable estimators.

The distinction between the average and the marginal return is an economically very important one, and can be illustrated by the following example. Suppose we consider schooling choices which can take only two values ($S = 0$ or $S = 1$) and let R be the absolute dollar return and C be the dollar cost of going to school. Assume that R varies in the population but everyone faces the same C. Individuals decide to enroll in school ($S = 1$) if $R - C > 0$. Figure 1 plots the density of R, $f(R)$, and also presents the cost everyone faces, C. Individuals who have values of R to the right of C choose to enroll in school, while those to the left choose not to enroll. The average return for the individuals who choose to go to school, $E(R \mid R \geq C)$, is computed with respect to the part of $f(R)$ that is to the right of C. The marginal return (the return for individuals at the margin), is exactly equal to C. Figure 1 presents the average and the marginal return for this example. Suppose that there is a policy such as a tuition subsidy that changes the cost of attending school from C to C'. Those individuals who are induced to enroll in school by this policy have R below C (they were not enrolled in school before the policy) and R above C' (they decide to enroll after the policy), and the average return for these individuals is $E(R|C' < R \leq C)$. In this example, the marginal entrant into college has a lower return than the average entrant. The return for the average student is not the relevant return to evaluate the policy. The goal of this paper is to

[1] See Heckman (2001) for a summary of the evidence from this literature.

estimate marginal and average returns when there is self-selection into economic sectors.

The method of instrumental variables (IV) is the most commonly used method to control for the econometric problems of endogeneity and self selection. In the standard regression model for outcome $\ln Y$ as a function of scalar S,

$$\ln Y = \alpha + S\beta + U$$

where α, β are parameters and where S is correlated with mean zero error U, least squares estimators of β are biased and inconsistent. Economists since Haavelmo (1943) have defined the "causal effect," or "effect" of S on $\ln Y$, as β. This corresponds to a manipulation of S holding U fixed - what Marshall (1890) called a "ceteris paribus" effect of S on $\ln Y$. 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. Both 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. 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 may be different from the average β. In this case, there is no single "effect" of S on $\ln Y$ and different estimators produce different scalar summary measures of the distribution of β. In the empirical work reported in this paper, S is schooling, and $\ln Y$ is log earnings, so in the example motivating Figure 1, $R = \left(e^{\beta} - 1\right) e^{\alpha + U}$. We estimate marginal and average returns to schooling when β varies in the population and is correlated with S. Our methods apply more generally to the estimation of a wide variety of returns including those to migration, unionism, and medical care, and outcomes may be discrete or continuous.[2]

We compare economically motivated parameters with the estimands produced by instrumental variable estimators and find that conventional IV does not, in general, answer well posed economic questions, although by accident it may sometimes do so. Only if the instrument is the same as the policy being studied, and the policy is exogenously imposed, does the instrument identify the effect

[2] Carneiro, Hansen and Heckman (2003) estimate entire distributions of returns to schooling.

of the exogenously imposed policy on the outcome being studied. Different valid instruments define different parameters, all of which can be called "effects" of S, but which only rarely answer well-posed economic problems. We show how to use the economic theory of choice to combine multiple instruments into a scalar instrument. We use this instrument to improve on conventional IV to estimate economically interpretable average and marginal returns and to estimate conventional treatment parameters.

We use the Marginal Treatment Effect (MTE), introduced in Björklund and Moffitt (1987) and extended in Heckman and Vytlacil (1999, 2000), to construct estimates of marginal and average returns, to construct policy relevant parameters and to characterize what instrumental variable methods estimate. Our empirical analysis of the returns to schooling is of interest in its own right. In it, we establish that (a) comparative advantage or self-selection is an empirically important feature of schooling choice, (b) marginal college attendees earn less than average attendees and the fall off in their returns is sharp, (c) OLS ("Mincer") and conventional IV estimators substantially underestimate the average marginal return and policy relevant effects, (d) support problems (limitations on the ranges of instrumental variables) compromise the ability of analysts to estimate conventional summary measures of returns, such as the average return to schooling in the population, but not marginal returns which in general are economically more interesting, and (e) many of the instruments used in the recent literature on estimating the returns to schooling are questionable, given the absence of ability measures in most data sets.

The plan of the paper is as follows. Section 1 contrasts two basic models that are currently used in the empirical literature: a common coefficient model and a random (or variable) coefficient model. These models motivate the empirical work we report in this paper. In this section, we also define the Benthamite policy parameter estimated in this paper. Section 2 characterizes the nonparametric framework and assumptions that we will use for the rest of the paper. The framework we use is the one developed by Heckman and Vytlacil (1999, 2000, 2001b, 2004a,b). Section 3 presents the policy relevant treatment effect introduced in Heckman and Vytlacil (2001b) that is a central object of attention in this paper. Section 4 asks and answers the question "What Does The Instrumental Variable Estimator Estimate?" Section 5 shows how to estimate

the marginal treatment effect (MTE) which is the building block for all of our analyses. Section 6 discusses estimates of marginal and average returns to schooling. Section 7 discusses limitations of the empirical literature using instrumental variables to estimate the returns to schooling when ability measures are not available and documents the empirical importance of ability bias. Section 8 concludes.

1 Models with Heterogeneous Returns to Schooling

The familiar semilog specification of the earnings-schooling equation popularized by Mincer (1974), and used in the introduction, writes log earnings $\ln Y$ as a function of S. The framework developed in this paper applies to a general class of models analyzing the consequences of economic choice. In this paper, S will be binary corresponding to two schooling (or treatment) levels ($S = 0$ "high school" or $S = 1$ "college") to simplify the exposition and connect to the empirical work reported in Section 6. For simplicity throughout this paper we suppress explicit notation for dependence of the parameters on the covariates X unless it is clarifying to make this dependence explicit. Under special conditions discussed in Willis (1986) and Heckman, Lochner and Todd (2001), β is the rate of return to schooling.[3] Our methods apply more generally to analyzing returns to unionism, migration, job training, medical interventions, and the like, and the outcomes may be discrete or continuous.

When β is a constant for all persons (conditional on X), we obtain the conventional model. Measured S may be correlated with unmeasured U because of omitted ability factors and because of measurement error in S. Following Griliches (1977), many advocate using instrumental variable estimators for β to alleviate these problems. In this framework, because β is a constant, there is a unique effect of schooling. Indeed, β is "the" effect of schooling, and the marginal return is the same as the average return (conditional on X).

In terms of a model of counterfactual states or potential outcomes of the sort developed in the

[3] $R' = e^{\alpha+U}\left(e^{\beta}-1\right)$ is the absolute return and $R = \left[e^{\alpha+U}\left(e^{\beta}-1\right)\right]/e^{\alpha+U} = e^{\beta}-1 \doteq \beta$ is the rate of return to schooling, where $e^{\alpha+U}$ is earnings when S is fixed at 0.

Roy (1951) model, there are two potential outcomes $(\ln Y_0, \ln Y_1)$:

$$\ln Y_0 = \alpha + U, \quad \ln Y_1 = \alpha + \beta + U \tag{1}$$

and causal effect $\ln Y_1 - \ln Y_0 = \beta$ is a common effect, conditional on X.

From its inception, the modern literature on the returns to schooling has recognized that returns may vary across schooling levels and across persons of the same schooling level.[4] This early literature was not clear about the sources of variation in β. The Roy model, as applied by Willis and Rosen (1979), gives a more precise notion of why β varies and how it depends on S. In that framework, the potential outcomes are generated by two random variables (U_0, U_1) instead of one as in the common coefficient model:

$$\ln Y_0 = \alpha + U_0 \tag{2a}$$

$$\ln Y_1 = \alpha + \bar{\beta} + U_1 \tag{2b}$$

where $E(U_0) = 0$ and $E(U_1) = 0$ so α $(= E(\ln Y_0))$ and $\alpha + \bar{\beta}$ $(= E(\ln Y_1))$ are the mean potential outcomes for $\ln Y_0$ and $\ln Y_1$ respectively. The causal effect of educational choice $S = 1$ is

$$\beta = \ln Y_1 - \ln Y_0 = \bar{\beta} + U_1 - U_0.$$

There is a distribution of returns across individuals.

Observed earnings are

$$\ln Y = S \ln Y_1 + (1 - S) \ln Y_0 = \alpha + \beta S + U_0 = \alpha + \bar{\beta} S + \{U_0 + S(U_1 - U_0)\}. \tag{3}$$

In the Roy framework, the choice of schooling is explicitly modeled. In its simplest form

$$\begin{array}{rcl} S & = & 1 \text{ if } \ln Y_1 \geq \ln Y_0 \Longleftrightarrow \beta \geq 0 \\ & = & 0 \text{ otherwise.} \end{array} \tag{4}$$

If agents know or can partially predict β at the time they make their schooling decisions, there is dependence between β and S in equation (3). This justifies the title "correlated random coefficient model" that is often applied to general versions of (3). Decision rules similar to (4) characterize other economic choices.

[4] See Becker and Chiswick (1966), Chiswick (1974) and Mincer (1974).

In this setup there are three sources of potential econometric problems; (a) S is correlated with U_0; (b) β is correlated with S (*i.e.*, $U_1 - U_0$ is correlated with S); (c) β is correlated with U_0. Source (a) arises in ability bias or measurement error models. Source (b) arises if agents partially anticipate β when making schooling decisions so that $\Pr(S = 1 \mid X, \beta) \neq \Pr(S = 1 \mid X)$. In this framework, β is an ex post causal effect. Ex ante agents may not know β. In the case where decisions about S are made in the absence of information about β, β is independent of S. ($\beta \perp\!\!\!\perp S$ where " $\perp\!\!\!\perp$"denotes independence).

Source (c) arises from the possibility that the gains to schooling (β) may be dependent on the level of earnings in the unschooled state as in the Roy model. The best unschooled (those with high U_0) may have the lowest return to schooling.

When β varies in the population, the return to schooling is a random variable and there is a distribution of causal effects. There are various ways to summarize this distribution and, in general, no single statistic will capture all aspects of the distribution.

Many summary measures of the distribution of β are used. Among them are

$$\begin{aligned} E(\beta \mid X = x) &= E(\ln Y_1 - \ln Y_0 \mid X = x) \\ &= \bar{\beta}(x) \end{aligned}$$

the return to the population average person given characteristics $X = x$. This quantity is sometimes called "the" causal effect of S.[5] Others report the return for those who attend school:

$$\begin{aligned} E(\beta \mid S = 1, X = x) &= E(\ln Y_1 - \ln Y_0 \mid S = 1, X = x) \\ &= \bar{\beta}(x) + E(U_1 - U_0 \mid S = 1, X = x).^{6} \end{aligned}$$

This is the parameter emphasized by Willis and Rosen (1979) where $E(U_1 - U_0 \mid S = 1, X = x)$ is the sorting gain, how people who take $S = 1$ differ from a randomly sampled person.

Other parameters are the return for those who are currently not going to school:

$$\begin{aligned} E(\beta \mid S = 0, X = x) &= E(\ln Y_1 - \ln Y_0 \mid S = 0, X = x) \\ &= \bar{\beta}(x) + E(U_1 - U_0 \mid S = 0, X = x). \end{aligned}$$

[5] It is the Average Treatment Effect (*ATE*) parameter. Card (1999, 2001) defines it as the "true causal effect" of education. See also Angrist and Krueger (2001).

Angrist and Krueger (1991) and Meghir and Palme (2001) estimate this parameter. In addition to these "effects" is the effect for persons indifferent between the two levels of schooling, which in the simple Roy model is

$$E\left(\ln Y_1 - \ln Y_0 \mid \ln Y_1 - \ln Y_0 = 0\right) = 0.$$

A more general expression for this marginal effect incorporating discounting and tuition costs is given in the next section.

Depending on the conditioning sets and the summary statistics desired, a variety of "causal effects" can be defined. Different causal effects answer different economic questions. As noted by Heckman and Robb (1986;2000), and Heckman (1997), under two conditions

I: $U_1 = U_0$ (common effect model)
or more generally
II: $Pr(S = 1 \mid X = x, \beta) = Pr(S = 1 \mid X)$ (conditional on X, β does not affect choices)

all of the mean treatment effects conditional on X collapse to the same parameter. Otherwise there are many candidates for the title of causal effect and this has produced considerable confusion in the empirical literature as different analysts use different definitions in reporting empirical results so the different estimates are not strictly comparable.[7]

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, where β is a constant, $\bar{\beta}$, the mean change in welfare is

$$E(\ln Y \mid \text{Alternative Policy}) - E(\ln Y \mid \text{ Baseline Policy}) = \bar{\beta}(\Delta P) \tag{5}$$

[7] For example, Heckman and Robb (1985) note that in his survey of the union effects on wages, Lewis (1986) confuses these different "effects." This is especially important in his comparison of cross section and longitudinal estimates where he inappropriately compares conceptually different parameters.

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 $\bar{\beta}$. This is also the mean change in log income if β is a random variable but independent of S so condition II applies.

In the general case, when agents partially anticipate β, and comparative advantage dictates schooling choices, none of the traditional treatment parameters plays the role of $\bar{\beta}$ in (5) or answers the stated economic question. Instrumental variables methods do not generally identify $\bar{\beta}$. Later in this paper, we develop the appropriate policy parameter, show how to estimate it, and contrast it with conventional treatment parameters and what IV estimates. We first introduce our framework and assumptions.

2 A Framework for Evaluating the Effects of Schooling

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, 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 costs which may include psychic costs, r is the discount rate, and lifetimes are assumed to be infinite to simplify the expressions. This is the prototypical discrete choice model applied to human capital investments. We follow Mincer (1974) and assume that earnings profiles in logs are parallel in experience. 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.

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 $C = \frac{C^*}{K}$, and define discount factor $\gamma = \frac{1}{(1+r)}$. Using growth rate g to relate potential earnings in the two schooling choices we may write $Y_1 = (1+g)Y_0$

where from equation (1), $\beta = \ln(1+g)$. Thus the decision to attend school ($S=1$) is made if

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

This is equivalent to

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

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

$$\beta \geq r + \frac{C}{Y_0}. \tag{6}$$

Ceteris paribus, a higher r or 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 and leads people to attend college. Equation (6) generalizes decision rule (4) by adding borrowing and tuition costs as determinants of schooling. Assuming $C=0$, the marginal return for those indifferent between going to school and facing interest rate r is $E(\beta|\beta=r)$. Below we introduce variables Z that shift costs and discount factors ($C=C(Z)$, $r=r(Z)$).

The conventional approach to estimating selection models postulates normality of (U_0, U_1) in equations 2(a) and 2(b), writes $\bar{\beta}$ and α as linear functions of X and postulates independence between X and (U_0, U_1). Parallel normality and independence assumptions are made for the unobservables and observables in selection equation (6). From estimates of the structural model, it is possible to answer a variety of economic questions and to construct the various treatment parameters and distributions of treatment parameters.[8] However these assumptions are often viewed as unacceptably strong.

A major advance in the recent literature in econometrics is the development of frameworks that relax conventional linearity, normality and separability assumptions to estimate various economic parameters. In this paper, we draw on the framework developed by Heckman and Vytlacil (1999, 2000).

[8] Willis and Rosen (1979) is an example of the application of the Roy model. Aakvik, Heckman and Vytlacil (2000), Heckman, Tobias and Vytlacil (2001;2003) derive all of the treatment parameters and distributions of treatment parameters for several parametric models including the normal. Carneiro, Hansen and Heckman (2003) estimate the distribution of treatment effects under semiparametric assumptions.

Using their setup we write

$$\ln Y_1 = \mu_1(X, U_1) \text{ and } \ln Y_0 = \mu_0(X, U_0). \tag{7}$$

The return to schooling is $\ln Y_1 - \ln Y_0 = \beta = \mu_1(X, U_1) - \mu_0(X, U_0)$, which is a general nonseparable function of (U_1, U_0). It is not assumed that $X \perp\!\!\!\perp (U_0, U_1)$ so X may be correlated with the unobservables in potential outcomes. Here and throughout this paper we use " $\perp\!\!\!\perp$" to denote statistical independence.

A latent variable model determines enrollment in school (this is the nonparametric analogue of decision rule (6)):

$$\begin{aligned} S^* &= \mu_S(Z) - U_S \\ S &= 1 \text{ if } S^* \geq 0. \end{aligned} \tag{8}$$

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, U_S) are unobserved. U_S 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.

Heckman and Vytlacil (2000, 2004b) establish that under the following assumptions, it is possible to develop a model that unifies different treatment parameters, that shows how the conventional IV estimand relates to these parameters and what policy questions IV answers. Those conditions are

(A-1) $\mu_S(Z)$ is a nondegenerate random variable conditional on X;

(A-2) The distribution of U_S is absolutely continuous with respect to Lebesgue measure;

(A-3) (U_0, U_1, U_S) is independent of Z conditional on X;

(A-4) $\ln Y_1$ and $\ln Y_0$ have finite first moments

and

(A-5) $1 > \Pr(S = 1 \mid X) > 0$.

Assumption (A-1) postulates the existence of an "instrument" - more precisely a variable or set of variables that are in Z but not in X, and thus shift S^* but not potential outcomes Y_0, Y_1 (which

are determinants of C and r in equation (6)). The recent empirical literature on the returns to schooling also assumes the existence of instruments. The literature on natural experiments and social experiments uses, respectively, natural experimental variation in variables or social experimental induced variation in treatment assignments as instruments. Assumption (A-2) is made for technical convenience and can be relaxed at greater cost of notation. Assumption (A-3) allows X to be arbitrarily dependent on the errors. X need not be "exogenous" in any conventional definition of that term. However, a no feedback condition is required for the interpretability of the estimates. Defining X_s as the value of X if S is set to s, a sufficient condition for interpretability is that $X_1 = X_0$ almost everywhere. This ensures that conditioning on X does not mask the effect of realized S on outcomes.[9] Assumption (A-4) is necessary for the definition of the mean parameters and assumption (A-5) ensures that in very large samples for each X there will be people with $S = 1$ and other people with $S = 0$, so comparisons of schooling and nonschooling outcomes can be made at all X values.

Denoting $P(z)$ as the probability of receiving treatment $S = 1$ conditional on $Z = z$, $P(z) \equiv \Pr(S = 1|Z = z) = F_{U_S}(\mu_S(z))$. Without loss of generality we may write $U_S \sim$ Unif[0,1] so $\mu_S(z) = P(z)$. Thus with no loss of generality if $S^* = \nu(Z) - V_S$, we can always reparameterize the model so $\mu_S(Z) = F_V(\nu(Z))$ and $U_S = F_V(V)$. Vytlacil (2002) establishes that the model of equations (7), (8) and (A-1) - (A-5) is equivalent to the *LATE* model of Imbens and Angrist (1994)[10]

The index structure produced by assumptions (A-1) - (A-5) joined with the model of equations (7) and (8) allows us to define a new treatment effect: the marginal treatment effect (*MTE*)

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

[9] However this condition is not strictly required. If imposed, it produces the "total effect" of S on Y. See Pearl (2000). Heckman and Vytlacil (2004a,b) relax this condition.

[10] These conditions impose testable restrictions on (Y, S, Z, X). See Heckman and Vytlacil (1999, 2000, 2004b). One restriction is index sufficiency, $\Pr(\ln Y_j \in A \mid X = x, Z = z, S = j) = \Pr(\ln Y_j \in A \mid X = x, P(Z) = P(z), S = j)$ for $j = 0, 1$. This says that Z enters the conditional distribution of $\ln Y_1$, $\ln Y_0$, only through the index $P(Z)$. Another restriction is a monotonicity restriction. Let $g(\cdot)$ denote any function such that $g(\ln Y) > 0$ w.p.1. Then $E(Sg(\ln Y)|X = x, P(Z) = p)$ is strictly increasing in p and $E((1 - S)g(\ln Y)|X = x, P(Z) = p)$ is strictly decreasing in p. For example, if $\ln Y > 0$ w.p.1, then taking $g(\cdot)$ to be the identity function we have that $E(S \ln Y|X = x, P(Z) = p)$ is strictly increasing in p and $E((1 - S) \ln Y|X = x, P(Z) = p)$ is strictly decreasing in p.

This is the mean gain to schooling for individuals with characteristics $X = x$ just indifferent between taking schooling or not at level of unobservable $U_S = u_S$. It is a willingness to pay measure for people at the margin of indifference for schooling given X and U_S.[11] The *LATE* parameter of Imbens and Angrist (1994) may be written in this framework as

$$\Delta^{LATE}(x, u'_S, u_S) = E(\beta \mid X = x, u_S \leq U_S \leq u'_S)$$

where $u_S \neq u'_S$. *MTE* is the limit of *LATE* as $u'_S \to u_S$.[12]

Heckman and Vytlacil (1999, 2000) establish that under assumptions (A-1) - (A-5) 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. We discuss the other weights in later sections.

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), then all of these mean treatment parameters conditional on X are the same. This arises in cases I and II analyzed in Section 1 where, respectively, there is no heterogeneity (β is constant given X) or agents do not act on it.[13]

3 Policy Relevant Treatment Effects

With the framework of Section 2 in hand, we can answer the policy question framed at the end of Section 1, when β is heterogeneous and people act on β in making decisions about S.[14] We focus on this parameter in the empirical work we report below. We consider a class of policy interventions that affect P but not $(\ln Y_1, \ln Y_0)$. This is the standard assumption in the partial equilibrium treatment effect literature.[15]

[11] Björklund and Moffitt (1987) introduced this parameter in the context of the parametric normal Roy model.

[12] Assumptions (A-2) to (A-4) imply that the limit exists (a.e.).

[13] All of these parameters in Tables 1A and 1B can be defined even if (a) $U_S \not\perp\!\!\!\perp Z$ or (b) for $S = 1(\Omega(Z, U_S) \geq 0)$ there is no additively separable version of Ω in terms of U_S, Z or (c) $Z = X$ (no instrument). However, the conditions presented in the text are required to identify the *MTE*. See Heckman and Vytlacil (2000).

[14] Ichimura and Taber (2000) present a related analysis of policy evaluation. Their framework does not use the *MTE* to unify estimators and policy counterfactuals.

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

Let P be the baseline probability of $S = 1$ with density f_P. We keep the conditioning on X implicit. Define P^* as the probability produced under an alternative policy regime with density f_{P^*}. Then we can write

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

where $\omega(u) = F_P(u) - F_{P^*}(u)$ where F_P and F_{P^*} denote the cdf of P and P^*, respectively.[16]

To define a parameter comparable to $\bar{\beta}$ in equation (5), we normalize the weights by ΔP, the change in the proportion of people induced into the program, conditional on $X = x$. Thus if we use the weights

$$\tilde{\omega}(u) = (\omega(u))/\Delta P$$

we produce the gain in the outcome for the people induced to change into (or out of) schooling by the policy change. These weights define the Policy Relevant Treatment Effect ($PRTE$).

Observe that these weights differ from the weights for the conventional treatment parameters. Knowing TT or ATE does not answer a variety of well posed policy questions except in special cases (Heckman and Smith, 1998). We next show that in the general case where β varies among individuals conditional on X and people make schooling decisions based on it, IV weights MTE differently than the weighting required for the $PRTE$ or required to generate the conventional treatment parameters.

4 What Does The Instrumental Variable Estimator Estimate?

The intuition underlying the application of instrumental variables to the common coefficient model is well understood. It is misleading in the more general case where β varies in the population and choices of S are made on the basis of it.

Let W denote a potential instrument. For example, in our framework, W might be an element of Z or any function of Z. In the common coefficient model (1) the econometric problem is that $Cov(U, S) \neq 0$. If there is an instrument W with the properties (a) $Cov(U, W) = 0$ and (b)

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

$Cov\,(W,S) \neq 0$ then we may identify (consistently estimate) β by IV even though OLS is biased and inconsistent. Thus

$$\text{plim}\hat{\beta}_{IV} = \frac{Cov(W, \ln Y)}{Cov(W,S)} = \beta + \frac{Cov(W,U)}{Cov(W,S)} = \beta.$$

This intuition breaks down in the more general case of equation (3):

$$\ln Y = \alpha + \bar{\beta} S + \{(U_1 - U_0)S + U_0\}.$$

Finding an instrument W correlated with S but not U_0 or $U_1 - U_0$ is not enough to identify $\bar{\beta}$, or $\bar{\beta} + E(U_1 - U_0 \mid S = 1)$, or other conventional treatment parameters.[17] Simple algebra reveals that

$$\text{plim}\hat{\beta}_{IV} = \frac{Cov(W, \ln Y)}{Cov(W,S)} = \bar{\beta} + \frac{Cov(W,U_0)}{Cov(W,S)} + \frac{Cov(W, S(U_1 - U_0))}{Cov(W,S)}.$$

By the standard IV condition (a), the second term vanishes ($Cov(W,U_0) = 0$). But in general the third term does not:

$$Cov(W, S(U_1 - U_0))/Cov(W,S) =$$
$$P\Big\{Cov[W, (U_1 - U_0)\,|S = 1] + [E\,(W|S = 1) - E(W)]E(U_1 - U_0|S = 1)\Big\}/Cov(W,S)$$

where $P = \Pr(S = 1)$. If $U_1 - U_0 \equiv 0$ (a common coefficient model, condition I) or more generally if $U_1 - U_0$ is independent of S and W (condition II) this term vanishes.[18] But in general $U_1 - U_0$ is dependent on S and the term does not vanish.[19]

To see why, consider the schooling choice model of equation (6) when $C = 0$ and r depends on Z ($r = Z\gamma$). Take the instrument to be an element of Z, say $W = Z_k$ where Z_k is the kth element of Z. Then

$$S = 1 \Longleftrightarrow \bar{\beta} + U_1 - U_0 \geq Z\gamma,$$

and $Cov(Z_k, U_1 - U_0 \mid S = 1) = Cov(Z_k, U_1 - U_0 \mid \bar{\beta} + U_1 - U_0 \geq Z\gamma)$. One can show that the sign of $Cov(Z_k, U_1 - U_0 \mid \bar{\beta} + U_1 - U_0 \geq Z\gamma)$ is the same as the sign of γ_k, so that this covariance

[17] Recall that we keep the conditioning on X implicit.

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

[19] See Heckman and Robb (1985, 1986; 2000) and Heckman (1997).

will be nonzero for any element of Z with a nonzero γ coefficient. Intuitively, suppose that Z is a scalar with a positive effect on the discount rate r. Then if an individual selects into college despite having a high Z (and thus a high discount rate), then the individual must have a large direct gain from going to college to have chosen to attend college despite the high discount rate. Thus, even if $Z \perp\!\!\!\perp (U_1 - U_0)$, Z is not independent of $U_1 - U_0$ conditional on $S = 1$.

Another way to make this general point is to explore what we estimate by using an instrument based on compulsory schooling. Compulsory schooling is sometimes viewed as an ideal instrument (see Angrist and Krueger 1991). But when returns are heterogeneous, and agents act on that heterogeneity in making schooling decisions, compulsory schooling as an instrument identifies only one of many possible treatment parameters. Define $P(x) = \Pr(S = 1 \mid X = x)$ as the probability of attending school conditional on $X = x$ if there is no compulsion. Let $T = 1$ if the individual is in the regime with compulsion, and $T = 0$ otherwise. We assume that T is exogenous, in the sense that $T \perp\!\!\!\perp (U_S, U_0, U_1)|X$.

Compulsory schooling selects at random persons who ordinarily would not be schooled ($S = 0$) and forces them to be schooled. Observed earnings for individuals in the compulsory schooling regime (conditional on X) are

$$\begin{aligned} &E(\ln Y|X = x, T = 1) \\ &\qquad = \; E(\ln Y_1 \mid X = x, S = 1)P(x) + E(\ln Y_1 \mid X = x, S = 0)(1 - P(x)), \end{aligned}$$

and for individuals in the regime with no compulsion

$$\begin{aligned} &E(\ln Y|X = x, T = 0) \\ &\qquad = \; E(\ln Y_1 \mid X = x, S = 1)P(x) + E(\ln Y_0 \mid X = x, S = 0)(1 - P(x)). \end{aligned}$$

From the difference in conditional means we can identify:

$$E(\ln Y|X = x, T = 1) - E(\ln Y|X = x, T = 0) = (1 - P(x))E(\ln Y_1 - \ln Y_0 \mid X = x, S = 0).$$

Since in a non-compulsory schooling regime we identify $P(x)$, we can identify treatment on the untreated:

$$E(\ln Y_1 - \ln Y_0 \mid X = x, S = 0) = E(\beta|X = x, S = 0)$$

but not $ATE = E(\ln Y_1 - \ln Y_0) = \bar{\beta}$ or treatment on the treated $TT = E(\ln Y_1 - \ln Y_0 \mid X = x, S = 1) = E(\beta | X = x, S = 1)$. However under the two special cases I and II of Section 1, we identify all three treatment parameters because

$$E(\ln Y_0 \mid X = x, S = 0) = \alpha(x), E(\ln Y_1 \mid X = x, S = 0) = \alpha(x) + \bar{\beta}(x)$$

and $TT = ATE = MTE = LATE = PRTE$ because $\Delta^{MTE}(x, u_S)$ does not vary with u_s.

Treatment on the untreated answers an interesting policy question. It is informative about the earnings gains for a policy directed toward those who ordinarily would not attend schooling and who are selected into schooling at random from this pool. If the policy we want to evaluate is compulsory schooling then the instrumental variable estimand and the policy relevant treatment effect coincide. More generally, if the instrumental variable we use is exactly the policy we want to evaluate, then the IV estimand and the policy relevant parameter coincide. But whenever that is not the case, the IV estimand does not identify the effect of the policy when returns vary among people and they make choices of treatment based on those returns. For example, if the policy we want to consider is a tuition subsidy directed toward the very poor within the pool, then an instrumental variable estimate based on compulsory schooling will not be the relevant return to evaluate the policy.[20]

So what exactly does linear IV estimate? Heckman and Vytlacil (2000) establish that linear IV using $P(Z)$ as an instrument (conditional on $X = x$) identifies a weighted average of MTE parameters.

$$\text{plim}\hat{\beta}_{IV} = \Delta^{IV}(x) = \int_0^1 \Delta^{MTE}(x, u) h_{IV}(x, u) du$$

where

$$h_{IV}(x, u) = \frac{(E(P(Z) - E(P(Z)) \mid P(Z) \geq u, X = x)) \Pr(P(Z) \geq u, X = x)}{Var(P(Z) \mid X = x)}$$

and $\int_0^1 h_{IV}(x, u) du = 1$. These weights do not, in general, coincide with the policy weights of Section 3 or the weights for the treatment parameters presented in Table 1B.

[20] Heckman and Vytlacil (2004b) show that for every policy it is possible in principle to define an instrumental variable that generates the correct policy relevant treatment effect. However, such an instrument may not be feasible in any given data set because of support problems. Different policies define different policy relevant instrumental variables.

A closer look at these weights reveals that

$$h_{IV}(x,u) = \frac{\int_u^1 (p - E(P(Z) \mid X = x)) f_P(p \mid X = x) dp}{Var(P \mid X = x)}$$

where $f_P(\cdot|X = x)$ is the density of $P(Z)$ conditional on $X = x$.[21] Using this expression, one can easily see the following properties of h_{IV}. For any given x, h_{IV} is a proper weight as a function of u, $h_{IV}(x,u) \geq 0$ for all $u \in [0,1]$ and $h_{IV}(x,\cdot)$ integrates to one,

$$\int_0^1 h_{IV}(x,u)du = 1.$$

For any given x, $h_{IV}(x,\cdot)$ achieves a maximum value at $E(P(Z)|X = x)$, so that for any x,

$$h_{IV}(x, E(P(Z) \mid X = x)) \geq h_{IV}(x,u) \quad \forall u \in [0,1]$$

with the inequality being strict if $f_P(p|X = x) > 0$ for p in a neighborhood of $E(P(Z) \mid X = x)$. $h_{IV}(x,\cdot)$ places zero weight outside of the support of the distribution of $P(Z)$ conditional on $X = x$,

$$h_{IV}(x, P_x^{Max}) = 0 = h_{IV}(x, P_x^{Min})$$

and

$$h_{IV}(x,p) = 0 \qquad \text{for} \qquad p \leq P_x^{Min}, \qquad p \geq P_x^{Max},$$

where P_x^{Max} and P_x^{Min} are the maximum and minimum of the support of the distribution of $P(Z)$ conditional on $X = x$. For proofs, see Heckman and Vytlacil (2000).[22,23]

We can also fit *OLS* into this framework. Table 1B gives the exact weights for OLS. The *OLS* weights are not guaranteed to be positive or to integrate to one.[24]

[21] For this expression we are assuming that the distribution of $P(Z)$ conditional on X has a density with respect to Lebesgue measure.

[22] Take a more general instrument J, and recenter J so that $E(J) = 0$. Keeping conditioning on X implicit, $\text{plim}\hat{\beta}_{IV} = \frac{E(JY)}{E(JS)}$ where $\text{plim}\hat{\beta}_{IV}(J) = \int MTE(u)h(u;J)du$ and $h(u;J) = \frac{E(J|P(Z)\geq u)\Pr(P(z)\geq u)}{E(JP)}$, $\int_0^1 h(u;J)du = 1$. We have the following properties: (i) $h(u;J)$ non-negative iff $E(J \mid P \geq p)$ weakly increasing in p, (ii) Support $h(u;J) \subseteq [P^{Min}, P^{Max}]$ (Support of P); (iii) defining $T(p) = E(J \mid P = p)$, we have $h(u;J) = h(u;T(P))$. (See Heckman and Vytlacil, 2000, 2004b).

[23] The idea of interpreting IV as a weighted average of the limit of *LATE* can also be found in Card (1999, 2001) (weighted average of the distribution of return to schooling), Angrist, Graddy and Imbens (2000) (weighted average of Wald estimators) and Yitzhaki (1996, 1999). However, these authors do not relate the weights to those of the policy relevant treatment effects or to the weights required to estimate the conventional treatment parameters.

[24] Moreover, they are also not defined for values of u_S where $MTE(x,u_S) = 0$.

Observe that if Z is a vector, then given our conditions (A-1) to (A-5) we have that any element of Z that is not an element of X is a valid instrument according to the traditional definition of an instrument. And yet each possible element of Z will result in an IV estimand that is weighting MTE differently and hence estimating a different object than the IV estimand corresponding to any other element of Z. We illustrate this point in the empirical analysis reported in section 6.

This dependence of estimated parameters on the choice of instruments is a central feature of a model that fails condition I or II - a correlated random coefficient model.[25] This highlights a central point of this paper. When returns vary in the population, and are correlated with the choice of activity (S), different summary measures (in this case different instrumental variable estimators) of the distribution exist.

Summarizing the paper thus far, under assumptions (A-1) - (A-5) and the model of equations (7) and (8), the IV estimand, the policy relevant treatment effect, and the conventional treatment parameters are all weighted averages of the MTE. Using the MTE we unify the estimation, treatment effect and policy evaluation literatures as generating parameters or estimands as integrals of MTE using different weights:

$$\text{Estimand } j \text{ or parameter } j \text{ (given } X) = \int_0^1 \Delta^{MTE}(x, u_S)\omega_j(x, u_S)du_S$$

where different estimands or different treatment parameters correspond to different weights $\omega_j(x, u_S)$.

Table 1 summarizes a central result of this paper and the various weights for the different estimands and parameters. The treatment effect parameters weight MTE differently than what is required to produce the policy relevant treatment effect. Thus the conventional treatment parameters do not, in general, coincide with the policy relevant parameters. The weighting for the OLS or IV estimand do not correspond to the weights required to generate the policy relevant treatment parameters.

Figure 2A plots the MTE and the weights used to form ATE, TT and TUT for a generalized Roy model (with tuition costs) with the parameter values displayed at the base of Table 2.[26]

[25] Angrist, Graddy and Imbens (2000) also emphasize the dependence of the IV estimand on the choice of instruments in a random coefficient framework.

[26] The form of the Roy model we use assumes additive separability and generates U_0,U_1 and U_S from a common unobservable ε. Thus the distribution of $U_1 - U_0$ given U_S is degenerate.

This is the model of equation (3) with decision rule (6). TT overweights the MTE for persons with low values of U_S who, *ceteris paribus,* are more likely to attend school. TUT overweights the MTE for persons with high values of U_S who are less likely to attend school. ATE weights MTE evenly. The decline in MTE reveals that the gross return (β) declines with U_S. Those more likely to attend school (based on lower U_S) have higher gross returns. Not surprisingly, in light of the shape of MTE and the shape of the weights, $TT > ATE > TUT$. See Table 2. There is a positive sorting gain ($E(U_1 - U_0 \mid X = x, S = 1)$) and a negative selection bias ($E(U_0 \mid X = x, S = 1) - E(U_0 \mid X = x, S = 0)$). Figure 2B displays the MTE and the OLS and IV weights using $P(Z)$ as the instrument. IV weights the MTE more symmetrically and in a different fashion than ATE, TUT or TT. OLS weights MTE very differently.

The most direct way to produce the policy relevant treatment parameters is to estimate MTE directly and then generate all of the treatment effect parameters using the appropriate weights. We develop a strategy for doing this next.[27]

5 Using Local Instrumental Variables to Estimate the MTE

Using equation (3) the conditional expectation of log Y given Z is

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

where we keep the conditioning on X implicit. By the exclusion condition for Z, (A-1), and the index sufficiency assumption embodied in (A-3) and (8), we may write this expectation as

$$E(\ln Y \mid Z = z) = E(\ln Y_0) + E(\beta \mid P(z) \geq U_S, P(Z) = P(z))P(z).$$

Given our assumptions, we have the following index sufficiency restriction:

$$E(\ln Y \mid Z = z) = E(\ln Y \mid P(Z) = P(z)).$$

Applying the Wald estimator for two different values of Z, z and z' assuming $P(z) \neq P(z')$,

[27] We note parenthetically that the method of matching assumes that $\beta \perp\!\!\!\perp S|X$ or $\beta \perp\!\!\!\perp S|X, Z$ where the variables after "|" denote the conditioning sets (see Heckman and Navarro, 2003). It assumes that for all X, or for all X, Z, the marginal return equals the average return and begs the stated question of interest in this paper.

we obtain the IV formula:

$$\begin{aligned}
&\frac{E(\ln Y \mid P(Z)=P(z))\text{-}E(\ln Y \mid P(Z)=P(z'))}{P(z)\text{-}P(z')} \\
= \;& \bar{\beta}+\frac{E(U_1\text{-}U_0 \mid P(z)\geq U_S)P(z)\text{-}E(U_1\text{-}U_0 \mid P(z')\geq U_S)P(z')}{P(z)-P(z')} \\
= \;& \Delta^{LATE}(P(z),P(z')),
\end{aligned}$$

where Δ^{LATE} was defined in Section 2. When $U_1 \equiv U_0$ or $(U_1 - U_0) \perp\!\!\!\perp U_S$, corresponding to the two special cases in the literature, IV based on $P(Z)$ estimates ATE $(= \bar{\beta}\ )$ because the second term on the right hand side of this expression vanishes. Otherwise IV estimates an economically difficult-to-interpret combination of MTE parameters as discussed in the last section.

Another representation of $E(\ln Y \mid P(Z)=P(z))$ that reveals the index structure underlying this model more explicitly writes

$$E(\ln Y \mid P(Z)=P(z)) = \alpha+\bar{\beta}P(z)+\int_{-\infty}^{\infty}\int_{0}^{P(z)}(U_1-U_0)f(U_1-U_0 \mid U_S=u_S)du_S d(U_1-U_0). \quad (9)$$

We can differentiate with respect to $P(z)$ and obtain MTE:

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

IV estimates $\bar{\beta}$ if $\Delta^{MTE}(u_S)$ does not vary with u_S. Under this condition $E(\ln Y \mid P(Z)=P(z))$ is a linear function of $P(z)$. Thus, under our assumptions, a test of the linearity of the conditional expectation of $\ln Y$ in $P(z)$ is a test of the validity of linear IV for $\bar{\beta}$. It is also a test for the validity of conditions I and II.

More generally, a test of the linearity of $E(\ln Y \mid P(Z)=P(z))$ in $P(z)$ is a test of whether or not the data are consistent with a correlated random coefficient model and is also a test of comparative advantage in the labor market for educated labor. If $E(\ln Y|P(z))$ is linear in $P(z)$, standard instrumental variables methods identify "the" effect of S on $\ln Y$. In contrast, if $E(\ln Y|P(z))$ is nonlinear in $P(z)$, then there is heterogeneity in the return to college attendance, individuals act at least in part on their own idiosyncratic return, and standard linear instrumental variables methods will not in general identify the average treatment effect or any other of the

treatment parameters defined earlier. This test is simple to execute and interpret and we apply it below.

We consider $E(\ln Y|P(Z) = P(z))$ and differentiate this conditional expectation to obtain MTE. We also could 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 of either Z or individual components of Z has several advantages. By examining derivatives of $E(\ln Y|P(Z) = P(z))$, we are able to identify the MTE function for a broader range of values than would be possible by examining derivatives of $E(\ln Y|Z_k = z_k)$ while removing the ambiguity of which element to condition upon. Also, by connecting the MTE to $E(\ln Y|P(Z) = P(z))$, 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.[28] See Heckman (2001) and Heckman and Vytlacil (2001b, 2004b).

It is straightforward to estimate the levels and derivatives of $E(\ln Y \mid P(Z) = P(z))$ and standard errors using the methods developed in Heckman, Ichimura, Smith and Todd (1998). The derivative estimator of MTE is the local instrumental variable (LIV) estimator of Heckman and Vytlacil (1999, 2000).

This framework can be extended to consider multiple treatments, which in this case can be either multiple years of schooling, or multiple types or qualities of schooling. These can be either continuous (see Florens, Heckman, Meghir and Vytlacil, 2002) or discrete (see Carneiro, Hansen and Heckman, 2003, Carneiro and Heckman, 2003, and Heckman and Vytlacil 2004a).

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

In this section we report estimates of the MTE using a sample of white males from the National Longitudinal Survey of Youth. The data are described in the appendix. $S = 1$ denotes college

[28] 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.

attendance. In our data set there are 713 high school graduates who never attend college and 731 individuals who attend any type of college.[29] Table 3 documents that individuals who attend college have on average a 32% higher wage than those who do not attend college. They also have one year less of work experience since they spend more time in school.[30] 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.[31] Persons who only attend high school come from larger families and have less educated parents than individuals who attend college. They also live in counties where tuition is higher, and they live farther away from a college, two measures of direct costs of schooling. Those who do not go on to college live in counties where local wages for unskilled labor are higher, a measure of the opportunity cost of schooling. The wage equations include, as variables in X, experience, experience squared and schooling-adjusted AFQT. Our instruments are the number of siblings, parental education, distance to college, tuition, local wage and local unemployment variables.[32] AFQT enters the schooling choice equation (and therefore the Z vector) but it does not play the role of an instrument since it is included in the X vector as well.

We use a probit model for schooling choice with $\mu_s(z) = z\gamma$, $U_S \sim \mathrm{N}(0,1)$, and thus $P(z) = \Phi(z\gamma)$ and $S = 1[\Phi(U_S) \leq P(z)]$ where $\Phi(\cdot)$ is the standard normal cdf. Alternative functional form specifications for the choice model produce very similar results to the ones reported here. 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

[29] These are white males, in 1992, with either a high school degree or above and with a valid wage observation, as described in appendix A. We average over wage observations in adjacent years. We obtain comparable results for adjacent years of the data. For these results and for other results using different data sets, see Carneiro (2002).

[30] Wages are constructed as an average of all nonmissing wages between 1990 and 1994 for each individual. Actual work experience (not potential experience) is measured in 1992. Since individuals in the NLSY are born between the years of 1957 and 1964, in 1992 they are 28 to 35 years of age.

[31] We use a measure of this score corrected for the effect of schooling attained by the participant at test date, since at the date the test was taken, in 1981, different individiduals 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 (2003). We perform this correction for all demographic groups in the population and then standardize the AFQT to have mean 0 and variance 1.

[32] Our basic empirical results are barely changed if we include family background variables in both the outcome and schooling choice equations and so do not use these variables as instruments. We discuss results excluding family background measures below.

unobservables.

Table 4 gives estimates of γ and the corresponding average marginal derivatives. The Z variables are strong predictors of schooling. An exception is "distance to college at 14" which appears with a positive sign in the choice equation, but the effect of this variable is very imprecisely estimated.[33] Our tuition effects conform to the ones found in the literature that measures enrollment-tuition responses in the US: a \$1000 reduction in (four year college) tuition leads to an increase in enrollment of 5% (see Kane, 1994 or Cameron and Heckman, 2001 for summaries of the literature).[34]

The support of the estimated $P(Z)$ is shown in Figure 3 and it is almost the full unit interval,[35] although at the extremes of the interval the cells of data become very thin. The sparseness of data in the tails results in a large amount of noise (variability) in the estimation of $E(Y|X, P(Z) = p)$ for values of p close to zero or one, which in turn makes estimation of the parameters defined over the full support of U_S (and thus requiring estimation of $E(Y|X, P(Z) = p)$ over the full unit interval) problematic. We discuss these problems below. Note, however that the MTE can be estimated pointwise for a wide range of evaluation points without full support.

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 results in a feasible semiparametric estimation problem. In particular, we assume linearity in X and separability between X and U_1 and U_0 in the outcome equations, $\ln Y_1 = \alpha_1 + X\theta_1 + U_1$ and $\ln Y_0 = \alpha_0 + X\theta_0 + U_0$. In addition to reducing the dimensionality of the estimation problem, these restrictions also make our empirical results comparable to those obtained from specifications of schooling equations estimated in the preceding literature. Our linearity assumptions on the outcome equations imply that the return to college attendance can

[33] Use of slightly different samples or a slightly different measure of distance to college leads to reversals of this sign, although the estimated effect is never very strong.

[34] These are partial equilibrium estimates of the effects of tuition. Heckman, Lochner and Taber (1998, 1999) show that partial and general equilibrium analyzes of tuition policy can lead to very different conclusions.

[35] Formally, for nonparametric analysis, we need to investigate the support of $P(Z)$ conditional on X. However, the partially linear structure that we will impose below implies that we only need to investigate the marginal support of $P(Z)$.

be written as a linear function of observables (X) and unobservables ($U_1 - U_0$):

$$\beta = \alpha_1 - \alpha_0 + X(\theta_1 - \theta_0) + U_1 - U_0.$$

Thus the outcome equation can be written as

$$\ln Y = \alpha_0 + X\theta_0 + S[\alpha_1 - \alpha_0 + X(\theta_1 - \theta_0)] + U_0 + S(U_1 - U_0) \tag{10}$$

with $(U_0, U_1, U_S) \perp\!\!\!\perp (X, Z)$.[36] Combining the model for S with the model for Y implies a partially linear model for the conditional expectation of Y:

$$E(\ln Y | X, P(Z)) = \alpha_0 + X\theta_0 + P(Z)(\alpha_1 - \alpha_0) + P(Z)X(\theta_1 - \theta_0) + K(P(Z)) \tag{11}$$

where

$$K(P(Z)) = E(U_1 - U_0 | P(Z), S = 1)P(Z) = E(U_1 - U_0 | \Phi(U_S) \leq P(Z))P(Z)$$

where $\Phi(\cdot)$ is the standard normal cdf. No parametric assumption is imposed on the distribution of (U_0, U_1), and thus $K(\cdot)$ is an unknown function that must be estimated nonparametrically. In general, unless P has full support in the unit interval, it is not possible to separately identify the intercept of the regression (α_0), the intercept term in $(P(Z)(\alpha_1 - \alpha_0))$ and the intercept of the function $K(P)$.[37] However the MTE can still be identified at U_S evaluation points within the support of $P(Z)$ since

$$\begin{aligned} \Delta^{MTE}(x, p) &= \frac{\partial E[\alpha_0 + P(\alpha_1 - \alpha_0) + P(z)x(\theta_1 - \theta_0) + K(P)]}{\partial P}|_{P=p} \\ &= (\alpha_1 - \alpha_0) + x(\theta_1 - \theta_0) + E(U_1 - U_0 | U_S = p). \end{aligned}$$

The semiparametric, partially linear, form for the conditional expectation has several advantages in conducting empirical work. It imposes a dimension reduction compared to a fully nonparametric model, while not restricting the form of the K function and thus allowing greater

[36] (A-3) only requires that $(U_0, U_1, U_S) \perp\!\!\!\perp Z | X$, so we do not estimate the most general possible model within our framework.

[37] This is the "identification at infinity" point made by Heckman (1990).

flexibility than traditional parametric approaches.[38,39] For simplicity (and in accordance with the traditional Mincer model and the model of Willis and Rosen, 1979), we restrict the coefficients on experience and experience squared to be the same in the high school and in the college outcome equations ($\theta_1^{\text{experience}} = \theta_0^{\text{experience}}$, $\theta_1^{\text{experience}^2} = \theta_0^{\text{experience}^2}$)[40]. AFQT is the only X variable that influences the return to schooling ($\theta_1^{\text{AFQT}} \neq \theta_0^{\text{AFQT}}$).[41]

The coefficient on the interaction between $P(Z)$ and X $(= \theta_1 - \theta_0)$ indicates whether ability affects returns to schooling. Simple least squares regressions of log wages on schooling, ability measures, and interactions of schooling and ability (ignoring selection arising from uncontrolled unobservables) have been widely estimated in this and other data sets and generally show that cognitive ability is an important determinant of the returns to schooling[42]. We include AFQT in the model as an observable determinant of the returns to schooling and of the decision to go to college. In the absence of such a measure of cognitive ability, selection arising from unobservables should be important. Most data sets that are used to estimate the returns to education (such as the Current Population Survey or the Census) lack such ability measures. We discuss the empirical consequences of omitting ability in section 7.

We can test for selection on the individual returns to attending college by using equation (9) to check whether $E(\ln Y|X, P)$ is a linear or a nonlinear function of P. Nonlinearity in P 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). A simple way to implement this test is to approximate $K(P)$ with a third order polynomial in P and test whether

[38] The partially linear model was introduced by Robinson (1988).

[39] Imposing the partially linear model weakens the support condition that otherwise would be required for $P(Z)$. In particular, fully nonparametric analysis of all treatment parameters and policy counterfactuals would require that the support of the distribution of $P(Z)$ conditional on X be the full unit interval. In contrast, the analysis with the partially linear model requires that X be full rank conditional on $P(Z)$ and that the marginal distribution of $P(Z)$ have support equal to the full unit interval, without requiring that distribution of $P(Z)$ conditional on X have support equal to the full unit interval.

[40] Allowing $\theta_1^{\text{experience}} \neq \theta_0^{\text{experience}}$ and $\theta_1^{\text{experience}^2} \neq \theta_0^{\text{experience}^2}$ produces some instability in the estimates of these and other parameters of the regression. Our main conclusions reported below are robust when we use the more general specification but the estimates are less precise.

[41] Results where AFQT^2 and AFQT^3 are added to the model are available on request. They are qualitatively similar to the ones we present in this paper.

[42] See Blackburn and Neumark (1993), Bishop (1991), Grogger and Eide (1995), Heckman and Vytlacil (2001a), Murnane, Levy and Willett (1995), Meghir and Palme (2001), Carneiro (2002) and Table A3 in the appendix of the paper.

the coefficients in the second and third order terms are statistically significant.[43] We reject the null hypothesis that these coefficients are jointly equal to zero (p-value = 0.0564). Nonlinearity of $E(Y|X, P(Z))$ in $P(Z)$ implies that the MTE is not constant in u_S and that the IV estimate of the return to schooling is not an estimate of $\bar{\beta}(x) = ATE$.[44]

Following Heckman, Ichimura, Smith and Todd (1998), we estimate the partially linear model using a double residual regression procedure involving the use of local linear regression.[45] We use a biweight kernel with a bandwidth of 0.3[46] and all the standard errors we present are bootstrapped.[47] Figure 4 plots the estimated function for $E(\ln Y|P = p)$ as a general function of P (along with a model which imposes linearity of this expectation in P). There is a substantial departure from linearity.

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

$$\begin{aligned} MTE(x, u_S) &= E(\ln Y_1 - \ln Y_0|X = x, U_S = u_S) \\ &= \alpha_1 - \alpha_0 + x(\theta_1 - \theta_0) + E(U_1 - U_0|U_S = u_S). \end{aligned}$$

The component dependent on X is a linear function of AFQT. Table 5 reports the coefficients on the X variables. The effect of AFQT on returns is positive and quantitatively important but is imprecisely estimated.[48] The local IV estimate is close to the OLS estimate but with larger standard errors. Individuals with higher AFQT have a higher return to schooling. (See also Carneiro, 2002 for further evidence.)[49] Figure 5 plots the component of the MTE that depends

[43] The results from this test are reported in Table A2. The tests are based on a bootstrap procedure with 51 bootstrap replications.

[44] The standard errors used to perform this test are adjusted to account for parameter estimation in $P(Z)$.

[45] In a first stage we estimate a local linear regression of each variable in the X vector (as well as all interactions between X and P) on P. Then we compute the residuals corresponding to each of these regressions and regress wages ($\ln Y$) on each of the residuals of this first stage to estimate θ_0, $\alpha_1 - \alpha_0$ and $\theta_1 - \theta_0$. Finally, we compute the residual of this latter regression and regress it (using a local linear regression) on P to estimate $K(P(Z))$. An alternative to estimating $K(P(Z))$ nonparametrically would be to use polynomials in $P(Z)$ or splines. Carneiro (2002) shows that using polynomials of degree three and four in $P(Z)$ generates basically the same results as those presented in this paper.

[46] The results are robust to variations in the bandwidth between 0.15 and 0.35, with some sensitivity in the tails due to the sparseness of the data in the tails.

[47] We use 51 bootstrap replications. In each iteration of the bootstrap we reestimate $P(Z)$ so that all standard errors account for the fact that $P(Z)$ is itself an estimated object.

[48] The effect of AFQT on the return, $(\theta_1 - \theta_0)$, is the coefficient on $P(Z)$AFQT in Table 5.

[49] Since the measure of AFQT ranges from -2.6 to 2.7 the difference in the return to college between two individuals with the same level of U_S, one with an AFQT score of 2.7 and the other with an AFQT score of -2.6 is 46.95% (dividing by 3.5, the difference in the averafference in the return of 13.41% per year of college).

on U_S but not on X $(= \alpha_1 - \alpha_0 + E(U_1 - U_0 | U_S = u_S))$, derived from Figure 4 using the formula of equation (9).[50] We approximate the derivative of $K(P(Z))$ by taking discrete differences:

$$\frac{\partial K(P)}{\partial P} \doteq \frac{K(P+h) - K(P)}{h}$$

where $h = 0.01$. $E(U_1 - U_0 \mid U_S = u_S)$ is declining in u_S for values of u_S up to 0.4 and then it is rising.[51] Returns are annualized to reflect the fact that college goers attend 3.5 years of college. The most college worthy persons in the sense of having high gross returns are more likely to go to college. They have low values of u_S, the "cost" of college. But for high values of u_S (above 0.4) the estimated MTE is increasing in u_S indicating that individuals not likely to go to college (in terms of their unobservables) would also benefit substantially from attending college. The lowest returns are for individuals in the middle ranges of u_S.[52] The magnitude of the heterogeneity in returns is substantial: returns can vary from slightly above 5% to above 40% per year of college. The rising portion of $E(U_1 - U_0 | U_S = u_S)$ indicates that other factors besides financial returns determine the decision to go to college since individuals with high returns are choosing not to attend college.

Carneiro, Hansen and Heckman (2003) estimate that a major determinant of college attendance is the psychic cost of going to school. In their framework, psychic cost is a function of a measure of cognitive ability, but they also allow the psychic cost to depend on other unobservables. They show that substantial changes in the ex-ante distribution of financial returns (perceived by the agent at the time he is deciding whether or not to enroll in college) have trivial effects on college attendance, precisely because psychic cost plays such an important role in this decision relative to the role of financial returns. Therefore, individuals with high levels of u_S may well have high financial returns to college (although not as high as the returns for those with low values of u_S) but still decide not to attend college because their (psychic) costs are very high.[53]

[50] For better visualization of the pointwise estimates of the MTE, in appendix figure A1 we plot the same curve as in figure 4 without the standard errors.

[51] Note that the decision rule in (8) is $S = 1$ if $\mu_S(Z) - U_S \geq 0$ so, for a given Z, individuals with a higher U_S are less likely to go to college.

[52] Figure A2 (in the appendix) plots both components of the marginal treatment effect: returns are highest for individuals with a high level of AFQT (X in the figure) and a high level of u_S, and are lowest for individuals with a low level of AFQT and with values of u_S close to 0.4.

[53] This pattern is also consistent with the existence of credit constraints affecting a segment of the population.

Table 6 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 a \$1000 tuition subsidy are obtained in the following way. 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, which depends on the parameter we want to compute:

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

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

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

where Z and Z' are the values of the instruments under the baseline regime and under the new policy regime, respectively.[54] These densities are also weights, but instead of weighting functions of U_S they weight functions of X (see also Carneiro, 2002).

Instead of high psychic costs, individuals with high u_S may face high borrowing costs which discourage college attendance This pattern is also consistent with high rates of time preference.

[54] PRT is defined conditional on two policies. The baseline policy is the current policy in the data, for which each individual has his or her actually observed random vector Z. The policy experiment is to shift Z to Z'. In our example, the policy experiment leaves each element of Z unchanged except for tuition, and reduces tuition by 1000. Thus, $Z'_k = Z_k$ for all elements k of Z that do not correspond to the tuition variable, and $Z'_k = Z_k - 1000$ for the element k of Z that does correspond to the tuition variable. Both Z and Z' are assumed to be nondegenerate random vectors.

The limited support of P near the boundary values of $P = 0$ and $P = 1$ creates a practical problem for the computation of the treatment parameters such as ATE, TT, and $PRTE$, since we cannot evaluate MTE for values of U_S outside the interval $[0.01, 0.96]$. 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. The numbers presented in Table 6 are constructed after restricting the weights to be defined only over the region $[0.01, 0.96]$. These can be interpreted as the parameters defined in the empirical support of $P(Z)$, which is close to the full unit interval. The close to full support for $P(Z)$ in this paper is in marked contrast to the limited support found in Heckman, Ichimura, Smith and Todd (1998), 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.

Alternative ways to deal with the problem of limited support are to construct bounds for the parameters or to use a parametric extrapolation outside of the observed support. We report various bounds and extrapolations in Table A-4 in the appendix. Bounds on the treatment effects are generally wide even though the support is almost full. Parametric extrapolation outside of the support is potentially sensitive to the choice of extrapolation model. Estimates based on locally adapted extrapolations show much less sensitivity than do estimates based on global approximation schemes.

The sensitivity of estimates to lack of support in the tails ($P = 0$ or $P = 1$) is important for parameters, such as ATE or TT, that put substantial weight on the tails of the MTE distribution. 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 highly sensitive to imprecise estimation or extrapolation error for $E(Y|X, P(Z) = p)$ for values of p close to 0 or 1. Even though empirical economists often seek to identify these parameters, often they are not easily estimated nor are they always the economically interesting ones. In contrast, $PRTE$ 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.[55] $AMTE$ places weight $f_P(u|X=x)$ on MTE where $f_P(\cdot|X=x)$ is the density of $P(Z)$ conditional on X.[56] This implies that (1) if the distribution of $P(Z)$ has a density with respect to Lebesgue measure, then identification of $AMTE$ does not require a support condition on $P(Z)$ since $AMTE$ only weights MTE where the density of $P(Z)$ is positive;[57] and (2) $AMTE$ will put the most weight on MTE where there is the most data and thus the most precise estimates and the least weight on MTE where there is the least data and thus the least precise estimates. $AMTE$ and $PRTE$ are thus much easier to estimate because they place little weight on the tails of the MTE.[58] As demonstrated in Table A-4, these parameters are much less sensitive to alternative methods for extrapolating MTE than are TT, ATE and TUT.

Integrating only over the support of the distribution of $P(Z)$, $[0.01, 0.96]$, Table 6[59] reports estimates of the average annual return to college for a randomly selected person in the population ATE of 18.70%, which is between the annual return for the average individual who attends college (TT), 20.69%, and the average return for high school graduates who never attend college (TUT), 16.77%. The average marginal individual ($AMTE$) has an annual return of 15.95% which is below the annual return for the average person (TT). These estimates are slightly above the range of the instrumental variables estimates of returns to schooling reported by Card (1999, 2001) in his surveys of literature, which range from 6% to 16% per year of schooling.[60] Our linear IV estimate

[55] However, we could also define a policy that affects people at either tail of the MTE and hence reverse this conclusion.

[56] We are assuming for the $AMTE$ weights that the distribution of $P(Z)$ conditional on X has a density with respect to Lebesgue measure.

[57] Formally, the identification condition is that there be no isolated points in the support of the distribution of $P(Z)$ so that local variation in $P(Z)$ identifies $E(\ln Y|X, P(Z)=p)$ at each p in the support of the distribution of $P(Z)$ conditional on X. The assumption that the distribution of $P(Z)$ conditional on X has a density with respect to Lebesgue measure implies that there are no isolated points in the support of the distribution of $P(Z)$ conditional on X.

[58] This has an interesting consequence. We can formally reject that $AMTE = ATE$ (and that $PRTE = ATE$) at a 10% significance level, but we cannot formally reject that $PRTE = TT$ nor that $PRTE = TUT$. Both TT and TUT substantially weight sections of the MTE (in the tails of the support of P) where it is very imprecisely estimated, while $AMTE$ and ATE place a much smaller weight on the tails. Therefore the standard errors of the estimates of the latter two parameters are much smaller than the standard errors of the estimates of the former two parameters.

[59] The numbers presented in the second column of Table A4 in the appendix are constructed by restricting the weights to only integrate over the region $[0.05, 0.90]$.

[60] However most of the estimates reported in these papers are based on samples constructed from earlier years, in which we expect the returns to schooling to be lower than in the more recent dataset we are using. Furthermore, none of these papers estimates all of the parameters reported in Table 6. The averaging of wages across five different years of data also leads to an increase of return. If we restrict ourselves to 1992 wages (instead of averaging wages

of 12.5% is in the range of the IV estimates reported in the literature. None of these numbers corresponds to the average annual return to college for those individuals induced to enroll in college by a \$1000 tuition subsidy ($PRTE$), which is 15.92%, although this estimate is very close to the return for the average marginal person.[61] This is the relevant return for evaluating this specific policy using a Benthamite welfare criterion. It is below TT, which means that the marginal entrant induced to go to college by this specific policy has an annual return well below (five log points) that of the average college attendee. Figure 6 graphs the weights for $E(Y_1 - Y_0|U_S = u_S)$ for ATE, TT and $PRTE$. ATE gives a uniform weight to all U_S[62], while TT overweights individuals with low levels of U_S (and therefore very likely to have enrolled in college) and $PRTE$ puts more weight on individuals in middle ranges of U_S. Figure 7 presents these weights for $E(Y_1 - Y_0|AFQT)$. The $PRTE$ places more weight at the center of the distribution of $AFQT$ than does TT or ATE. Figure 8 presents the joint $(U_S, AFQT)$ policy weights.[63] Individuals attracted into college by a tuition subsidy differ from the average individual who attends college both in terms U_S and in terms of AFQT. There is a tradeoff in AFQT and cost (U_S). Low cost people attracted into college by the subsidy have lower AFQT.[64]

We next compare all of these 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 (10). It estimates ATE if S and X are orthogonal to $U_0 + S(U_1 - U_0)$. The IV estimate is derived

between 1990 and 1994) then TT is 15.90%, ATE is 18.52%, TUT is 13.27%. $AMTE$ is 12.91% and $PRTE$ is 12.88%. This sample has 1280 individuals.

[61] We do not correct AFQT for effect of schooling at test we obtain ATE 0.1862, TT 0.2275, TUT 0.1478, $AMTE$ 0.1596 and $PRTE$ 0.1584. The only sizeable effects are on TT and TUT.

[62] Since the density of U_S is uniform in the population, this corresponds to weighting $E(U_1 - U_0|U_S)$ by the density of U_S.

[63] $E(Y_1 - Y_0|AFQT)$ is plotted in this figure. It is a straight line. The slope of this line is given by the coefficient on the interaction of P and AFQT in the regression reported above. $E(Y_1 - Y_0|AFQT)$ is scaled to fit in the figure. The joint $(U_S, AFQT)$ weights of figure 8 apply to the MTE graphed in appendix figure A2.

[64] When we exclude family background variables in X, but these variables appear in the outcome equation, the instruments become tuition, distance and local labor market variables. We included number of siblings and father's education in levels, but not in returns, in the wage equation. The main patterns of the findings just described in the text do not change very much. The estimated parameters are: $ATE = 0.1842$, $TT = 0.2045$, $TUT = 0.1645$, $AMTE = 0.1570$. When we include family background variables both in levels and in returns the function $E(Y_1 - Y_0|U_S)$ has roughly the same shape as the one presented but all the estimates become more imprecisely estimated (standard errors increase substantially).

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 for individuals induced to enroll in college by a \$1000 tuition subsidy[65], so that we can compare these estimates with the policy relevant treatment effect. The OLS estimate of the return to a year of college is 7.60% while the IV estimate is 12.56%, well below the policy relevant treatment effect. 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, $PRTE$ weights high values more both in the initial declining segment and in the final rising segment of MTE. IV places greater weights relatively on the lower values of MTE at the middle of the figure. Therefore, in this sample (and for this instrument), the IV estimate is below the $PRTE$. 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.[66]

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.[67] 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 for this fact is that returns are heterogeneous and IV estimates the return for the marginal person[68] and 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 (Card, 2001). However in this section we show that the marginal person has a return substantially below the return for the average person, and still $\beta_{IV} > \beta_{OLS}$.

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 (keeping the conditioning on X

[65] This is obtained by integrating X with respect to $f_X(x|PRT) = f_X(x|Z\gamma - U_S < 0, Z'\gamma - U_S \geq 0)$.

[66] The weight for the marginal individual in this sample is very close to the weight for $PRTE$. Appendix Figure A3 contrasts the weights for the marginal person ($AMTE$) with the weights for the average person (TT).

[67] 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 $E(Y_1|X, U_S = p) = \frac{\partial E(SY|X,P)}{\partial P}|_{P=p}$ and $E(Y_0|X, U_S = p) = -\frac{\partial E[(1-S)Y|X,P]}{\partial P}|_{P=p}$. These derivatives are estimated using the same procedure we described for the estimation of $E(Y_1 - Y_0|X, U_S = p) = \frac{\partial E(Y|X,P)}{\partial P}|_{P=p}$.

[68] It estimates the return for the "switchers".

implicit)

$$E(\ln Y \mid S=1) - E(\ln Y \mid S=0) = E(\beta \mid S=1) + [E(U_0 \mid S=1) - E(U_0 \mid S=0)]$$
$$= \bar{\beta} + E\left(U_1 - U_0 \mid S=1\right) + \left[E\left(U_0 \mid S=1\right) - E\left(U_0 \mid S=0\right)\right].$$

In a model without variability in the returns to schooling, $E(\beta \mid S=1) = E(\beta) = \overline{\beta}$ 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). It suggests that OLS may provide an upward biased estimated 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. People who go to college may be the worst persons in the Y_0 distribution, *i.e.* $E(U_0|S=1) - E(U_0|S=0) < 0$ even though they could be the best persons in the Y_1 distribution. This could offset the positive $E(U_1 - U_0|S=1)$ and make the OLS estimate below that of the IV estimate, even if the IV estimate is below the return for the average person ($E\left(\beta|S=1\right)$).[69] Thus the evidence reported in the recent literature comparing OLS and IV is not informative on the comparison of average and marginal returns.

A major advantage of our approach to instrumental variables over the approach adopted in the recent literature is that it enables us to use the economic theory of choice to combine multiple instruments into one scalar instrument $P\left(z\right) = \Pr\left(S=1|Z=z\right)$. In the general case when conditions I and II of section 1 do not apply, each instrument defines a different parameter. Table 7 compares the conventional IV estimates for each of the instruments used in $P\left(z\right)$ in this paper. The estimates range all over the map. They are different from each other and from the estimate generated by using $P\left(z\right)$. None of these numbers is of intrinsic economic interest and none is close to the policy relevant treatment effect or the average marginal treatment effect. Using local instrumental variables (LIV), we can identify the MTE and construct economically interpretable parameters that answer precisely posed policy questions.

[69] Carneiro and Heckman (2002) develop this argument further.

7 Ability Bias and the Validity of the Conventional Instruments

Except for the OLS estimates reported in this paper, all of our estimates rely on instrumental variables. The instruments used in this paper are those conventionally used in the literature estimating the returns to schooling. (See, e.g., Card 1999, 2001.) In this section, we examine the validity of the conventional instruments. Many data sets on earnings and schooling do not possess measures of cognitive ability. For example, the CPS and many other data sets used to estimate the returns to schooling do not report measures of cognitive ability. In this case, ability becomes part of U_1, U_0 and U_S instead of being in X.

The assumption of independence (between the instrument and U_1 and U_0) implies that the instruments have to be independent of cognitive ability. However, the instruments that are commonly used in the literature are correlated with AFQT. The first column of Table 8A shows the correlations between different instruments (Z) and college attendance (S), denoted by $\rho_{Z,S}$.[70] With the exception of local unemployment rate, all candidate instruments are strongly correlated with schooling. The second column of this table presents the correlation between instruments and AFQT scores (A), denoted by $\rho_{Z,A}$. It shows that most of the candidates for instrumental variables in the literature are also correlated with cognitive ability. Therefore, in data sets where cognitive ability is not available most of these variables are not valid instruments since they violate assumption (A-3). Notice that the local wage for unskilled workers and the local unemployment rate are not strongly correlated with AFQT. However, they are weakly correlated with college attendance as well. In the third column of Table 8A we present the F-statistic for the test of the hypothesis that the coefficient on the instrument is zero in a regression of schooling on the instrument. Staiger and Stock (1997) suggest using an F-statistic of 10 as a threshold for separating weak and strong instruments[71]. The table shows that the local wage and local unemployment

[70] When constructing this table we include all white males individuals with with nonmissing observations for each pairwise correlation, so the sample sizes for each correlation are larger than in the sample used in the previous section (in particular because we do not need wage observations to construct this table). We obtain a similar set of results if we restrict ourselves to the sample used in the previous section.

[71] In a recent paper Stock and Yogo (2003) propose a different test. However they still find that the rule of thumb first proposed in Staiger and Stock (1997) works well in general.

variables have F statistics well below 10 which suggests that they are weak instruments. Therefore either the candidate instrumental variable is correlated with ability or it is weakly correlated with schooling.

Table 8B presents partial correlations between instruments, schooling and ability, after controlling for family background variables (number of siblings and parental education).[72] Conditioning on family background weakens the correlation between AFQT and the instruments. However the F-test for a regression of schooling on the residualized instrument is low by Staiger-Stock standards. Residualizing on family background attenuates the correlation between the instruments and ability but also between the instruments and schooling. This correlation is reported in the third column of Table 8B. The instrument we use in the empirical work reported in this paper is $P(Z)$. If we regress schooling on experience, experience squared, corrected AFQT (the variables we include in the wage regression) and $P(Z)$, the F-statistic of the coefficient on P is 160. If we add number of siblings and parental education to the regression the F-statistic on this same coefficient becomes 154.[73] By including AFQT in the wage regression we avoid the problem of using invalid instruments. By using an index of instruments instead of a single instrument, we overcome the weak instrument problem. Furthermore, using an index of instruments instead of a single instrument tends to reduce support problems for any instrument. Even if one instrument has limited support, other instruments can sometimes augment the support of P.

Our use of ability measures in the empirical work reported in this paper makes our estimates more credible. When we exclude ability from the estimating equation, or use data sets for years comparable to those used in this analysis that exclude an ability measure, the estimates of most measures of returns are often implausibly large (see Carneiro, 2002). See the substantial increase in the IV estimate in Table 9 when AFQT is omitted from the model. Ability bias is an important empirical phenomenon and failure to control for it leads to substantial upward biases in estimated

[72] We exclude the family background variables from this table since we want to use these variables as controls.

[73] Because $P(Z)$ is a nonlinear function of the instruments these high $F-$statistics may be driven by this nonlinearity. When we only include tuition, distance, local wage and local unemployment rates in $P(Z)$ and then we residualize schooling by AFQT and family background, the reported $F-$statistic on P becomes 16.41. If we construct P as the predicted value of a linear regression of schooling on tuition, distance, local wage and local unemployment then this $F-$statistic becomes 16.28. This is equivalent to using a standard linear IV procedure where the instruments are tuition, distance, local wage and local unemployment.

returns.

8 Summary and Conclusions

This paper presents a framework for estimating marginal and average returns to economic choices when returns differ among individuals and persons select into economic activities based in part on their return to them. We show that different conventional average return parameters and IV estimators are weighted averages of the marginal treatment effect (MTE). Different instruments define different parameters. 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 perform out-of-sample policy forecasts. 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.

Using this framework 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. This confirms in a semiparametric setting a central finding of the parametric Willis and Rosen (1979) analysis. Individuals sort into schooling on the basis of both observed and unobserved gains where the observer is the economist analyzing the data.

Instrumental variables are not guaranteed to estimate policy relevant treatment parameters or conventional treatment parameters. Different instruments define different parameters, and in our empirical analysis produce wildly different "effects" of schooling on earnings. In our empirical analysis, IV understates the Policy Relevant return by three log points.

The marginal return is substantially below the average return to college for those who attend it. Controlling for ability greatly reduces the estimated marginal and average returns to schooling.

Ability bias is an important empirical phenomenon. Most of the standard instrumental variables used to estimate returns to schooling are not valid if ability is not properly accounted for.

References

[1] Aakvik, A. and J. Heckman, E. Vytlacil (2000), "Treatment Effects For Discrete Outcomes When Responses to Treatment Vary Among Observationally Identical Persons: An Application to Norwegian Vocational Rehabilitation Programs," forthcoming, *Journal of Econometrics*, 2003.

[2] Angrist, J., K. Graddy, and G. Imbens (2000), "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *Review of Economic Studies*, 67:499-527.

[3] Angrist, J. and A. Krueger (1991), "Does Compulsory School Attendance Affect Schooling and Earnings," *Quarterly Journal of Economics*, 106:979-1014.

[4] Angrist, J. and A. Krueger (2001)," Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives* 15(4): 69-85.

[5] Becker, G. and B. Chiswick (1966), "Education and the Distribution of Earnings," *American Economic Review*, 56:358-69.

[6] Bishop, J. (1991). "Achievement, Test Scores, and Relative Wages." in M. Kosters, ed. *Workers and their wages: Changing patterns in the United States*. AEI Studies, no. 520 (Washington, D.C.: AEI Press). p. 146-86

[7] Björklund, A. and R. Moffitt (1987), "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics*, 69:42-49.

[8] Blackburn, M. and D. Neumark (1993)."Omitted-Ability Bias and the Increase in the Return to Schooling," *Journal of Labor Economics*, 11(3):521-544.

[9] Bureau of Labor Statistics (2001), *NLS Handbook,* 2001, Washington, DC: US

[10] Cameron, S. and J. Heckman (2001), "The Dynamics of Educational Attainment for Black, Hispanic, and White Males," *Journal of Political Economy* 109(3): 455-99.

[11] Card, D. (1999), "The Causal Effect of Education on Earnings," Orley Ashenfelter and David Card, (editors), Vol. 3A, *Handbook of Labor Economics,* (Amsterdam: North-Holland).

[12] Card, D. (2001), "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69(5): 1127-60.

[13] Carneiro, P. (2002), "Heterogeneity in the Returns to Schooling – Implications for Policy Evaluation." Ph.D. dissertation, University of Chicago.

[14] Carneiro, P. and J. Heckman (2002), "The Evidence on Credit Constraints in Post-secondary Schooling," *Economic Journal* 112(482): 705-34.

[15] Carneiro, P. and J. Heckman (2003), "Empirical Estimates of Rates of Return to Schooling", forthcoming in E. Hanushek and F. Welch, (eds.), *Handbook of Economics of Education*, (North-Holland:Amsterdam).

[16] Carneiro, P., K. Hansen and J. Heckman (2003), "Estimating Distributions of Counterfactuals with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on Schooling Choice," *International Economic Review*, 44(2): 361-422.

[17] Chiswick, B. (1974), *Income Inequality: Regional Analyses Within a Human Capital Framework*, (New York; National Bureau of Economic Research).

[18] Florens, J., J. Heckman, C. Meghir and E. Vytlacil (2002), "Instrumental Variables, Local Instrumental Variables and Control Functions", CEMMAP working paper CWP15/02.

[19] Griliches, Z. (1977), "Estimating The Returns to Schooling: Some Econometric Problems," *Econometrica*, 45(1):1-22.

[20] Grogger, J. and E. Eide. (1995). "Changes in College Skills and the Rise in the College Wage Premium." *Journal of Human Resources* 30(2): 280-310.

[21] Haavelmo, T. (1943). "The Statistical Implications of a System of Simultaneous Equations," *Econometrica,* 11(1): 1-12.

[22] Hansen, K., K. Mullen and J. Heckman (2003), "The Effect of Schooling and Ability on Achievement Test Scores," forthcoming in *Journal of Econometrics.*

[23] Heckman, J. (1990), "Varieties of Selection Bias," *American Economic Review*, 80: 313-318.

[24] Heckman, J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources*, 32(3):441-462.

[25] Heckman, J. (2001), " Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy,* 109(4): 673-748.

[26] Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017-1098.

[27] Heckman, J., L. Lochner and C. Taber (1998), "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents," *Review of Economic Dynamics*, 1(1): 1-58.

[28] Heckman, J., L. Lochner and C. Taber (1999), "General equilibrium cost benefit analysis of education and tax policies," in *Trade, Growth and Development: Essays in Honor of T. N. Srinivasan*, Ranis, G., & Raut, L. K. (eds.).

[29] Heckman, J., L. Lochner and P. Todd (2001), "Fifty Years of Mincer Earnings Functions," unpublished manuscript, University of Chicago, Presented at Royal Economic Society Meetings, Durham, England, April, 2001.

[30] Heckman, J. and S. Navarro (2003), "Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models," forthcoming, *Review of Economics and Statistics.*

[31] Heckman, J. and R. Robb (1985), "Alternative Methods for Estimating the Impact of Interventions," in J. Heckman and B. Singer, (eds.), *Longitudinal Analysis of Labor Market Data*, (New York: Cambridge University Press), 156-245.

[32] Heckman, J. and R. Robb (1986, 2000), "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," in H. Wainer, (ed.), *Drawing Inference from Self-Selected Samples*, (Mahwah, NJ: Lawrence Erlbaum Press).

[33] Heckman, J. and J. Smith (1998), "Evaluating the Welfare State," in *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial*, Econometric Monograph Series, ed. by S. Strom, Cambridge, UK: Cambridge University Press.

[34] Heckman, J., J. Tobias and E. Vytlacil (2001), "Four Parameters of Interest in the Evaluation of Social Programs," *Southern Economic Journal*, 68(2): 211-223.

[35] Heckman, J., J. Tobias and E. Vytlacil (2003), "Simple Estimators for Treatment Parameters in a Latent Variable Framework," forthcoming, *Review of Economics and Statistics.*

[36] Heckman, J. and E. Vytlacil (1998), "Instrumental Variables Methods for the Correlation Random Coefficient Model," *Journal of Human Resources*, 33(4):974-1002.

[37] Heckman, J. and E. Vytlacil (1999), "Local Instrumental Variable and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96:4730-4734.

[38] Heckman, J. and E. Vytlacil (2000), "Local Instrumental Variables," in C. Hsiao, K. Morimune, and J. Powells, (eds.), *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, (Cambridge: Cambridge University Press, 2000), 1-46.

[39] Heckman, J. and E. Vytlacil (2001a). "Identifying the Role of Cognitive Ability in Explaining the Level of and Change in the Return to Schooling," *Review of Economics and Statistics*, 83(1):1-12.

[40] Heckman, J. and E. Vytlacil (2001b), "Policy Relevant Treatment Effects," *American Economic Review Papers and Proceedings,* 91(2): 107-111.

[41] Heckman, J. and E. Vytlacil (2004a), "Econometric Evaluations of Social Programs," forthcoming in J. Heckman and E. Leamer, (eds.), *Handbook of Econometrics*, Volume 5, (North-Holland:Amsterdam).

[42] Heckman, J. and E. Vytlacil (2004b), "Structural Equations, Treatment, Effects and Econometric Policy Evaluation," forthcoming in *Econometrica*.

[43] Ichimura, H., and C. Taber (2000), "Direct Estimation of Policy Impacts " Northwestern University and University College London, unpublished manuscript.

[44] Imbens, G. and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2):467-475.

[45] Kane, T. (1994), "College Entry by Blacks Since 1970: The Role of College Costs, Family Background and the Returns to Education," *Journal of Political Economy,* 102(5): 878-911.

[46] Lewis, H. G. (1986), *Union Relative Wage Effects : A Survey.* (Chicago : University of Chicago Press).

[47] Marshall, A. (1890), *Principles of Economics,* First Edition, (London and New York, MacMillan and Co.).

[48] Meghir, C. and M. Palme (2001). "The Effect of a Social Experiment in Education," The Institute for Fiscal Studies working paper W01/11.

[49] Mincer, J. (1974), *Schooling, Experience and Earnings* (New York: Columbia University Press).

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

[51] National Center for Education Statistics (2003), *Digest of Education Statistics*, 2002, US Department of Education.

[52] Pearl, J. (2000), *Causality*. Cambridge, England: Cambridge University Press.

[53] Robinson, P., (1988), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931-954.

[54] Roy, A. (1951), "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3:135-146.

[55] Staiger, D. and J. Stock, (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica* 65(3): 557-86.

[56] Stock, J. and M. Yogo (2003), "Testing for Weak Instruments in Linear IV Regression," Working paper, Harvard University.

[57] Vytlacil, E. (2002), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*.70(1): 331-41

[58] Willis, R. (1986), "Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions," in O. Ashenfelter and R. Layard (eds.), *Handbook of Labor Economics,* (Amsterdam: North-Holland).

[59] Willis, R. and S. Rosen (1979), "Education and Self-Selection," *Journal of Political Economy*, 87(5):Pt2:S7-36.

[60] Yitzhaki, S. (1996), "On Using Linear Regression in Welfare Economics," *Journal of Business and Economic Statistics*, 14:478:486.

[61] Yitzhaki, S. (1999), "The Gini Instrumental Variable, or 'The Double IV Estimator'," unpublished manuscript, Hebrew University.

Figure 1
Density of Absolute Returns

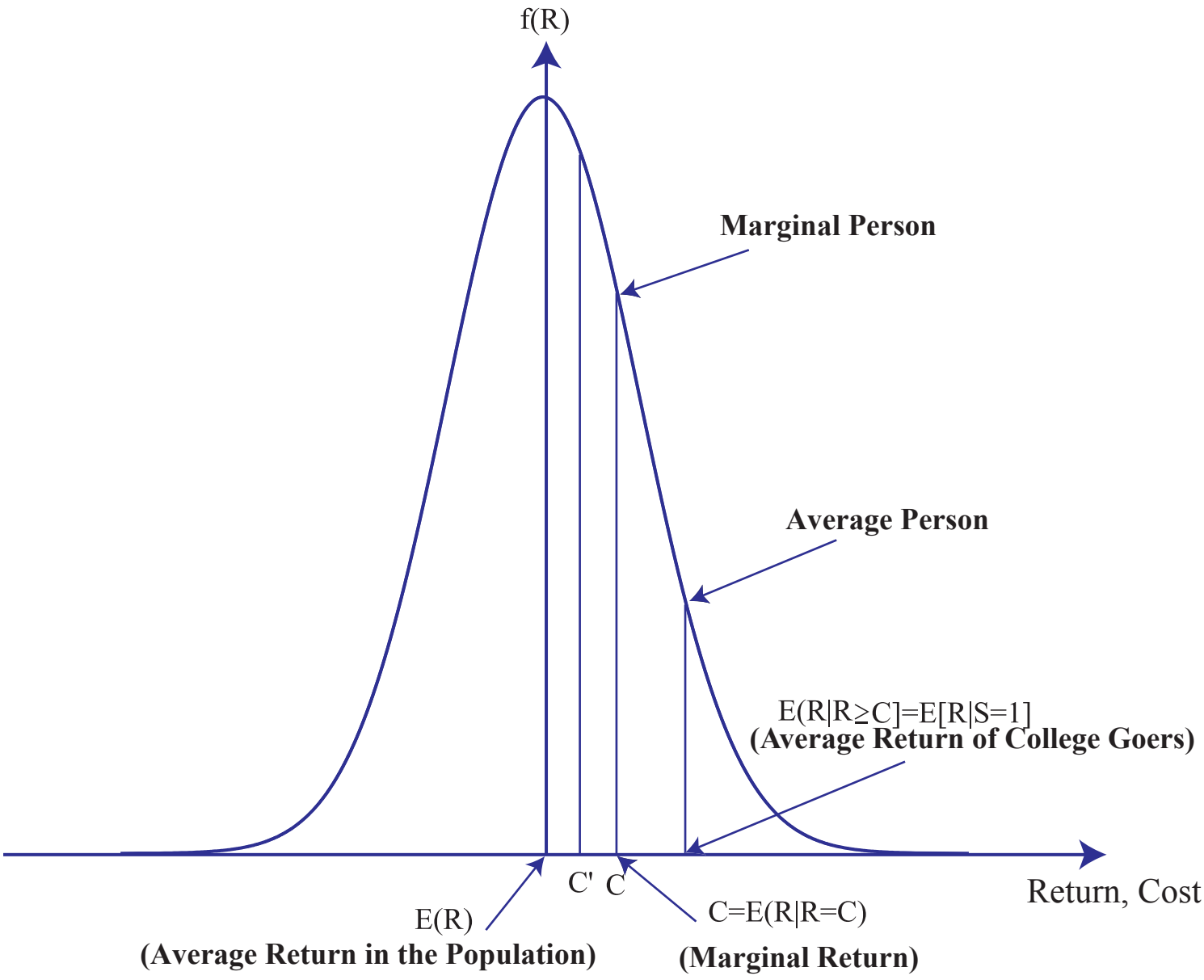


Table 1A
Treatment Effects and Estimands as Weighted Averages of the Marginal Treatment Effect

$$ATE(x) = \int_0^1 MTE(x, u_S) du_S$$

$$TT(x) = \int_0^1 MTE(x, u_S) h_{TT}(x, u_S) du_S$$

$$TUT(x) = \int_0^1 MTE(x, u_S) h_{TUT}(x, u_S) du_S$$

$$AMTE(x) = \int_0^1 MTE(x, u_S) h_{AMTE}(x, u_S) du_S$$

$$\text{Policy Relevant Treatment Effect } (x) = \int_0^1 MTE(x, u_S) h_{PRT}(x, u_S) du_S$$

$$\text{IV}(x) = \int_0^1 MTE(x, u_S) h_{\text{IV}}(x, u_S) du_S$$

$$\text{OLS}(x) = \int_0^1 MTE(x, u_S) h_{\text{OLS}}(x, u_S) du_S$$

Table 1B
Weights

$$h_{ATE}(x, u_S) = 1$$

$$h_{TT}(x, u_S) = \left[\int_{u_S}^1 f_P(p \mid X = x) dp \right] \frac{1}{E(P \mid X = x)}$$

$$h_{TUT}(x, u_S) = \left[\int_0^{u_S} f_P(p \mid X = x) dp \right] \cdot \frac{1}{E((1 - P) \mid X = x)}$$

$$h_{AMTE}(x, u_S) = f_P(p \mid X = x)$$

$$h_{PRT}(x, u_S) = \left[\frac{F_{P^*,X}(u_S) - F_{P,X}(u_S)}{\Delta P} \right]$$

$$h_{\text{IV}}(x, u_S) = \left[\int_{u_S}^1 (p\text{-}E(P \mid X = x)) f(p \mid X = x) dp \right] \frac{1}{Var(P \mid X = x)}$$

$$h_{\text{OLS}} = \frac{E(U_1 \mid X = x, U_S = u_S) h_1(x, u_D) \text{-} E(U_0 \mid X = x, U_S = u_S) h_0(x, u_S)}{MTE(x, u_S)}$$

$$h_1(x, u_S) = \left[\int_{u_S}^1 f_P(p \mid X = x) dp \right] \frac{1}{E(P \mid X = x)}$$

$$h_0(x, u_S) = \left[\int_0^{u_S} f_P(p \mid X = x) dp \right] \frac{1}{E((1 - P) \mid X = x)}$$

Figure 2A

Weights for the Marginal Treatment Effect for Different Parameters

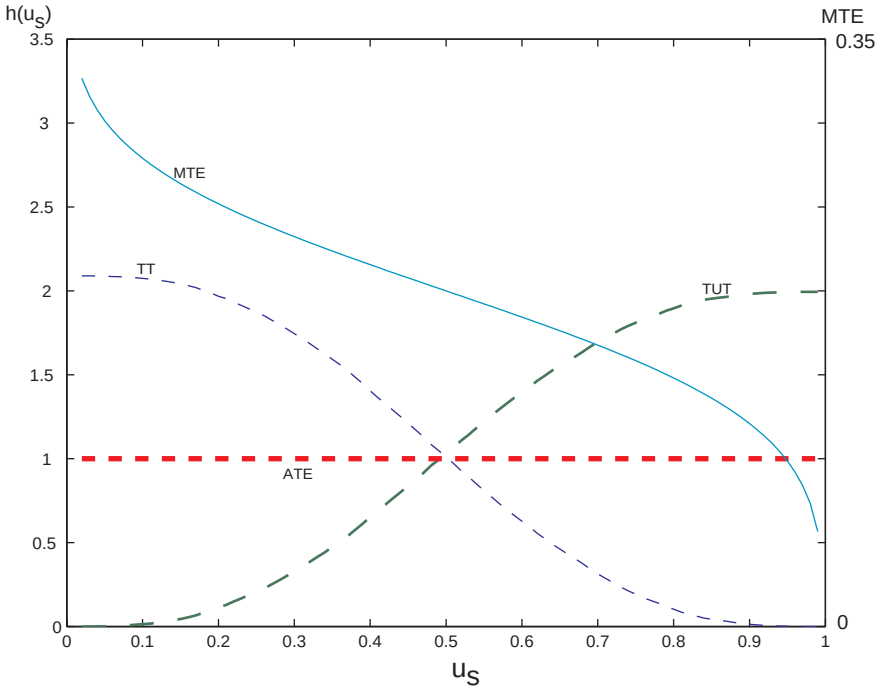
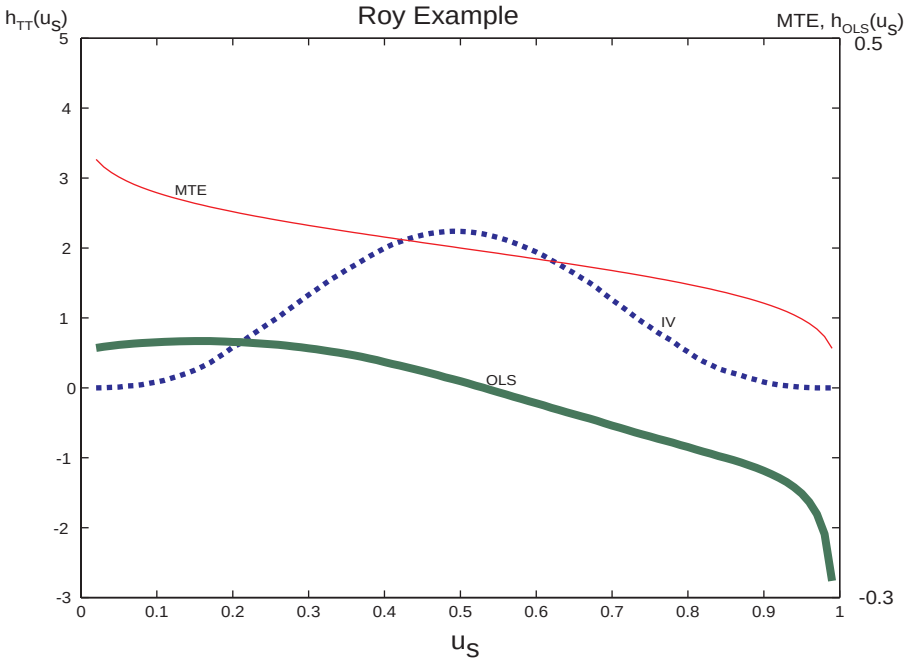


Figure 2B

Marginal Treatment Effect vs Linear Instrumental Variables and Ordinary Least Squares Weights



$$\ln Y_1 = \alpha + \bar{\beta} + U_1 \qquad U_1 = \sigma_1 \varepsilon \qquad \alpha = 0.67$$

$$\ln Y_0 = \alpha + U_0 \qquad U_0 = \sigma_0 \varepsilon \qquad \bar{\beta} = 0.2$$

$$S = 1 \text{ if } Z - U_S > 0 \qquad U_S = \sigma_S \varepsilon \qquad \varepsilon \tilde{} N(0,1)$$

$$Z \tilde{} N(-0.0026, 0.27)$$

$$\sigma_1 = 0.012$$

$$\sigma_0 = -0.05$$

$$\sigma_S = -1$$

Table2
Treatment Parameters in the Generalized Roy Example

Ordinary Least Squares	0.1735
Treatment on the Treated	0.2442
Treatment on the Untreated	0.1570
Average Treatment Effect	0.2003
Sorting Gain (1)	0.0402
Selection Bias (2)	-0.0708
Linear Instrumental Variables (3)	0.2017

(1) E[U1-U0|S=1]=TT-ATE

(2) E[U0|S=1]-E[U0|S=0]=OLS-TT

(3) Using Propensity Score as the Instrument

Table 3
Sample Statistics

	S=1 (N=731)	S=0 (N=713)
Years of Schooling	15.5007	11.9677
	(1.7543)	(0.5087)
Log Wage	2.6123	2.2994
	(0.4797)	(0.4067)
Years of Experience	9.4083	10.3410
	(2.9711)	(2.8365)
Corrected AFQT	1.0137	0.1408
	(0.7083)	(0.8494)
Number of Siblings	2.6484	3.0477
	(1.7590)	(1.8407)
Father's Years of Schooling	13.9959	11.6255
	(3.1373)	(2.7710)
Average County Tuition at 17	19.2496	20.9525
(in $100)	(8.0055)	(8.0905)
Distance to College at 14	2.9811	4.5303
	(9.7070)	(10.7643)
Average State Blue Collar Wage at 17	6.6217	6.8129
(in dollars)	(1.1994)	(1.2137)
County Unemployment Rate at 17	6.2778	6.3077
(in %)	(1.6625)	(1.6967)

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, 2003; see also the data appendix of this paper). The tuition variable corresponds to average tuition of four-year colleges in the county of residence. Distance to College at 14 is measured in miles and measures distance to the nearest college at age 14. Local wage at 17 measures average blue collar wages in the state of residence at 17, and the unemployment variable corresponds to average unemployment in the county of residence at 17. High School dropouts are excluded from this sample. Wages are constructed to be weighed averages of all nonmissing wages between the years of 1990 and 1994. Work experience is measured in 1992. We use only white males from the NLSY79, excluding the oversample of poor whites and the military sample. Standard deviations are in parenthesis

Table 4
Probit Coefficients and
Average Derivatives for the College Decision Model

	b	E(dF/dZ)
Corrected AFQT	0.7595	0.2222
	(0.0500)	(0.0107)
Number of Siblings	-0.0369	-0.0108
	(0.0216)	(0.0075)
Father's Years of Schooling	0.1217	0.0356
	(0.0134)	(0.0035)
Average County Tuition at 17	-0.0138	-0.0040
(in $100)	(0.0049)	(0.0012)
Distance to College at 14*	0.0471	0.0138
	(0.3710)	(0.1028)
Average State Blue Collar Wage at 17	-0.1275	-0.0373
(in dollars)	(0.0348)	(0.0103)
County Unemployment Rate at 17	0.0650	0.0190
(in %)	(0.0261)	(0.0076)

(*) Distance in hundreds of miles.

The first column of the table reports the coefficients of a probit regression of college attendance (a dummy variable that is equal to 1 if an individuals has ever attended college and equal to 0 if he has never attended college but has graduated from high school) on the set of variables listed in the table. The second column corresponds to average marginal derivatives. 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. Standard errors are in parenthesis and boostrapped.

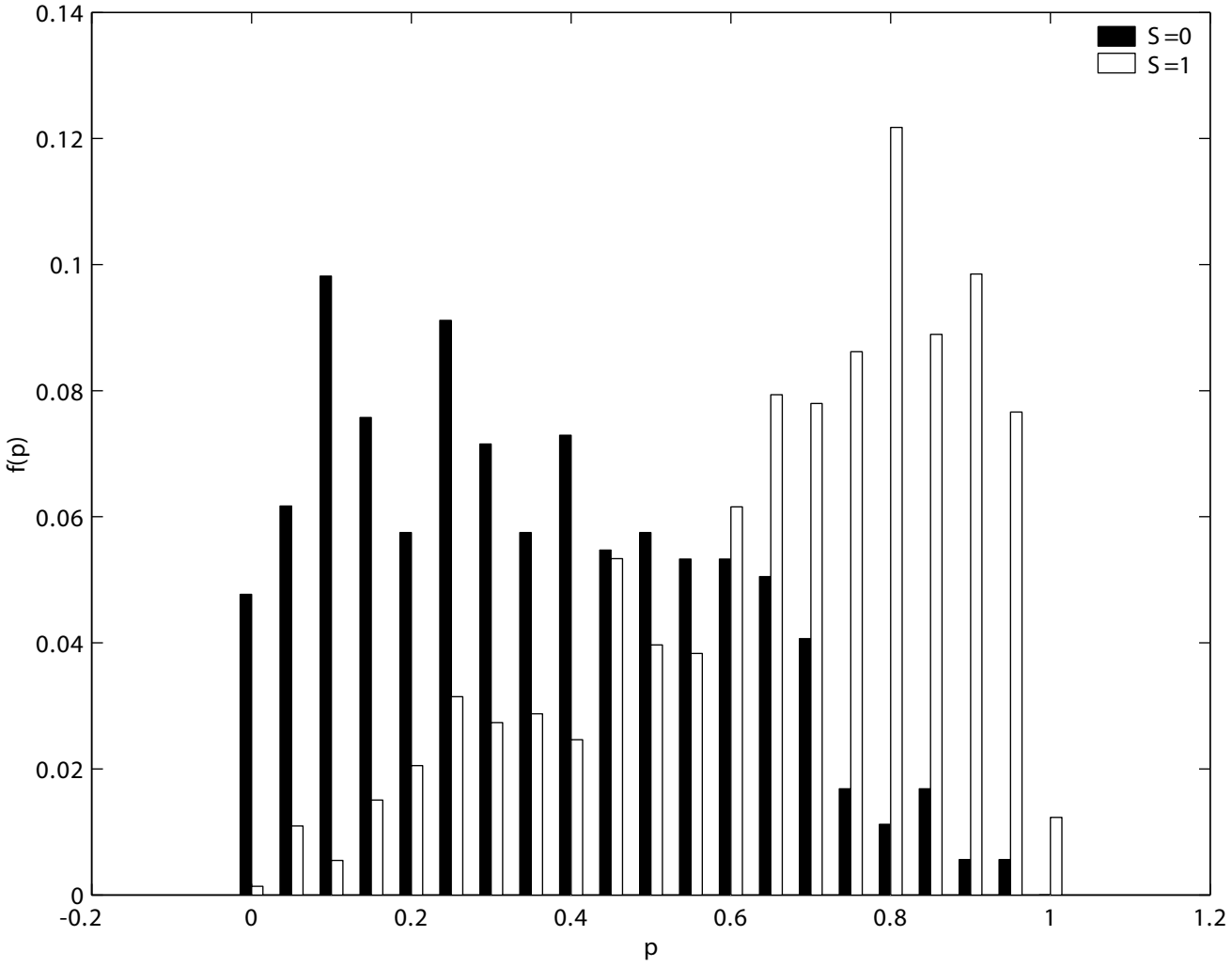
Table 5
Coefficients from Partial Linear Regression of Wages on Experience, Experience Squared, AFQT, P*AFQT and K(P)

Years of Experience	0.0771
	(0.0175)
Years of Experience Squared	-0.0022
	(0.0009)
AFQT (θ_0)	-0.0055
	(0.0624)
Phat*AFQT ($\theta_1 - \theta_0$)	0.0887
	(0.1029)

The estimates reported in this table come from a regression of log wages on experience, experience squared, corrected AFQT, P*AFQT (where P, or Phat in the table, is the predicted probability of attending college), and K(P), a nonparametric function of P, K(P) is estimated by local linear regression. We report the coefficients on the remaining variables in the regression.

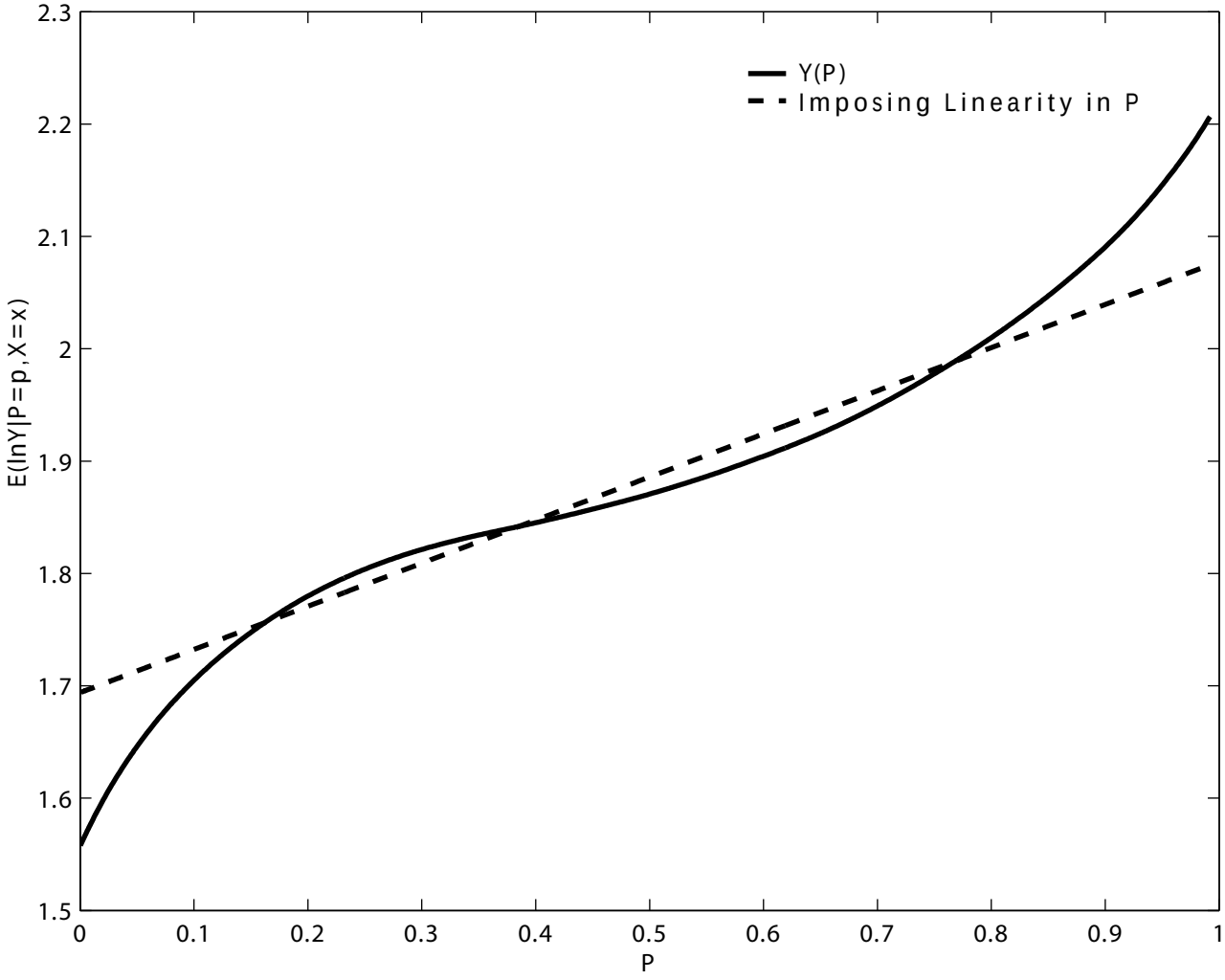
Standard errors are in parenthesis (standard errors are boostrapped) ans account for estimation of P.

Figure 3
Histogram of the Predicted Probability of College Attendance
For Individuals Who only Attended High School (S=0) and Who Attend College (S=1)



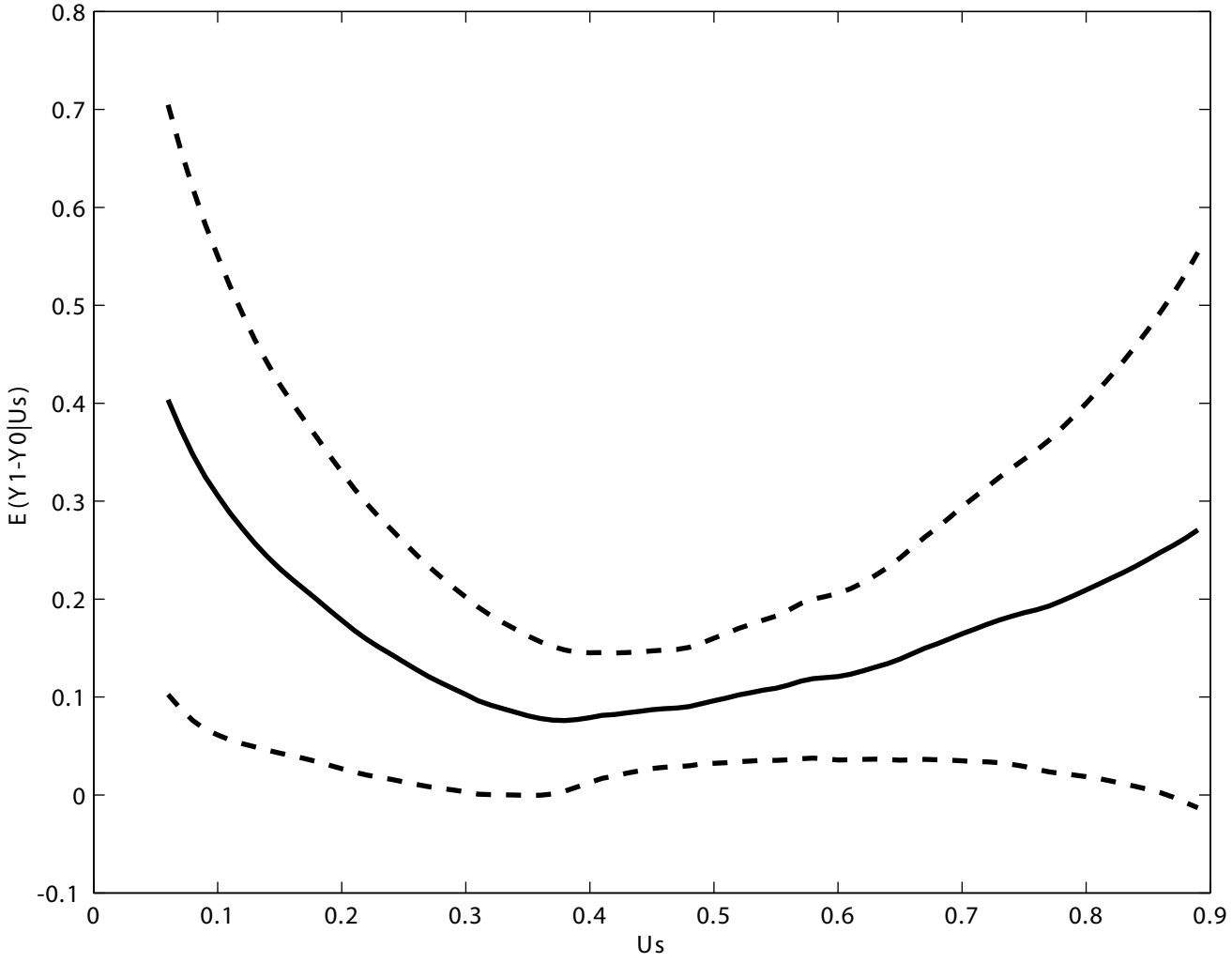
P is the estimated probability of going to college. It is estimated from a (probit) regression of college attendance on corrected AFQT, father's education, number of siblings, tuition, distance to college, local wage and local unemployment.

Figure 4
Estimate of E(lnY|P=p,X=x) using Local Linear Regression



The estimated nonlinear function in this figure comes from a regression of log wages on experience, experience squared, corrected AFQT, P*AFQT (where P is the predicted probability of attending college), and Y(P), a nonparametric function of P. E(lnY|P=p,X=x) is estimated by local linear regression and graphed above. The straight line is generated by imposing that E(lnY|P=p,X=x) is linear in P.

Figure 5
Estimate of E(Y1-Y0|Us)



The estimated function in this figure comes from a regression of log wages on experience, experience squared, corrected AFQT, P*AFQT (where P is the predicted probability of attending college), and K(P), a nonparametric function of P. K(P) is estimated by local linear regression. The function graphed above is E(Y1-Y0|Us) and it is estimated in the following way. First we compute the first derivative of K(P) with respect to P. Then we add a constant term to this function which is simply the average AFQT in the population multiplied by the coefficient on P*AFQT. E(Y1-Y0|Us) is divided by 3.5 to account for the fact that individuals that attend college have on average 3.5 more years of schooling than those who do not. Therefore these correspond to estimates of returns to one year of college.
We could evaluate this function at a different level of AFQT. These affect the level of the MTE function but not the curve of the function. Standard errors are bootstrapped.

Table 6
Estimates of Various Returns to one Year of College

	$0.01 \leq P \leq 0.96$
Average Treatment Effect	0.1870
	(0.0427)
Treatment on the Treated	0.2069
	(0.0722)
Treatment on the Untreated	0.1677
	(0.0586)
Average Marginal Treatment Effect	0.1595
	(0.0351)
Policy Relevant Treatment Effect	0.1592
($1000 tuition subsidy)	(0.0346)
Ordinary Least Squares	0.0760
	(0.0068)
Instrumental Variables	0.1256
	(0.0027)

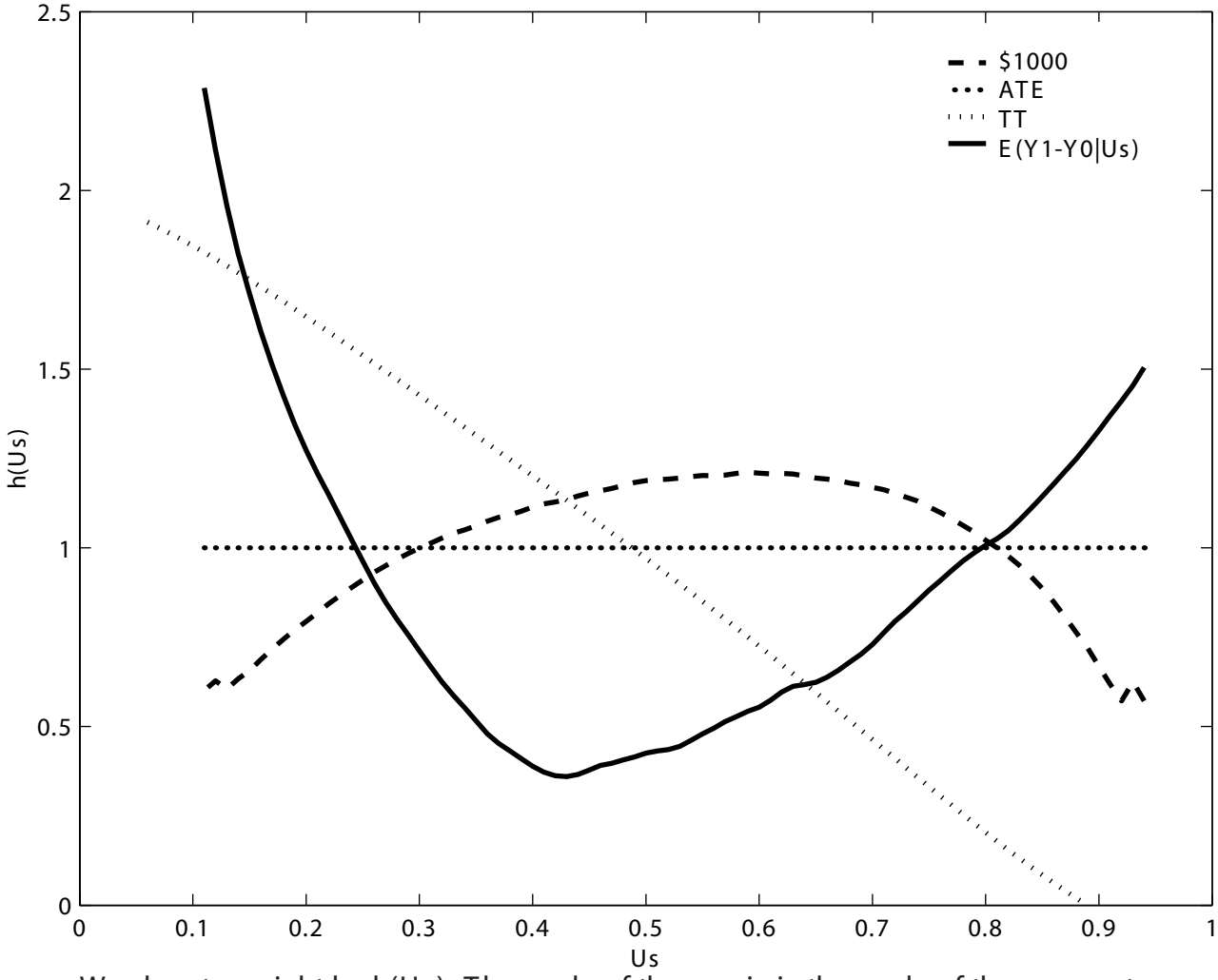
To compute the first four parameters in this table we first estimate the marginal treatment effect (using local linear regression) and then we weight it using the appropiate weights developed in the paper. The OLS and IV estimates are evaluated at the same average value of AFQT as the PRTE estimate (i.e., the average AFQT score for individuals induced to enroll in college by $1000 tuition subsidy).

Table 7
Linear Instrumental Variable Estimates of the Returns to Schooling

Instrumental Variable	
Number of Siblings	0.1100
	(0.2242)
Father's Years of Schooling	0.1748
	(0.0484)
Average County Tuition at 17	0.0901
	(0.1060)
Distance to College at 14	0.4610
	(0.4773)
Average State blue Collar Wage at 17	0.0311
	(1.0541)
County Unemployment Rate at 17	-0.0522
	(0.3457)
Phat	0.1285
	(0.0197)

This table presents linear IV estimates of the returns to one year of college using different instruments. We regress log hourly wages on schooling (college attendence), experience, experience squared, AFQT and an interaction of schooling and AFQT Since the returns to college depend directly on AFQT, we evaluate them at mean level of AFQT for white males, which is 0.58. Phat is the predicted value of a probit regression of college attendance on corrected AFQT, number of siblings, father's years of schooling, tuition, distance, local wage and local unemployment.

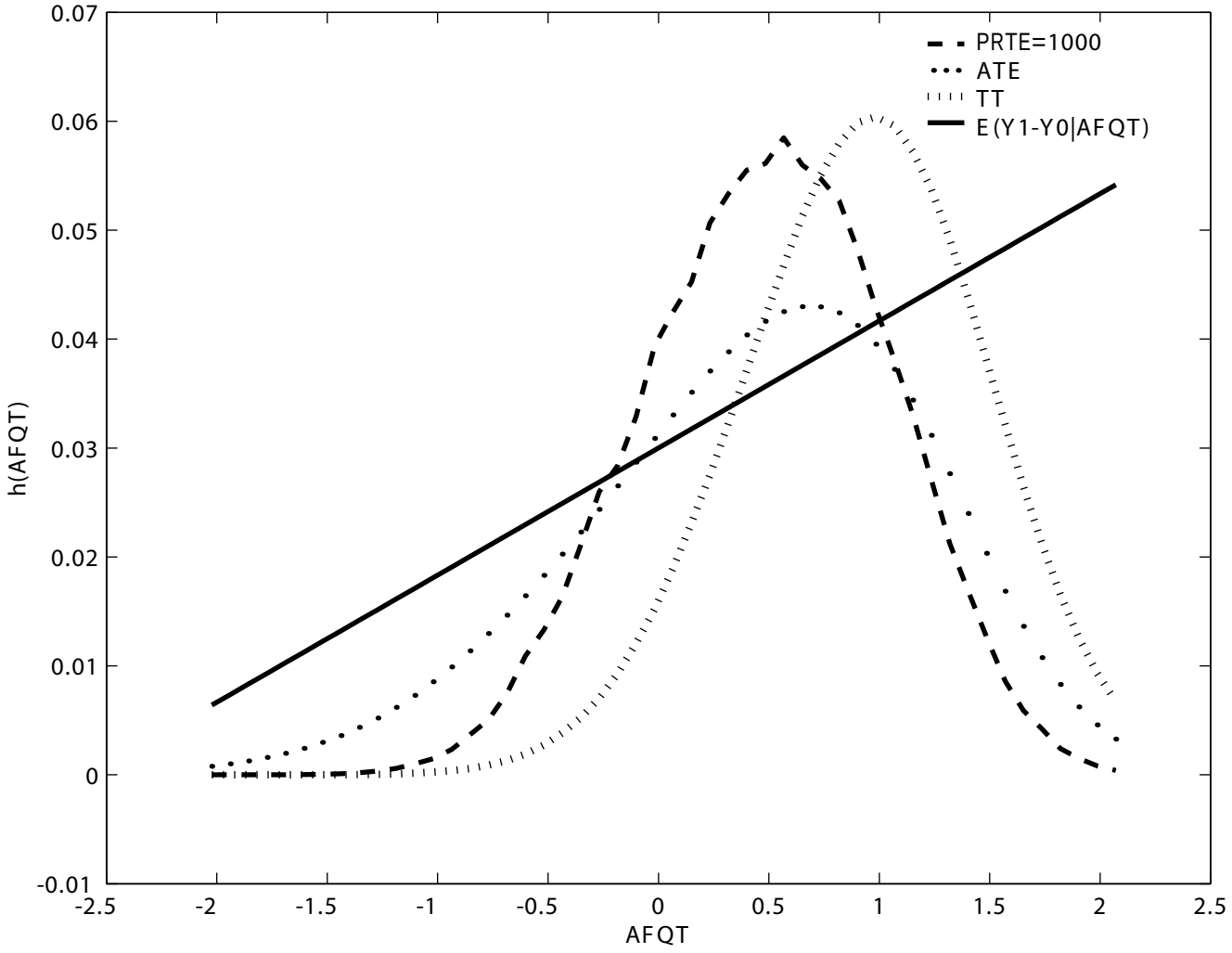
Figure 6
Weights for E(Y1-Y0|Us) (evaluated at mean AFQT)
for Average Treatment Effect, Treatment on the Treated
and Policy Relevant Treatment Effect ($1000 Tuition subsidy)



We denote weight by h(Us). 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.

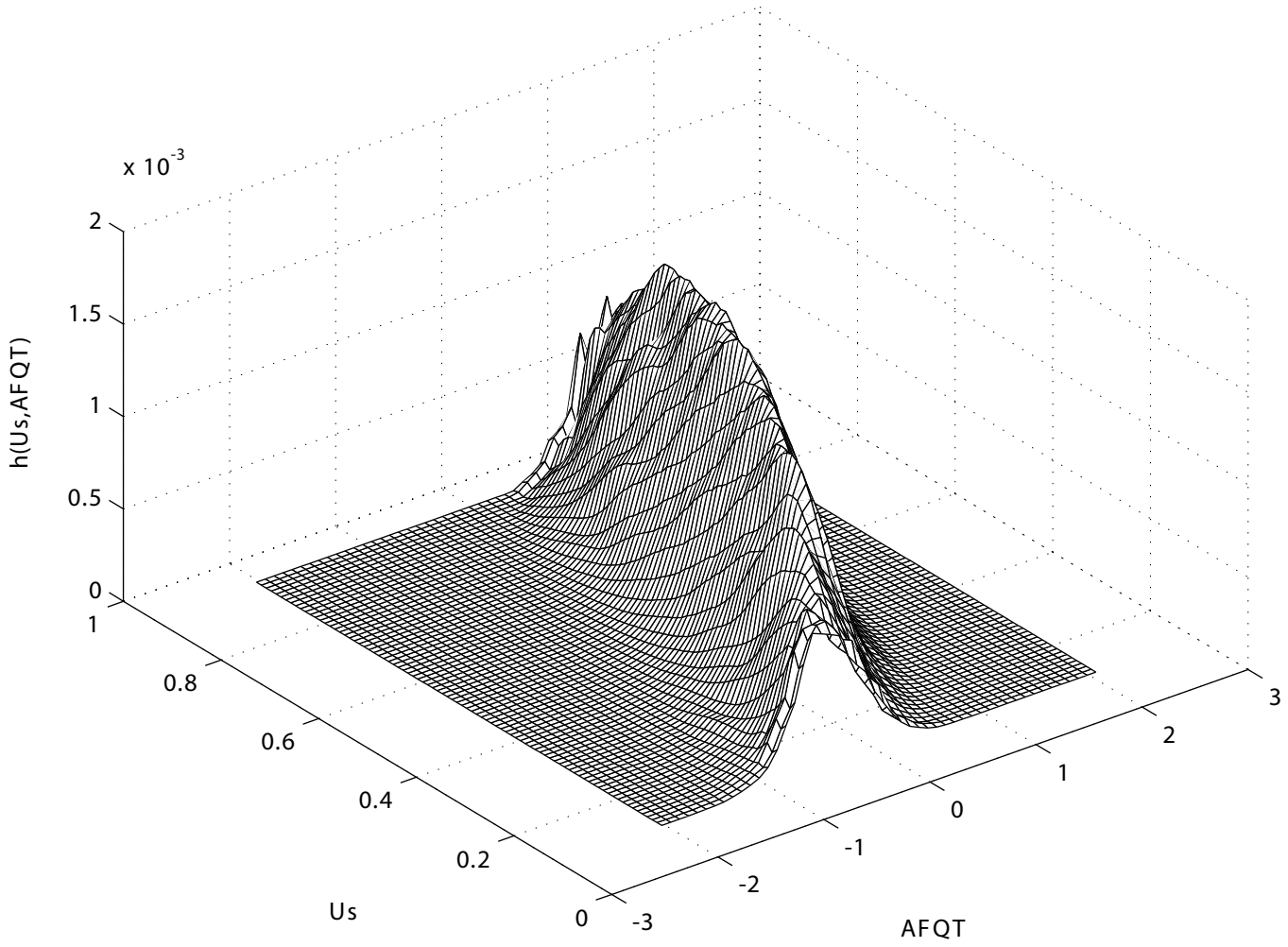
Figure 7

Weights for E(Y1-Y0|Us) (evaluated at mean AFQT)
for Average Treatment Effect, Treatment on the Treated
and Policy Relevant Treatment Effect ($1000 Tuition Subsidy)



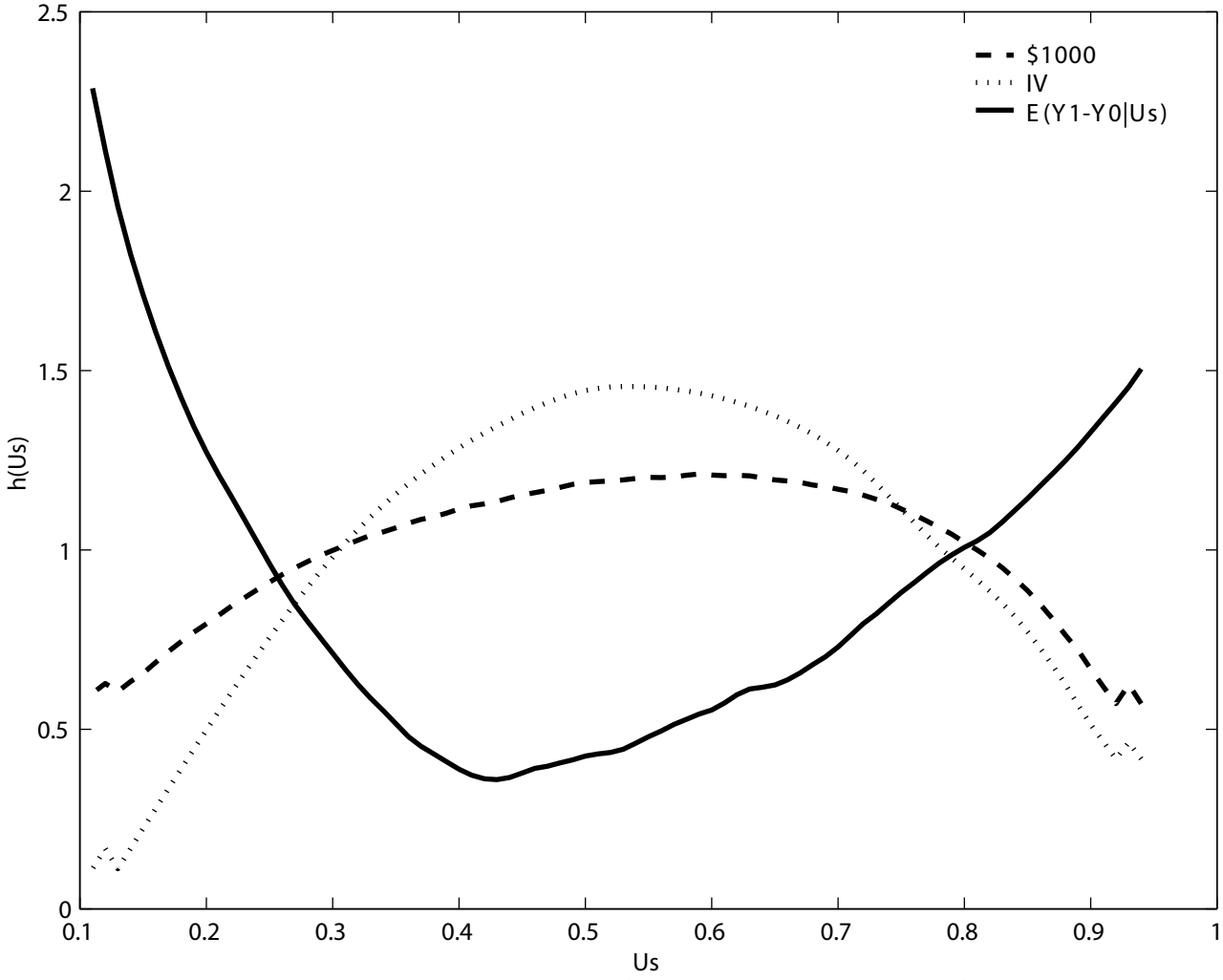
We denote weight by h(AFQT). The scale of the y-axis is the scale of the parameter weights, not the scale of the MTE. MTE is scaled to fit the picture.

Figure 8
Weights for $1000 Tuition Subsidy (Policy Relevant Treatment Effect)



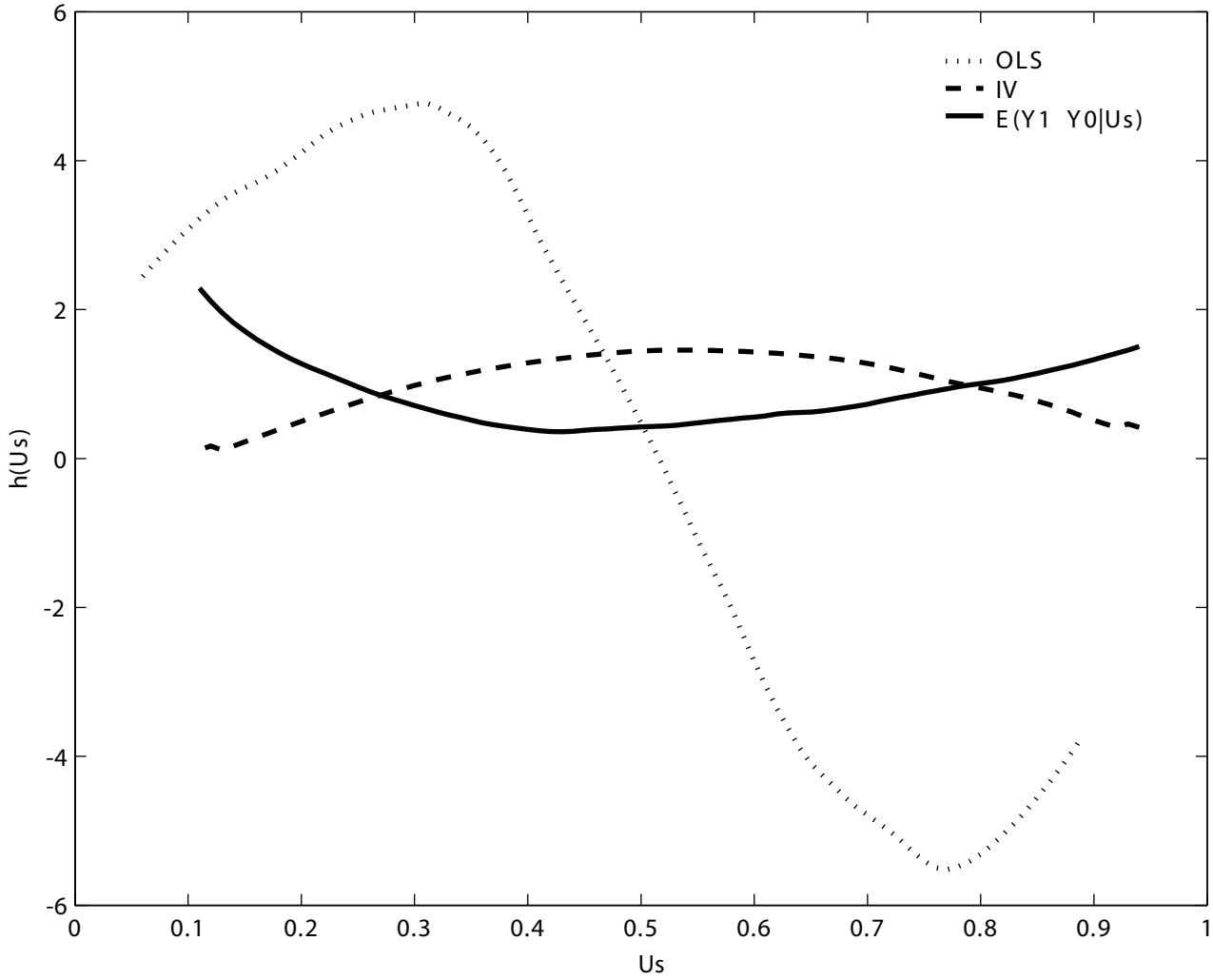
This figure shows the joint (Us,AFQT) policy weights. This is the joint density of (Us,AFQT) for individuals induced to attend college by the tuition subsidy. Carneiro (2002) presents the exact expression for the weight.

Figure 9
Weights for E(Y1-Y0|Us) (evaluated at mean AFQT)
for Instrumental Variables and Policy Relevant Treatment Effect ($1000 Tuition subsidy)



We denote weight by h(Us). 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
Weights for E(Y1-Y0|Us) (evaluated at mean AFQT)
for Ordinary Least Squares and Instrumental Variables



We denote weight by h(Us). 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.

Table 8A
Correlation of the Instrumental Variables (Z) with Schooling (S) and AFQT(A)

Instruments	ρz,s	ρz,a	F-statistic
Number of Sibling	-0.1159	-0.1016	26.05
	(0.0206)	(0.0221)	
Father's Education	0.3821	0.3527	312.76
	(0.0175)	(0.0185)	
Average County Tuition at 17	-0.0956	-0.0420	17.64
(in $100)	(0.0228)	(0.0249)	
Distance to College at 14	-0.0705	-0.0795	8.92
	(0.0233)	(0.0250)	
Average State Blue Collar Wage at 17	-0.0560	0.0194	5.86
(in dollars)	(0.0212)	(0.0223)	
County Unemployment Rate at 17	-0.0097	-0.0015	0.18
(in %)	(0.0231)	(0.0228)	

The first column presents the correlation between the instruments and college attendance.The second column presents the correlation between the instrument and measured ability (AFQT). For each correlation in the first and second column I use all white males in NLSY with nonmissing values for the variables needed (instrument,college attendance and AFQT).GED recipients and high school dropouts are excluded. Standard errors are bootstrapped, the number of replications is 100

Table 8B
Residualized Correlation of the Instrumental Variables (Z) with Schooling (S) and AFQT(A)

Instruments	ρz,s	ρz,a	F-statistic
Average County Tuition at 17	-0.0519	0.0108	4.94
(in $100)	(0.0234)	(0.0236)	
Distance to College at 14	-0.0313	-0.0407	1.67
	(0.0249)	(0.0260)	
Average State Blue Collar Wage at 17	-0.0508	0.0369	4.59
(in dollars)	(0.0218)	(0.0226)	
County Unemployment Rate at 17	0.0061	0.0124	0.07
(in %)	(0.0237)	(0.0245)	

Schooling Instruments and Test Scores are first residualized on number of siblings and father's education using linear regression. Then the first column presents the correlation between the instruments and college attendance.The second column presents the correlation between the instrument and measured ability (AFQT). For each correlation in the first and second column I use all white males in NLSY with nonmissing values for the variables needed (instrument,college attendance and AFQT).GED recipients and high school dropouts are excluded. Standard errors are bootstrapped, the number of replications is 100

Table 9
Estimates of Returns to One Year of College

	Model Without Including AFQT	Model Including AFQT
Average Marginal Treatment Effect	0.1702	0.1502
	(0.0214)	(0.0345)
OLS	0.0989	0.0760
	(0.0187)	(0.0067)
IV	0.1758	0.1256
	(0.0176)	(0.0272)

Both regressions estimate a random coefficient model: ln $Y = \alpha + \beta S + \varepsilon$ where β varies among individuals. The only difference is that for the model underlying the numbers in the column on the right of the table AFQT influences both α and β , while in the model for the column on the left of the table AFQT is excluded from the model. The OLS and IV estimates in the column on the right are evaluated at the average level of AFQT for the marginal individual. The instrumental variable we use to compute the IV estimate is the predicted value of a probit regression of college attendance on corrected AFQT, number of siblings, father's years of schooling, tuition, distance, local wage and local unemployment.

Appendix A Description of the Data

We restrict the NLSY[1] sample to white males with a high school degree or above. We define high school graduates as individuals having high school degree, or having completed 12 grades and never reporting college attendance. We define participation in college as having ever gone to college or having complete 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 and who never reported college attendance are excluded from the sample. The wage variable that is used is an average of all deflated (to 1983) no missing hourly wages from 1990 to 1994. Experience is actual work experience in weeks accumulated from 1979 to 1992 (annual weeks 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 number of siblings, father's years of schooling, schooling corrected AFQT, year of birth dummies, average deflated (to 1993) tuition of the colleges in the county the individual lives in at 17 (we simulate the policy change by decreasing this variable by \$1000 for each individual), distance to the nearest college at 14, average local blue collar wage in state of residence at 17 (or in 1979, for individuals entering the sample at ages older than 17) and local unemployment rate in county of residence in 1979. For the construction of the tuition variable see Cameron and Heckman (2001). Distance to college is constructed by matching college location data in the Higher Education General Information Survey (HEGIS)[2] with county of residence in NLSY. State average blue collar wages are constructed using data from the BLS by matching state blue collar wages with state of residence at 17 (or in 1979, for individuals entering the sample at ages older than 17). For a description of the NLSY sample see BLS (2001). The NLSY79 has an oversample of poor whites which we exclude from this analysis. We also exclude the military sample. We are left with 2439 white males, and 1916 of these have a high school degree (or equivalent) or above. Of these 1916 white males, 1718 have a valid hourly wage observation for 1992, 1813 have a nonmissing corrected AFQT, 1831 have nonmissing father's education, 1790 have nonmissing distance, and 1862 have nonmissing local labor market observations (there are no missing values for the variables not mentioned). The main variables causing the reduction in the sample to 1444 are the wage variable and the distance variable. Many individuals report having a bachelors degree or more and at the same time having 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 the estimated returns in the paper by $3.5/3.2 = 1.09$. To remove the effect of schooling on AFQT we implement the following procedure (based on Hansen, Heckman and Mullen, 2003). Let T be test score, S_T schooling at test date and A cognitive ability. Assume that T is an additive function of S_T, A and ε:

$$T = \phi(S_T) + \delta(A)$$

For exposition, assume $\phi(S_T) = \theta S_T$ and $\delta(A) = A$ (the scale of the test score does not have intrinsic meaning and therefore we normalize the scale of ability to be the scale of the test score). Assume $A = \bar{A} + \nu$, where $\bar{A}$ is mean ability in the population. The goal is to recover A, and for that we need to estimate θ. Individuals have to take the test in 1981.

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

[2] For a description, see National Center for Education Statistics (2003).

Since they were born in different years, they start school at different dates and in 1981 they have different amounts of schooling. If everybody was still in school in 1981, there was no grade repetition and no school interruptions (nobody in the sample dropped out of school in some year prior to 1981 and then return in a subsequent year, also prior to 1981), then we could assume that $A \perp\!\!\!\perp S_T$ since S_T is random. However, if in 1981 some individuals dropped out of school already then

$$E(T|S_T) = \theta S_T + E(A|S_T)$$

where A and S_T are likely to be correlated if individuals with low A drop out of school earlier. If $COV(A, S_T) > 0$ then

$$\text{plim}\hat{\theta}_{OLS} = \theta + \frac{COV(S_T, A)}{V(S_T)} > \theta$$

and therefore $\hat{A} = T - \hat{\theta}_{OLS} S_T < A$. Restricting the exercise to individuals who have not completed schooling at the test date does not solve the problem because we get choice-based sampling bias. However, we observe individuals in the NLSY into their adult years and therefore we know the completed level of schooling for everyone. Let S_C denote completed schooling. Then,

$$E(T|S_T, S_C) = \theta S_T + E(A|S_T, S_C)$$

Suppose completed schooling is a function of ability:

$$S_C = \lambda(A) + \eta$$

where the η are other factors that influence schooling attainment and are uncorrelated with A. Then for those individuals who have completed their schooling at test date, $S_T = S_C$ and therefore $E(A|S_T, S_C) = E(A|S_C)$. For those who have not completed schooling in 1981, the date of the test is exogenous. In particular, consider the set of individuals with $S_C = s_C$. For this group of people, we will assume that S_T and A are independent precisely because the date of the test is exogenous. Everybody completes the same level of schooling at the end and therefore we do not need to worry about the fact that those who had less schooling at test date dropped out of school early because of low ability. In other words, we assume that $E(A|S_T, S_C) = E(A|S_C)$. Therefore, the solution to the problem is to run a regression of test scores on schooling at test date within groups of individuals with the same level of completed schooling. Another way to state this is

$$E(T|S_T, S_C) = \theta S_T + E(A|S_C)$$

and therefore all we need to do is to include a general function of S_C in the regression (what is usually called a control function). In this paper, we group individuals in 7 levels of schooling at test date ($\leq$8, 9, 10, 11, 12, 13-15, 16+) and 4 levels of completed schooling (high school dropout, high school graduate, some college, college graduate) and run a regression of AFQT on dummy variables for schooling at test date (we do not assume that $\phi(S_T)$ is linear in S) and dummy variables for completed schooling ($E(A|S_C)$, the control function). Then we use the coefficients on the former set of dummy variables to correct AFQT. These coefficients are presented in Table A1. Finally, after correcting AFQT (not only for white males, but for all race and gender groups) we standardize it to have mean zero and variance one.

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)

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

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

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
Formal Test of Selection On Unobservable Returns

Phat Squared	-3.0879
	(1.3461)
Phat Cubed	2.1155
	(0.8817)
F test	F(2,1390)=2.8806
p-value	0.0564

The last two lines refer to a test of joint significance of the coefficients on phat squared and phat cubed. The estimates on this table come from a regression of log wages on experience, experience squared, corrected AFQT, P (or Phat in the table, th predicted probability of attending college), P*AFQT, P squared and P cubed. We report the coefficients on the last two variables. Standard errors are in parenthesis.

Standard errors are bootstrapped to account for the fact that P is an estimated object.The regression is only run over the relevant support of P, i.e., for P between 0.001 and 0.96

Table A3
OLS Estimates of a Regression of Log Wage on Experience, AFQT and College Attendance

Experience	0.0756
	(0.0185)
Experience Square	-0.0020
	(0.0010)
CAFQT	0.0595
	(0.0188)
College	0.2088
	(0.0318)
College*CAFQT	0.0862
	(0.0292)
Intercept	1.7392
	(0.0849)

Estimates of Return to College at Different Values of AFQT

	Min(AFQT)=-2.66	Mean(AFQT)=0.58	Max(AFQT)=2.72
β	-0.0060	0.0740	0.1268

The estimated regression is:

$ln\ Y=\alpha_0+X\theta_0+S[\alpha_1-\alpha_0+X(\theta_1-\theta_0)]+U$, where U is the error term of the regression. Therefore:

$\beta=\alpha_1-\alpha_0+X(\theta_1-\theta_0)$ where $\alpha_1-\alpha_0$ is estimated from the coefficient on College and $\theta_1-\theta_0$ is the coefficient on College*AFQT (we assume $\theta_1-\theta_0=0$ for experience and experience squared). Since the difference in the average years of schooling of high school graduates and individuals that attend college is 3.5 we divide β by this number in order to generate returns per year of schooling. Those are reported in the bottom panel of the table for different values of AFQT

Table A4
Returns to College Under Different Extrapolations and Definitions of Support

	$0.01 < P < 0.96$		$0.05 < P < 0.90$		E1	E2	E3 (P^3)	E4 (P^5)	E5 (P^5)	E6 (P^6)
ATE	0.1870		0.1625		0.2073	0.2086	0.2082	0.2077	0.2721	0.2663
TT	0.2069		0.1807		0.2213	0.2233	0.1886	0.2406	0.2993	0.3225
TUT	0.1677		0.1456		0.1931	0.1935	0.2282	0.1742	0.2443	0.2090
AMTE	0.1595		0.1502		0.1616	0.1617	0.1561	0.1565	0.1671	0.1662
PRTE	0.1592		0.1505		0.1612	0.1612	0.1568	0.1558	0.1660	0.1650
Bounds for ATE	Lower	Upper	Lower	Upper						
	0.1779	0.2437	0.0878	0.2852						

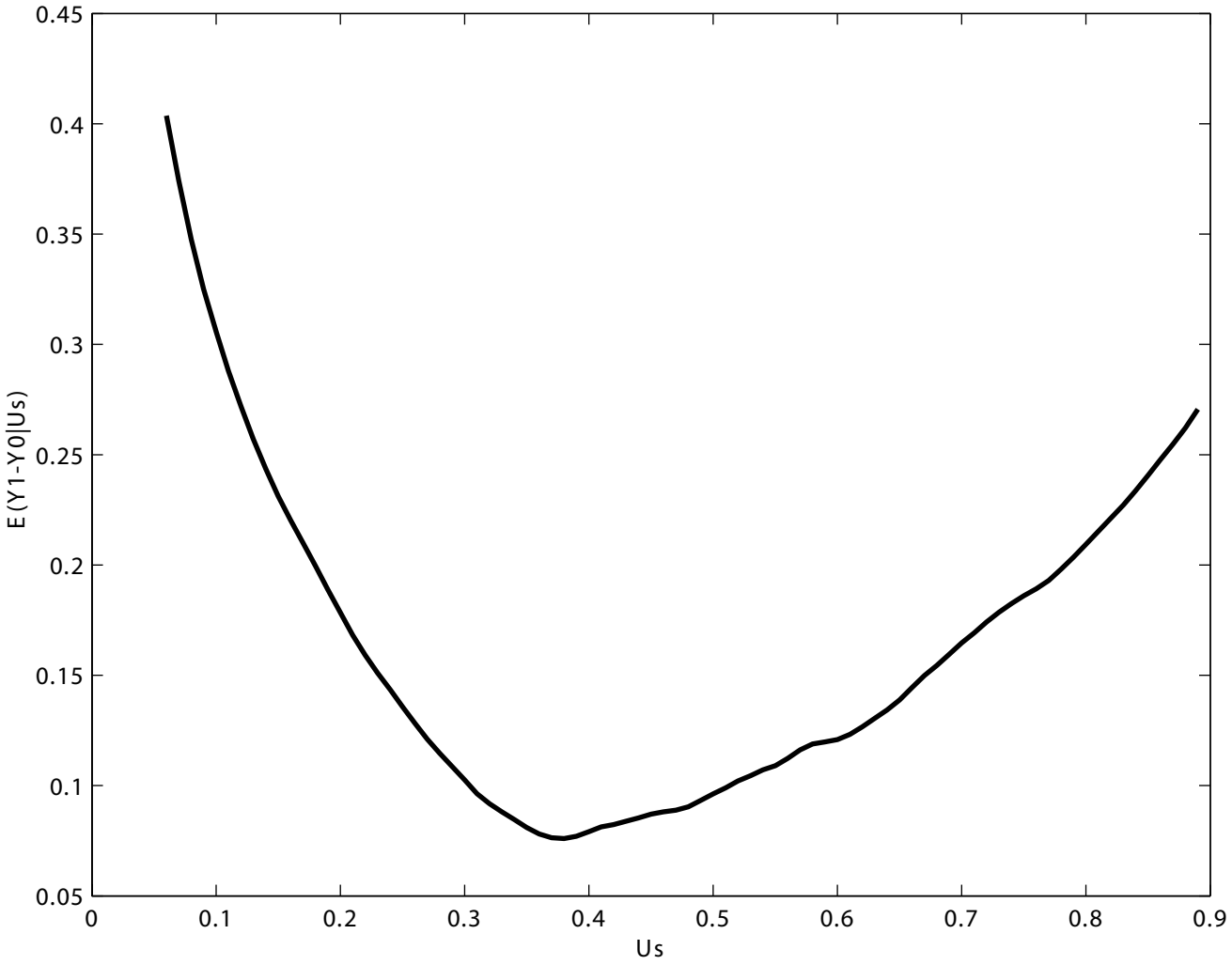
The bounds estimated in the bottom panel of this table are developed in Heckman and Vytlacil (2000). Let $ATE(x) = E(Y_1 - Y_0|X)$, $p_x^{\max}$ and $p_x^{\min}$ be the minimum and maximum values of P in the sample, and y_x^u and y_x^l be the upper and lower bounds for the values of log wages. We assume these variables do not depend on x and choose $y^u = \ell n(100)$ and $y^l = \ell n(1)$. Then these bounds are the following:

$$\begin{aligned} ATE(X=x) &\leq p_x^{\max} E[\ell n Y_1 | X=x, P(Z)=p_x^{\max}, D=1] + (1-p_x^{\max}) y_x^u \\ &\quad - \left(1-p_x^{\min}\right) E\left[\ell n Y_0 | X=x, P(Z)=p_x^{\min}, D=0\right] - p_x^{\min} y_x^l \\ ATE(X=x) &\geq p_x^{\max} E[\ell n Y_1 | X=x, P(Z)=p_x^{\max}, D=1] + (1-p_x^{\max}) y_x^l \\ &\quad - \left(1-p_x^{\min}\right) E\left[\ell n Y_0 | X=x, P(Z)=p_x^{\min}, D=0\right] - p_x^{\min} y_x^u \end{aligned}$$

We evaluate these bounds at $x = E(X)$.

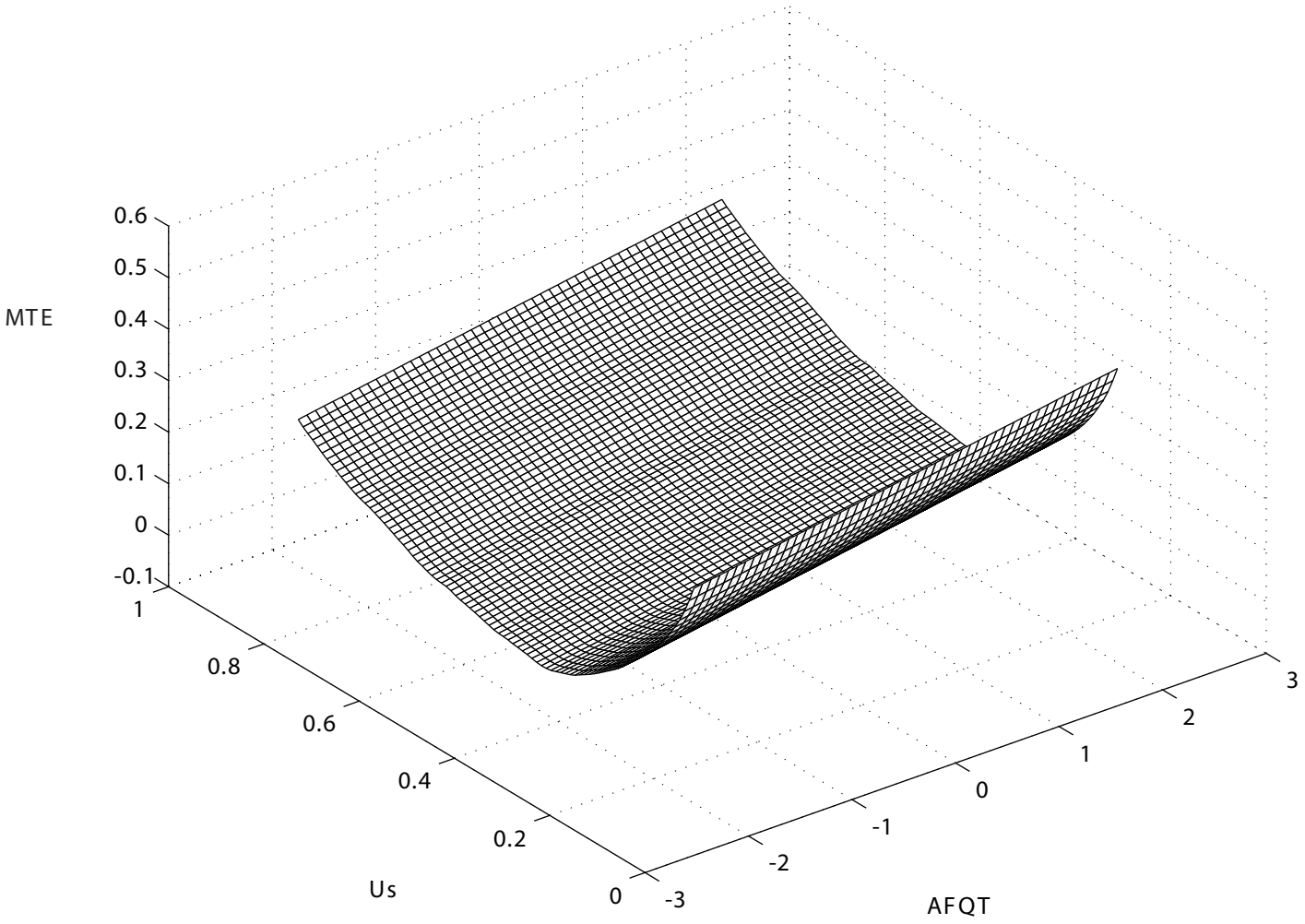
For extrapolation 1 (E1) we fit a linear regression at the extremes of MTE (last 10% on each tail: $0 \leq P \leq 0.1$ and $0.9 \leq P \leq 1$) and extrapolate. Then I do calculations assuming full support. For extrapolation 2 (E2) we fit a quadratic regression on each tail. For E3-E6 we estimate K(P) using polynomials in P and assuming support from 0 to 1 (full support). The label in the column indicates the degree of the polynomial: P^3 means that we fit a cubic ($K(P) = \alpha + \beta P + \gamma P^2 + \theta P^3$).

Figure A1
Estimate of E(Y1-Y0|Us)



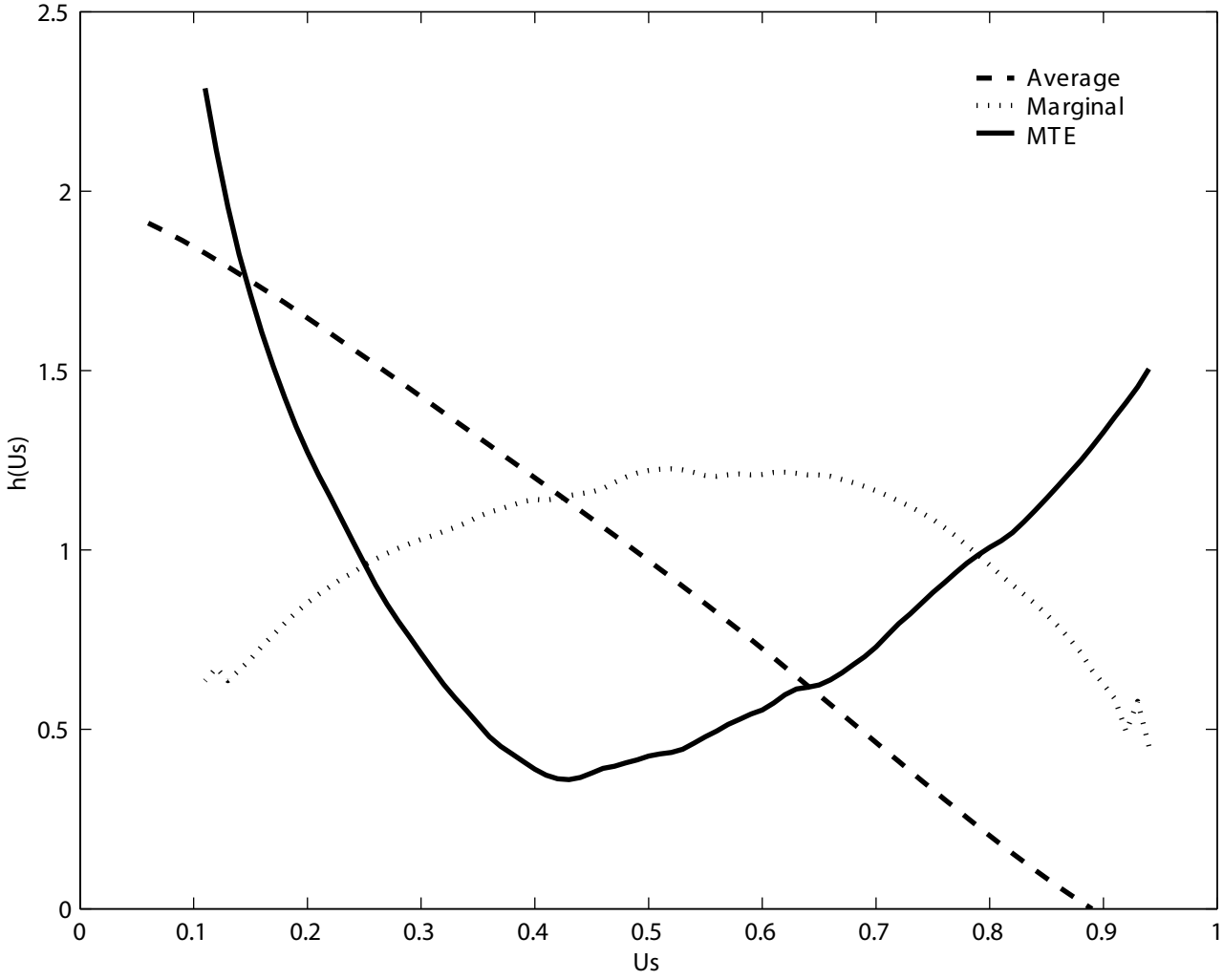
The estimated function in this figure comes from a regression of log wages on experience, experience squared, corrected AFQT, P*AFQT (where P is the predicted probability of attending college), and K(P), a nonparametric function of P. K(P) is estimated by local linear regression. The function graphed above is E(Y1-Y0|Us) and it is estimated in the following way. First we compute the first derivative of K(P) with respect to P. Then we add a constant term to this function which is simply the average AFQT in the population multiplied by the coefficient on P*AFQT. E(Y1-Y0|Us) is divided by 3.5 to account for the fact that individuals that attend college have on average 3.5 more years of schooling than those who do not. Therefore these correspond to estimates of returns to one year of college.
We could evaluate this function at a different level of AFQT. Different levels of AFQT shift this function in a parallel fashion.

Figure A2
Estimate of the Marginal Treatment Effect: E(Y1-Y0|AFQT,Us)



The estimated function in this figure comes from a regression of log wages on experience, experience squared, corrected AFQT, P*AFQT (where P is the predicted probability of attending college), and K(P), a nonparametric function of P. K(P) is estimated by local linear regression. Let gamma be the coefficient on P*AFQT. Then the function graphed above is E(Y1-Y0|AFQT,Us) = gamma*AFQT + E(U1-U0|Us) (in this case the only relevant X variable is AFQT), where E(U1-U0|Us) is equal to the first derivative of K(P) with respect to P. E(Y1-Y0|AFQT,Us) is divided by 3.5 to account for the fact that individuals that attend college have on average 3.5 more years of schooling than those who do not. Therefore these correspond to estimates of returns to one year of college

Figure A3
Weights for E(Y1-Y0|Us) (evaluated at mean AFQT)
for Average and Marginal Person Weights



We denote weight by h(Us) . 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.