Estimating The Return to Education When It Varies Among Individuals

Pedro Carneiro
University of Chicago

James J. Heckman[1]
University of Chicago and The
American Bar Foundation

Edward Vytlacil
Stanford University

October, 2000
Revised, April, 2001
Revised, May, 2001

[1]This research was supported by NSF 97-09-873 and NICHD-40-4043-000-85-261. We have benefitted from comments received at the Applied Price Theory Workshop, especially those of Kevin Murphy, and from participants at the Royal Economic Society, Durham, England, April 10, 2001; especially those of Richard Blundell, Partha Dasgupta and Costas Meghir.

In response to increases in the measured economic return to higher education, there has been substantial interest in estimating the "true" rate of return. In approaching this empirical measurement problem, two fundamentally different views of the labor market have been taken. These two views lead to different econometric estimation strategies.

The first view, associated with Griliches (1977), adopts an efficiency units view of the labor market. Human capital is homogeneous but people possess different amounts of it. This literature focuses on ability bias and measurement error as the central problems in estimating "the" return to schooling. Unmeasured or erroneously measured human capital components summarized by "ability" play a central role and the econometric problem is one of omitted variable bias in an earnings equation. The favored estimator in this literature is the method of instrumental variables.

The second view, associated with Roy (1951), Willis and Rosen (1979) and Willis (1986), focuses more explicitly on the choice of schooling and emphasizes comparative advantage in a labor market with heterogeneous human capital as a guiding principle.[2] This viewpoint is at odds with an efficiency units point of view. Formally, the econometric model is one of a correlated random coefficient model. Econometric models of self selection of the type originally used to model the choice of labor supply and wages are typically applied to estimate this model.

The recent literature on estimating the return to schooling favors instrumental variables estimators over selection models on the grounds that instrumental variables estimators are more robust (Moffitt, 1999; Krueger, 2000).[3] While robust methods are always preferred to

[2]See Sattinger, 1993 for a survey of the evidence on comparative advantage in the labor market.

[3]Thus Krueger writes: "The econometric methods and strategies commonly employed in labor economics research have changed.... Based on his survey of econometric techniques used in articles on labor

non-robust methods, the recent literature emphasizes properties of estimators rather than the economic content of what is being estimated. It neglects a crucial feature of models of comparative advantage when some components of gains are observed by the agents being studied but not by the econometrician studying them. In such models there is a distribution of rates of return that cannot, in general, be summarized by a single parameter. There is no single average rate of return that qualifies as "the" rate of return, and different estimators and instruments define different parameters.

This paper seeks to redress the balance in the literature. In it, we define and estimate parameters that answer well posed economic questions. We compare the 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.

Specifically, this paper makes six main points.

- In the presence of heterogeneous responses to schooling on which individuals act (select into schooling) a variety of "effects" of schooling can be defined and conventional intuitions about instrumental variables estimators break down.

- The empirical evidence from several data sets in the U.S. suggests that heterogeneity in the response to schooling on which individuals act is a central feature of the data.

economics published in eight leading economics journals in 1985-87 and 1995-97, Robert Moffitt (1999) writes, "Selection bias methods of all types have shown a marked decline over the period. This includes the two-step methods as well as full-information maximum likelihood methods....Moving in the opposite direction are methods using IV [instrumental variables] or two-stage least squares (2SLS), which have grown enormously." He interprets these trends as evidence that econometric practice in labor economics is shifting toward techniques that are, or at least can be argued to be, less restrictive and more robust than some of those used in the past." Krueger (2000)

This evidence challenges the validity of conventional instrumental variable methods for providing the estimates required to answer well posed policy evaluation problems.

- The contrast often made in the empirical literature between IV and selection models is a false one. Recently developed IV methods are special cases of nonparametric selection models.

- The Marginal Treatment Effect (MTE) introduced in Björklund and Moffitt (1987) and developed by Heckman and Vytlacil (1999, 2000) provides a link between the two literatures.

- Evidence that instrumenting schooling raises the coefficient on schooling above the value produced by OLS says nothing about credit constraints governing schooling decisions. Rather, it just reveals that IV weights the MTE differently than OLS.

- The evidence from several data sets using robust semiparametric methods suggests that there is comparative advantage in the labor market and that selection bias is empirically important in estimating the economic returns to schooling.

The plan of the rest of the paper is as follows. Section 1 presents the two basic models of schooling that are used in the current literature: a common coefficient model and a random coefficient model. Section 2 presents the policy parameters of interest studied in this paper. Section 3 characterizes two approaches to estimating the random coefficient schooling model the classical parametric approach and a nonparametric approach developed by Heckman and Vytlacil (1999, 2000, 2001a, 2002). Section 4 presents the policy relevant treatment effect introduced in Heckman and Vytlacil (2001b) that is the central object of attention

in this paper. Section 5 asks and answers the question "What Does The Instrumental Variable Estimator Estimate?". Section 6 shows how to estimate the marginal treatment effect (MTE) which is the building block for all of our analyses. Section 7 presents our estimates. Section 8 compares our analysis and estimates with those reported in the recent literature. Section 9 concludes.

1 Models with Heterogeneous Returns to Schooling

Consider the familiar semilog specification of the earnings equation popularized by Mincer (1958; 1974):

$$\ell nY = \alpha + \beta S + U \qquad E(U) = 0. \tag{1}$$

For much of this paper, S will be binary corresponding to two schooling levels ($S = 0$ or $S = 1$) to simplify the exposition and connect to the empirical work reported in Section 7. For simplicity we suppress explicit notation for dependence on the covariates X unless it is helpful to make this dependence explicit. Under special conditions explicitly discussed in Willis (1986) and Heckman, Lochner and Todd (2001), β estimates the rate of return to schooling.[4]

When β is a constant for all persons (conditional on X), we obtain the efficiency units model of Griliches (1977). In his paper, measured S may be correlated with unmeasured U because of omitted ability factors and because of measurement error in S. He advocates instrumental variable estimators for β to alleviate these problems. In his framework, because β is a constant, there is a unique effect of schooling ("the" effect of schooling).

[4]Heckman, Lochner and Todd (2001) demonstrate that these conditions applied in U.S. data for the period 1940-1970 but no longer apply.

In terms of a model of counterfactual states or potential outcomes, we have two potential outcomes $(\ell nY_0, \ell nY_1)$:

$$\ell nY_0 = \alpha + U, \quad \ell nY_1 = \alpha + \beta + U$$

and $\ell nY_1 - \ell nY_0 = \beta$ a constant 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. Variation in returns to schooling are used to partially account for variation over time in aggregate inequality.[5]

The 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 is dependent 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:

(2a) $$\ell nY_0 = \alpha + \quad U_0$$

(2b) $$\ell nY_1 = \alpha + \overline{\beta} + U_1$$

where $E(U_0) = 0$ and $E(U_1) = 0$ so α and $\alpha + \bar{\beta}$ are the mean potential outcomes for ℓnY_0 and ℓnY_1 respectively. The causal effect of education is

$$\beta = \ell nY_1 - \ell nY_0 = \overline{\beta} + U_1 - U_0$$

and we may write observed income as

[5]See Becker and Chiswick (1966), Chiswick (1974) and Mincer (1970,1974, 1995).

$$\ell nY = S\ell nY_1 + (1-S)\ell nY_0 = \alpha + \beta S + U_0 \tag{3}$$
$$= \alpha + \bar{\beta} S + \{U_0 + S(U_1 - U_0)\}$$

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

$$S = 1 \text{ if } \ell n\ Y_1 \geq \ell n\ Y_0 \Longleftrightarrow \beta \geq 0 \tag{4}$$
$$= 0 \text{ otherwise.}$$

If agents know or can partially predict β at the time they make schooling decisions, there is dependence between β and S in equation (3).

In this setup there are three sources of potential econometric problems: (a) S is correlated with U_0; (b) β 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. Thus $\Pr(S = 1 \mid X, \beta) \neq \Pr(S = 1 \mid X)$.

In this framework, β is an ex post causal effect. Ex ante agents may not know β. In the case where schooling decisions 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 $U_1 - U_0 \neq 0$, so 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.

Economists influenced by the biostatistical literature focus on the population average return:

$$E(\beta \mid X = x) = E(\ell nY_1 - \ell nY_0 \mid X = x) = \overline{\beta}(x)$$

which is the return to a random selection of the population for given characteristics $X = x$. This is sometimes called "the" causal effect.[6] Others look at the return for those who attend school:

$$E(\beta \mid S{=}1, X = x){=}E(\ell nY_1\text{-}\ell nY_0 \mid S{=}1, X = x){=}\overline{\beta}(x){+}E(U_1\text{-}U_0 \mid S{=}1, X = x).^7$$

Other parameters can be formulated such as

$$E(\beta \mid S{=}0, X = x){=}E(\ell nY_1\text{-}\ell nY_0 \mid S{=}0, X = x){=}\overline{\beta}(x){+}E(U_1\text{-}U_0 \mid S{=}0, X = x).$$

This is the gain that those who do not go to school would experience if they went to school.

Observe that

$E(\beta - \overline{\beta} \mid S = 1, X = x) = E(U_1 - U_0 \mid S = 1, X = x) - E(U_1 - U_0 | X = x)$ ("Sorting Gain")

is the sorting gain that arises from purposive selection into schooling. In a model based on decision rule (4), this sorting gain is positive. In more general models, it may be of either sign.

Depending on the conditioning sets and the summary statistics desired, one can define a variety of causal "effects". Different causal effects answer different economic questions. Notice that if

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

[7] It is the Treatment on the Treated parameter as discussed by Heckman and Robb (1985, 1986).

I. $U_1 \equiv U_0$ (common effect model)

or

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 collapse to the same parameter and there is a single "causal effect" of schooling. 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.[8]

2 Policy Parameters of Interest

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. Suppose that we adopt a standard welfare framework. Aggregate per capita income under one policy is to be compared with aggregate per capita income under another. One of the policies may be no policy at all. For utility criterion $V(Y)$, a standard welfare analysis compares

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

Adopting the common coefficient view, and a log utility specification ($V(Y) = \ell nY$) and ignoring general equilibrium effects, the mean change in welfare is

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

[8]Heckman and Robb (1985) noted that Lewis' survey of the union effects on wages (1986) confused different "effects". This is especially important in his comparison of cross section and longitudinal estimates.

where $\overline{\beta}$ is "the" causal effect of schooling and (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. This is also the mean change in log income if case II applies. In that case, β is distributed independently of S and we obtain a standard random coefficient model of the sort analyzed by Becker and Chiswick (1996) and Mincer (1974).

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 $\overline{\beta}$ in (5). We present the appropriate parameter and compare it with what IV estimates and the conventional treatment parameters after reviewing some basic results from our previous research that are required to make this comparison. We consider two approaches to estimating the distribution of the returns to schooling, or some features of it.

3 Two Approaches to Estimating the Schooling Model

Consider a standard model of schooling. 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 1 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. Follow Mincer (1974) and assume that earnings profiles in logs are parallel in experience. Thus $Y_1(t) = Y_1 e(t)$ and $Y_0(t) = Y_0 e(t)$.

The agent attends school if

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

where $e(t)$ is a post-school experience growth factor. 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 income in the two schooling choices we may write $Y_1 = (1+g)Y_0$ where from equation (2), $\beta = \ell n(1+g)$. Thus the decision to attend school ($S = 1$) is made if

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

This is equivalent to

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

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

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

Equation (6) generalizes decision rule (4) by adding borrowing and tuition costs as determinants of schooling. Below we introduce variables Z that shift costs and discount factors ($C = C(Z)$, $r = r(Z)$). Notice that

$$E(\beta \mid S = 1, r, C_0, Y_0) = E\left(\beta \middle| \beta \geq r + \frac{C}{Y_0}, r, C_0, Y_0\right)$$

so the average return to schooling conditional on taking schooling is increasing in r and C and decreasing in Y_0 if β is independent of these variables.

The conventional approach to estimating selection models postulates normality of (U_0, U_1) in equations 2(a) and 2(b), writes $\overline{\beta}(X)$ and $\alpha(X)$ as linear functions of X and postulates

independence between X and (U_0, U_1). 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.[9]

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

Using their setup we write

$$\ell nY_1 = \mu_1(X, U_1) \text{ and } \ell nY_0 = \mu_0(X, U_0). \tag{7}$$

The treatment effect is $\ell nY_1 - \ell nY_0 = \beta = \mu_1(X, U_1) - \mu_0(X, U_0)$, 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.

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

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

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. The Z vector may include some or all of the components of X.

Heckman and Vytlacil (2000, 2001) establish that under the following assumptions, it is possible to develop a model that unifies different treatment parameters, that shows how

[9] Aakvik, Heckman and Vytlacil (2000) and Heckman, Tobias and Vytlacil (2000) derive all of the treatment parameters and distributions of treatment parameters for several parametric models.

the conventional IV estimand relates to these parameters and what policy questions IV estimates. Those conditions are

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

(A-2) 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) $\ell n Y_1$ and $\ell n Y_0$ have finite first moments

and

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

Assumption (A-1) assumes 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. (These are determinants of C and r in equation (6)). Assumption (A-2) is made for technical convenience and can be relaxed. 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 interpretability. Defining X_s as the value of X if S 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. 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 $S = 0$ (existence of treatments and controls).

Denoting $P(z)$ as the probability of receiving treatment conditional on $Z = z$, $P(z) \equiv \Pr(S = 1 \mid Z = z) = F_{U_S}(\mu_S(z))$. 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) including for the case of continuous instruments.

This does impose testable restrictions on (Y, S, Z, X). Those are:

(i) Index Sufficiency

$$\Pr(\ell n Y_j \in A \mid Z = z, S = j) = \Pr(\ell n Y_j \in A \mid P(Z) = P(z), S = j).$$

for $j = 0, 1$. This says that Z enters the conditional distribution of $\ln Y_1$, $\ln Y_0$, only through the index $P(Z)$.

A second testable implication is a monotonicity condition:

(ii.a) If $g(\cdot)$ is a function such that $\Pr[g(\ln Y_0) > 0] = 1$, then

$E[g(\ln Y_0)(1 - S) \mid X, P(Z) = p]$ is decreasing in p.

(ii.b) If $g(\cdot)$ is a function such that $\Pr[g(\ln Y_1) > 0] = 1$, then

$E[g(\ln Y_1)S \mid X, P(Z) = p]$ is increasing in p.

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*)

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

This is the marginal gain to schooling of a person 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 by people at the margin of indifference for schooling given X and U_S.[10] The *LATE* parameter of Imbens and Angrist (1994) may be written in this framework.

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

where $u_S \neq u'_S$. *MTE* is the limit of *LATE*, if the limit exists.

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:

$$\Delta^{ATE}(x) = \int_0^1 E(\beta \mid X = x, U_S = u)du.$$

$$\Delta^{TT}(x, P(z), S = 1) = \tfrac{1}{P(z)} \int_0^{P(z)} E(\beta \mid X = x, U_S = u)du$$

$$\Delta^{LATE}(x, u_S = P(z), u'_S = P(z')) = \left[\int_{P(z)}^{P(z')} E(\beta \mid X = x, U_S = u)du \right] \frac{1}{P(z') - P(z)}$$

$$\Delta^{TT}(x, S = 1) = \int_0^1 E(\beta \mid X = x, U_S = u) g_x(u)du$$

$$g_x(u) = \frac{1 - F_{P(Z)|X}(u \mid x)}{\int_0^1 (1 - F_{P(Z)|X}(t \mid x))dt} = \frac{S_{P(Z)}(u \mid x)}{E(P \mid x)}$$

where $S_{P(Z)}(u \mid x)$ is the survivor function for $P(Z)$ given X evaluated at u.

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

Notice that if

1. β is a constant

   or

2. $E(\beta \mid X = x, U_S = u_s) = E(\beta \mid X)$

   (β mean independent of U_S)

then $MTE = ATE = TT = LATE$. This corresponds to the two cases (I and II) in Section 2 where there is no heterogeneity (β constant, case I) or agents don't act on it (Case II).[11]

Throughout this paper, we assume that $P(Z)$ has support equal to the full unit interval, [0,1]. This is found in the data used in this paper. Bounds for the parameters and estimands when the support is less than full are presented in Heckman and Vytlacil (2000, 2001).

4 Policy Relevant Treatment Effects

With the framework of Section 3 in hand, we can answer the policy question framed in Section 2 when β is heterogeneous and people act on β in making schooling decisions - the case of a nonconstant *MTE*. We consider a class of policy interventions that affect P but not $(\ell nY_1, \ell nY_0)$.

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

[11] All of these parameters 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).

with density f_{P^*}. Then we can write

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

where $\omega(u) = F_P(u) - F_{P^*}(u)$ where F_P and F_{P^*} denote the cdf of P and P^*, respectively.[12]

To define a parameter comparable to $\bar{\beta}$ in equation (7), 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.

Notice that these weights differ from the weights on the conventional treatment parameters. Knowing TT or ATE does not answer a well posed policy question except in extreme 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 is required for policy analysis or is required to generate the conventional treatment parameters.

[12] Keeping the conditioning on X implicit, we have

$$E(V(Y) \mid \text{baseline}) = \int_0^1 E(V(Y) \mid P(Z) = p)dF_P(p) =$$

$$\int_0^1 \left[\int_0^1 \mathbf{1}_{[0,p]}(u)E(V(Y_1) \mid U = u) + \mathbf{1}_{(p,1]}(u)E(V(Y_0) \mid U = u)du \right] dF_P$$

$$= \int_0^1 \left[F_P(u)E(V(Y_1) \mid U = u) + (1 - F_P(u))E(V(Y_0) \mid U = u) \right] du$$

where $\mathbf{1}_{\mathcal{A}}(u)$ is an indicator function for the event $u \in \mathcal{A}$. Thus comparing the baseline to the new regime

$$E_{P^*}(V(Y)) - E_P(V(Y))$$

$$= \int_0^1 E(\Delta_V \mid U = u)(F_P(u) - F_{P^*}(u))du.$$ See Heckman and Vytlacil (2001).

5 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 also misleading in the more general case where β varies among the population and schooling choices are made on the basis of it.

In the common coefficient model (1) the econometric problem is that $Cov(U, S) \neq 0$. If there is an instrument Z with the properties (a) $Cov(U, Z) = 0$ and (b) $Cov(Z, S) \neq 0$ then we may identify (consistently estimate) β by IV even though OLS is biased and inconsistent. Thus

$$\text{plim } \hat{\beta}_{IV} = \frac{Cov(Z, \ell nY)}{Cov(Z, S)} = \beta + \frac{Cov(Z, U)}{Cov(Z, S)} = \beta.$$

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

$$\ell nY = \alpha + \overline{\beta} S + \{(U_1 - U_0)S + U_0\}.$$

Finding an instrument Z correlated with S but not U_0 or $U_1 - U_0$ is not enough to identify $\overline{\beta}$ or $\overline{\beta} + E(U_1 - U_0 \mid S = 1)$ or other conventional treatment parameters.[13] Simple algebra reveals that

$$\text{plim } \hat{\beta}_{\text{IV}} = \frac{Cov(Z, \ell nY)}{Cov(Z, S)} = \overline{\beta} + \frac{Cov(Z, U_0)}{Cov(Z, S)} + \frac{Cov(Z, S(U_1 - U_0))}{Cov(Z, S)}.$$

By the standard argument, the second term vanishes ($Cov(Z, U_0) = 0$). But in general the third term does not:

[13] We keep the conditioning on X implicit.

$$\frac{Cov(Z, S(U_1 - U_0))}{Cov(Z, S)} = \frac{Cov(Z, U_1 - U_0 \mid S = 1)P}{Cov(Z, S)} \neq 0$$

where $P = \Pr(S = 1)$. If $U_1 - U_0 \equiv 0$ (common coefficient model) or if $U_1 - U_0$ is independent of S and Z this term vanishes.[14] But in general $U_1 - U_0$ is dependent on S and the term does not vanish.[15]

To see why, consider the schooling choice model of equation (6) when $C = 0$ and r depends on Z ($r = Z\gamma$). Then

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

and $Cov(Z, U_1 - U_0 \mid S = 1) = Cov(Z, U_1 - U_0 \mid \overline{\beta} + U_1 - U_0 \geq Z\gamma)$. 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 an instrument based on compulsory schooling estimates. Compulsory schooling is sometimes viewed as an ideal instrument. But when returns are heterogeneous, and agents act on that heterogeneity in making schooling decisions, compulsory schooling as 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 with compulsion regime, and $T = 0$ otherwise. We assume that T is exogenous, in particular, that $T \perp\!\!\!\perp (U_S, U_0, U_1)|X$.

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

[14] If $U_1 - U_0$ is independent of Z, we have that $U_1 - U_0$ will be independent of (S, Z) if $U_1 - U_0$ does not determine S conditional on Z.

[15] This point was first made by Heckman and Robb (1985, 1986) and Heckman (1997).

$$E(\ell nY \mid X = x, T = 1) = E(\ell nY_1 \mid X = x) = E(\ell nY_1 \mid X = x, S = 1)P(x)$$
$$+E(\ell nY_1 \mid X = x, S = 0)(1 - P(x)),$$

and for individuals in the regime with no compulsion

$$E(\ell nY \mid X = x, T = 0) = E(\ell nY_1 \mid X = x, S = 1)P(x)$$
$$+E(\ell nY_0 \mid X = x, S = 0)(1 - P(x)).$$

>From the difference in conditional means we can identify:

$E(\ell nY \mid X{=}x, T{=}1)$-$E(\ell nY \mid X{=}x, T{=}0)$=(1-$P(x)$)$E(\ell nY_1$-$\ell nY_0 \mid X{=}x, S{=}0)$.

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

$$E(\ell nY_1 - \ell nY_0 \mid X = x, S = 0) = E(\beta \mid X = x, S = 0)$$

but not $ATE = E(\ell nY_1 - \ell nY_0) = \bar{\beta}$ or treatment on the treated

$$TT = E(\ell nY_1 - \ell nY_0 \mid X = x, S = 1) = E(\beta \mid X = x, S = 1).$$

However under cases I and II of Section 2 we identify all three treatment parameters because $E(\ell nY_0 \mid X = x, S = 0) = \alpha, E(\ell nY_1 \mid X = x, S = 0) = \alpha(x) + \overline{\beta}(x)$ and $TT = ATE = MTE = LATE$ because $\Delta^{MTE}(x, u_S)$ does note vary with u_s.

Treatment on the untreated answers an interesting policy question. It is informative about the income gains for a policy directed toward those who ordinarily would not attend schooling and who are selected into schooling at random from this pool. However, it does not in general identify the effect of other policies (*e.g.* tuition subsidies directed toward the very poor within the pool).

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

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

where

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

and $\int_0^1 h_x(u)du = 1$. These weights do not, in general, coincide with the policy weights of Section 4 or the weights for the treatment parameters defined in Section 3.

A closer look at these weights reveals that

$$h_x(u) = \frac{\int_u^1 (p - E(P(Z) \mid X = x)) f(p \mid X = x) dp}{Var(P \mid X)}$$

where $h_x(u) \geq 0$ which achieve a maximum value at $u = E(P(Z) \mid X = x)$ and $h_x(0) = h_x(1) = 0$ and

$$\int_0^1 h_x(u)du = 1.$$

The weights center at the mean of the P data: $h_x(P^{Max}) = 0 = h_x(P^{Min})$ and

$$h_x(p) = 0 \qquad p \leq P^{Min} \qquad p \geq P^{Max}.$$

For proofs, see Heckman and Vytlacil (2000).[16]

The idea of interpreting IV as a weighted average of the limit of $LATE$ can also be found in Card (1999, 2000) (weighted average of the distribution of return to schooling),

[16] Take a more general instrument J, assuming we recenter J so that $E(J) = 0$. Keeping conditioning on X implicit, $\hat{\beta}_{IV} = \frac{E(JY)}{E(JS)}$ where $\hat{\beta}_{IV}(J) = \int MTE(u)h(u;J)du$ and $h(u;J) = \frac{E(J \mid P(Z) \geq u) \Pr(P(z) \geq u)}{E(JP)}$, $\int_0^1 h(u;J)du = 1$. We have the following properties: (i) $h(u;J)$ non-negative iff $E(J \mid P \geq p)$ weakly increasing in p, (ii) Support $h(u;P) = [\underline{p}, \bar{p}]$ (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).

Angrist, Graddy and Imbens (2000) (weighed average of Wald estimators) and Yitzhaki (1996, 1999).

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 all are weighted averages of the *MTE*. Thus we unify the estimation, treatment effect and policy evaluation literatures as generating parameters or estimands as integrals of *MTE* using different weights:

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

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

We can also fit *OLS* into this framework. It is straightforward to show that conditional on $X = x$

$$\begin{aligned}
\text{plim } \hat{\beta}_{OLS} &= \overline{\beta}(x) + E(U_1 \mid X = x, S = 1) - E(U_0 \mid X = x, S = 0) \\
&= [\overline{\beta}(x) + E(U_1 - U_0 \mid X = x, S = 1)] \\
&+ E(U_0 \mid X = x, S = 1) - E(U_0 \mid X = x, S = 0) \\
&= TT(x) + E(U_0 \mid X = x, S = 1) - E(U_0 \mid X = x, S = 0).
\end{aligned}$$

So we may write

$$\hat{\beta}_{OLS} = \int_0^1 \Delta^{MTE}(x, u_S)h_{OLS}(x, u_S)du_S$$

where

$$h_{OLS}(x, u_S) = 1 + \frac{E(U_1 \mid X = x, U_S = u_S)h_1(x, u_S) - E(U_0 \mid X = x, U_S = u_S)h_0(x, u_S)}{MTE(x, u_S)}$$

and under our assumptions,

$$E(U_1 \mid X = x, S = 1) = \int_0^1 E(U_1 \mid X = x, U_S = u_S) h_1(x, u_S) du_S$$

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

$$E(U_0 \mid X = x, S = 0) = \int_0^1 E(U_0 \mid X = x, U_S = u_S) h_0(x, u_S) du_S$$

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

The *OLS* weights are not guaranteed to be positive or to integrate up to one.[17]

Table 1 summarizes the central results 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 1A plots the *MTE* and the weights used to form *ATE*, *TT* and *TUT* for the Roy model with the parameter values displayed at the base of Table 2.[18] *TT* overweights the *MTE* for persons with low values of U_S who, *ceteris paribus,* are more likely to attend school. (See equation (8)). *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 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

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

[18] 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.

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 1B 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 weights are negative and do not necessarily integrate to one. The *IV* estimate is higher than *OLS* estimate even though the marginal persons attracted into schooling by a subsidy have a lower gross return than those who attend school.

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 h_{PRT} weights. We turn to this topic next.

6 Using Local Instrumental Variables to Estimate the MTE

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

$$E(\ell nY \mid Z = z) = E(\ell nY_0 \mid Z = z) + E(\ell nY_1 - \ell nY_0 \mid Z = z, S = 1) \Pr(S = 1 \mid Z = z)$$

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

$$E(\ell nY \mid Z = z) = E(\ell nY_0) + E(\beta \mid P(z) \geq U_S, P(Z) = P(z))P(z).$$

Recall that Z enters the model only through $P(z)$ so we use $P(z)$ as the instrument:

$$E(\ell nY \mid Z = z) = E(\ell nY \mid P(Z) = P(z)).$$

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

$$\frac{E(\ell nY \mid P(Z) = P(z))\text{-}E(\ell nY \mid P(Z) = P(z'))}{P(z)\text{-}P(z')}$$
$$= \overline{\beta} + \frac{E(U_1\text{-}U_0 \mid P(z) \geq U_S)P(z)\text{-}E(U_1\text{-}U_0 \mid P(z') \geq U_S)P(z')}{P(z) - P(z')}$$
$$= \Delta^{LATE}(x, P(z), P(z')), \text{ where } \Delta^{LATE} \text{ was defined above in section 3. When}$$

$U_1 \equiv U_0$ or $(U_1 - U_0) \perp\!\!\!\perp U_S$, corresponding to cases I and II of Section 2 respectively, IV based on $P(Z)$ estimates ATE $(= \overline{\beta}\ )$ because the second term on the right hand side of this expression vanishes. Otherwise IV estimates a difficult to interpret combination of MTE parameters as discussed in the last section.

Another representation of $E(\ell nY \mid P(Z) = P(z))$ that reveals the index structure more explicitly writes

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

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

$$\frac{\partial E(\ell nY \mid P(Z) = P(z))}{\partial P(z)} = \overline{\beta} + \int_{-\infty}^{\infty} (U_1\text{-}U_0) f(U_1\text{-}U_0 \mid U_S = P(z)) d(U_1\text{-}U_0) = MTE.$$

IV estimates $\overline{\beta}$ if $\Delta^{MTE}(x, u_s)$ does not vary with s, and $\Delta^{MTE}(x, u_s)$ not varying with s implies that $E(\ell nY \mid P(Z) = p)$ is a linear function of $P(Z)$. Thus, given that our model and assumptions are correct, a test of the linearity of the conditional expectation of $\log Y$ in $P(Z)$ is a test of the validity of linear IV for $\overline{\beta}$. More generally, a test of the linearity of $E(\ell nY \mid P(Z) = p)$ in $P(Z)$ is a test of whether or not the data are consistent with a correlated random coefficient model. Nonlinearity in $P(Z)$ is consistent with comparative advantage in the labor market for educated labor (*i.e.* it is a test of the null hypothesis that Case I or Case II of Section 2 describe the data).

It is straightforward to estimate the levels and derivatives of $E(\ell nY \mid P(Z) = P(z))$ and standard errors using the methods surveyed in Ichimura and Todd (2002). The derivative estimator of MTE is the local instrumental variable (LIV) estimator of Heckman and Vytlacil (1999, 2000).

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

In this section we estimate the returns to college. We estimate the MTE using the PSID data described in Appendix A. We obtain comparable results for adjacent years of the data and many, but not all, of our qualitative conclusions also hold up when we estimate models based on the NLSY. The role of the instrument Z is played by family background and tuition variables. We impose a probit model for the determination of D,

$$S = \mathbf{1}[Z\gamma - U_S \geq 0]$$

with U_S independent of Z and distributed standard normal. See Appendix Table B-1 for the estimates of the γ vector. The support of the estimated $P(Z)$ is basically the full unit interval.[19] See Figure B-1.

We impose a linear equation for the outcome equation

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

with $(U_0, U_1, U_S) \perp\!\!\!\perp (X, Z)$. Combining with the model for S with the model for Y implies

[19] The estimated support of $P(Z)$ is $[0.03, 1]$.

a partially linear model for the conditional expectation of Y:

$$E(Y|X,Z,S) = X\alpha + \bar{\beta}S + (1-S)K_0(P(Z)) + SK_1(P(Z))$$

where

$$K_0(P(Z)) = E(U_0|P(Z), D=0) = E(U_0|\Phi(U_S) \leq P(Z))$$

$$K_1(P(Z)) = E(U_1|P(Z), D=1) = E(U_1|\Phi(U_S) > P(Z))$$

where $\Phi(\cdot)$ is the standard normal cdf. No parametric assumption is imposed on the distribution of (U_0, U_1), and thus $K_1(\cdot)$ and $K_0(\cdot)$ are unknown functions that must be estimated nonparametrically. This semiparametric, partially linear form for the conditional expectation has several advantages for empirical work.[20] It imposes a dimension reduction compared to a fully nonparametric model, while not restricting the form of the g_0 and g_1 functions and thus allowing greater flexibility than traditional parametric approaches. In addition, imposing the partially linear model weakens the support condition that otherwise would be required on $P(Z)$. In particular, fully nonparametric analysis of all treatment parameters and policy counterfactuals would require the condition that the support of $P(Z)$ conditional on X to 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 $P(Z)$ have support the full unit interval, without requiring that $P(Z)$ conditional on X have support the full unit interval. We estimate the partially linear model using a double residual regression procedure using local linear regression for the requisite nonparametric regression steps, following Heckman, Ichimura, Smith and Todd (1998).

[20] The partially linear model was introduced by Robinson (1988) in the context of a seemingly unrelated regression framework.

Figure 2 plots the estimated $E(\ell n\ Y \mid X, P(Z) = p)$ as a function of p. The coefficients of the linear terms of the partially linear are presented in Appendix B. (See Table B-2). All regressors enter linearly except the control function terms ($K_1(p) = E(U_1 \mid P(Z) = p, S = 1)$ and $K_0(p) = E(U_0 \mid P(Z) = p, S = 0)$) which are nonparametric functions of $P(Z)$. The tests for index sufficiency and monotonicity described in Section 3 are passed. The index model describes the data and the monotoncity tests are satisfied for thresholds above the minimum wage. See Appendix Tables B-3 and B-4 and Figures B-2 and B-3.

Observe that $E(\ell n\ Y \mid X, P(Z) = p)$ is nonlinear in $P(Z)$. Formal tests for linearity reject the null hypothesis of no linearity.[21] As noted in the preceding section, nonlinearity in $P(Z)$ implies that the *MTE* is not constant in U_S and that *IV* does not estimate $\overline{\beta}(x) = ATE$. Thus the data are consistent with comparative advantage in the labor market.

Figure 3 plots the *MTE* derived from Figure 2 using the formula of equation (9). The mean gross return ($E(\beta \mid U_S = u_S)$) is declining in u_S.[22] The most college worthy persons in the sense of having high gross returns are more likely to go to college. Figure 4 plots the *MTE* weight for *IV* and the *MTE* weight for *OLS* on the same scale. Because of the large negative components of the *OLS* weight, it is not surprising that the *OLS* estimate is lower than *IV* even though people induced into schooling by changes in Z have lower gross returns to schooling. (Compare the *OLS* estimate in Table 3 with the *IV* estimate in that table). Figure 5 plots the *MTE* weights for *TT, IV* and *ATE* and also displays the *MTE*.

It is not surprising, in view of the pattern of these weights, to find that $TT = E(\beta \mid$

[21] A test of whether polynomials in $P(Z)$ above order one enter the model indicates that these terms are statistically significant.

[22] 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.

$S = 1) = \overline{\beta} + E(U_1 - U_0 \mid S = 1) > ATE = \overline{\beta} > OLS$. There is a positive sorting gain ($E(U_1 - U_0 \mid S = 1)$ of .058 log points. At the same time there is negative selection bias ($E(U_0 \mid S = 1) - E(U_0 \mid S = 0) = OLS - \text{TT}$). The story that emerges in these data is one of comparative advantage in the labor market. The unobservables in the no schooling state for those who go to school are dramatically lower than the unobservables in the no schooling state for those who do not go to school. This is the story of Willis and Rosen (1979) which holds up in our semiparametric setting. Notice that IV underestimates TT and ATE.

Figure 6 plots the weight for MTE for the policy relevant treatment effect corresponding to the partial equilibrium policy of reducing tuition by \$500, and for the sake of comparison reproduces the weight for IV. The two systems of weights are roughly comparable but by no means identical. The IV estimand understates - by 0.58 log points - the policy relevant treatment effect. The marginal person attracted into schooling has a gross return very close to the policy relevant treatment effect.[23] The rough agreement between the IV weights and the weights for the tuition reduction policy does not hold up for other policies. See Figure 7. Only by accident does IV identify policy relevant treatment effects when the MTE is not constant in U_S.

The picture of the labor market for educated workers that emerges from our analysis is one of comparative advantage and not the conventional model of efficiency units that guides much of the recent research on the economics of education. Not only is there comparative advantage ex post but ex ante agents select into schooling based on the comparative ad-

[23] The marginal person is defined by the limit of the policy relevant treatment effect as the policy change becomes arbitrarily small.

vantage. (Recall our substantial estimated sorting effect of almost six log points). This evidence is in agreement with a broad body of evidence on comparative advantage in the labor market summarized by Sattinger (1993).

8 Relationship of our Work To The Current Literature

In his chapter in the Handbook of Labor Economics, Card (1999), provides a survey of the recent literature on the estimation of returns to schooling. He starts with studies that use instrumental variables based on institutional features of the school system. His favored instruments are compulsory schooling laws, distance to the nearest college and tuition. A general pattern emerges in all of these studies: IV estimates tend to be larger than OLS estimates of the return to schooling.[24] This is true across studies that use different data sets (different cohorts, different countries) and different instruments. By going from OLS to IV the estimates often increase by more than 30% and in some cases by close to 100%[25]. Card proposes different explanations for this phenomenon. The principal parameter to which estimates are compared is ATE. And it is supposed that we should expect that OLS is upward biased for ATE. Based on this, one explanation, suggested by Bound and Jaeger (1996), is that IV estimates are even further upward biased for ATE than are OLS estimates. Unobserved differences between the treatment and comparison groups implicit in the use of IV may be accentuating the bias instead of atenuating it. As Card (2000)

[24] Another pattern is that IV estimates are also more imprecise than OLS estimates of the return to schooling.

[25] Across these studies OLS estimates are between 0.05 and 0.09 and IV estimates between 0.06 and 0.15.

puts it, this is analogous to comparing a micro level regression to a grouped regression (where there are only two groups). A second explanation, attributed to Griliches (1977) and Angrist and Krueger (1991), is that OLS is biased downwards relative to *ATE* due to measurement error. *IV* exceeds *OLS* just because of measurement error. The third explanation is by Ashenfelter, Harmon and Oorstebeek (1999) and is one of publication bias that leads to a positive *OLS-IV* gap. Finally there is Card's favored explanation for the phenomenon which is as follows. In a correlated random coefficients model, *IV* estimates the average return to schooling for individuals induced to go to school by the policy. The individuals induced to go to school by supply side interventions (the source of variation in his instruments) are also individuals with little schooling. And because $IV > OLS$ and $OLS > ATE$, $IV > ATE$, we can also conclude that these are individuals with high returns to schooling. The reason they are not going to school has to be that they are facing high costs of schooling since they have relatively high returns compared to those who go to school without the intervention. A related paper by Kling (2000) states the same idea. He also finds that *IV* exceeds *OLS* and his interpretation of this result is similar to Card's. Cameron and Taber (2000) argue that the evidence that *IV* exceeds *OLS* should not be interpreted as evidence for borrowing constraints. They point out that *IV* standard errors are large so the evidence for $IV > OLS$ is weak. They also argue that if one includes a local income variable in a regression that uses distance to college as an instrument for schooling *IV* does not exceed *OLS.*

Card also surveys studies that use family background as an instrument. And in these studies *IV* also tends to exceed *OLS*, often by 30% or more. Card argues that these are not valid instruments because they proxy for ability and that is the main reason *IV* estimates

based on them exceed *OLS*.

Finally Card also presents some estimates of the returns to schooling based on studies with twins. Again in those studies *IV* estimates tend to exceed *OLS*, although in some studies *IV* estimates are lower than *OLS* or not much different than *OLS*.

In summary, Card presents the following argument for explaining why *IV* estimates are generally higher than *OLS* estimates:

- $OLS > \bar{\beta}$ (= "True Causal Effect");
- Instruments (supply side interventions) affect people with small amounts of schooling;
- $IV > OLS$

  (This has been found in many studies since Griliches, 1977).

- *IV* estimates the return for people who change in response to the *IV*

  (This has been established in Section 3 and in the previous literature)

- $\therefore IV > OLS > \bar{\beta}$

  (People who would have not gone to school without the intervention have higher rates of returns than those who go. Therefore they have high costs of schooling which is evidence of credit constraints)

This paper establishes that, contrary to the presumption in the literature,

- $\bar{\beta} > OLS$.
- Instruments affect people everywhere in the distribution of U_S, but oversample persons with middle to high U_S values

- $IV > OLS$
- IV is an average of many Wald estimators, all of the which are returns to schooling for people who change.
- $\bar{\beta} > IV > OLS$

But this says nothing whatsoever about credit constraints.

9 Summary and Conclusions

This paper presents a framework for uniting the literatures on policy evaluation, treatment effects and instrumental variable estimation. Different parameters and estimands are weighted averages of the marginal treatment effect (MTE). We show how to identify and estimate the MTE using a robust nonparametric selection model.

Using this framework we estimate the returns to college using a sample of white males extracted from the Panel Survey of Income Dynamics (PSID). We propose and implement a test for the presence of a correlated random coefficient model of schooling - *i.e.* a test for comparative advantage in the labor market.

The data suggest that comparative advantage is an empirically important phenomenon governing schooling choices and that naive efficiency units models of the labor market are empirically inappropriate. Individuals sort into schooling on the basis of both observed and unobserved gains where the observer is the economist analyzing the data. Instrumental variables estimate policy relevant treatment effects only by accident. Different instruments estimate different parameters. Evidence that IV estimates exceed OLS estimates says nothing about the empirical importance of credit constraints and is consistent with the view

that persons induced into schooling by a change in the instrument have, on average, lower gross returns to schooling than those who would attend school without the intervention characterized by the instrument.

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, unpublished manuscript, University of Chicago.

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

[3] Angrist, J. and A. Krueger (2000), "Empirical Strategies in Labor Economics," in Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics*, Vol. 3A (North Holland: Amsterdam and New York), 1277-1365.

[4] 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.

[5] Ashenfelter, 0., C. Harmon and H. Oosterbeek (1999), "A Review of Estimates of the Schooling/Earnings Relationship, with Tests of Publication Bias," *Labour Economics*, 6:453-470.

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

[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] Bound, J. and D. Jaeger (1996), "On the Validity of Season of Birth as an Instrument in Wage Equations: A Comment on Angrist and Krueger's Does Compulsory School Attendance Affect Schooling and Earnings?," NBER Working Paper #5835.

[9] Cameron, S. and C. Taber (2000), "Borrowing Constraints and The Returns to Schooling," NBER working paper, #7761.

[10] Card, D. (1995), "Earnings, Schooling, and Ability Revisited," *Research in Labor Economics,* 14:23-48.

[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. (2000), "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," NBER Working Paper, #7769.

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

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

[15] Heckman, J. (2001b), "Microdata, Heterogeneity and Econometric Policy Evaluation," Nobel Memorial Lecture in Economic Sciences, forthcoming *Journal of Political Economy*, August, 2001.

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

[17] 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.

[18] 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.

[19] Heckman, J. and R. Robb (1986), "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*, (NY: Springer-Verlag), 63-107. Republished by Lawrence Erlbauum Press, Mahwah, New Jersey, 2000.

[20] 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.

[21] Heckman, James, Tobias, J. and E. Vytlacil (2000), "Simple Estimators for Alternate Treatment Parameters in a Latent Variable Framework with an Application to Estimating the Returns to Schooling," NBER Working Paper No. W7950, under review.

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

[23] 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.

[24] 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.

[25] Heckman, J. and E. Vytlacil (2001a), "Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effect," in M. Lechner and F. Pfeiffer (eds.), *Econometric Evaluations of Active Labor Market Policies in Europe*, (Berlin: Springer-Verlag).

[26] Heckman, J. and E. Vytlacil (2002), "Econometric Evaluations of Social Programs," forthcoming in J. Heckman and E. Leamer, (eds.), *Handbook of Econometrics*, Volume 5, (North-Holland:Amsterdam).

[27] Ichimura, H. and P. Todd (2002), "Implementing Nonparametric and Semiparametric Estimators," forthcoming in J. Heckman and E. Leamer, (eds.), *Handbook of Econometrics*, Volume 5, (North-Holland:Amsterdam).

[28] Kling, J. (1999), "Interpreting Instrumental Variables Estimates of the Returns to Schooling," Princeton University Industrial Relations Section Working Paper, No. 415.

[29] Krueger, A. (2000), *Labor Policy and Labor Research Since the 1960s, in Economic Events, Ideas and Policies: The 1960s and After*, edited by George Perry and James Tobin, (Washington DC, Brookings Institution Press).

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

[31] Mincer, J. (1958), "Investment in Human Capital and Personal Income Distribution," *Journal of Political Economy,* 66:281-302.

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

[33] Moffitt, R. (1999), "New Developments in Econometric Methods for Labor Market Analysis," in O. Ashenfelter and D. Card (editors), *Handbook of Labor Economics*

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

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

[36] Sattinger, M. (1993), "Assignment Models of the Distribution of Earnings," *Journal of Economic Literature*, 31:831-880.

[37] Vytlacil, E. (2002), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," forthcoming, *Econometrica*.

[38] Willis, R. (1986), "Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions," in 0. Ashenfelter and R. Layard (eds.), *Handbook of Labor Economics,* (Amsterdam: North-Holland).

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

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

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

Appendix A

This paper uses data from the Panel Study of Income Dynamics (PSID). For our analysis we restricted the sample to 743 white males who are either high school graduates or have one or more year of college completed and receive a positive wage in 1991. We only use heads of households. We exclude GED recipients from the sample.

The variables we use as follows:

Definitions of Variables - PSID91

Southern Residence when Growing Up	=1 if the respondent grew up in the South
Broken Home	=1 if the respondent lived with both natural parents most of the time until age 16
Urban Residence When Growing Up	=1 if the respondent did not grew up in a farm, rural area or in the country
Number of Siblings	Number of brothers and sisters
Highest Grade Completed of Mother	Highest grade completed by mother (in years)
Highest Grade Completed of Father	Highest grade completed by father (in years)
Family Income at Age 17	Family income measured at age 17.
Tuition for 4-Year College	Four-year college tuition for in-state students in the individual's state
Year of Birth Dummies	=1 if the respondent was born in a given year
College	=1 if the respondent completed one or more years of college

Definitions of Variables - PSID91

Log Hourly Wage	Log hourly wage (computed by dividing total labor income by total hours worked in a given year)
Experience	Years of full time experience worked since age 18

Individuals who have attended some college have on average higher wages and less experience than high school graduates. They come from more advantaged backgrounds as measured from parental education and family income at age 17 (measured in 1983 dollars). When we look at the probits we estimate, along with number of siblings, these are the most important predictors of the decision of going to college. For the individual with the average demographics in our sample an increase in one year of education for either of her parents is associated with an increase in the probability of going to college of 4 to 5 percentage points. All the variables except the indicator for growing up in a southern region have the expected sign in the probit, although some of them are not significantly different from zero. Year of birth effects are in general not important except for being born in 1958 which has a positive and signficant coefficient in the probit.

Table 1A

Treatment Effects and Estimands as Weighted Averages of the MTE

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

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

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

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

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

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

Table 1B

Weights

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

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

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

$$h_{TUT}(x,u_S) = \left[\int_{0}^{u_S} f(p \mid X)dp\right]\frac{1}{E((1-P) \mid X = x)}$$

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

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

$$h_0(x,u_S) = \left[\int_{0}^{u_S} f(p \mid X = x)dp\right]\frac{1}{E((1-P) \mid X = x)}$$

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

Table 2

Roy Example

OLS	0.1735
TT	0.2442
TUT	0.1570
ATE	0.2003
Sorting Gain[1]	0.0402
Selection Bias[2]	-0.0708
Linear IV[3]	0.2017

[1] $E[U_1 - U_0|D = 1] = \text{TT} - \text{ATE}$

[2] $E[U_0|D = 1] - E[U_0|D = 0] = \text{OLS} - \text{TT}$

[3]Using Propensity Score as the instrument

Table 3

Local Linear Regression

PSID91, White Males

Bootstrapped

Baseline Model Without Interactions

OLS	0.0981
	(0.0156)
TT	0.2738
	(0.0808)
ATE	0.2156
	(0.0619)
Sorting Gain[1]	0.0582
	(0.0307)
Selection Bias[2]	-0.1757
	(0.0757)
Linear IV[3]	0.1932
	(0.0511)
Policy III[4]	0.2005
	(0.0557)
Marginal Person	0.2002
	(0.0575)

[1] $E[U_1 - U_0|S = 1]$ = TT - ATE

[2] $E[U_0|S = 1] - E[U_0|S = 0]$ = OLS - TT

[3]Using Propensity Score as the instrument

[4]Tuition Subsidy = $500

Table A-1
Sample Statistics*
PSID91, White Males

	High School	College
Log Wage	2.2735	2.6102
	(.5583)	(.6036)
Years of Work Experience	6.8621	5.1934
	(3.2476)	(3.5008)
Years of Work Experience Squared	57.6019	39.1981
	(48.3230)	(45.6022)
Lived in South	.0846	.0920
	(.2788)	(.2893)
Lived in Urban Area	.7931	.8986
	(.4057)	(.3022)
Broken Family	.1536	.1274
	(.3611)	(.3338)
Number of Siblings	3.1129	2.75
	(2.0109)	(1.8499)
Mother's Education	11.7492	13.2193
	(1.9441)	(2.2081)
Father's Education	11.3573	13.5236
	(2.5124)	(2.7598)
Family Income	20188.61	28015.97
	(13751.1)	(21055.82)
Tuition for 4-Year College	17.5765	17.8199
	(5.3649)	(5.0104)
Born in 1956	.1160	.1226
	(.3207)	(.3284)
Born in 1957	.1223	.0967
	(.3281)	(.2959)
Born in 1958	.0502	.1179
	(.2186)	(.3229)
Born in 1959	.1129	.0967
	(.3169)	(.2959)
Born in 1960	.1160	.1203
	(.3207)	(.3257)
Born in 1961	.0878	.1108
	(.2834)	(.3143)
Born in 1962	.0940	.0825
	(.2923)	(.2755)
Born in 1963	.0909	.0731
	(.2879)	(.2606)
Born in 1964	.0784	.0825
	(.2692)	(.2755)
Born in 1965	.0721	.0495
	(.2591)	(.2172)
Number of Observations	319	424

*Standard deviations in parentheses

Table B-1

Estimated Marginal Effects from Probit

Dependent Variable $S = 1$, College Attendance

PSID91, White Males

	Estimate	Standard Error
Lived in South	.0368	.0677
From Broken Family	-.0295	.0558
Lived in Urban Area	.1477	.0570
Number of Siblings	-.0166	.0102
Mother's Education	.0420	.0111
Father's Education	.0485	.0085
Family Income	.000003	.000001
Tuition for 4-Year College	-.0003	.1162
Born in 1956	.1135	.0920
Born in 1957	.0414	.0990
Born in 1958	.2529	.0775
Born in 1959	-.0039	.1027
Born in 1960	.0692	.0961
Born in 1961	.0951	.0953
Born in 1962	.0100	.1052
Born in 1963	.0155	.1071
Born in 1964	.0826	.1013
Born in 1965	-.0318	.1162

Table B-2

Coefficients from Local Linear Regression

PSID91, White Males

Bootstrapped

Baseline Model Without Interactions

	Estimate	Standard Error
Years of Work Experience	0.0195	0.0269
(Years of Work Experience)^2	-0.0029	0.0018
Lived in South at Age 14	0.0917	0.0959
Lived in Urban Area at Age 14	0.1689	0.0651
Born in 1956	0.4981	0.1462
Born in 1957	0.4367	0.1149
Born in 1958	0.3358	0.1407
Born in 1959	0.3425	0.1208
Born in 1960	0.3204	0.1174
Born in 1961	0.2684	0.1178
Born in 1962	0.2613	0.1094
Born in 1963	0.1583	0.1158
Born in 1964	0.0969	0.1039
Born in 1965	0.2201	0.1210

Table B-3

Index Sufficiency Test
PSID91 - From Broken Home

P	P-Value
.300	.771
.386	.931
.471	.948
.557	.803
.643	.544
.729	.122
.814	.407
.900	.457
Joint	.679

Index Sufficiency Test
PSID91 - Mother's Education

P	P-Value
.300	.962
.386	.123
.471	.948
.557	.958
.643	.961
.729	.962
.814	.961
.900	.959
Joint	.998

Index Sufficiency Test
PSID91 - Father's Education

P	P-Value
.300	.895
.386	.955
.471	.988
.557	.993
.643	.982
.729	.976
.814	.973
.900	.971
Joint	.991

Table B-4

Monotonicity Test

PSID91

P	$\frac{\partial E(DU\|P)}{\partial P}$	t-statistic
0.1	1.5497	1.7725
0.2	2.0338	4.2698
0.3	2.3733	7.5027
0.4	2.5555	10.0992
0.5	2.5656	13.6236
0.6	2.4756	14.9740
0.7	2.3451	10.9541
0.8	2.1318	6.5213
0.9	1.8654	3.8163

Monotonicity Test

PSID91

P	$\frac{\partial E[(1-D)U\|P]}{\partial P}$	t-statistic
0.1	0.2347	0.2361
0.2	-0.7821	-1.5563
0.3	-1.4030	-4.0743
0.4	-1.7979	-6.8025
0.5	-2.0007	-10.4063
0.6	-2.0456	-13.0357
0.7	-1.9773	-11.0669
0.8	-1.8087	-7.4912
0.9	-1.5506	-4.6676

Figure 1A
Weights for MTE for Different Treatment Effects

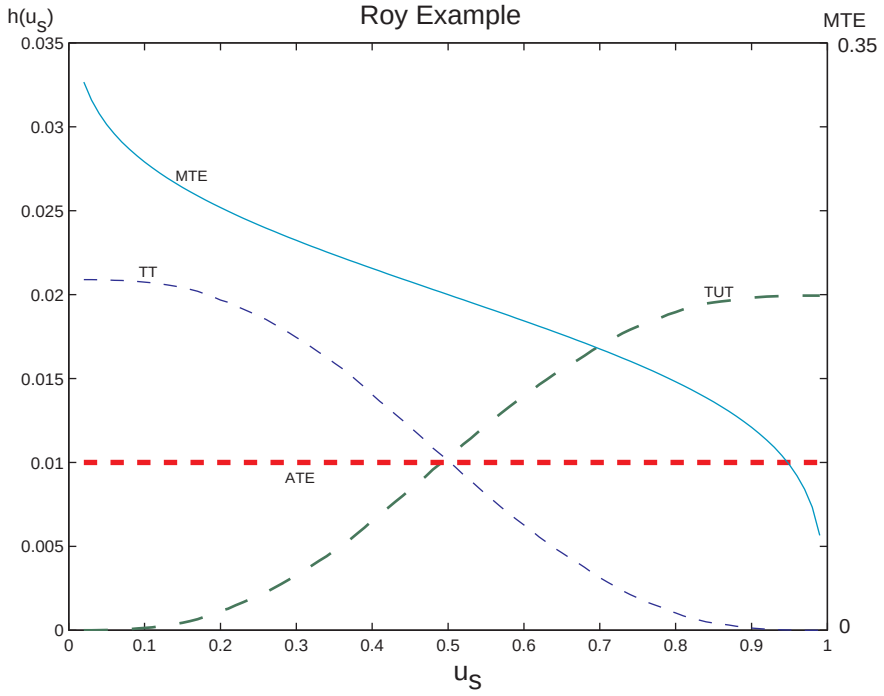
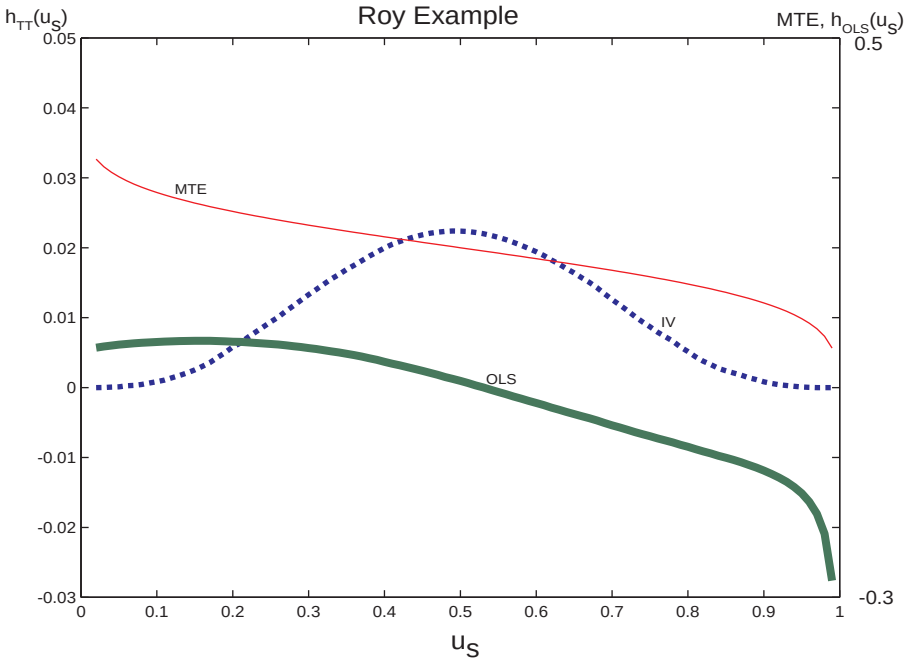


Figure 1B
Marginal Treatment Effect vs Linear IV and OLS Weights



$$
\begin{aligned}
\ln Y_1 &= \alpha + \bar{\beta} + U_1 \\
\ln Y_0 &= \alpha + U_0 \\
S &= 1 \text{ if } Z - U_S > 0
\end{aligned}
\qquad
\begin{aligned}
U_1 &= \sigma_1 \varepsilon \\
U_0 &= \sigma_0 \varepsilon \\
U_S &= \sigma_S \varepsilon
\end{aligned}
\qquad
\begin{aligned}
\alpha &= 0.67 \\
\bar{\beta} &= 0.2 \\
&\varepsilon \tilde{} N(0,1) \\
&Z \tilde{} N(-0.0026, 0.27) \\
\sigma_1 &= 0.012 \\
\sigma_0 &= -0.05 \\
\sigma_S &= -1
\end{aligned}
$$

Figure 2
Local Linear Regression of Log Wage on P

(bootstrapped)

White Males, PSID91, Baseline Model Without Interactions

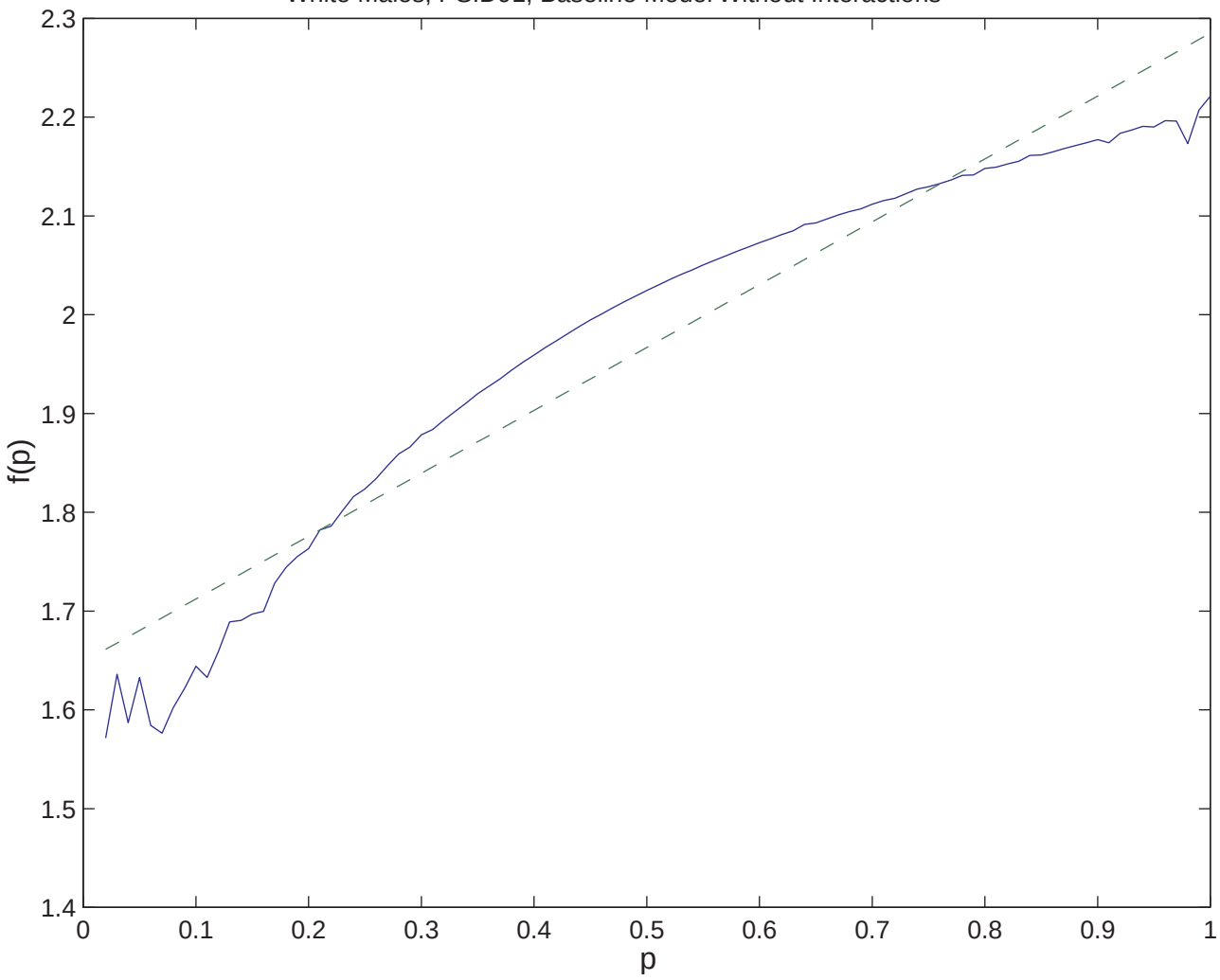


Figure 3

Marginal Treatment Effect

White Males, PSID91, Baseline Model Without Interactions

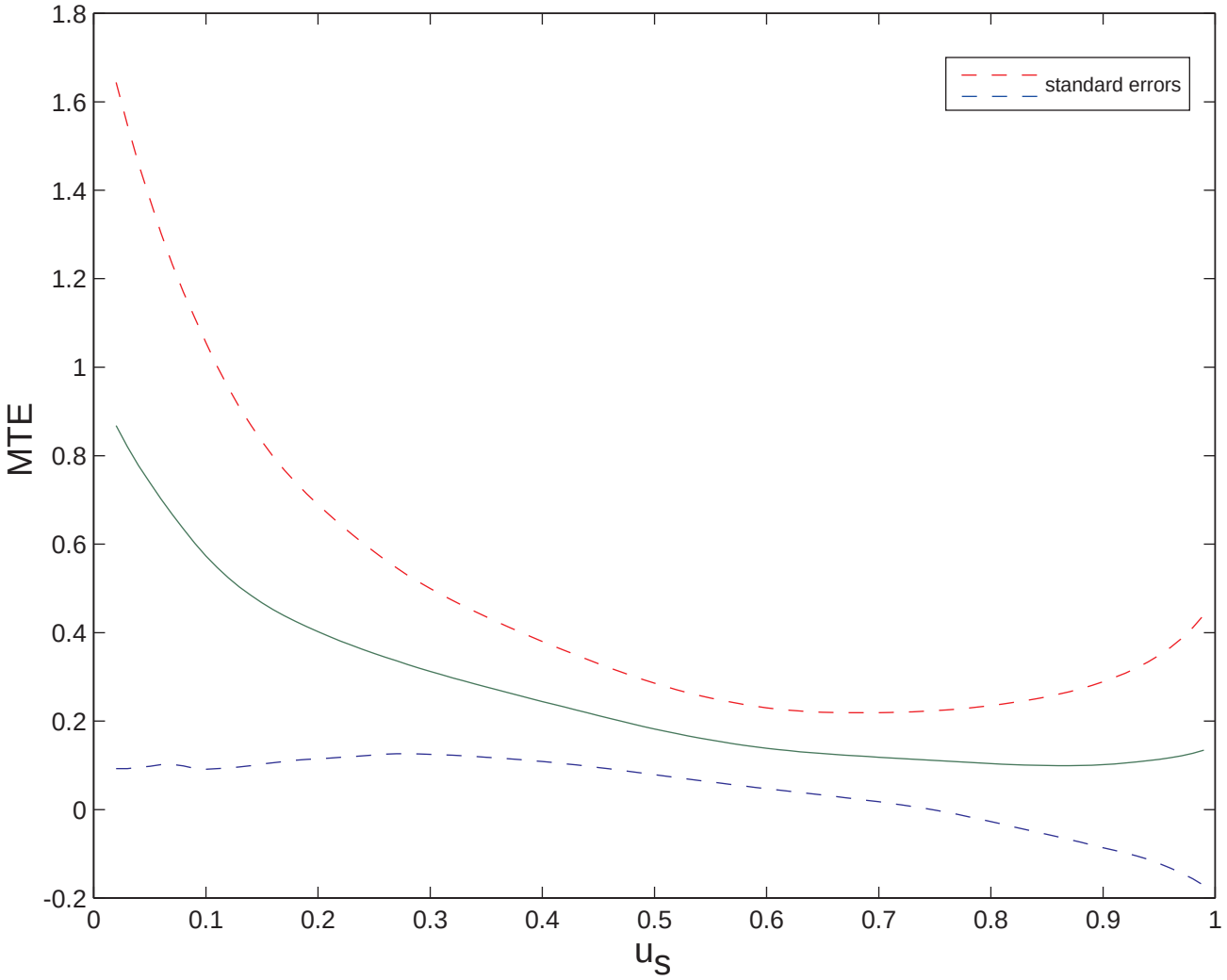


Figure 4

Marginal Treatment Effect vs Linear IV and OLS Weights

From PSID 1991

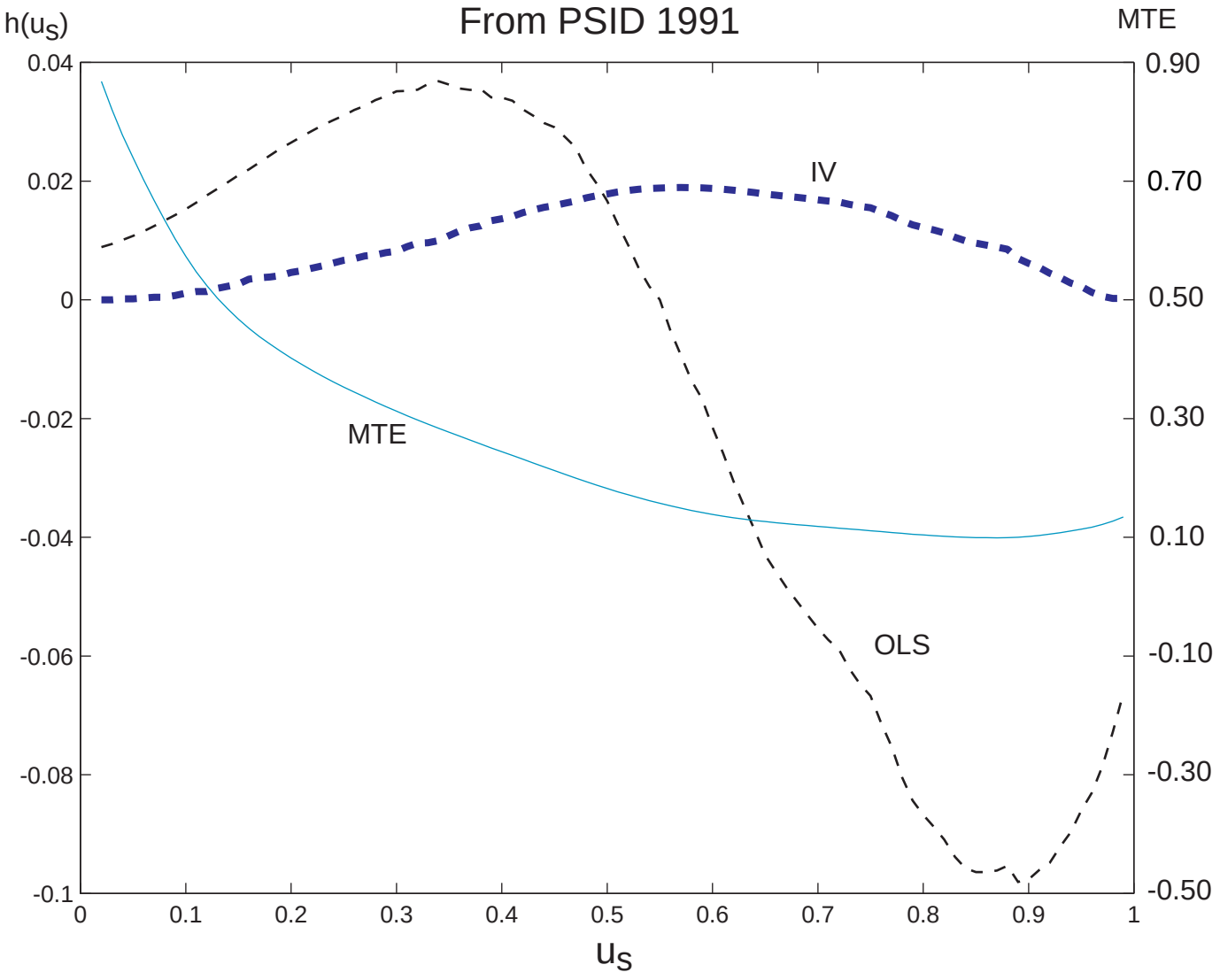


Figure 5

Marginal Treatment Effect vs Linear IV, TT, and ATE Weights

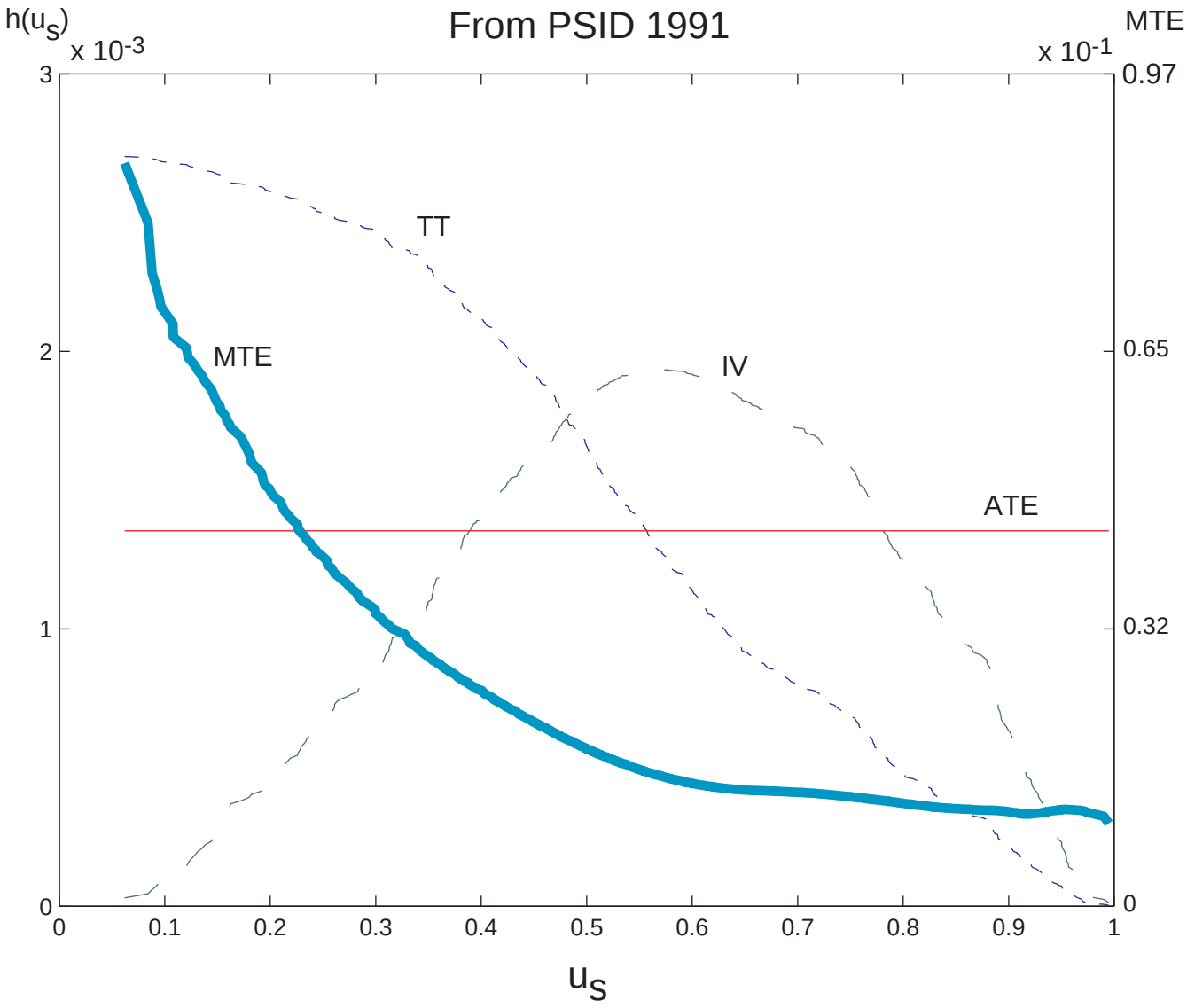
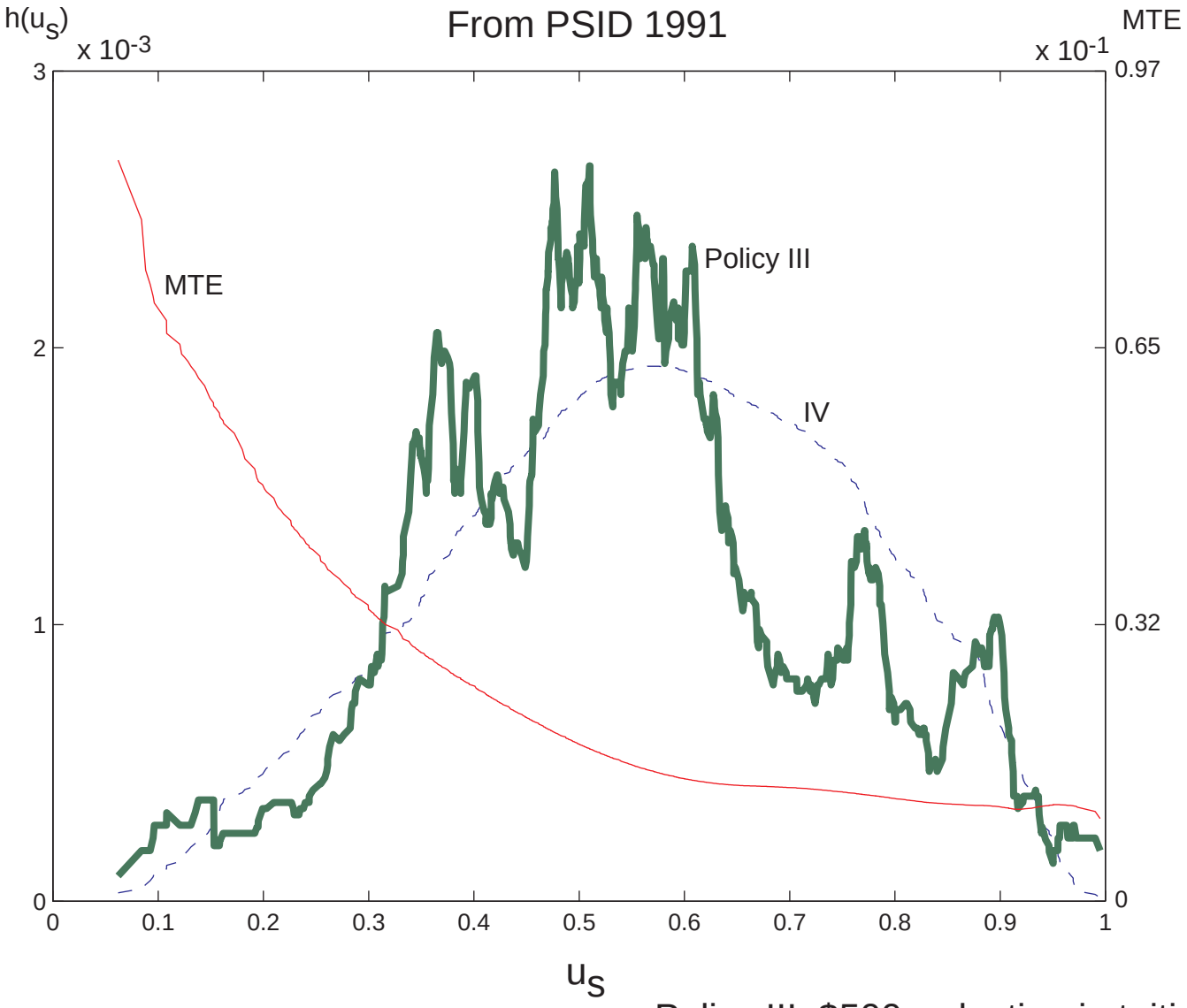
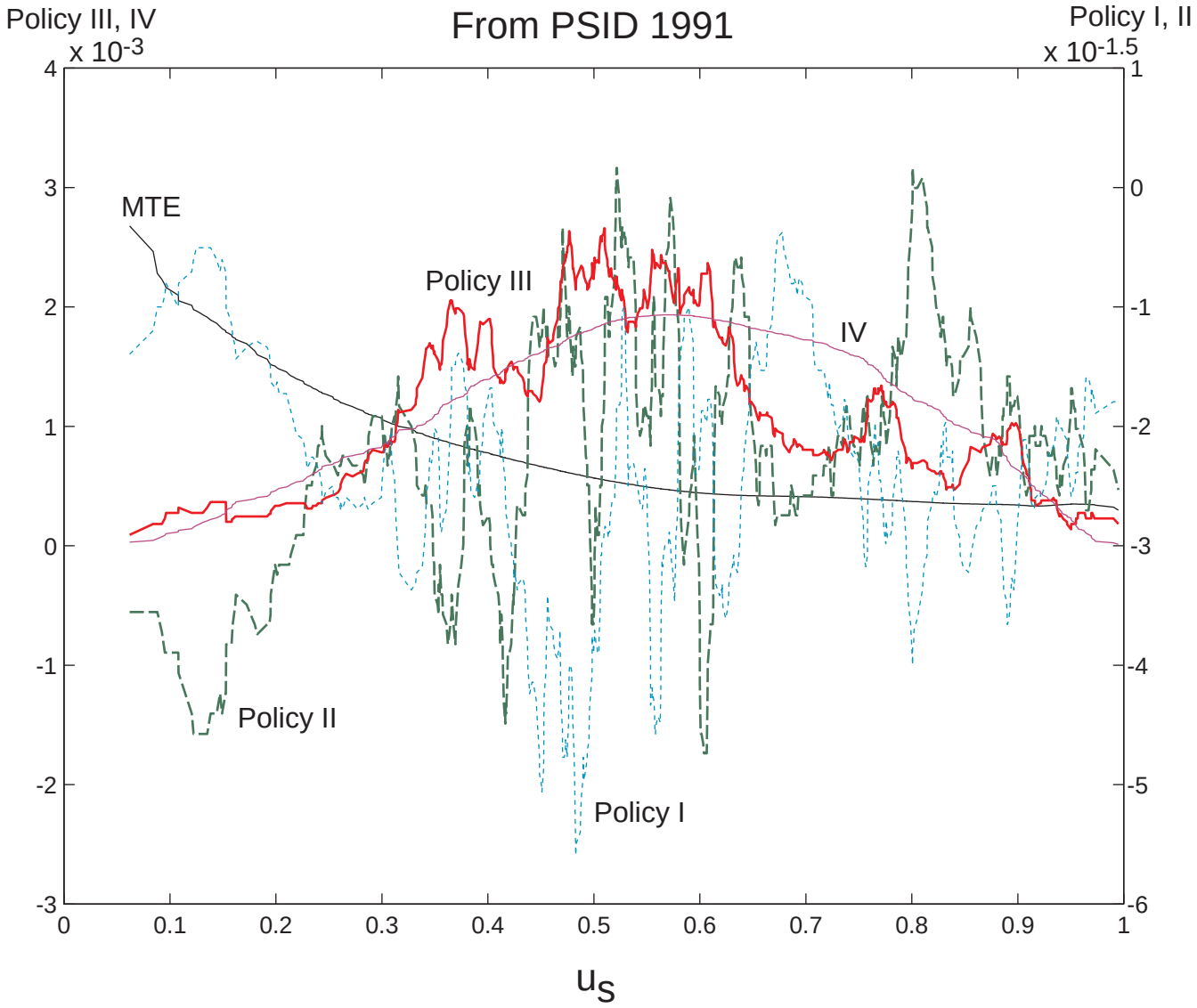


Figure 6
Marginal Treatment Effect vs Policy III and Linear IV Weights
From PSID 1991



Policy III: $500 reduction in tuition

Figure 7
Marginal Treatment Effect vs Policy and Linear IV Weights



Policy I: Pushing people to extremes of tuition distribution
Policy II: Pushing people to mean of tuition distribution
Policy III: $500 reduction in tuition

Figure B1
Propensity Score Distribution

PSID 91, White Males, Probability of College Attendance

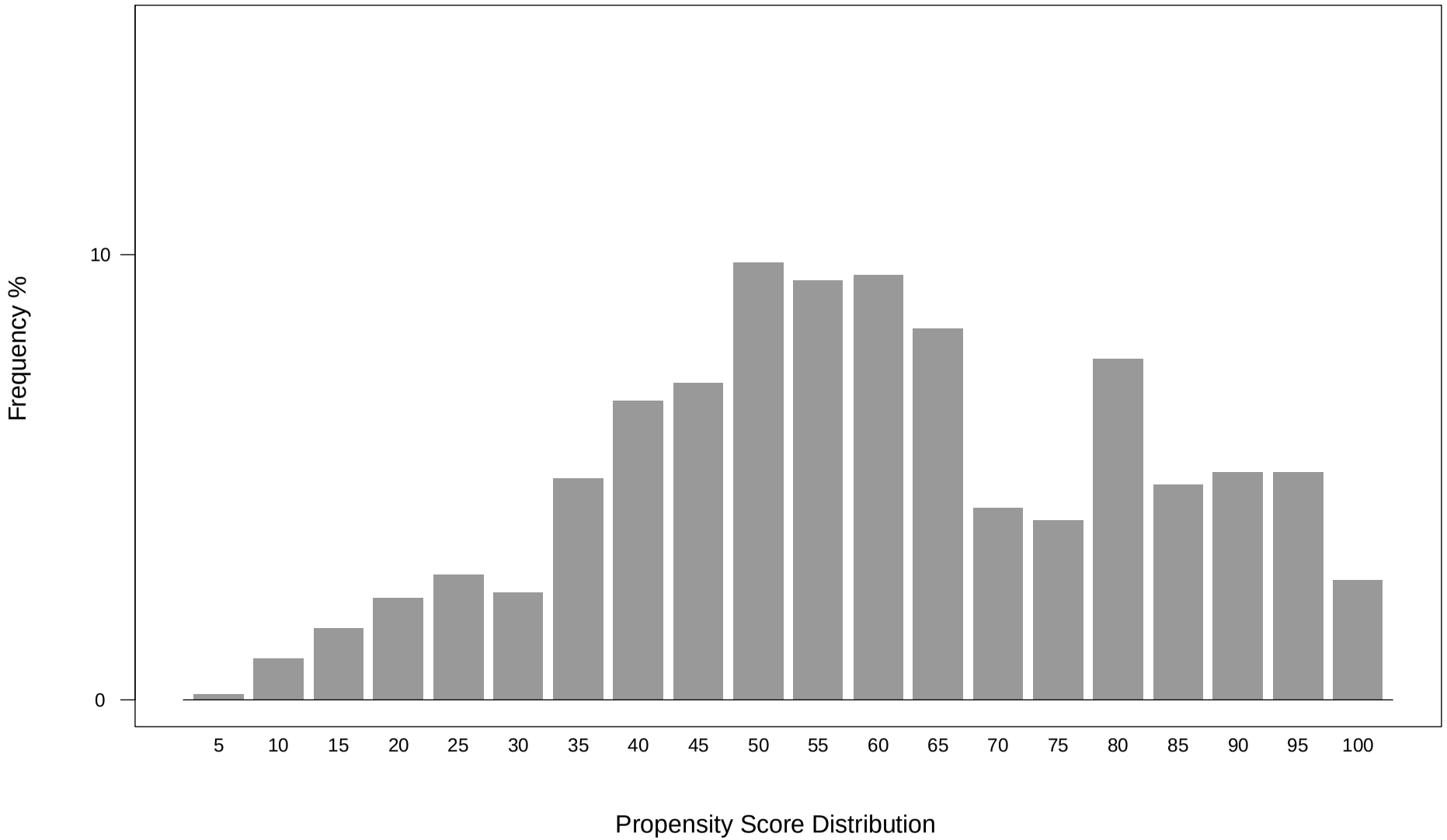


Figure B2
Monotonicity Test

PSID91, White Males

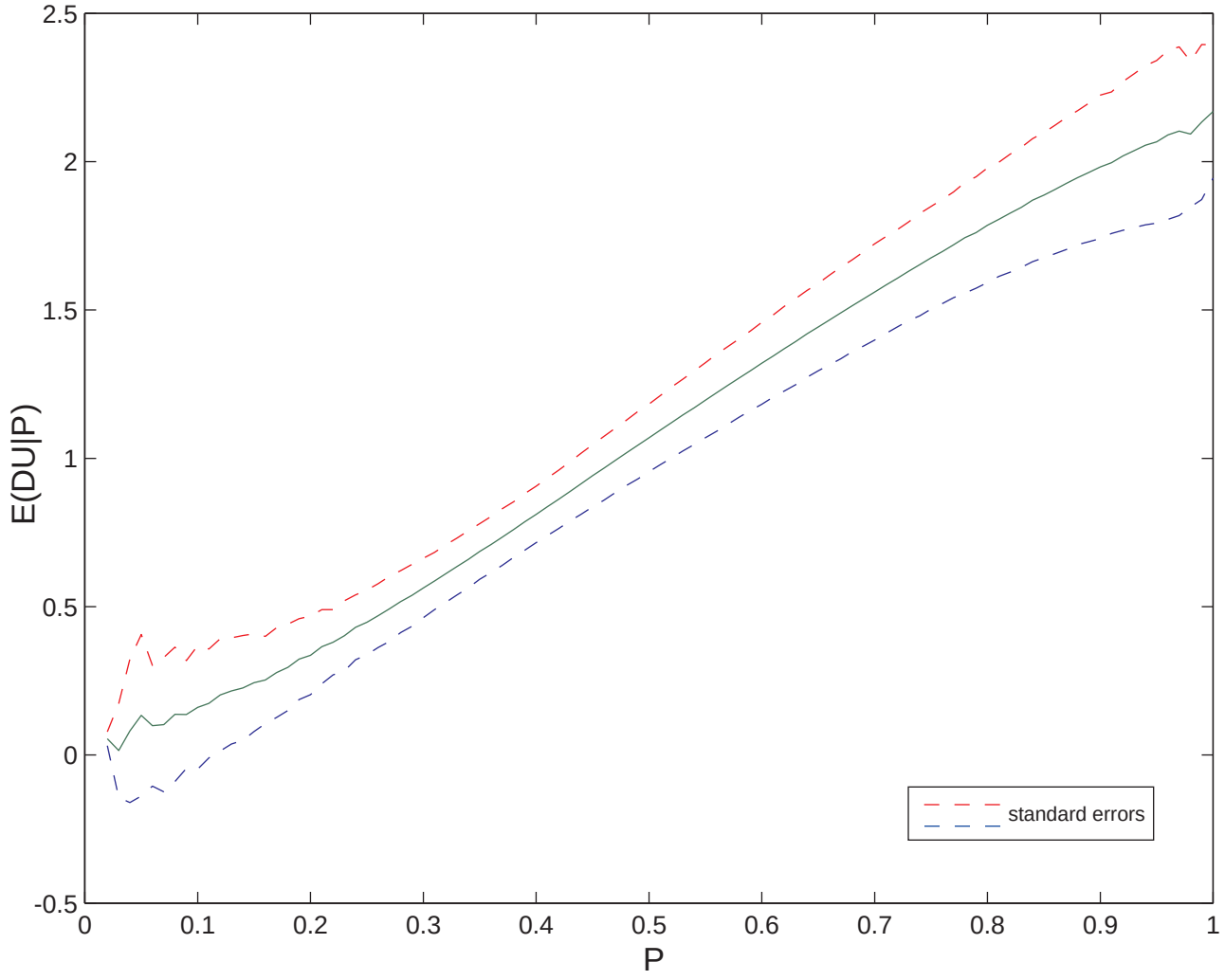


Figure B3
Monotonicity Test

PSID91, White Males