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

*This research was supported by NSF 97-09-873, NSF-SES-0099195 and NICHD-40-4043-000-85-261. Carneiro benefited from support from Fundação Ciência e Tecnologia and Fundação Calouste Gulbenkian. We have benefitted from comments received at the Applied Price Theory Workshop, from comments received from Jaap Abbring, Mark Duggan, Michael Greenstone, Larry Katz, Steve Levitt, Kevin Murphy, Derek Neal and from participants at the Royal Economic Society, Durham, England, April 10, 2001; especially those of Richard Blundell, Partha Dasgupta 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. 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
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. 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 with the following example. Suppose we consider schooling choices which can take only two values \((S = 0 \text{ or } S = 1)\) 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, \(f(R \mid R ≥ C)\), 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 there is a policy that changes costs of attending school from \(C\) to \(C'\) (such as a tuition subsidy). 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

\(^{1}\text{See Heckman (2001) for a summary of the evidence from this literature.}\)
the policy), and the average return for these individuals is just \( E(R | C' < R \leq C) \). 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 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 \( \alpha, \beta \) are parameters and where \( S \) is correlated with mean zero error \( U \), least squares estimators of \( \beta \) are biased and inconsistent. Economists since Haavelmo (1943) have defined the “causal effect,” or “effect” of \( S \) on \( \ln Y \), as \( \beta \). 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 \( U \), \( \beta \) can be identified, at least in large samples. Both valid social experiments and valid natural experiments can be interpreted as 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 \( \beta \) varies in the population and people sort into economic sectors on the basis of at least partial knowledge of \( \beta \), then the marginal \( \beta \) may be different from the average \( \beta \). 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 \( \beta \). In the empirical work reported in this paper, \( S \) is schooling, and \( \ln Y \) is log earnings, so in the example motivating Figure 1, \( R = (e^\beta - 1) e^{\alpha + U} \). We estimate marginal and average returns to schooling. 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.\(^2\)

\(^2\)Carneiro, Hansen and Heckman (2003) estimate entire distributions of returns to schooling.
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 of the policy. 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. We use this instrument to improve on conventional IV to estimate economically interpretable average and marginal returns.

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 variables 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 misestimate 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. This example motivates the empirical work we report in this paper. In this section, we also define the Benthamite policy parameter estimated in this
Section 2 characterizes two approaches to estimating the variable coefficient model: the classical parametric approach and a nonparametric approach developed by Heckman and Vytlacil (1999, 2000, 2001a, 2004b). 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 presents our estimates and 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 we suppress explicit notation for dependence of the parameters on the covariates $X$ unless it is helpful to make this dependence explicit. Under special conditions discussed in Willis (1986) and Heckman, Lochner and Todd (2001), $\beta$ is the rate of return to schooling. 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 $\beta$ 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

$$3R = e^{\alpha+U} (e^{\beta} - 1)$$

is the absolute return and

$$R = \left[ e^{\alpha+U} (e^{\beta} - 1) \right] / e^{\alpha+U} = e^{\beta-1} = \beta$$

is the rate of return to schooling, where $e^{\alpha+U}$ is earnings when $S$ is fixed at 0.
instrumental variable estimators for $\beta$ to alleviate these problems. In this framework, because $\beta$ is a constant, there is a unique effect of schooling. Indeed, $\beta$ 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 Roy (1951) model, there are two potential outcomes ($\ln Y_0, \ln Y_1$):

$$\begin{align*}
\ln Y_0 &= \alpha + U, \\
\ln Y_1 &= \alpha + \beta + U
\end{align*}$$

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 $\beta$. The Roy model, as applied by Willis and Rosen (1979), gives a more precise notion of why $\beta$ 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:

$$\begin{align*}
\ln Y_0 &= \alpha + U_0 \\
\ln Y_1 &= \alpha + \bar{\beta} + U_1
\end{align*}$$

where $E(U_0) = 0$ and $E(U_1) = 0$ so $\alpha$ 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

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

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

\[ S = \begin{cases} 1 & \text{if } \ln Y_1 \geq \ln Y_0 \iff \beta \geq 0.4 \\ 0 & \text{otherwise.} \end{cases} \]  

(2)

If agents know or can partially predict \( \beta \) at the time they make their schooling decisions, there is dependence between \( \beta \) and \( S \) in equation (2). This justifies the title “correlated random coefficient model” that is often applied to general versions of (2). Decision rules similar to (2) characterize other economic choices.

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

Source (c) arises from the possibility that the gains to schooling \( (\beta) \) 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 \( \beta \) 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 \( \beta \) are used. Among them are

\[ E(\beta \mid X = x) = E(\ln Y_1 - \ln Y_0 \mid X = x) = \bar{\beta}(x) \]

the return to the population average person given characteristics \( X = x \). This is sometimes
called “the” causal effect of \( S \).\(^5\) Others report the return for those who attend school:

\[
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).\(^6\)
\]

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:

\[
E(\beta \mid S = 0, X = x) = E(\ln Y_1 - \ln Y_0 \mid S = 0, X = x) = \bar{\beta}(x) + E(U_1 - U_0 \mid S = 0, X = x).
\]

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

\[
E(\ln Y_1 - \ln Y_0 \mid \ln Y_1 - \ln Y_0 = 0) = 0.
\]

A more general expression 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

II: \( Pr(S = 1 \mid X = x, \beta) = Pr(S = 1 \mid X) \) (conditional on \( X, \beta \) 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.\(^7\)

\(^5\)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).

\(^6\)It is the Treatment on the Treated parameter as discussed by Heckman and Robb (1985, 2000).

\(^7\)For example, Heckman and Robb (1985) noted 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.
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 earnings under one policy are 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 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 $\beta$ is a constant, $\bar{\beta}$, the mean change in welfare is

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

(5)

where $(\Delta P)$ 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 $\beta$ is a random variable but independent of $S$ if conditions I or II apply.

In the general case, when agents partially anticipate $\beta$, and comparative advantage dictates schooling choices, none of the traditional treatment parameters plays the role of $\bar{\beta}$ in (??) or answers the stated economic question. Instrumental variables methods do not generally identify $\bar{\beta}$ unless the instrument is the imposed policy. We develop the appropriate parameter and show to estimate it and contrast it with conventional treatment parameters and what IV estimates. We first consider two approaches to estimating the distribution of the returns to schooling.

2 Two Approaches to Estimating the Schooling Model

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 \(Y_0(t) = Y_0 e(t)\).

The agent attends school if
\[
(1 + r)^{\infty} t=0 \frac{Y_1 - Y_0}{(1 + r)^t} e(t) \geq C^*,
\]
where \(e(t)\) is a post-school experience growth factor. Let \(K = \sum_{t=0}^{\infty} e(t) (1 + r)^t\) 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 (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
\[
\beta \geq r + \frac{C}{Y_0}.
\]
Ceteris paribus, a higher \(r\) or \(C\) lowers the likelihood that \(S = 1\). A higher opportunity cost \((Y_0)\) lowers the likelihood of going to school. Equation (??) generalizes decision rule (??) 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 \(\alpha\) 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.\(^8\) 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).

Using their setup we write

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

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 \(X \perp(U_0, U_1)\) so \(X\) may be correlated with the unobservables in potential outcomes. Here and throughout this paper we use “\(\perp\)” to denote statistical independence.

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

\[
S^* = \text{sgn}(Z) - U_S8
\]
\[
S = 1 \text{ if } S^* \geq 0.
\]

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 \(X\).

\(^8\)Willis and Rosen (1979) is an example of the application of the Roy model. Aakvik, Heckman and Vytlacil (2000), Heckman, Tobias and Vytlacil (2001) 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.
Heckman and Vytlacil (2000, 2001b) 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) $U_S$ 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 $Y_0, Y_1$. (These are determinants of $C$ and $r$ in equation (??)). The recent empirical literature on the returns to schooling also assumes the existence of 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 $S$ 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 $S = 0$, so comparisons of schooling and nonschooling outcomes can be made at all $X$ values.

However this condition is not strictly required. If imposed, it produces the “total effect” of $S$ on $Y$. See Pearl (2000). Heckman and Navarro (2003) and Heckman and Vytlacil (2004a) relax this condition.
Denoting \( P(z) \) as the probability of receiving treatment \( S = 1 \) 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 \( U_S \sim \text{Unif}[0,1] \) 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 (??), (8) and (A-1) - (A-5) is equivalent to the LATE model of Imbens and Angrist (1994) extended to allow for continuous instruments.\(^{10}\)

The index structure produced by assumptions (A-1) - (A-5) joined with the model of equations (??) and (8) allow 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 for 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 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
\[
\Delta^{LATE}(x, u'_S, u_S) = E(\ln Y \mid X = x, u_S \leq U_S \leq u'_S)
\]
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. See Table 1A for the treatment parameters expressed in terms of \( MTE \) and Table 1B for the weights.

If \( \beta \) is a constant or \( E(\beta \mid X = x, U_S = u_s) = E(\beta \mid X = x) \), (\( \beta \) mean independent of \( U_S \)), then all of these mean treatment parameters are the same. This arises in cases I and

\(^{10}\)These conditions impose testable restrictions on \( (Y, S, Z, X) \). See Heckman and Vytlacil (1999, 2000, 2001b). The primary restriction is index sufficiency \( \Pr(\ln Y_j \in A \mid Z = z, S = j) = \Pr(\ln Y_j \in A \mid P(Z) = P(z), S = j) \) for \( j = 0,1 \). This says that \( Z \) enters the conditional distribution of \( \ln Y_1, \ln Y_0 \), only through the index \( P(Z) \).

\(^{11}\)Björklund and Moffitt (1987) introduced this parameter in the context of the parametric normal Roy model.
II analyzed in Section 1 where, respectively, there is no heterogeneity (β constant) or agents do not act on it.\textsuperscript{12}

The standard treatment parameters assume that \( P(Z) \) has support equal to the full unit interval, \([0,1]\). The marginal treatment effect is more basic and does not require full support. This makes it a more easily identified parameter. Bounds for the parameters and estimands when the support of \( P(Z) \) is less than full are presented in Heckman and Vytlacil (2000, 2001b).

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 \( \beta \) is heterogeneous and people act on \( \beta \) in making decisions about \( S \). 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.\textsuperscript{13}

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

\[
E(V(Y) \mid \text{Alternative Policy}^*) - E(V(Y) \mid \text{Baseline Policy}) = \int \omega(u) MTE(u) du
\]

where \( \omega(u) = F_P(u) - F_{P^*}(u) \) where \( F_P \) and \( F_{P^*} \) denote the cdf of \( P \) and \( P^* \), respectively.\textsuperscript{14}

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

Thus if we use the weights

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

\textsuperscript{12}All of these parameters in Tables 1A and B can be defined even if (a) \( U_S \perp \perp Z \) or (b) For \( S = 1(\Omega(Z, U_S) \geq 0) \) there is no additively separable version of \( \Omega \) 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).

\textsuperscript{13}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 effect the choice of \( S \) and hence \( P \) and also \( (\ln Y_1, \ln Y_0) \) if changes in aggregate schooling participation affect skill prices.

\textsuperscript{14}For a proof see Heckman and Vytlacil (2001b). Other criteria produce different weights.
we produce the gain in the outcome for the people induced to change into (or out of) schooling by the policy change.

Observe that these weights differ from the weights for 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 $\beta$ varies among individuals conditional on $X$ and people make schooling decisions based on it, IV weights $MTE$ differently than the weighting required for policy analysis or required to generate the conventional treatment parameters. Standard IV methods including natural and social experiments in general do not answer well-posed policy problems. The exception to this rule occurs when the instrument is the policy and it is exogenously implemented.

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 $\beta$ varies in the population and choices of $S$ are made on the basis of it.

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

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

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

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

Finding an instrument $Z$ correlated with $S$ but not $U_0$ or $U_1 - U_0$ is not enough to identify $\bar{\beta}$, or $\bar{\beta} + E(U_1 - U_0 \mid S = 1)$, or other conventional treatment parameters.\footnote{Recall that we keep the conditioning on $X$ implicit.} Simple algebra
reveals that
\[ \text{plim} \hat{\beta}_{IV} = \frac{\text{Cov}(Z, \ln Y)}{\text{Cov}(Z, S)} = \tilde{\beta} + \frac{\text{Cov}(Z, U_0)}{\text{Cov}(Z, S)} + \frac{\text{Cov}(Z, S(U_1 - U_0))}{\text{Cov}(Z, S)}. \]

By standard IV condition (a), the second term vanishes (\( \text{Cov}(Z, U_0) = 0 \)). But in general the third term does not:
\[ \frac{\text{Cov}(Z, S(U_1 - U_0))}{\text{Cov}(Z, S)} = \frac{PCov[Z (U_1 - U_0) | S = 1] + P[E (Z|S = 1) - E(Z)]E(U_1 - U_0|S = 1)}{\text{Cov}(Z, S)} \neq 0 \]
where \( P = \text{Pr}(S = 1) \). If \( U_1 - U_0 \equiv 0 \) (a common coefficient model, condition I) or if \( U_1 - U_0 \) is independent of \( S \) and \( Z \) (condition II) this term vanishes.\(^{16}\) But in general \( U_1 - U_0 \) is dependent on \( S \) and the term does not vanish.\(^{17}\)

To see why, consider the schooling choice model of equation (??) when \( C = 0 \) and \( r \) depends on \( Z \) \( (r = Z\gamma) \). Then
\[ S = 1 \iff \tilde{\beta} + U_1 - U_0 \geq Z\gamma, \]
and \( \text{Cov}(Z, U_1 - U_0 | S = 1) = \text{Cov}(Z, U_1 - U_0 | \tilde{\beta} + U_1 - U_0 \geq Z\gamma) \). Even if \( Z \perp (U_1 - U_0) \), \( Z \) 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 (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 \( P(x) = \text{Pr}(S = 1 | X = x) \) 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 \( T \perp (U_S, U_0, U_1) | X \).

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

\(^{16}\)If \( U_1 - U_0 \) is independent of \( Z \) and if \( U_1 - U_0 \) does not determine \( S \) conditional on \( Z \), then \( U_1 - U_0 \) will be independent of \((S, Z)\).

\(^{17}\)See Heckman and Robb (1985, 2000) and Heckman (1997).
schooling regime (conditional on $X$) are

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

and for individuals in the regime with no compulsion

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

From the difference in conditional means we can identify:

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

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

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

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

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

and $TT = ATE = MTE = LATE$ 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 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. If the policy we want to evaluate is compulsory schooling then the instrumental variable 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 estimate and the policy relevant parameter coincide. But whenever that is not the case, the IV estimand does not identify the effect of the policy.
For example, if the policy we want to consider is tuition subsidies 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.\footnote{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 possible in any given dataset because of support problems. Different policies define different policy relevant instrumental variables.}

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.

$$plim \hat{\beta}_{IV} = \Delta^{IV} = \int_{0}^{1} \Delta^{MTE}(x, u)h_x(u)du$$

where

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

and $\int_{0}^{1} h_x(u)du = 1$. These weights do not, in general, coincide with the policy weights of Section 3 or the weights for the treatment parameters presented in Table 1B.

A closer look at these weights reveals that

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

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

$$h_x(p) = 0 \quad p \leq P^{Min} \quad p \geq P^{Max}.$$
For proofs, see Heckman and Vytlacil (2000).\textsuperscript{19,20}

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.\textsuperscript{21}

Observe that different choices of instruments $Z$, all satisfying conditions (A-1) and (A-3), generate different weights and hence estimate different objects. 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. 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 (??) 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:

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

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

Table 1 summarizes a central result of this literature and the various weights for the different estimands and parameters. The treatment effect parameters weight $MTE$ differently

\textsuperscript{19}Take a more general instrument $J$, and recenter $J$ so that $E(J) = 0$. Keeping conditioning on $X$ implicit, $\hat{\beta}_{IV} = \frac{E(JY)}{E(JS)}$ where $\hat{\beta}_{IV}(J) = MTE(u) h(u; J) du$ and $h(u; J) = \frac{E(J | P(Z) \geq u) \Pr(P(z) \geq u)}{E(JP)}$\textsuperscript{20} if $E(J | P \geq p)$ weakly increasing in $p$, (ii) Support $h(u; P) = [p, \bar{p}]$ (Support of $P$); (iii) defining $T(p) = E(J | P = p)$, we have $h(u; J) = h(u; T(P))$. (See Heckman and Vytlacil, 2000, 2004b).

\textsuperscript{20}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) (weighed average of Wald estimators) and Yitzhaki (1996, 1999). However, they do not relate the weights to those of the policy relevant treatment effects or to the weights required to estimate the conventional treatment parameters.

\textsuperscript{21}Moreover, they are also not defined for values of $u_S$ where $MTE(x, u_S) = 0$. 

18
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.\(^{22}\) $TT$ 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 ($\beta$) 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.\(^{23}\)

5 Using Local Instrumental Variables to Estimate the MTE

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

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

\(^{22}\)The form of the Roy model we use assumes additive separability and generates $U_0, U_1$ and $U_S$ from a
common unobservable $\varepsilon$. Thus the distribution of $U_1 - U_0$ given $U_S$ is degenerate.

\(^{23}\)We note parenthetically that the method of matching assumes that $\beta \perp \perp S \mid X$ or $\beta \perp \perp S \mid X, Z$ 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 in this paper.
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

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

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

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

Use of $P(Z)$ as an instrument, rather than the individual components of $Z$ separately, resolves the issue of the multiplicity of instruments (and parameters) that is a fundamental problem in the application of the method of instrumental variables to the correlated random coefficient model. By combining different instruments into a single index motivated by economic theory, we avoid the ambiguity of parameters defined by each instrument and extend the support of any one instrument. This extension of the support of any instrument is useful in 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 $Z$.


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

$$\frac{E(\ln Y | P(Z) = P(z)) - E(\ln Y | P(Z) = P(z'))}{P(z) - P(z')} = \beta + \frac{E(U_1 - U_0 | P(z) \geq U_S)P(z) - E(U_1 - U_0 | P(z') \geq U_S)P(z')}{P(z) - P(z')} = \Delta^{LATE}(x, P(z), P(z')),$$

where $\Delta^{LATE}$ was defined in Section 2. When $U_1 \equiv U_0$ or $(U_1 - U_0) \perp U_S$, corresponding to the two special cases in the literature, IV based on $P(Z)$ estimates $ATE (= \beta)$ because

---

24Thus if $\mu(Z) = Z\gamma$, we can use the variation in the other $Z$s to substitute for the missing variation in $Z_1$ given identification of the $\gamma$ up to a common scale.
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 $E(\ln Y \mid P(Z) = P(z))$ that reveals the index structure underlying this model more explicitly writes

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

(9)

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

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

IV estimates $\bar{\beta}$ if $\Delta^{MTE}(x, u_s)$ does not vary with $U_s$. Under this condition $E(\ln Y \mid P(Z) = p)$ 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 $E(\ln Y \mid P(Z) = p)$ 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 $E(\ln Y \mid P(Z))$ is linear in $P(z)$, standard instrumental variables methods identify “the” effect of $S$ on $\ln Y$. This test is simple to execute and interpret and we apply it below.

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 Vtlacil 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. $S$ is college attendance. In our data set there are 713 high school graduates who never attend college and 731 individuals who attend any type of college.\(^{25}\) Table 3 documents that individuals who attend college have on average a 32% higher wage than those who do attend college. They also have one year less of work experience since they spend more time in school.\(^{26}\) 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.\(^{27}\) 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, two measures of the opportunity costs 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.\(^{28}\) AFQT enters the schooling choice equation (and therefore the $Z$

---

\(^{25}\)These are white males, in 1992, with either a high school degree or above (GED holders are included as high school graduates) 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 datasets, see Carneiro (2002).

\(^{26}\)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.

\(^{27}\)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 individuals have different amounts of schooling and the effect of schooling on AFQT scores is important. We use a version of the 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.

\(^{28}\)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 these results below.
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 $\mu_s(z) = z\gamma$. 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 depend on arbitrary functional form assumptions about unobservables.

Table 4 gives estimates of $\gamma$ 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. Our tuition effects conform to the ones found in the literature that measures enrollment-tuition responses in the US: a $1000 \text{ reduction in (four year college)}$ tuition leads to an increase in enrollment of 5% (see Kane, 1994 or Cameron and Heckman, 2001 for summaries of the literature). The support of the estimated $P(Z)$ is shown in Figure 3 and it is almost the full unit interval, although at the extremes of the interval the cells of data become very thin. This makes estimation of the parameters defined over the full support of $P$ problematic. However, the $MTE$ can be estimated pointwise and does not require full support.

In order to simplify the estimation procedure and make our results comparable to specifications of schooling equations estimated in the literature, 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

---

29 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.

30 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.

31 The marginal treatment effect can only be identified for values of $P$ in the overlap of the supports of $P$ for these two populations. If $S = 0$ and $S = 1$, the exact support of $P$ for $S = 0$ individuals goes from 0.00005 to 0.9644 while for $S = 1$ individuals it goes from 0.0132 to 0.9917. The support of these two populations overlaps in the interval between 0.0132 and 0.9644.
\[ \ln Y_0 = \alpha_0 + X\theta_0 + U_0. \] This implies that the return can be written as

\[ \beta = \alpha_1 - \alpha_0 + X(\theta_1 - \theta_0) + U_1 - U_0 \]

so 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 \((U_0, U_1, U_S) \perp \perp (X, Z)\).\(^{32}\) Combining the model for \(S\) with the model for \(Y\) implies a partially linear model for the conditional expectation of \(Y\):

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

where

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

with \(\Phi(\cdot)\) 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 \((\alpha_0)\), the intercept term in \((P(Z)(\alpha_1 - \alpha_0))\) and the intercept of the function \(K(P)\). However the \(MTE\) can still be identified since it does not depend on these intercepts:

\[ \frac{\partial E[\alpha_0 + P(\alpha_1 - \alpha_0) + K(P)]}{\partial P}_{P=p} = \alpha_1 - \alpha_0 + E(U_1 - U_0|U_S = p). \]

This 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 flexibility than traditional parametric approaches.\(^{33}\) For simplicity (and in

\(^{32}\text{(A-3) only requires that } (U_0, U_1, U_S) \perp \perp Z|X, \text{ so we do not estimate the most general possible model within our framework.}

\(^{33}\text{The partially linear model was introduced by } \text{Robinson (1988). Imposing the partially linear model weakens the support condition that otherwise would be required for } P(Z). \text{ In particular, fully nonparametric analysis of all treatment parameters and policy counterfactuals would require that the support of } P(Z) \text{ conditional on } X \text{ be the full unit interval. In contrast, the analysis with the partially linear model requires that } X \text{ be full rank conditional on } P(Z) \text{ and that } P(Z) \text{ have support on the full unit interval, without requiring that } P(Z) \text{ conditional on } X \text{ have support on the full unit interval.}\)
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 ($\theta_{1}^{\text{experience}} = \theta_{0}^{\text{experience}}, \theta_{1}^{\text{experience}^2} = \theta_{0}^{\text{experience}^2}$). $^{34}$ AFQT is the only $X$ variable that influences the return to schooling ($\theta_{1}^{\text{AFQT}} \neq \theta_{0}^{\text{AFQT}}$). $^{35}$

Selection on observables ($X$) is important if the coefficient on the interaction between $P(Z)$ and $X (= \theta_{1} - \theta_{0})$ is precisely estimated. Simple least squares regressions of log wages on schooling, ability measures, and interactions of schooling and ability (ignoring selection on 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$^{36}$. 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 can test for selection or comparative advantage using equation (??) by checking whether $\ln Y$ is a linear or a nonlinear function of $P$. Nonlinearity in $P$ means that selection is important. A simple way to implement this test is to approximate $K(P)$ with a third order polynomial in $P$ and test whether the coefficients in the second and third order terms are statistically significant.$^{37}$ We reject the null hypothesis that these coefficients are jointly equal to zero (p-value = 0.0564). Nonlinearity 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 $\tilde{\beta}(x) = ATE$. $^{38}$

Following Heckman, Ichimura, Smith and Todd (1998), we estimate the partially linear

$^{34}$ 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.

$^{35}$ 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.


$^{37}$ The results from this test are reported in Table A2.

$^{38}$ The standard errors used to perform this test account for parameter estimation in $P(Z)$. 
model using a double residual regression procedure using local linear regression.\textsuperscript{39} We use a bandwidth of 0.3\textsuperscript{40} and all the standard errors we present are bootstrapped.\textsuperscript{41} 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 $u_S$:

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

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 quantitatively important but is imprecisely estimated. 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).\textsuperscript{42} Figure 5 plots the component of the $MTE$ that depends 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 (??).\textsuperscript{43} We approximate the derivative of $K(P(Z))$ by taking discrete differences:

$$\frac{\partial K(P)}{\partial P} = \frac{K(P + h) - K(P)}{h}.$$
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.\(^{44}\) Returns are annualized to reflect the fact that college goers attend 3.5 years of school. The most college worthy persons in the sense of having high gross returns are more likely to go to college (have low values of $u_S$, the “cost” of college). But for high values of $u_S$ (above 0.4) this curve 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 at middle ranges of $u_S$.\(^{45}\) 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 $E(U_1 - U_0 \mid U_S = u_S)$ indicates that other factors besides financial returns determines 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 (psychic) cost is very high.\(^{46}\)

Table 6 presents estimates of different summary measures of returns to one year of college.

\(^{44}\)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. The function is only plotted between the values of $U_S = 0.05$ and $U_S = 0.90$. In fact the support of $P$ extends from 0.01 to 0.96 but we opted to trim the extremes because the cells of data become very thin at the extremes of the distribution of $P$.

\(^{45}\)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.

\(^{46}\)This pattern is also consistent with the existence of credit constraints affecting a segment of the population. 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.
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:

$$
\Delta^{ATE} = \int \Delta^{ATE}(x) f_X(x) \, dx
$$

$$
\Delta^{TT} = \int \Delta^{TT}(x) f_X(x|S = 1) \, dx
$$

$$
\Delta^{TUT} = \int \Delta^{TUT}(x) f_X(x|S = 0) \, dx
$$

$$
\Delta^{AMTE} = \int \Delta^{AMTE}(x) f_X(x|\text{Marginal}) \, dx
$$

$$
\Delta^{PRT} = \int \Delta^{PRT}(x) f_X(x|\text{PRT}) \, dx
$$

where $f_X(x|\text{PRT})$ is the density of $X$ for individuals induced to go to college by the policy.

The schooling choice equation is: $S = 1[Z\gamma - U_S \geq 0]$, so

$$
f_X(x|S = 1) = f_X(x|Z\gamma - U_S \geq 0)
$$

$$
f_X(x|S = 0) = f_X(x|Z\gamma - U_S < 0)
$$

$$
f_X(x|\text{Marginal}) = f_X(x|Z\gamma = U_S)
$$

$$
f_X(x|\text{PRT}) = f_X(x|z\gamma - U_S < 0, z'\gamma - U_S \geq 0)
$$

where $z$ and $z'$ are the values of the instruments ($PRT$ is defined conditional on a policy) under the baseline regime and under the new policy regime, respectively. These densities are also weights, but instead of weighting functions of $U_S$ they weight functions of $X$ (see also Carneiro, 2002). We estimate that the average annual return to college for a randomly selected person in the population ($ATE$) is 16.26% which is between the annual return for the average individual who attends college ($TT$), 18.08%, and the average return for high
school graduates who never attend college ($TUT$), 14.56%. The average marginal individual
(AMTE) has an annual return of 15.02% 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.\textsuperscript{47} 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 15.05%, although this estimate is very close to
the return for the average marginal person. This is the relevant return for evaluating this
specific policy using the 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 (three 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
$U_S$\textsuperscript{48}, 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 $E(Y_1 - Y_0 | AFQT)$ and Figure 8 presents the joint
($U_S, AFQT$) policy weights.\textsuperscript{49} 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 $AFQT$.\textsuperscript{50}

The limited support of $P$ creates a practical problem for the computation of the param-

\textsuperscript{47}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. Our linear IV
estimate of 12.5% is in the range of the IV estimates in the literature.

\textsuperscript{48}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 $U_S$.

\textsuperscript{49}$E(Y_1 - Y_0 | AFQT)$ 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.

\textsuperscript{50}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: $ATE = 0.1600$,
$TT = 0.1786$, $TUT = 0.1427$, $AMTE = 0.1478$. 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).
eters presented in this table, since we cannot evaluate $MTE$ for values of $U_S$ outside the interval $[0.01, 0.96]$. Furthermore, at the extremes of this interval there do not exist much data to accurately estimate the $MTE$. The numbers presented in Table 6 are constructed by assuming that every individual in the population has a $U_S$ inside the interval $[0.05, 0.9]$. Therefore, we ignore $MTE$ outside this interval since every parameter gives weight zero to observations outside this interval. An alternative way to deal with the problem of limited support is to construct bounds for the parameters, but these bounds are generally wide. However, the problem of limited support only affects those parameters, such as $ATE$ or $TT$, that put substantial weight on the tails of the $MTE$ which cannot be estimated with limited support. Even though empirical economists often seek to identify these parameters, often they are not the economically interesting ones. The $PRTE$ or the $AMTE$ are both economically more interesting and much easier to estimate because they place little weight on the tails of the $MTE$.

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 (??). It estimates $ATE$ if $S$ and $X$ are orthogonal to $U_0 + S(U_1 - U_0)$. The $IV$ estimate is derived from the same equation. Since the returns estimated by $OLS$ and by $IV$ both depend on $X$ (in this case, AFQT), we evaluate the $OLS$ and $IV$ returns at the average

---

51 In the upper panel of appendix Table A4 we replicate Table 6 assuming that the support of $U_S$ in the population is $[0.01, 0.96]$ (the support of $P$ in this sample). We also present the numbers in Table 6 for comparison.

52 As an illustration of how wide these bounds can be, bounds for $ATE$ (as proposed in Heckman and Vytlacil, 2000) are presented at the bottom of Table A4, for two different supports for the data.

53 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 $TUT$ weight substantially 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.

54 However, we could also define a policy that affects people at either tail of the $MTE$. 

---

30
value of $X$ for individuals induced to enroll in college by a $1000 tuition subsidy\textsuperscript{55}, 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 $U_S$.\textsuperscript{56}

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.\textsuperscript{57} 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\textsuperscript{58} 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 $\beta_{IV} > \beta_{OLS}$.

Notice that the least squares estimator does not identify the return to the average person $E(\beta|S = 1) = E(\ln Y_1 - \ln Y_0|S = 1)$. Rather it identifies (keeping the conditioning on $X$

\textsuperscript{55}This is obtained by integrating $X$ over $f_X(x|PRT) = f_X(x|z\gamma - U_S < 0, z'\gamma - U_S \geq 0)$.

\textsuperscript{56}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 ($TT$).

\textsuperscript{57}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 $E(Y_1 - Y_0|X,U_S = p) = \frac{\partial E(Y|X,P)}{\partial p}|_{p=p}$.

\textsuperscript{58}It estimates the return for the “switchers”.

31
\[
E(\ln Y | S = 1) - E(\ln Y | S = 0) = E(\beta | S = 1) + [E(U_0 | S = 1) - E(U_0 | S = 0)]
\]
\[
= \bar{\beta} + E(U_1 - U_0 | S = 1) + [E(U_0 | S = 1) - E(U_0 | S = 0)].
\]

In a model without variability in the returns to schooling, \(E(\beta | S = 1) = E(\beta) = \bar{\beta}\) is the same constant for everyone, so it is plausible that if \(U_0\) is ability, the second term in parentheses 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 \((E(\beta | S = 1))\). Thus the evidence reported in the recent literature comparing OLS and IV is not informative on the comparison of the returns to the average and the marginal person.

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 \(P(z) = \Pr(S = 1|Z = z)\). 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 \(P(z)\) in this paper. The estimates range all over the map. They are different from each other and from the estimate generated by using \(P(z)\). 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.

\[59\] 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 \( \rho_{Z,S} \). 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\(^{61}\).

\(^{60}\)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.

\(^{61}\)In a recent paper Stock and Yogo (2003) propose a different test. However they still find that the rule
The table shows that for the local wage and local unemployment variables these statistics are 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). 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. 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.

Our use of ability measures 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

---

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

63 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.

34
implausibly large (see Carneiro, 2002). See the substantial increase in the IV estimate in Table 9 when AFQT is omitted from the model. Indeed, all of the estimates reported in Table 9 increase substantially when AFQT is eliminated. Ability bias is an important empirical phenomenon and failure to control for it leads to substantial upward biases in estimated 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 (valid) 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 the 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. Different instruments define different parameters, and in our empirical analysis produce wildly different “effects” of schooling on earnings. In our empirical analysis, $IV$ understates our stated policy relevant return by two 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


