American Economic Review 101 (October 2011): 2754–2781
http://www.aeaweb.org/articles.php?doi=10.1257/aer.101.6.2754

Estimating Marginal Returns to Education†

By PEDRO CARNEIRO, JAMES J. HECKMAN, AND EDWARD J. VYTLACIL*

Estimating marginal returns to policies is a central task of economic cost-benefit analysis. A comparison between marginal benefits and marginal costs determines the optimal size of a social program. For example, to evaluate the optimality of a policy that promotes expansion in college attendance, analysts need to estimate the return to college for the marginal student and compare it to the marginal cost of the policy.

This is a relatively simple task (i) if the effect of the policy is the same for everyone (conditional on observed variables) or (ii) if the effect of the policy varies across individuals given observed variables but agents either do not know their idiosyncratic returns to the policy, or if they know them, they do not act on them. In these cases, individuals do not choose their schooling based on their realized idiosyncratic individual returns, and thus the marginal and average ex post returns to schooling are the same.[1]

Under these conditions, the mean marginal return to college can be estimated using conventional methods applied to the following Mincer equation:

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

where Y is the log wage, S is a dummy variable indicating college attendance, β is the return to schooling (which may vary among persons), and ε is a residual. The standard problem of selection bias (S correlated with ε) may be present, but this problem can be solved by a variety of conventional methods (instrumental variables (IV), regression discontinuity, and selection models).

* Carneiro: Department of Economics, University College London, Gower Street, London WC1E 6BT, United Kingdom, Institute for Fiscal Studies and Centre for Microdata Methods and Practice (e-mail: p.carneiro@ucl.ac.uk); Heckman: Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637, American Bar Foundation, Geary Institute, and University College Dublin (e-mail: jjh@uchicago.edu); Vytlacil: Department of Economics and Cowles Foundation, Yale University, Box 208281, New Haven, CT 06520-8281 (e-mail: edward.vytlacil@yale.edu). Carneiro acknowledges the support of ESRC RES-000-22-2542 and RES-589-28-0001 through the Centre for Microdata Methods and Practice. Heckman thanks the National Institutes of Health (R01-HD054702), the JB and MK Pritzker Family Foundation, the Buffett Early Childhood Fund, the American Bar Foundation, and the Committee for Economic Development with a grant from The Pew Charitable Trusts and the Partnership for America's Economic Success. Heckman also thanks the Cowles Foundation at Yale University, which supported a visit that facilitated completion of this research. Vytlacil thanks NSF SES-0832845 and Hitotsubashi University, at which he was a visiting professor while the research in part was conducted. We thank Michael Greenstone and Ben Williams for comments. The views expressed in this paper are those of the authors and not necessarily those of these funders.

† To view additional materials, visit the article page at http://www.aeaweb.org/articles.php?doi=10.1257/aer.101.6.2754.

[1] See Heckman and Vytlacil (2007b).

The recent literature shows how to empirically test the conditions that justify application of conventional methods (Heckman, Daniel Schmierer, and Sergio Urzua 2010; Heckman and Schmierer 2010). Applying these methods on data from the National Longitudinal Survey of Youth of 1979 (NLSY), we find that returns vary (i.e., β is random) and, furthermore, agents act as if they possess some knowledge of their idiosyncratic returns (i.e., β is correlated with S). Selection on gains complicates the estimation of marginal returns.

Under assumptions presented in Guido W. Imbens and Joshua D. Angrist (1994), an instrumental variable estimator identifies a local average treatment effect (LATE), which measures the return to college for individuals induced to go to school by the change in the instrument. Unfortunately, the people induced to go to school by a change in an instrument need not be the same as the people induced to go to school by a given policy change, and the returns to the two groups of people can differ substantially, as we illustrate in this paper. We show how to use a local version of instrumental variables introduced in Heckman and Vytlacil (1999, 2005, 2007b) to estimate the marginal returns to alternative ways of producing marginal expansions in college attendance without requiring that the variation in the available instruments correspond exactly to the variation induced by a policy.

For a sample of white males from the NLSY, we establish that marginal expansions in college attendance attract students with lower returns than those enjoyed by persons currently attending college. The contrast between what conventional IV measures and the marginal return to a policy can be stark. For example, while the conventional IV estimate is 0.0951, the estimated marginal return to a policy that expands each individual's probability of attending college by the same proportion is only 0.0148. This policy induces students who should not attend college to attend it.[2]

The methods used in this paper improve on LATE by identifying what sections of an economically interpretable mean marginal benefit surface are identified by different instruments. It is thus possible to compare on a common scale the different margins identified by different instruments. Furthermore, our methods allow the evaluation of policy changes that do not directly correspond to current variation in any particular instrument and are thus not directly identified by standard IV procedures.

The plan of this paper is as follows. In the next section, we present the empirical framework used in this paper. The following section discusses the estimation method and presents empirical estimates and robustness checks. The final section concludes.

I. Methods for Estimating Marginal Returns

The generalized Roy model is a basic choice-theoretic framework for policy analysis.[3] Let Y_1 be the potential log wage if the individual were to attend college, and Y_0

[2] For this policy change, the analysis of Charles Murray (2008a, b) that too many students go to college appears to be correct.

[3] The model originates in the work of A. D. Roy (1951) and Richard E. Quandt (1958, 1972). See, e.g., Heckman and Vytlacil (2007a) for a discussion of the model, its origins, and its wide uses in economics.

the potential log wage if the individual were not to attend college.[4] Define potential outcomes as

$$Y_1 = \mu_1(\mathbf{X}) + U_1 \quad \text{and} \quad Y_0 = \mu_0(\mathbf{X}) + U_0, \tag{2}$$

where $\mu_1(\mathbf{x}) \equiv E(Y_1 | \mathbf{X} = \mathbf{x})$ and $\mu_0(\mathbf{x}) \equiv E(Y_0 | \mathbf{X} = \mathbf{x})$. The return to schooling is $Y_1 - Y_0 = \beta = \mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + U_1 - U_0$, so that the average treatment effect conditional on $\mathbf{X} = \mathbf{x}$ is given by $\overline{\beta}(\mathbf{x}) = E(\beta | \mathbf{X} = \mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$ and the average effect of treatment on those who choose to attend college conditional on $\mathbf{X} = \mathbf{x}$ is given by $E(\beta | \mathbf{X} = \mathbf{x}, S = 1) = \overline{\beta}(\mathbf{x}) + E(U_1 - U_0 | S = 1, \mathbf{X} = \mathbf{x})$. $\mathbf{X}$ need not be statistically independent of (U_0, U_1). We condition on $\mathbf{X}$ throughout.

A standard latent variable discrete choice model represents the individual's decision to enroll in college (see, e.g., Willis and Rosen 1979). Let I_S be the net benefit to the individual of enrolling in college, which depends on observed $(\mathbf{Z})$ and unobserved (V) variables:

$$I_S = \mu_S(\mathbf{Z}) - V, \tag{3}$$

$$S = 1 \text{ if } I_S \geq 0; \quad S = 0, \text{ otherwise.}$$

A person goes to college $(S = 1)$ if $I_S \geq 0$; otherwise, $S = 0$. In this notation, the analyst observes $(\mathbf{Z}, \mathbf{X})$ but not (U_0, U_1, V). V is assumed to be a continuous random variable with a strictly increasing distribution function F_V. V may depend on U_1 and U_0 in a general way. The $\mathbf{Z}$ vector may include some or all of the components of $\mathbf{X}$, but it also includes variables excluded from equation (2) (i.e., excluded from $\mathbf{X}$). We assume that (U_0, U_1, V) is statistically independent of $\mathbf{Z}$ given $\mathbf{X}$. The additive separability between $\mathbf{Z}$ and V in the latent index plays an essential role in the instrumental variables literature. Model (3) with $\mathbf{Z}$ statistically independent of (U_0, U_1, V) given $\mathbf{X}$ implies and is implied by the Imbens-Angrist independence and "monotonicity" assumptions (see Vytlacil 2002 and the discussion in Heckman 2010).

Let $P(\mathbf{z})$ denote the probability of attending college $(S = 1)$ conditional on $\mathbf{Z} = \mathbf{z}$, $P(\mathbf{z}) \equiv \Pr(S = 1 | \mathbf{Z} = \mathbf{z}) = F_V(\mu_S(\mathbf{z}))$, where we keep the conditioning on $\mathbf{X}$ implicit. $P(\mathbf{z})$ is sometimes called the propensity score. Define $U_S = F_V(V)$. It is uniformly distributed by construction, and different values of U_S correspond to different quantiles of V. We can rewrite (3) using $F_V(\mu_S(\mathbf{Z})) = P(\mathbf{Z})$ so that $S = 1$ if $P(\mathbf{Z}) \geq U_S$. $P(\mathbf{Z})$ is the mean scale utility function in discrete choice theory (Daniel McFadden 1974).

The marginal treatment effect (MTE), defined by

$$\text{MTE}(\mathbf{x}, u_S) \equiv E(\beta | \mathbf{X} = \mathbf{x}, U_S = u_S),$$

[4] We reduce schooling choices to two levels, as in Robert J. Willis and Sherwin Rosen (1979), Christopher R. Taber (2001), or Robert A. Moffitt (2008). Heckman and Vytlacil (2007b) and Heckman, Urzua, and Vytlacil (2006, 2008) present methods for analyzing multiple schooling levels. We annualize our estimates by dividing them by the difference in the average years of schooling of individuals in each group. Since this is a potentially important restriction, in the empirical work reported below we present results using alternative definitions of schooling.

is central to our analysis of how to go from (local) IV estimates to policy effects. This parameter was introduced in the literature by Anders Björklund and Moffitt (1987) and extended in Heckman and Vytlacil (1999, 2001a, 2005, 2007b). It is the mean return to schooling for individuals with characteristics $\mathbf{X} = \mathbf{x}$ and $U_S = u_S$. Recall that U_S has been normalized to be unit uniform, so that tracing MTE over u_S values shows how the returns to schooling vary with different quantiles of the unobserved component of the index of the desire to go to college. Alternatively, it is the mean return to schooling for persons indifferent between going to college or not who have mean scale utility value $P(\mathbf{Z}) = u_S$.

The MTE can be estimated by the method of local instrumental variables proposed by Heckman and Vytlacil (1999, 2001a, 2005), which we implement in this paper. It is identified by differentiating $E(Y|\mathbf{X} = \mathbf{x}, P(\mathbf{Z}) = p)$ with respect to p, which can be computed over the support of the distribution of $P(\mathbf{Z})$. Using the equations in (2), observed earnings are

$$
\begin{aligned}
(4) \qquad Y &= (S)Y_1 + (1 - S)Y_0 \\
&= \mu_0(\mathbf{X}) + [\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + U_1 - U_0]S + U_0 \\
&= \mu_0(\mathbf{X}) + [\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})]S + \{U_0 + S(U_1 - U_0)\}.
\end{aligned}
$$

The conditional expectation of Y given $\mathbf{X} = \mathbf{x}$ and $P(\mathbf{Z}) = p$ is

$$
\begin{aligned}
E(Y|\mathbf{X} = \mathbf{x}, P(\mathbf{Z}) = p) &= E(Y_0|\mathbf{X} = \mathbf{x}, P(\mathbf{Z}) = p) \\
&\quad + E(Y_1 - Y_0|\mathbf{X} = \mathbf{x}, S = 1, P(\mathbf{Z}) = p)p.
\end{aligned}
$$

Using choice equation (3), this expression can be written as

$$
\begin{aligned}
(5) \quad & E(Y|\mathbf{X} = \mathbf{x}, P(\mathbf{Z}) = p) \\
&= \mu_0(\mathbf{x}) + [\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})]p \\
&\quad + \int_{-\infty}^{\infty} \int_0^p (u_1 - u_0) f(u_1 - u_0|\mathbf{X} = \mathbf{x}, U_S = u_S)\, du_S\, d(u_1 - u_0),
\end{aligned}
$$

where $f(u_1 - u_0|\mathbf{X} = \mathbf{x}, U_S = u_S)$ is the conditional density of $U_1 - U_0$.[5] Simplifying the expression,

$$
E(Y|\mathbf{X} = \mathbf{x}, P(\mathbf{Z}) = p) = \mu_0(\mathbf{x}) + \int_0^p \mathrm{MTE}(\mathbf{x}, u_S)\, du_S.
$$

[5] We are assuming in this derivation that $U_1 - U_0$ is a continuous random variable. However, the result holds under more general conditions. See Heckman and Vytlacil (2001a, 2005) and Heckman, Urzua, and Vytlacil (2006).

The left-hand side of this expression can be consistently estimated from sample data. Differentiating with respect to p, we obtain the MTE:

$$\frac{\partial E(Y|\mathbf{X} = \mathbf{x}, P(\mathbf{Z}) = p)}{\partial p} = \text{MTE}(\mathbf{x}, p). \tag{6}$$

Applied to sample data, this is the local instrumental variable (LIV) estimator of Heckman and Vytlacil (1999). We can recover the return to schooling for persons indifferent between $S = 1$ and $S = 0$ at all margins of U_S within the empirical support of $P(\mathbf{Z})$ (conditional on $\mathbf{X}$). Notice that persons with a high mean scale utility function $P(\mathbf{Z})$ identify the return for those with a *high* value of U_S, i.e., a value of U_S that makes persons *less* likely to participate in schooling. Marginal increases in $P(\mathbf{Z})$ starting from high values of $P(\mathbf{Z})$ induce those individuals with high U_S values into schooling. Those with low values of U_S are already in school for such values of $P(\mathbf{Z})$ so that a marginal increase in $P(\mathbf{Z})$ starting from a high value has no effect on them. We can identify returns at all quantiles of U_S within the support of the distribution of $P(\mathbf{Z})$. Thus, we can determine which persons (identified by the quantile of the unobserved component of the desire to go to college, U_S) are induced to go into college ($S = 1$) by a marginal change in $P(\mathbf{Z})$.

As noted by Heckman, Urzua, and Vytlacil (2006), the probability of selection ($P(\mathbf{Z})$), sometimes called the "propensity score," plays a central role in instrumental variable models that satisfy the independence and monotonicity conditions of Imbens and Angrist (1994) (or equivalently that are characterized by the latent variable discrete choice model of equation (3)). Aggregating the instruments into the scalar index $P(\mathbf{Z})$ enlarges the range of values over which we can identify MTE in comparison to using each instrument one at a time.[6] One can also determine the contribution of each instrument to identifying different regions of the MTE function.

Heckman and Vytlacil (1999, 2001a, 2005, 2007a, b) establish that standard summary measures of the return to college, such as the average return to college in the population ($E(\beta|\mathbf{X})$) and the average return to college among those who attend college ($E(\beta|\mathbf{X}, S = 1)$), can be expressed as different weighted averages of the MTE. They show that treatment parameter j, $\Delta_j(\mathbf{x})$, can be written as a weighted average of the MTE:

$$\Delta_j(\mathbf{x}) = \int_0^1 \text{MTE}(\mathbf{x}, u_S)\, h_j(\mathbf{x}, u_S)\, du_S. \tag{7}$$

See Table A-1B in the online Appendix for the weights for different treatment parameters in the literature. Heckman and Vytlacil also show that the standard IV estimator

[6] In the case of multiple instruments when β is correlated with S, the common practice of using one instrument at a time to identify the marginal effect of that instrument (e.g., David Card 1999, 2001) is fraught with danger. It is necessary to account for the variation in other instruments associated with the variation in the instrument used in order to isolate the ceteris paribus effect of any particular instrument. Thus, if $\mathbf{Z} = (Z_1, \ldots, Z_K)$ and $K \geq 2$, computing the LATE for Z_1 not controlling for $Z_2, \ldots, Z_K$ produces the direct effect of Z_1 on S as it affects Y, as well as the effect produced by varying $Z_2, \ldots, Z_K$ and their effects on S and hence Y if they covary with Z_1. Aggregating all $\mathbf{Z}$ into $P(\mathbf{Z})$, and looking at the effect of variations in $P(\mathbf{Z})$ on outcomes, avoids this problem. From the economics of the problem, all $\mathbf{Z}$ enter choices through their effects on $P(\mathbf{Z})$. See Heckman (2010) for a discussion of this issue.

is a weighted average of the MTE with estimable weights, i.e., for instrument Z^k,

$$\text{IV}_k(\mathbf{x}) \equiv \frac{\text{Cov}(Y, Z^k | \mathbf{X} = \mathbf{x})}{\text{Cov}(S, Z^k | \mathbf{X} = \mathbf{x})} = \int_0^1 \text{MTE}(\mathbf{x}, u_S) h_k(\mathbf{x}, u_S) \, du_S, \tag{8}$$

where, again, the weights can be consistently estimated from sample data.[7] Observe that if MTE $(\mathbf{x}, u_S)$ does not depend on u_S, all instruments estimate the same parameter which is $\beta(\mathbf{x})$, the average treatment effect. In this case, marginal and average ex post returns are equal, and all IVs estimate a common policy effect that is the same no matter how $P(\mathbf{Z})$ is varied by policy shifts. The instrumental variable weights arise because different values of the instruments identify different segments of the MTE and IV averages out the different values using weight $h_k(\mathbf{x}, u_S)$. The explicit formula for the IV weight is given in online Appendix Table A-1B.[8] In general, the IV weights are different from the treatment parameter weights.

A. *Policy Relevant Treatment Effects and Marginal Policy Relevant Treatment Effects*

Following Heckman and Vytlacil (2001b, 2005, 2007b), we consider a class of policies that change $P(\mathbf{Z})$, the probability of participation in the program, but that do not affect potential outcomes or the unobservables related to the selection process, (Y_0, Y_1, V).[9] An example from the literature on the economic returns to schooling would be policies that change tuition or distance to school, but that do not directly affect potential wages (Card 2001).[10]

Let S^* be the treatment choice that would be made after the policy change. Let P^* be the corresponding probability that $S^* = 1$ after the policy change. S^* is defined by $S^* = 1[P^* \geq U_S]$. Let $Y^* = S^* Y_1 + (1 - S^*) Y_0$ be the outcome under the alternative policy. Heckman and Vytlacil (2005, 2007b) show that the mean effect of going from a baseline policy to an alternative policy per net person shifted is the policy relevant treatment effect (PRTE), defined when $E(S) \neq E(S^*)$ as

$$\frac{E(Y | \text{Alternative Policy}) - E(Y | \text{Baseline Policy})}{E(S^* | \text{Alternative Policy}) - E(S | \text{Baseline Policy})} = \frac{E(Y^*) - E(Y)}{E(S^*) - E(S)} = \int_0^1 \text{MTE}(u_S) \omega_{\text{PRTE}}(u_S) \, du_S,$$

where

$$\omega_{\text{PRTE}}(u_S) = \frac{F_P(u_S) - F_{P^*}(u_S)}{E_{F_{P^*}}(P) - E_{F_P}(P)},$$

[7] The weights integrate to one, no matter what the instrument, but they need not be positive at all values of u_S.

[8] The online Appendix shows the relationship between LATE and MTE. The former is the integral of the latter.

[9] This restriction can be relaxed to a weaker policy invariance assumption for the distribution of (Y_0, Y_1, V); see Heckman and Vytlacil (2005, 2007b).

[10] We ignore general equilibrium effects.

where F_{P^*} and F_P are the distributions of P^* and P, respectively, and we suppress $\mathbf{X}$ to simplify notation. The condition $E(S) \neq E(S^*)$ is consistent with the policy having a nonmonotonic effect on participation, as long as the fraction switching into treatment is not exactly offset by the fraction switching out of treatment. The PRTE parameter gives the normalized effect of a change from a baseline policy to an alternative policy and depends on the alternative being considered.[11]

As shown by the above equation, PRTE depends on the policy change only through the distribution of P^* after the policy change. In other words, given our assumptions, F_{P^*} is sufficient to summarize everything about the proposed policy change that is relevant for calculating the average effect of the policy change. The PRTE maps the proposed policy change (corresponding to a distribution of P^*) to the resulting per-person change in outcomes. In general, the PRTE for a proposed policy change of interest will differ from the probability limit of an IV estimator. The change in the distribution of P induced by the policy in general differs from the change induced by an instrument, unless the instrument is the policy change. For example, policy variations in tuition will not, in general, have the same effects as instrument variations in distance to college.

The PRTE is defined for a discrete change from a baseline policy to a fixed alternative. As noted in Carneiro, Heckman, and Vytlacil (2010), identifying it in any sample can be a challenging task because it often requires that the support of $P(\mathbf{Z})$ be the full unit interval. Below, we show that our estimated $P(\mathbf{Z})$ does not satisfy this requirement. Instead, we estimate a marginal version of the PRTE parameter (MPRTE) that corresponds to a marginal change from a baseline policy. It is less empirically demanding to estimate, yet answers an economically interesting question. The marginal version of the PRTE depends on the nature of the perturbation that defines the marginal change. For example, a policy change that subsidizes tuition by a fixed amount and a policy change that subsidizes tuition so that the probability of college-going expands proportionately will, in general, have different limits for infinitesimally small subsidies (Carneiro, Heckman, and Vytlacil 2010).

More formally, we define the MPRTE as follows. Consider a sequence of policies indexed by a scalar variable α, with $\alpha = 0$ denoting the baseline, status quo policy. We associate with each policy α the corresponding fitted probability of schooling P_α, where $P_0 = P(\mathbf{Z})$, the baseline propensity score. For each policy α we define the corresponding PRTE parameter for going from the baseline status quo to policy α. We define the MPRTE as the limit of such a sequence of PRTEs as α goes to zero. We will consider the following examples of such sequences of policies: (i) a policy that increases the probability of attending college by an amount α, so that $P_\alpha = P_0 + \alpha$ and $F_\alpha(t) = F_0(t - \alpha)$; (ii) a policy that changes each person's probability of attending college by the proportion $(1 + \alpha)$, so that $P_\alpha = (1 + \alpha)P_0$ and $F_\alpha(t) = F_0(t/(1 + \alpha))$; and (iii) a policy intervention that has an effect similar to a shift in one of the components of $\mathbf{Z}$, say $Z^{[k]}$, so that $Z_\alpha^{[k]} = Z^{[k]} + \alpha$ and $Z_\alpha^{[j]} = Z^{[j]}$ for $j \neq k$. For example, the kth element of $\mathbf{Z}$ might be college tuition, and the policy

[11] The PRTE can be interpreted as an economically more explicit version of James Stock's (1989) nonparametric policy analysis parameter for a class of policy interventions with explicit agent preferences, where the policies evaluated operate solely on agent choice sets. The condition $E(S) \neq E(S^*)$ is not required if we define PRTE using aggregates not normalized to a per capita basis (see Heckman and Vytlacil 2001b).

TABLE 1—WEIGHTS FOR MPRTE

Measure of distance for people near the margin	Definition of policy change	Weight
$\lvert \mu_S(\mathbf{Z}) - V \rvert < e$	$Z^k_\alpha = Z^k + \alpha$	$h_{MPRTE}(\mathbf{x}, u_S) = \frac{f_{P\mid\mathbf{X}}(u_S) f_{V\mid\mathbf{X}}(F^{-1}_{V\mid\mathbf{X}}(u_S))}{E(f_{V\mid\mathbf{X}}(\mu_S(\mathbf{Z}))\mid\mathbf{X})}$
$\lvert P - U \rvert < e$	$P_\alpha = P + \alpha$	$h_{MPRTE}(\mathbf{x}, u_S) = f_{P\mid\mathbf{X}}(u_S)$
$\lvert \frac{P}{U} - 1 \rvert < e$	$P_\alpha = (1+\alpha)P$	$h_{MPRTE}(\mathbf{x}, u_S) = \frac{u_S f_{P\mid\mathbf{X}}(u_S)}{E(P\mid\mathbf{X})}$

Source: Carneiro, Heckman, and Vytlacil (2010).

under consideration subsidizes college tuition by the fixed amount α. In each of these three cases, we consider the corresponding PRTE for going from the status quo to policy α, and consider the limit of such PRTEs as α goes to zero. These limits differ from IV estimates in general. Just as IV is a weighted average of the MTE, as in equation (8), there is a similar expression for average marginal policy changes that weights up the MTE by the proportion of persons induced to change by the policy. In general, the weights are different for IV and MPRTE. (Compare the IV weights in Table A-1B in the online Appendix and the weights in Table 1 in the text.)

The MPRTE is the appropriate parameter with which to conduct cost-benefit analysis of marginal policy changes. In our empirical work we contrast our estimates of the MPRTE with conventional IV estimates of the returns to college.

Carneiro, Heckman, and Vytlacil (2010) relate the MPRTE to the average marginal treatment effect (AMTE): the mean benefit of treatment for people indifferent between participation in treatment and nonparticipation. There are technical issues that arise in identifying the marginal gain to persons in indifference sets that arise from the thinness of the indifference sets. For conventional economic models, the probability that anyone is indifferent is exactly zero. Different ways to approximate the indifference set $P(\mathbf{Z}) = U_S$ through limit operations determine different values of the AMTE. The effect of a marginal policy change for a particular perturbation of $P(\mathbf{Z})$ is the same as the average effect of treatment for those who are arbitrarily close to being indifferent between treatment or not, using a metric $m(P, U_S)$ measuring the distance between $P(\mathbf{Z})$ and U_S. This parameter is defined as $AMTE = \lim_{e \to 0} E[Y_1 - Y_0 \mid m(P, U_S) \leq e]$. For the three examples of MPRTE previously discussed, the corresponding metrics defining the AMTE are, respectively: (i) $m(P, U_S) = \lvert F_V^{-1}(P) - F_V^{-1}(U_S) \rvert = \lvert \mu_S(\mathbf{Z}) - V \rvert$; (ii) $m(P, U_S) = \lvert P - U_S \rvert$; (iii) $m(P, U_S) = \lvert (P/U_S) - 1 \rvert$. Table 1 shows the different weights associated with the different definitions of the AMTE and the associated MPRTE.

B. *Practical Issues in Estimating the MPRTE*

In practice, it is very difficult to condition on $\mathbf{X}$ nonparametrically, especially when $\mathbf{X}$ includes many variables (as is the case in our empirical work). Therefore, in our empirical work we proceed by imposing an additional assumption, which is standard in the literature applying both IV and selection models to estimate the returns to schooling: that $(\mathbf{X}, \mathbf{Z})$ is independent of (U_0, U_1, U_S) (as opposed to the weaker assumption that (U_0, U_1, U_S) is independent of $\mathbf{Z}$ given $\mathbf{X}$). Under this assumption the MTE is additively separable in $\mathbf{X}$ and U_S. One important consequence of imposing

this assumption is that the MTE is identified over the unconditional support of P, as opposed to the support of P conditional on $\mathbf{X}$. We discuss this assumption further in Section IIC below.

In our empirical analysis, we work with linear-in-the-parameters versions of $\mu_1(\mathbf{X})$, $\mu_0(\mathbf{X})$, and $\mu_S(\mathbf{Z})$: $\mu_1(\mathbf{X}) = \mathbf{X}\delta_1$, $\mu_0(\mathbf{X}) = \mathbf{X}\delta_0$, $\mu_S(\mathbf{Z}) = \mathbf{Z}\gamma$. In this case,

$$E(Y|\mathbf{X} = \mathbf{x}, P(\mathbf{Z}) = p) = \mathbf{x}\delta_0 + p\mathbf{x}[\delta_1 - \delta_0] + K(p), \tag{9}$$

where $K(p) = E(U_1 - U_0 | S = 1, P(\mathbf{Z}) = p)$ can be estimated nonparametrically. It is straightforward to estimate the levels and derivatives of $E(Y|\mathbf{X} = \mathbf{x}, P(\mathbf{Z}) = p)$ and their standard errors using the methods developed in Heckman et al. (1998).

An alternative to the semiparametric model just described is to invoke parametric assumptions on the joint distribution of (U_0, U_1, V) and derive the expression for the MTE (see Heckman, Justin L. Tobias, and Vytlacil 2001; Arild Aakvik, Heckman, and Vytlacil 2005). Below, we also present estimates based on the assumption that (U_0, U_1, V) is jointly normally distributed and independent of $(\mathbf{X}, \mathbf{Z})$. Following conventions in discrete choice analysis (McFadden 1974), we normalize the variance of V to 1. In this case:

$$\begin{aligned} \text{MTE}(\mathbf{x}, u_S) &= \mathbf{x}(\delta_1 - \delta_0) + E(U_1 - U_0 | U_S = u_S) \\ &= \mathbf{x}(\delta_1 - \delta_0) + E\big(U_1 - U_0 | V = \Phi^{-1}(u_S)\big) \\ &= \mathbf{x}(\delta_1 - \delta_0) - (\sigma_{1V} - \sigma_{0V})\Phi^{-1}(u_S), \end{aligned} \tag{10}$$

where $\sigma_{1V} = \text{Cov}(U_1, V)$, $\sigma_{0V} = \text{Cov}(U_0, V)$, and where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cumulative distribution function. The parameters $(\delta_0, \delta_1, \sigma_{1V}, \sigma_{0V})$ and their standard errors can be estimated by maximum likelihood with the resulting parameter estimates plugged into equation (10) to form an estimate of the MTE.

II. Estimates of the MTE and Comparison of Marginal Returns, Policy Relevant Returns, and IV Estimands

A. *Data*

This section reports estimates of the MTE using a sample of white males from the NLSY. The data are described in the online Appendix. Individuals are separated into two groups: $S = 0$ (high school dropouts and high school graduates) and $S = 1$ (individuals with some college, college graduates, and postgraduates). Below, we study the sensitivity of our estimates to alternative definitions of schooling groups, alternative specifications of the model, and alternative samples. Schooling is measured in 1991 (individuals are between 28 and 34 years of age in 1991).

The variables Y, S, $\mathbf{X}$, and $\mathbf{Z}\backslash\mathbf{X}$ (the instruments or identifying exclusion restrictions) are presented in Table 2, together with the main papers that previously used these instruments. The instruments $(\mathbf{Z}\backslash\mathbf{X})$ are (i) the presence of a four-year college in the county of residence at age 14 as a measure of distance to college, (ii) local wage in the county of residence at age 17, (iii) local unemployment in the state

TABLE 2—DEFINITIONS OF THE VARIABLES USED IN THE EMPIRICAL ANALYSIS

Variable	Definition
Y	Log wage in 1991 (average of all nonmissing wages between 1989 and 1993)
$S = 1$	If ever enrolled in college by 1991; zero otherwise
$\mathbf{X}$	AFQT,[a] mother's education, number of siblings, average log earnings 1979–2000 in county of residence at 17, average unemployment 1979–2000 in state of residence at 17, urban residence at 14, cohort dummies, years of experience in 1991, average local log earnings in 1991, local unemployment in 1991
$\mathbf{Z}\backslash\mathbf{X}$[b]	Presence of a college at age 14 (Card 1995; Stephen V. Cameron and Christopher Taber 2004), local earnings at 17 (Cameron and Heckman 1998; Cameron and Taber 2004), local unemployment at 17 (Cameron and Heckman 1998), local tuition in public four-year colleges at 17 (Thomas J. Kane and Cecilia E. Rouse 1995)

[a]We use a measure of this score corrected for the effect of schooling attained by the participant at the date of the test, since at the date the test was taken, in 1981, different individuals have different amounts of schooling and the effect of schooling on AFQT scores is important. We use a correction based on the method developed in Karsten T. Hansen, Heckman, and Kathleen J. Mullen (2004). We take the sample of white males, perform this correction, and then standardize the AFQT to have mean 0 and variance 1 within this sample. See Table A-2 in the online Appendix.
[b]The papers in parentheses are papers that previously used these instruments.

of residence at age 17, and (iv) average tuition in public four-year colleges in the county of residence at age 17.[12]

Distance to college was first used as an instrument for schooling by Card (1995) and was subsequently used by Thomas J. Kane and Cecilia Elena Rouse (1995), Jeffrey R. Kling (2001), Janet Currie and Enrico Moretti (2003), and Cameron and Taber (2004). Cameron and Taber (2004) and Carneiro and Heckman (2002) show that distance to college in the NLSY79 is correlated with a measure of ability (Armed Forces Qualification Test (AFQT)). In this paper, we include this measure of ability in the outcome equation.

Cameron and Heckman (1998, 2001) and the papers they cite emphasize the importance of controlling for local labor market characteristics (see also Cameron and Taber 2004). If local unemployment and local earnings at age 17 are correlated with the unobservables in the earnings equations in the adult years, our measures of local labor market conditions would not be valid instruments. To mitigate this concern, we have included measures of permanent local labor market conditions (which we define as the average earnings and unemployment between 1973 and 2000 for each location of residence at 17) both in the selection and outcome equations. Effectively, we use only the innovations in the local labor market variables as instruments. This is similar to the procedure used by Cameron and Taber (2004). Further, in the outcome equations we also include the average log earnings in the county of residence in 1991, and the average unemployment rate in the state of residence in 1991.[13]

Tuition is used to predict college attendance in Cameron and Heckman (1998, 2001) and Kane and Rouse (1995). We control for AFQT and maternal education

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

[13]As Cameron and Taber (2004) argue, the sign of the total impact of these variables on schooling choice is theoretically ambiguous. Local labor market conditions can influence schooling through two possible channels. On the one hand, better labor market conditions increase the opportunity costs of schooling and reduce educational attainment. On the other hand, better labor market conditions lead to an increase in the resources of credit constrained households and, therefore, promote educational attainment.

in all of our models. These variables are likely to be highly correlated with college quality. We use these variables to account for any correlation between our measure of tuition (which corresponds only to four-year public colleges) and college quality. In order to examine the sensitivity of our estimates to the choice of instruments, we estimate models with and without tuition.

Included among $\mathbf{X}$ and $\mathbf{Z}$ in the linear-in-parameter representations are linear and quadratic terms in AFQT, mother's education, number of siblings, permanent local earnings, and permanent local unemployment, as well as a dummy variable indicating urban residence at age 14 and cohort dummies. The four exclusion restrictions or instruments enter $\mu_S(\mathbf{Z})$ but not the outcome equations $\mu_1(\mathbf{X})$ or $\mu_0(\mathbf{X})$. They are interacted with AFQT, maternal education, and number of siblings.[14] In addition, three variables are included in the outcome equations but not the selection equation: years of experience in 1991 (and its square), earnings in the county of residence in 1991, and unemployment in the state of residence in 1991. Given that these variables are realized only in 1991 and that we condition on earlier values of these variables, it is natural to assume the residualized variables do not enter the individuals' information sets at the time individuals make their college decision, which for most of them takes place more than 12 years earlier.[15]

Table 3 presents estimates of the parameters of a logit model for schooling choice. We provide estimates of the average marginal derivatives of each variable in the choice model (the coefficients are in Table A-4 in the online Appendix). The instruments are (jointly) strong predictors of schooling, as are mother's education, AFQT, number of siblings, and permanent local earnings in the county of residence at age 17.

The tests for selection on returns (β correlated with S in equation (1), also known as selection on gains), developed and applied in Heckman, Schmierer, and Urzua (2010), test whether the MTE is constant in u_S (β uncorrelated with S), or whether it varies with u_S (β correlated with S). Given equations (5) and (6), a simple test of selection on gains consists of estimating equation (9), specifying $K(P)$ to be a polynomial in P (P is estimated using a logit), and testing whether the coefficients on the polynomial terms of order higher than one are jointly equal to zero.

The results of this test are presented in Table 4, panel A. In each column of the table we specify a polynomial in P of orders 2 through 5. For each specification (i.e., each model defined by the highest order of the polynomial in P) we present the p-values of joint tests that the coefficients on the terms of order higher than one in each polynomial are equal to zero (rejection would indicate that $K(P)$ is a nonlinear function of P). We account for the fact that we test multiple hypotheses simultaneously (Joseph P. Romano and Michael Wolf 2005) by constructing an adjusted critical value, following

[14] The total sample size is 1,747 (882 with $S = 0$ and 865 with $S = 1$). Table A-3 in the online Appendix documents that individuals who attend college have, on average, a 34 percent higher wage than those who do not attend college. They also have 3.25 fewer years of work experience since they spend more time in school (on average, they have four additional years of completed schooling). The scores on a measure of cognitive ability, the AFQT, are much higher for individuals who attend college than they are for those who do not. Those who attend only high school have less educated mothers, come from larger families, and are less likely to have grown up in an urban area and in an area with a public four-year college than individuals who attend college. They face higher tuition in the areas where they grew up. Local labor market variables are not much different between these two groups of individuals, independently of the time of measurement.

[15] Below we also present results where instruments are not interacted with $\mathbf{X}$, and where these experience and local labor market conditions in 1991 are not excluded from the selection equation. The results are very similar to the ones we obtain in our main specification, although with larger standard errors in the first case.

TABLE 3—COLLEGE DECISION MODEL: AVERAGE MARGINAL DERIVATIVES

	Average derivative
Controls (**X**)	
Corrected AFQT	0.2826
	(0.0114)***
Mother's years of schooling	0.0441
	(0.0059)***
Number of siblings	−0.0233
	(0.0068)***
Urban residence at 14	0.0340
	(0.0274)
"Permanent" local log earnings at 17	0.1820
	(0.0941)**
"Permanent" state unemployment rate at 17	0.0058
	(0.0165)
Instruments (**Z**)	
Presence of a college at 14	0.0529
	(0.0273)**
Local log earnings at 17	−0.2687
	(0.1008)***
Local unemployment rate at 17 (in percent)	0.0149
	(0.0100)
Tuition in 4 year public colleges at 17 (in $100)	−0.0027
	(0.0017)*
Test for joint significance of instruments: *p*-value	0.0001

Notes: This table reports the coefficients and average marginal derivatives from a logit regression of college attendance (a dummy variable that is equal to one if an individual has ever attended college and equal to zero if he has never attended college but has graduated from high school) on polynomials in the set of variables listed in the table and on cohort dummies (not reported). For each individual, we compute the effect of increasing each variable by one unit (keeping all the others constant) on the probability of enrolling in college, and then we average across all individuals. Bootstrapped standard errors (in parentheses) are presented below the corresponding parameters (250 replications). At the bottom of the table we present *p*-values for the test of joint significance of coefficients on the instruments. Corrected AFQT corresponds to a standardized measure of the Armed Forces Qualifying Test score corrected for the fact that different individuals have different amounts of schooling at the time they take the test (see Hansen, Heckman, and Mullen 2004; see also the online Appendix). This variable is standardized within the NLSY sample to have mean 0 and variance 1. Local earnings and unemployment rates are averages across all individuals in the population residing in a given area (county for log earnings, state for unemployment), independent of age, gender, race, and skill level. For each location, "Permanent" local earnings and unemployment take the average of each variable between 1979 and 2000 (and then assign it to the location of residence at 17).

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

the procedure in Heckman, Schmierer, and Urzua (2010). We also account for the fact that P is estimated by using standard corrections to asymptotic standard errors (in Table A-5 of the online Appendix, we present alternative results where this correction is done using the bootstrap). Using this test, we reject that the MTE is constant in u_S —agents in our sample select in part on ex post gains to schooling.

Table 4, panel B, presents another test of the hypothesis that β is uncorrelated with S, that is developed in Heckman, Schmierer, and Urzua (2010). It tests whether LATEs are equal over different values of U_S within the support of $P(\mathbf{Z})$. Under the null, all LATEs are equal. We reject equality, supporting the inference drawn from Table 4, panel A. We discuss this test in greater detail in Section IIC.

TABLE 4— TEST OF LINEARITY OF $E(Y|\mathbf{X}, P = p)$ USING POLYNOMIALS IN P; AND TEST OF EQUALITY OF LATES OVER DIFFERENT INTERVALS (H_0: $LATE^j(U_S^{L_j}, U_S^{H_j}) - LATE^{j+1}(U_S^{L_{j+1}}, U_S^{H_{j+1}}) = 0$)

Panel A. Test of linearity of $E(Y|\mathbf{X}, P = p)$ using models with different orders of polynomials in P[a]

Degree of polynomial for model	2	3	4	5
p-value of joint test of nonlinear terms	0.035	0.049	0.086	0.122
Adjusted critical value		0.057		
Outcome of test		Reject		

Panel B. Test of equality of LATEs (H_0: $LATE^j(U_S^{L_j}, U_S^{H_j}) - LATE^{j+1}(U_S^{L_{j+1}}, U_S^{H_{j+1}}) = 0$)[b]

Ranges of U_S for $LATE^j$	(0, 0.04)	(0.08, 0.12)	(0.16, 0.20)	(0.24, 0.28)	(0.32, 0.36)	(0.40, 0.44)
Ranges of U_S for $LATE^{j+1}$	(0.08, 0.12)	(0.16, 0.20)	(0.24, 0.28)	(0.32, 0.36)	(0.40, 0.44)	(0.48, 0.52)
Difference in LATEs	0.0689	0.0629	0.0577	0.0531	0.0492	0.0459
p-value	0.0240	0.0280	0.0280	0.0320	0.0320	0.0520
Ranges of U_S for $LATE^j$	(0.48, 0.52)	(0.56, 0.60)	(0.64, 0.68)	(0.72, 0.76)	(0.80, 0.84)	(0.88, 0.92)
Ranges of U_S for $LATE^{j+1}$	(0.56, 0.60)	(0.64, 0.68)	(0.72, 0.76)	(0.80, 0.84)	(0.88, 0.92)	(0.96, 1)
Difference in LATEs	0.0431	0.0408	0.0385	0.0364	0.0339	0.0311
p-value	0.0520	0.0760	0.0960	0.1320	0.1800	0.2400
Joint p-value			0.0520			

[a]The size of the test is controlled using a critical value constructed by the bootstrap method of Romano and Wolf (2005) using a 10 percent significance level.

[b]In order to compute the numbers in this table, we construct groups of values of U_S and average the MTE within these groups, by computing $E(Y_1 - Y_0|\mathbf{X} = \bar{\mathbf{x}}, U_S^{L_j} \leq U_S \leq U_S^{H_j})$, where $U_S^{L_j}$ and $U_S^{H_j}$ are the lowest and highest values of U_S defined for interval j. Then we compare the average MTE across adjacent groups and test whether the difference is equal to zero using the bootstrap with 250 replications.

B. *Estimating the MTE and Marginal Policy Effects Using a Normal Selection Model*

The traditional approach to estimating the model of equations (2) and (3) specifies a parametric joint distribution for (U_0, U_1, V), usually that (U_0, U_1, V) are jointly normally distributed and independent of $(\mathbf{X}, \mathbf{Z})$, and estimates the outcome and choice equations together using the method of maximum likelihood (e.g., Björklund and Moffitt 1987). Although our primary empirical results are from a semiparametric method, the results based on a parametric normal model are a useful benchmark against which to compare our estimates from less functional form dependent estimators. The parametric specification is less flexible than our semiparametric specification, but the resulting estimates are much more precise.

Maximum likelihood estimates of the parameters $\delta_0, \delta_1, \gamma$ and their standard errors are presented in Table A-6 in the online Appendix. The MTE for this model is equation (10). A simple test of whether the slope of the MTE is zero (i.e., individuals do not select into college based on variability in β) is a test of whether $\sigma_{1V} - \sigma_{0V} = 0$. We estimate that $\sigma_{1V} - \sigma_{0V} = -0.2388$ with a standard error of 0.0982, so we reject this hypothesis (p-value = 0.0150). This supports the conclusions in Table 4 that do not impose the joint normality assumption.

Figure 1 plots the estimated MTE with 90 percent confidence bands, evaluated at mean values of $\mathbf{X}$ (we obtain annualized estimates of the returns to college by dividing the MTE by four, which is the average difference in years of schooling for those with $S = 1$ and those with $S = 0$). The MTE is declining and precisely estimated. The people with the highest gross returns are more likely to go to college (have low

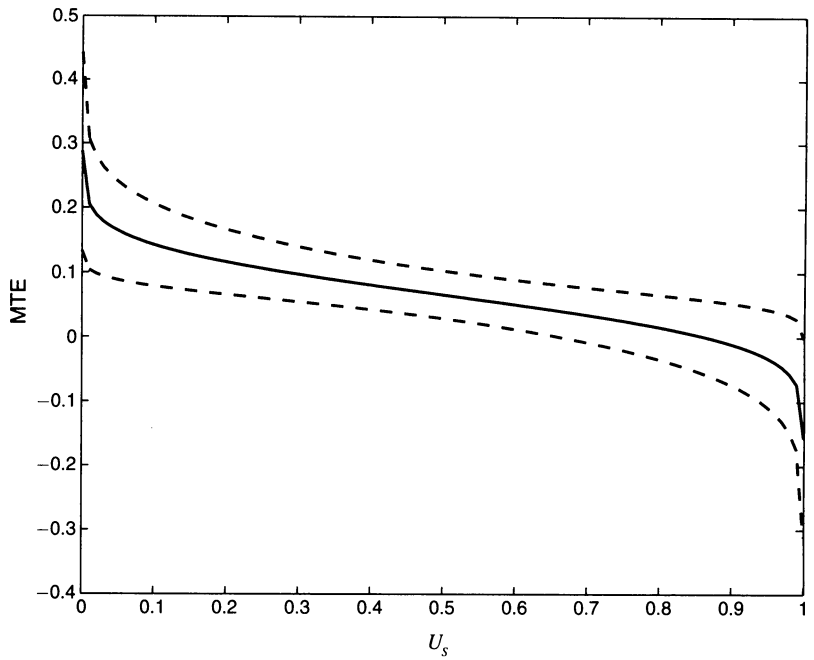

FIGURE 1. MTE ESTIMATED FROM A NORMAL SELECTION MODEL

Notes: To estimate the function plotted here, we estimate a parametric normal selection model by maximum likelihood. The figure is computed using the following formula:

$$\Delta^{\text{MTE}}(\mathbf{x}, u_S) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) - (\sigma_{1V} - \sigma_{0V})\Phi^{-1}(u_S),$$

where σ_{1V} and σ_{0V} are the covariances between the unobservables of the college and high school equation and the unobservable in the selection equation; and $\mathbf{X}$ includes experience, current average earnings in the county of residence, current average unemployment in the state of residence, AFQT, mother's education, number of siblings, urban residence at 14, permanent local earnings in the county of residence at 17, permanent unemployment in the state of residence at 17, and cohort dummies. We plot 90 percent confidence bands.

U_S). Individuals choose the schooling sector in which they have comparative advantage. The magnitude of the heterogeneity in returns on which agents select is substantial: returns can vary from −15.6 percent (for high U_S persons, who would lose from attending college) to 28.8 percent per year of college (for low U_S persons).[16] The magnitude of total heterogeneity is likely to be even higher since the MTE is the average gain at that quantile of desire to attend college. In general, there will be a distribution of returns centered at each value of the MTE. Furthermore, once we account for variation in $\mathbf{X}$ and its impact on returns through $\mathbf{X}(\delta_1 - \delta_0)$, we observe returns as low as −31.56 percent and as high as 51.02 percent.

Using the weights presented in online Appendix Table A-1B, we can construct the standard treatment parameters from the MTE. We present the results in the first column of Table 5 (standard errors are bootstrapped). These include marginal returns to the three different policies considered in Table 1 (MPRTE), which are all

[16] One unattractive feature of the normal model is that (for our estimates of σ_{1V} and σ_{0V}) $MTE(\mathbf{x}, 0) = +\infty$ and $MTE(\mathbf{x}, 1) = -\infty$. In order to get finite values at the extremes of the normal MTE, we restrict the support of U_S to be between 0.0001 and 0.9999.

TABLE 5—RETURNS TO A YEAR OF COLLEGE

Model		Normal	Semiparametric
$ATE = E(\beta)$		0.0670	Not identified
		(0.0378)	
$TT = E(\beta \mid S = 1)$		0.1433	Not identified
		(0.0346)	
$TUT = E(\beta \mid S = 0)$		−0.0066	Not identified
		(0.0707)	
MPRTE			
Policy perturbation	Metric		
$Z_\alpha^k = Z^k + \alpha$	$\lvert \mathbf{Z}\gamma - V \rvert < e$	0.0662	0.0802
		(0.0373)	(0.0424)
$P_\alpha = P + \alpha$	$\lvert P - U \rvert < e$	0.0637	0.0865
		(0.0379)	(0.0455)
$P_\alpha = (1 + \alpha)P$	$\lvert \frac{P}{U} - 1 \rvert < e$	0.0363	0.0148
		(0.0569)	(0.0589)
Linear IV (Using $P(\mathbf{Z})$ as the instrument)		0.0951	
		(0.0386)	
OLS		0.0836	
		(0.0068)	

Notes: This table presents estimates of various returns to college, for the semiparametric and the normal selection models: average treatment effect (ATE), treatment on the treated (TT), treatment on the untreated (TUT), and different versions of the marginal policy relevant treatment effect (MPRTE). The linear IV estimate uses P as the instrument. Standard errors are bootstrapped (250 replications). See online Appendix Table A-1 for the exact definitions of the weights. See Table 1 for the weights for MPRTE. For more discussion of MPRTE, see Carneiro, Heckman, and Vytlacil (2010).

below the return to the average student ($TT = E(\beta \mid S = 1)$), the average person ($ATE = E(\beta)$), and the IV estimate. But it is not clear if these estimates are reliable, given the strong normality assumption used to generate them. We next corroborate these estimates of marginal returns using a more robust semiparametric approach.

C. *Estimating the MTE and Marginal Policy Effects Using Local Instrumental Variables*

An alternative and more robust approach for estimating the MTE estimates $E(Y \mid \mathbf{X}, P(\mathbf{Z}) = p)$ semiparametrically and then computes its derivative with respect to p, as shown in the analysis of equations (5) and (6). If all we are willing to assume is that (U_0, U_1, V) is independent of $\mathbf{Z}$ given $\mathbf{X}$, then it is only possible to estimate the MTE over the support of P conditional on $\mathbf{X}$. Figure 2 plots $f(P \mid \mathbf{X})$, the density of P given $\mathbf{X}$ (P is estimated by a logit). Since $\mathbf{X}$ is multidimensional, we use an index of $\mathbf{X}$ ($\mathbf{X}[\delta_1 - \delta_0]$). It is striking how small the support of P is for each value of the $\mathbf{X}$ index. It is not possible to estimate MTE over the full unit interval, and as a consequence, it is not possible to estimate conventional treatment parameters such as the average treatment effect ($E(\beta)$) or the effect of treatment on the treated ($E(\beta \mid S = 1)$). It is still possible, however, to estimate MPRTE, since this parameter only puts positive weight over sections of the MTE that are identified within the support of $f(P \mid \mathbf{X})$.

Empirically, it is very difficult to apply the procedure described in Section I while conditioning on $\mathbf{X}$ nonparametrically. We first proceed by invoking the stronger assumption that $(\mathbf{X}, \mathbf{Z})$ is independent of (U_0, U_1, U_S). We relax it below. Under this

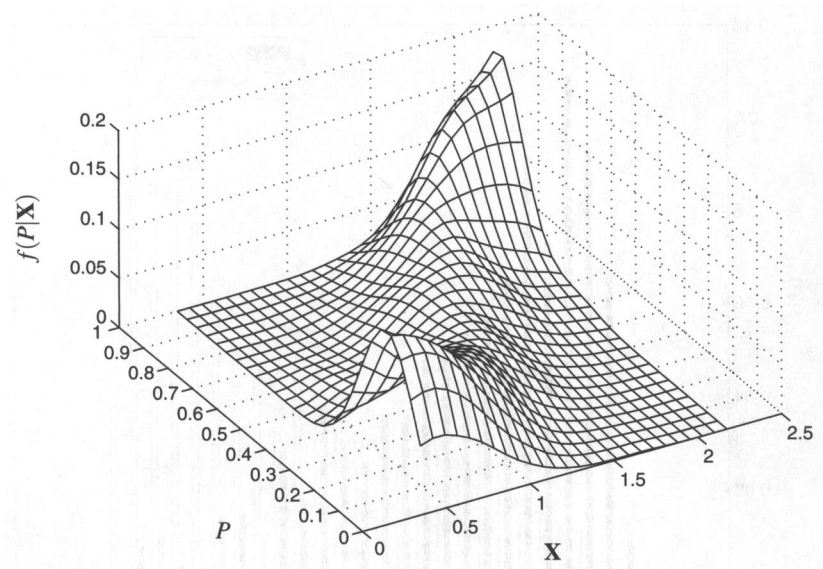

FIGURE 2. SUPPORT OF P CONDITIONAL ON $\mathbf{X}$

Notes: P is the estimated probability of going to college. It is estimated from a logit regression of college attendance on corrected AFQT, mother's education, number of siblings, urban residence at 14, permanent earnings in the county of residence at 17, permanent unemployment in the state of residence at 17, cohort dummies, a dummy variable indicating the presence of a college in the county of residence at age 14, average log earnings in the county of residence at age 17, and average state unemployment in the state of residence at age 17 (see Table 3). $\mathbf{X}$ corresponds to an index of variables in the outcome equation.

assumption, MTE is identified over the marginal support of $P(\mathbf{Z})$, and thus it is only necessary to investigate the marginal support of $P(\mathbf{Z})$ as opposed to the support of $P(\mathbf{Z})$ given $\mathbf{X}$. The support of the estimated $P(\mathbf{Z})$ is shown in Figure 3, and it is almost the full unit interval. We trim observations for which the estimated $P(\mathbf{Z})$ is below 0.0324 or above 0.9775, which are the minimum and maximum values of P for which we have common support.[17]

The parameters of equation (9) can be estimated by a partially linear regression of Y on $\mathbf{X}$ and $P(\mathbf{Z})$. We proceed in two steps. The first step is construction of the estimated $P(\mathbf{Z})$, and the second step is estimation of δ_1 and δ_0 using the estimated $P(\mathbf{Z})$. The first step is carried out using a logit regression of S on $\mathbf{Z}$. Our specification is quite flexible, and alternative functional form specifications for the choice model (e.g., probit) produce results similar to the ones reported here. In the second step we use the Peter M. Robinson (1988) method for estimating partially linear models as extended in Heckman, Hidehiko Ichimura, and Todd (1997).[18] Estimates of δ_1 and δ_0 are presented in online Appendix Table A-7.

[17] We define common support as the intersection of the support of $P(\mathbf{Z})$ given $D = 1$ and the support of $P(\mathbf{Z})$ given $D = 0$. Restricting our empirical estimates to the common support leads us to delete 67 observations, corresponding to 4.35 percent of the sample.

[18] We run kernel regressions of each of the regressors on P using a bandwidth of 0.05. We compute the residuals of each of these regressions and then run a linear regression of Y on these residuals. Our results are robust to choices of bandwidth between 0.01 and 0.2.

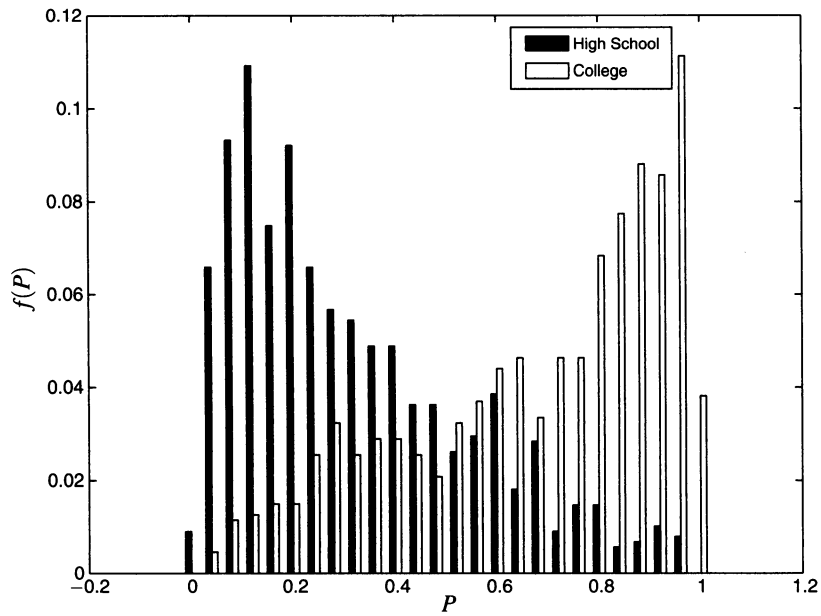

FIGURE 3. SUPPORT OF P FOR $S = 0$ AND $S = 1$

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

Next, consider estimation of $K(P(\mathbf{Z}))$. Equation (9) implies that

$$E(Y - \mathbf{X}\delta_0 - P(\mathbf{Z})\mathbf{X}[\delta_1 - \delta_0] \,|\, P(\mathbf{Z})) = K(P(\mathbf{Z})).$$

We thus use local polynomial regression of $Y - \mathbf{X}\hat{\delta}_0 - \hat{P}(\mathbf{Z})\mathbf{X}[\hat{\delta}_1 - \hat{\delta}_0]$ on $\hat{P}(\mathbf{Z})$ to estimate $K(P(\mathbf{Z}))$ and its partial derivative with respect to $P(\mathbf{Z})$. Local polynomial estimation not only provides a unified framework for estimating both a function and its derivative but also has a variety of desirable properties in comparison with other available nonparametric methods.[19]

Figure 4 plots the component of the MTE that depends on U_S, with 90 percent confidence bands computed from the bootstrap.[20] We fix the components of $\mathbf{X}$ at their

[19] Jianqing Fan and Irène Gijbels (1996) provide a detailed discussion of the properties of local polynomial estimators. In general, use of higher-order polynomials may reduce the bias but increase the variance by introducing more parameters. Fan and Gijbels suggest that the order π of the polynomial be equal to $\pi = \tau + 1$, where τ is the order of the derivative of the function of interest that we want to fit. That is, Fan and Gijbels recommend a local linear estimator for fitting a function and a local quadratic estimator for fitting a first-order derivative. Therefore, we use a local quadratic estimator of $\partial K(p)/\partial p$. We choose the bandwidth that minimizes the residual square criterion proposed in Fan and Gijbels, which gives us a bandwidth of 0.322. Our results are robust to the choice of bandwidths between 0.1 and 0.4.

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

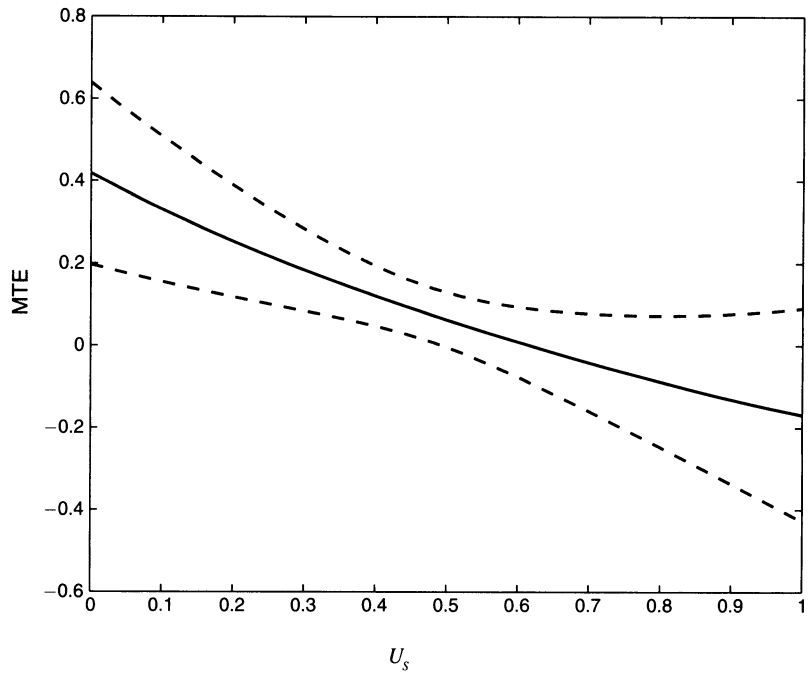

FIGURE 4. $E(Y_1 - Y_0|\mathbf{X}, U_S)$ WITH 90 PERCENT CONFIDENCE INTERVAL—LOCALLY QUADRATIC REGRESSION ESTIMATES

Notes: To estimate the function plotted here, we first use a partially linear regression of log wages on polynomials in **X**, interactions of polynomials in **X** and P, and $K(P)$, a locally quadratic function of P (where P is the predicted probability of attending college), with a bandwidth of 0.32; **X** includes experience, current average earnings in the county of residence, current average unemployment in the state of residence, AFQT, mother's education, number of siblings, urban residence at 14, permanent local earnings in the county of residence at 17, permanent unemployment in the state of residence at 17, and cohort dummies. The figure is generated by evaluating by the derivative of (9) at the average value of **X**. Ninety percent standard error bands are obtained using the bootstrap (250 replications).

mean values in the sample. As above, we annualize the MTE. Our estimates show that, in agreement with the normal model, $E(U_1 - U_0|U_S = u_S)$ is declining in u_S, i.e., students with high values of U_S have lower returns than those with low values of U_S.

Even though the semiparametric estimate of the MTE has larger standard errors than the estimate based on the normal model, we still reject the hypothesis that its slope is zero. We have already discussed the rejection of the hypothesis that MTE is constant in u_S, based on the test results reported in Table 4, panel A. But we can also directly test whether the semiparametric MTE is constant in u_S or not. We evaluate the MTE at 26 points, equally spaced between 0 and 1 (with intervals of 0.04). We construct pairs of nonoverlapping adjacent intervals (0–0.04, 0.08–0.12, 0.16–0.20, 0.24–0.28, ...), and we take the mean of the MTE for each pair. These are LATEs defined over different sections of the MTE. We compare adjacent LATEs. Table 4, panel B, reports the outcome of these comparisons. For example, the first column reports that

$$E(Y_1 - Y_0|\mathbf{X} = \bar{\mathbf{x}}, 0 \leq U_S \leq 0.04)$$

$$- E(Y_1 - Y_0|\mathbf{X} = \bar{\mathbf{x}}, 0.08 \leq U_S \leq 0.12) = 0.0689.$$

The p-value of the test of the hypothesis that this difference is equal to zero is reported below this number and is 0.0240, which implies that we reject this hypothesis at conventional levels of significance.[21] This table shows that the slope of the MTE is negative and statistically significant (at a 10 percent level of significance) for values of U_S up to 0.76 (p-values are reported in the bottom row of the table), and it remains negative but statistically insignificant after that. This is further evidence that individuals select into college based on heterogeneous returns in realized outcomes, although the rejection is only strong in the left tail of the estimated MTEs.[22] A joint test that the difference across all adjacent LATEs is different from zero has a p-value of 0.0520.

P only has support between 0.0324 and 0.9775, and thus it is not possible to estimate parameters that require full support, such as $E(\beta)$, $E(\beta|S=1)$ and $E(\beta|S=0)$. Estimation of such parameters is possible in the normal model only because of its parametric assumptions. Analysts often define such parameters as the objects of interest, even though they are very hard to identify,[23] and even though they are often not economically interesting.

In contrast, the MPRTE parameter not only answers interesting economic questions about the marginal gains of specific policies, but it is also identified without strong support assumptions, since it only requires estimating the MTE within the support of the data (see Carneiro, Heckman, and Vytlacil 2010). The second column of Table 5 presents estimates of three different versions of the MPRTE, where the policy considered is either a marginal change in tuition or a marginal change in P. The estimates are obtained in the following way. First we construct different weighted averages of the MTE by applying the weights presented in Table 1. Recall, however, that these weights are defined conditional on $\mathbf{X}$ and they define parameters conditional on $\mathbf{X}$. Therefore, after computing each of these parameters for each value of $\mathbf{X}=\mathbf{x}$, we need to integrate them against the appropriate distribution of $\mathbf{X}$.[24]

[21] Given that we bootstrap the nonparametric MTE, we implement the following test. For each pair of adjacent LATEs, the null is that the two LATEs are equal, and the alternative is that they are different. We construct 250 bootstrap replications of the MTE, evaluated at mean values of $\mathbf{X}$. The MTEs are evaluated on 26 equally spaced points on a grid (between 0 and 1). Let $LATE^j - LATE^{j+1}$ be the difference between two adjacent LATEs (j and $j+1$), and let $LATE_b^j$–$LATE_b^{j+1}$ be b^{th} bootstrap replication of this difference. Then we compute the following statistics: $T = |LATE^j - LATE^{j+1}|$ (the absolute value of the difference between two adjacent LATEs) and $T_b = |(LATE_b^j - LATE_b^{j+1}) - (LATE^j - LATE^{j+1})|$ (the recentered absolute value of the difference between two adjacent LATEs). The p-value of the test is the proportion of bootstrap replications for which $T_b > T$. A p-value for a joint test is also possible by constructing $C = \sum_{j=1}^{J-1} (LATE^j - LATE^{j+1})^2$ and $C_b = \sum_{j=1}^{J-1} [(LATE_b^j - LATE_b^{j+1}) - (LATE^j - LATE^{j+1})]^2$, where J is the number of LATEs taken (12 in our case). The p-value of the test is the proportion of bootstrap replications for which $C_b > C$.

[22] This and the other tests assume the validity of the instruments. If the instruments are not valid, these test results could be a consequence of the invalidity of the instruments.

[23] One alternative is to construct bounds, following Heckman and Vytlacil (2000). For example, applying the Heckman and Vytlacil bounds to our data, the estimated lower bound on ATE is 0.0411, and the estimated upper bound is 0.1043 (assuming a minimum value for the hourly wage of \$1 and a maximum value of \$100). We have also computed approximations to $E(\beta)$, $E(\beta|S=1)$, and $E(\beta|S=0)$, where we rescale the weights in Table 1 to integrate to one over the region [0.0324; 0.9775]. These estimates are respectively (bootstrapped standard errors in parenthesis): 0.0815 (0.0454), 0.2420 (0.0713), and −0.0662 (0.0819).

[24] Since $\mathbf{X}$ is a high dimensional vector, it is not computationally feasible to condition on it. Therefore, as an approximation, we condition instead on an index of $\mathbf{X}$: $\mathbf{X}(\delta_1 - \delta_0)$. A better approach would be to estimate the index using, for example, semiparametric least squares (SLS; see Ichimura 1993), exploiting the fact that $f(P|\mathbf{X}) = f(P|\mathbf{X}\theta) \Rightarrow E(P|\mathbf{X}) = E(P|\mathbf{X}\theta)$. Since this procedure is computationally very intensive, it is not feasible to

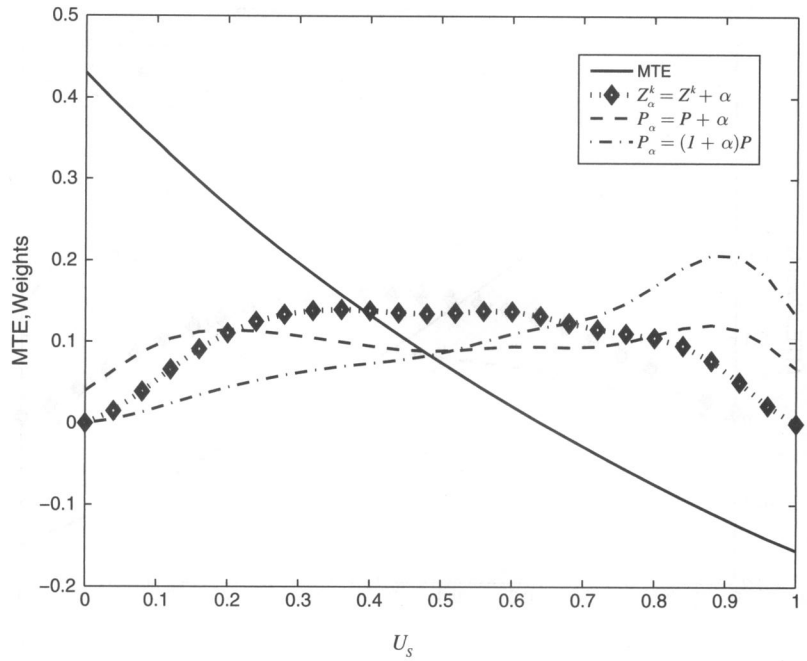

FIGURE 5. WEIGHTS FOR THREE DIFFERENT VERSIONS OF THE MPRTE

Note: The scale of the *y*-axis is the scale of the MTE, not the scale of the weights, which are scaled to fit the picture.

It is informative to visualize the weights underlying each parameter, which show us why some parameter values are higher or lower than others. Figure 5 graphs the weights on $E(Y_1 - Y_0 | \mathbf{X}, U_S = u_S)$ for the three MPRTE parameters with estimates reported in Table 5, all evaluated at the mean of $\mathbf{X}$. While the MPRTE weights for the first two policies ($Z^k_\alpha = Z^k + \alpha$ and $P_\alpha = P + \alpha$) weight mainly the middle section of the MTE, the third policy ($P_\alpha = (1 + \alpha)P$) overweights individuals with high levels of U_S because its effect on enrollment is larger for those with already high levels of P.[25]

The IV estimate, presented in the second to the last row of Table 5, is 0.0951. We use $P(\mathbf{Z})$ as the instrument, but it is possible to construct IV estimates for other combinations of instruments (see Table A-8 in the online Appendix). IV does not correspond to any of the MPRTE parameters that we consider, and it is particularly far from MPRTE in the case of the third policy. Figure A-1 in the online Appendix shows the sharp difference in the MTE weights for IV/LATE and the third of the MPRTE parameters, evaluated at mean $\mathbf{X}$.

Notice that both MPRTE and LATE correspond to some marginal effect (see Imbens 2010, for arguments for using LATE). However, LATE only estimates the policy effect of interest if the instrument variation corresponds exactly to the policy

compute standard errors from it. We show below that the point estimates we obtain when estimating this index by SLS are very similar to the ones we get with the index we use.

[25] Throughout the paper, we refer to the MPRTE of the policy $Z^k_\alpha = Z^k + \alpha$ as a shorthand for the limit as $\alpha \to 0$ of the PRTE parameters defined by $Z^k_\alpha = Z^k + \alpha$, and likewise with the MPRTE for $P_\alpha = P + \alpha$ and $P_\alpha = (1 + \alpha)P$.

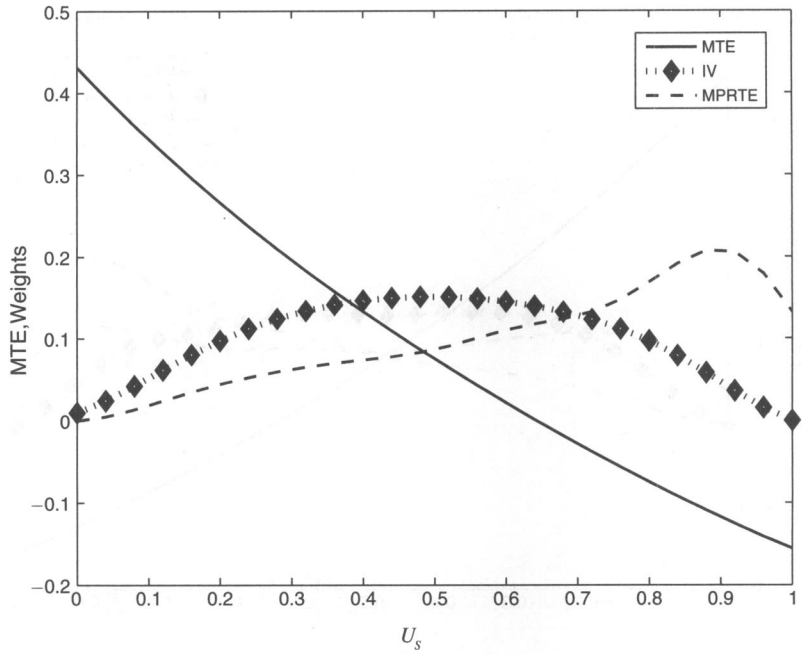

FIGURE 6. WEIGHTS FOR IV AND MPRTE

Note: The scale of the *y*-axis is the scale of the MTE, not the scale of the weights, which are scaled to fit the picture.

variation. For a specific policy of interest, LATE can be wildly off the mark. Figure 6 plots the IV weights (using P as the instrument) together with the MPRTE weights for the marginal policy defined by $P_\alpha = (1 + \alpha)P$. These two weights are dramatically different from each other.[26]

One benefit of our approach over the LATE approach is that it enables us to determine what portion of the MTE each instrument in $\mathbf{Z}\backslash\mathbf{X}$ identifies, i.e., it enables analysts to identify the quantiles of U_S that each instrument traces out. Thus, in our approach, the margin traced out by variation in each instrument is clearly identified. In the LATE approach that does not specify an explicit choice equation, the margin identified by variation in an instrument is not clearly defined.

Figure 7 shows the support of $P(\mathbf{Z})$ when we fix the variables in $\mathbf{X}$ at two different values and vary the instruments one at a time. This is the approach required to secure estimates if we condition on $\mathbf{X}$ and do not invoke independence between $\mathbf{X}$ and (U_0, U_1, V). This exercise informs us about what margin each instrument identifies under more general conditions. It also shows how far we expand the support of $P(\mathbf{Z})$ (and, therefore, the support of U_S over which we can estimate the MTE) by using multiple instruments simultaneously, as opposed to using them one at a time.

There are two curves in the picture corresponding to the MTE evaluated at different values of $\mathbf{X}$. The two lines correspond to the twenty-fifth (bottom) and seventy-fifth (top) percentiles of the distribution of $\mathbf{X}(\delta_1 - \delta_0)$. The curves have dashed and

[26] In fact, the IV estimate is a much better approximation to either of the other two MPRTE parameters, although the differences are still substantial (but not statistically significant). The approximation implicit in the estimates depends on the weights graphed in Figure A-1 in the online Appendix.

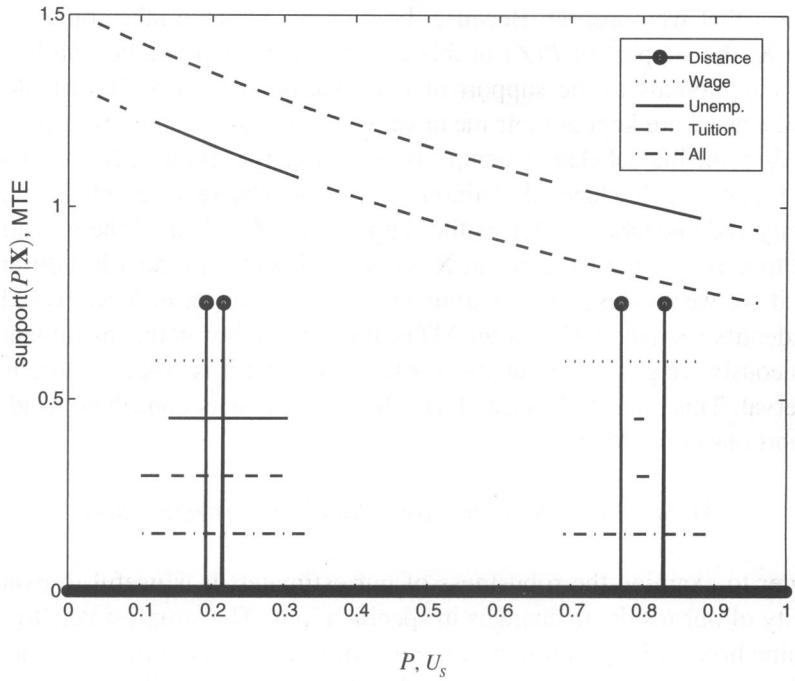

FIGURE 7. SUPPORT OF P FOR FIXED $\mathbf{X}$

Notes: This figures shows the support of P and the corresponding identified portion of the MTE when we fix the variables in $\mathbf{X}$ at a given value. We start by computing the average $\mathbf{X}$ for values of $\mathbf{X}(\delta_1 - \delta_0)$ at the twenty-fifth and seventy-fifth percentile of its distribution. The two curves in the figure represent the MTE evaluated at these two values of $\mathbf{X}(\delta_1 - \delta_0)$, with the solid component corresponding to the portion of the MTE that is identified if we do not assume independence between $\mathbf{X}$ and (U_0, U_1, V). In order to draw the line labeled "Distance," we not only fix $\mathbf{X}$ at the two values referred to above, but we also fix all the other instruments at the corresponding mean values. The line labeled "Wage" corresponds to the support of $P(\mathbf{Z})$ we obtain when all variables except local wage at 17 are kept at their mean values (conditional on a given percentile of $\mathbf{X}(\delta_1 - \delta_0)$), the line labeled "Unemp." is generated by varying only local unemployment at 17, and the line labeled "Tuition" is generated by varying only local tuition at 17. Finally, the line labeled "All" is the support of $P(\mathbf{Z})$ when all the instruments are allowed to vary and the variables in $\mathbf{X}$ are fixed.

solid segments. The solid segment represents the portion of the MTE we can identify at each value of $\mathbf{X}$ if we do not invoke independence between $\mathbf{X}$ and (U_0, U_1, V). To be precise, we find values of $\mathbf{X}$ for which $\mathbf{X}(\delta_1 - \delta_0)$ is at the twenty-fifth percentile of its distribution (we pick values of this index between the twenty-fourth and twenty-sixth percentile of its distribution and compute mean $\mathbf{X}$ in this interval). Then, we vary the instruments within this range, and we trace out the corresponding support of P for $\mathbf{X}$ fixed at this value. We then take our estimate of the MTE and select only the segment contained within the support of P for fixed $\mathbf{X}$. We do the same for values of $\mathbf{X}$ for which $\mathbf{X}(\delta_1 - \delta_0)$ is at the seventy-fifth percentile of its distribution. The dashed segments in each curve correspond to the additional portions of the MTE that we identify if, instead, we assume independence between $\mathbf{X}$ and (U_0, U_1, V).

It is informative to know not only what section of the MTE is identified at each value of $\mathbf{X}$, but also what section of the MTE is identified by varying each instrument one at a time. To generate the graph labeled "Distance," we fix not only $\mathbf{X}$ at the two values referred to above, but also fix all the other instruments at the corresponding mean values for each of the two percentiles of the distribution of

$\mathbf{X}(\delta_1 - \delta_0)$ that we consider. Because the distance variable takes only two values for each $\mathbf{X}$, the support of $P(\mathbf{Z})$ in this case has only two points. The line labeled "Wage" corresponds to the support of $P(\mathbf{Z})$ we obtain when all variables except local wage at 17 are kept at their mean values (conditional on a given percentile of $\mathbf{X}(\delta_1 - \delta_0)$), the line labeled "Unemp." is generated by varying only local unemployment at 17, and the line labeled "Tuition" is generated by varying only local tuition at 17. Finally, the line labeled "All" is the support of $P(\mathbf{Z})$ when all the instruments are allowed to vary and the variables in $\mathbf{X}$ are fixed. Each instrument has different support, so if we were to use each instrument in isolation at mean $\mathbf{X}$ we would only be able to identify a small section of the MTE. When we allow all the instruments to vary simultaneously we get larger support for the MTE, but it is still not close to the full unit interval. This analysis makes clear which instruments contribute to identifying which portions of the MTE.

D. *Sensitivity to Alternative Models and Specifications*

In order to examine the robustness of our estimates, it is useful to examine the sensitivity of our results to changes in specifications. The simplest way to do this is to examine how various summary measures of returns vary across specifications. It would be natural to present estimates of such parameters as the average treatment effect (ATE) ($= E(\beta)$), the effect of treatment on the treated (TT) $(= E(\beta | S = 1))$, and the effect of treatment on the untreated (TUT) $(= E(\beta | S = 0))$, which put most of their weight in very different sections of the MTE, and therefore should be sensitive to changes in the estimated MTE. Unfortunately, these parameters cannot be identified in our data, because the support of P is not quite the unit interval, even invoking assumptions about the independence of $\mathbf{X}$ and the unobservables. The support of P is close to the unit interval, however. We report estimates of what ATE, TT, and TUT would be if we restricted the weights to integrate to one in the support of the MTE ([0.0324;0.9775]). We call these parameters $\widetilde{ATE}$, $\widetilde{TT}$, and $\widetilde{TUT}$ (to distinguish them from ATE, TT, and TUT, which are not identified). In addition to these, we also consider (for parsimony, only) the first of the three definitions of MPRTE $(Z_\alpha^k = Z^k + \alpha)$, and the equivalent AMTE $(|\mathbf{Z}\gamma - V| < e)$, shown in Table 5.

Results are reported in Tables 6, panel A (where we focus mainly on choice of sample and specification of the outcome equations), and Table 6, panel B (where we focus mainly on the specification of the choice equation). Column 1 of Table 6, panel A, presents estimates of the baseline specification, in addition to the p-values for two simple tests of selection on returns: a test of the null that MPRTE $= \widetilde{TT}$ (does the marginal student attracted into schooling by the policy change have the same return as the average student in college?); and, for a test of the null, that there is no selection on returns in the normal model $(H_0 : \sigma_{1V} - \sigma_{0V} = 0)$.

Our baseline model has the important limitation that we restrict the schooling variable to take only two values. It is possible to extend this methodology to multiple levels of schooling (e.g., Heckman, Urzua, and Vytlacil 2006, 2008; Heckman and Vytlacil 2007b) but this requires multiple instrumental variables, one for each schooling transition. Even though the specification we use is common in the literature (e.g., Willis and Rosen 1979; Taber 2001; Carneiro and Heckman 2002; Moffitt 2008), it is still worthwhile to examine the robustness of our results to

TABLE 6—RETURNS TO A YEAR OF COLLEGE: SENSITIVITY TO DIFFERENT SAMPLES AND SPECIFICATION OF THE OUTCOME EQUATIONS; AND RETURNS TO A YEAR OF COLLEGE: SENSITIVITY TO THE SPECIFICATION OF THE CHOICE EQUATION

Panel A. Returns to a year of college: Sensitivity to different samples and specification of the outcome equations

	Baseline	No dropouts	Dummies for dropout and some college
$\widetilde{ATE}$	0.0815	0.1246	0.0995
	(0.0454)	(0.0555)	(0.0449)
$\widetilde{TT}$	0.2420	0.2605	0.2500
	(0.0713)	(0.0913)	(0.0712)
$\widetilde{TUT}$	0.0135	0.0274	−0.0388
	(0.0702)	(0.0879)	(0.0815)
$MPRTE\ (\lvert \mathbf{Z}\gamma - V \rvert < e)$	0.0802	0.1104	0.0988
	(0.0424)	(0.0514)	(0.0425)
p-values for:			
$\widetilde{TT} = AMTE$	0.0200	0.0565	0.0403
Normal—Selection test	0.0150	0.0018	0.0019

Panel B. Returns to a year of college: Sensitivity to the specification of the choice equation

	All $\mathbf{X}$ in $\mathbf{Z}$	No interactions with $\mathbf{Z}$	Cameron and Taber (2004)	No tuition	Use SLS for index of $\mathbf{X}$
$\widetilde{ATE}$	0.1409	0.1208	0.0851	0.0626	0.0871
	(0.0448)	(0.0703)	(0.0547)	(0.0560)	
$\widetilde{TT}$	0.2233	0.2125	0.2409	0.2056	0.2154
	(0.0713)	(0.0952)	(0.0895)	(0.0822)	
$\widetilde{TUT}$	0.0135	0.0350	−0.0570	−0.0682	−0.0337
	(0.0702)	(0.1003)	(0.0864)	(0.0924)	
$MPRTE\ (\lvert \mathbf{Z}\gamma - V \rvert < e)$	0.0802	0.1156	0.0821	0.0591	0.0799
	(0.0424)	(0.0666)	(0.0518)	(0.0528)	
p-values for:					
$\widetilde{TT} = AMTE$	0.0200	0.1694	0.0403	0.0605	
Normal—Selection test	0.0150	0.1080	0.0560	0.0120	

Notes: This table presents estimates of various returns to college for the semiparametric model estimated on several samples: average treatment effect (ATE), treatment on the treated (TT), treatment on the untreated (TUT), and the marginal policy relevant treatment effect. The ATE, TT, and TUT estimates are computed such that the weights integrate to one in the interval [0.0324; 0.9775]. The table also shows two tests of selection: the first tests whether TT = MPRTE in the semi-parametric model; the second tests whether $\text{COV}(U_1; V) = \text{COV}(U_0; V)$ in the normal model. Standard errors are bootstrapped (250 replications). The last column of panel B shows estimates of the main parameters when semiparametric least squares (SLS) is used to construct the index of $\mathbf{X}$ on which we condition to estimate $f(P|\mathbf{X})$ (estimates are presented without standard errors).

simple changes in it. To start with, note that other papers in the literature have either excluded high school dropouts from the sample (Willis and Rosen 1979; Moffitt 2008) or have imposed a simple model for the dropout decision (Taber 2001). In column 2 of Table 6, panel A, we present estimates for a model where high school dropouts are excluded from the sample. The main difference, relative to the model with estimates reported in column 1, is an increase in the standard errors and a small increase in the point estimates.

Column 3 presents estimates from a model where the sample is the same as in the baseline specification, but we include additional controls in the wage equations. The college equation includes a dummy for whether a person has some college but not a college degree (Willis and Rosen 1979), and the high school equation includes a dummy for whether a person has dropped out of high school (these dummies are

assumed to be exogenous). The estimates of the return to schooling are similar to the ones in the baseline model, although slightly higher. The main patterns remain the same, however.

Table 6, panel B, considers changes in the specification of the decision equation. We start by noting that in our baseline specification there are three variables that are included in the wage equation but not in the college decision equations: years of experience, local earnings in the county of residence in 1991, and local unemployment in the county of residence in 1991. The assumption is that (conditional on all other controls) the agent does not have information about the effect of these variables on earnings at the time of the college decision that is typically made 8 to 15 years before, and therefore they would have a zero coefficient in the selection equation. Alternatively, we could include these variables in the selection equation. The results from this exercise are shown in column 1 of panel B, Table 6. The magnitude and main pattern of the estimates resembles that of the baseline model.

Recall from Section IIA that the instruments (presence of college in area of residence, local wage and unemployment rate, local tuition) enter the college decision equation interacted with some of the control variables (AFQT, maternal education, number of siblings). It is natural to use this specification because the effect of each of these instruments is likely to vary across different families. Furthermore, such a specification is helpful in achieving lower standard errors. One may worry, however, that in such a specification what drives independent variation in college attendance is the set of controls (entering in a nonlinear way) as opposed to the instruments, even though we safeguard against this by having a very flexible specification of the wage equations. In column 2 of Table 6, panel B, we estimate a specification where none of the instruments is interacted with any other variable. The results are very similar to our baseline results, but with larger standard errors.

Columns 3 and 4 of Table 6, panel B, explore the sensitivity of our results to the choice of instruments. In column 3, we use only the instruments in Cameron and Taber (2004) (the presence of a college and local wage in the area of residence in adolescence), and in column 4 we exclude only tuition from the set of instruments in case it is correlated with college quality, even after controlling for AFQT and maternal education. The overall conclusion is that our main results are insensitive to the inclusion of tuition and of the local unemployment rate in the set of instruments, over and above the instruments used in Cameron and Taber (2004). The null hypothesis of no selection on returns is rejected in all of these samples, whether we use the normality test or the test that $\widetilde{TT} = \text{MPRTE}$.[27]

Finally, the last column of Table 6, panel B, shows the point estimates of the main parameters that result from using an alternative method to estimate the index of $\mathbf{X}$ on which to condition when estimating the density of P conditional on $\mathbf{X}$, $f(P \mid \mathbf{X})$. In particular, we estimate $E(P \mid \mathbf{X})$ assuming a linear index in $\mathbf{X}$ and using semiparametric least squares (SLS; see Ichimura 1993). We then use this index instead of the full $\mathbf{X}$ vector (and instead of an arbitrary index) to estimate $f(P \mid \mathbf{X})$, which we use to construct the parameter weights of Tables 1 and the online Appendix Table A-1B. The estimates we obtain using this method are very

[27] One exception is the model with estimates reported in column 2 of Table 6, panel B.

similar to the point estimates of our baseline specification. They are computationally very demanding, however, and it is not feasible to produce standard errors for our parameters using this method.

III. Summary and Conclusions

This paper estimates marginal returns to college when returns differ among individuals, and persons select into economic activities based in part on their idiosyncratic returns. Consistent with Heckman, Schmierer, and Urzua (2010), we find evidence that people select into schooling on the basis of realized returns to schooling. In general, marginal and average returns to college are not the same.[28] Conventional average return parameters and IV estimators are weighted averages of the marginal treatment effect (MTE). Unless the instruments are the policy changes of interest, these parameters do not answer well-posed policy questions. We show how to use the estimates from a local version of IV (LIV) to determine the marginal policy relevant treatment effect (MPRTE).

We show how to identify and estimate the MTE using a robust semiparametric selection model. Focusing on a policy-relevant question, we construct estimators based on the MTE to answer it, rather than hoping that a particular instrumental variable estimator happens to answer a question of economic interest.

It has been written "better LATE than nothing" (Imbens 2010). We say that in addressing policy questions, "better MPRTE than LATE." We use the MTE to identify the margins that different instruments identify.

We test for the importance of self-selection on ex post gains in the labor market. The data suggest that self-selection on gains (the correlation of β with S) is an empirically important phenomenon governing schooling choices, consistent with the analysis of Willis and Rosen (1979). Individuals sort into schooling on the basis of gains that are observed by the economist as well as unobserved (by the economist) variables. Some marginal expansions of schooling produce marginal gains that are well below average returns. For other policies associated with other marginal expansions, the marginal gains are substantial.

Comparing estimates from our semiparametric IV approach with estimates from a normal selection model of the sort used by Willis and Rosen (1979) and Björklund and Moffitt (1987), the semiparametric estimates produce an MTE with the same general shape consistent with diminishing returns to education for additional students, but with larger standard errors.

REFERENCES

Aakvik, Arild, James J. Heckman, and Edward J. Vytlacil. 2005. "Estimating Treatment Effects for Discrete Outcomes When Responses to Treatment Vary: An Application to Norwegian Vocational Rehabilitation Programs." *Journal of Econometrics*, 125(1–2): 15–51.

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

[28] This has implications for propensity score matching, which imposes the condition that marginal and average returns are equal. See Heckman and Vytlacil (2007b).

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

Cameron, Stephen V., and James J. Heckman. 2001. "The Dynamics of Educational Attainment for Black, Hispanic, and White Males." *Journal of Political Economy*, 109(3): 455–99.

Cameron, Stephen V., and Christopher Taber. 2004. "Estimation of Educational Borrowing Constraints Using Returns to Schooling." *Journal of Political Economy*, 112(1): 132–82.

Card, David. 1995. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling." In *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, ed. Louis N. Christofides, E. Kenneth Grant and Robert Swidinsky, 201–22. Toronto: University of Toronto Press.

Card, David. 1999. "The Causal Effect of Education on Earnings." In *Handbook of Labor Economics Volume 3*, Part 1, Handbooks in Economics, ed. Orley Ashenfelter and David Card, 1801–63. New York: Elsevier Science.

Card, David. 2001. "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems." *Econometrica*, 69(5): 1127–60.

Carneiro, Pedro, and James J. Heckman. 2002. "The Evidence on Credit Constraints in Post-Secondary Schooling." *Economic Journal*, 112(482): 705–34.

Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil. 2010. "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin." *Econometrica*, 78(1): 377–94.

Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil. 2011. "Estimating Marginal and Average Returns to Education: Dataset." *American Economic Review.* http://www.aeaweb.org/articles.php?doi=10.1257/aer.101.6.2754.

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

Fan, Jianqing, and Irène Gijbels. 1996. *Local Polynomial Modelling and Its Applications.* Vol. 66, Monographs on Statistics and Applied Probability. New York: Chapman and Hall.

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

Heckman, James J. 2010. "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy." *Journal of Economic Literature*, 48(2): 356–98.

Heckman, James J., and Daniel Schmierer. 2010. "Tests of Hypotheses Arising in the Correlated Random Coefficient Model." *Economic Modelling*, 27(6): 1355–67.

Heckman, James J., and Edward J. Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proceedings of the National Academy of Sciences,* 96(8): 4730–34.

Heckman, James J., and Edward J. Vytlacil. 2000. "The Relationship between Treatment Parameters within a Latent Variable Framework." *Economics Letters*, 66(1): 33–39.

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

Heckman, James J., and Edward J. Vytlacil. 2001b. "Policy-Relevant Treatment Effects." *American Economic Review*, 91(2): 107–11.

Heckman, James J., and Edward J. Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, 73(3): 669–738.

Heckman, James J., and Edward J. Vytlacil. 2007a. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In *Handbook of Econometrics,* Vol. 6B. Handbooks in Economics 2, ed. James J. Heckman and Edward E. Leamer, 4779–4874. New York: North-Holland.

Heckman, James J., and Edward J. Vytlacil. 2007b. "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments." In *Handbook of Econometrics,* Vol. 6B. Handbooks in Economics 2, ed. James J. Heckman and Edward E. Leamer, 4875–5144. New York: North-Holland.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "How Details Makes a Difference: Semiparametric Estimation of the Partially Linear Regression Model." Unpublished.

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

Heckman, James J., Justin L. Tobias, and Edward J. Vytlacil. 2001. "Four Parameters of Interest in the Evaluation of Social Programs." *Southern Economic Journal*, 68(2): 210–23.

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

Heckman, James J., Sergio Urzua, and Edward J. Vytlacil. 2008. "Instrumental Variables in Models with Multiple Outcomes: The General Unordered Case." *Annales d'Economie et de Statistique,* 2008(91–92): 151–74.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica*, 66(5): 1017–98.

Ichimura, Hidehiko. 1993. "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models." *Journal of Econometrics*, 58(1–2): 71–120.

Imbens, Guido W. 2010. "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature*, 48(2): 399–423.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–75.

Kane, Thomas J., and Cecilia Elena Rouse. 1995. "Labor-Market Returns to Two- and Four-Year College." *American Economic Review*, 85(3): 600–14.

Kling, Jeffrey R. 2001. "Interpreting Instrumental Variables Estimates of the Returns to Schooling." *Journal of Business and Economic Statistics*, 19(3): 358–64.

McFadden, Daniel. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers of Econometrics,* ed. Paul Zarembka, 105–42. New York: Academic Press.

Moffitt, Robert A. 2008. "Estimating Marginal Treatment Effects in Heterogeneous Populations." *Annales d'Economie et de Statistique*, 2008(91–92): 239–61.

Murray, Charles. 2008a. "Are Too Many People Going to College?" *The American,* Monday, Sept. 8 (http://www.american.com/archive/2008/september-october-magazine/are-too-many-people-going-to-college).

Murray, Charles. 2008b. *Real Education: Four Simple Truths for Bringing America's Schools Back to Reality.* New York: Three Rivers Press.

Quandt, Richard E. 1958. "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes." *Journal of the American Statistical Association,* 53(284): 873–80.

Quandt, Richard E. 1972. "A New Approach to Estimating Switching Regressions." *Journal of the American Statistical Association,* 67(338): 306–10.

Robinson, Peter M. 1988. "Root-N-Consistent Semiparametric Regression." *Econometrica*, 56(4): 931–54.

Romano, Joseph P., and Michael Wolf. 2005. "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing." *Journal of the American Statistical Association,* 100(469): 94–108.

Roy, A. D. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers New Series*, 3(2): 135–46.

Stock, James H. 1989. "Nonparametric Policy Analysis." *Journal of the American Statistical Association,* 84(406): 567–75.

Taber, Christopher R. 2001. "The Rising College Premium in the Eighties: Return to College or Return to Unobserved Ability?" *Review of Economic Studies*, 68(3): 665–91.

Vytlacil, Edward J. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica*, 70(1): 331–41.

Willis, Robert J., and Sherwin Rosen. 1979. "Education and Self-Selection." *Journal of Political Economy*, 87(5): S7–36.