The Causal Effects of Education on Earnings and Health*

James J. Heckman
University of Chicago
and the American Bar Foundation

John Eric Humphries
University of Chicago

Gregory Veramendi
Arizona State University

March 12, 2015

*James Heckman: Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637; phone: 773-702-0634; fax: 773-702-8490; email: jjh@uchicago.edu. John Eric Humphries: Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637; phone: 773-980-6575; email: johneric@uchicago.edu. Gregory Veramendi: Arizona State University, 501 East Orange Street, CPCOM 412A, Tempe, AZ 85287-9801; phone: 480-965-0894; email: gregory.veramendi@asu.edu. This paper was presented at the Becker Friedman Institute conference in honor of Gary Becker, October 30, 2014. It was also presented as the Sandmo Lecture at the Norwegian School of Economics, January 13, 2015. We thank Chris Taber for insightful comments. We also thank Ariel Pakes and other participants at a Harvard Labor Economics Workshop April, 2014, for helpful comments on a previous draft. We thank Eleanor Dillon and Mathew Wiswall for comments on this draft received at a seminar at Arizona State University, February, 2015. This research was supported in part by: the American Bar Foundation; the Pritzker Children’s Initiative; the Buffett Early Childhood Fund; NIH grants NICHD R37HD065072 and NICHD R01HD54702; an anonymous funder; Successful Pathways from School to Work, an initiative of the University of Chicago’s Committee on Education funded by the Hymen Milgrom Supporting Organization; and the Human Capital and Economic Opportunity Global Working Group, an initiative of the Center for the Economics of Human Development, affiliated with the Becker Friedman Institute for Research in Economics, and funded by the Institute for New Economic Thinking. Humphries acknowledges the support of a National Science Foundation Graduate Research Fellowship. The views expressed in this paper are solely those of the authors and do not necessarily represent those of the funders or the official views of the National Institutes of Health. The Web Appendix for this paper is https://heckman.uchicago.edu/eff-ed-earn-health.
Abstract

This paper estimates a robust dynamic model of the causal effects of different levels of schooling on earnings and health. Our framework synthesizes approaches used in the dynamic discrete choice literature with approaches used in the reduced form treatment effect literature. We estimate economically interpretable and policy relevant treatment effects and interpret the economic content of instrumental variables estimators. Cognitive and noncognitive endowments play important roles in explaining observed differences in earnings and health across education levels. Nonetheless, after controlling for them, there are substantial causal effects of education at all stages of schooling. Continuation values associated with dynamic sequential schooling choices are empirically important components of estimated causal effects. There is considerable heterogeneity in the effects of schooling on outcomes at different schooling levels and in these effects across persons. We find strong sorting on gains consistent with comparative advantage, but only at higher levels of schooling. This result is not imposed in our estimation procedure. We find that the estimated causal effects of education vary with the level of cognitive and noncognitive endowments. Estimates of causal effects using standard instrumental variables are often quite different from the economically interpretable and policy relevant treatment effects derived from our model.

Keywords: education, earnings, health, rates of return, causal effects of education

JEL codes: C32, C38, I12, I14, I21

James Heckman  
Department of Economics  
University of Chicago  
1126 East 59th Street  
Chicago, IL 60637  
Phone: 773-702-0634  
Email: jjh@uchicago.edu

John Eric Humphries  
Department of Economics  
University of Chicago  
1126 East 59th Street  
Chicago, IL 60637  
Phone: 773-980-6575  
Email: johneric@uchicago.edu

Gregory Veramendi  
Department of Economics  
Arizona State University  
501 East Orange Street, CPCOM  
412A  
Tempe, AZ 85287-9801  
Phone: 480-965-0894  
Email: gregory.veramendi@asu.edu
1 Introduction

In his pioneering research on human capital, Gary Becker (1962; 1964) identified the rate of return to education as a central policy parameter. He launched an active industry on estimating rates of return.\footnote{For surveys of this literature, see, e.g., Card (1999, 2001); Heckman, Lochner, and Todd (2006); Oreopoulos and Salvanes (2011); McMahon (2009), and Oreopoulos and Petronijevic (2013).}

Becker focused on internal rates of return that equate the discounted values of the earnings streams associated with different levels of education. He noted that the full return to schooling also includes nonmarket benefits and nonpecuniary costs. Individuals should continue schooling as long as their marginal internal rate of return exceeds their marginal opportunity cost of funds. If the social return exceeds the social opportunity cost of funds, there is aggregate under-investment in education.

Formidable challenges are faced in estimating internal rates of return: (a) lifetime earnings profiles are required; (b) observed earnings profiles are subject to selection bias; and (c) quantifying nonmarket benefits and nonpecuniary costs is a difficult task. In a neglected paper, Becker and Chiswick (1966) addressed challenge (a) and developed a tractable framework for measuring rates of return to schooling that utilizes cross-section synthetic cohort data on earnings to approximate life cycle earnings data. Mincer (1974) improved on this model by adding work experience. The “Mincer Equation” has become the workhorse of the empirical literature on estimating rates of return:

\[
\ln Y(S_i, X_i) = \alpha_i + \rho_i \frac{S_i}{\text{years of schooling}} + \phi \left( \frac{X_i}{\text{work experience}} \right)
\]

where \( Y(S_i, X_i) \) is the earnings of individual \( i \) with \( S_i \) years of schooling and work experience \( X_i \), \( \alpha_i \) is an “ability to earn” parameter that is common across all schooling levels and \( \rho_i \) is the “rate of return” to schooling for person \( i \) that is assumed to vary among individuals.

Equation (1) and its variants have become the standard framework for estimating rates of return.
of return.\textsuperscript{2} While $\rho_i$ is not, in general, an internal rate of return for individual $i$, it is the causal effect of an increase in one year of schooling on log earnings from any base state of schooling holding $\alpha_i$ and $X_i$ fixed.\textsuperscript{3}

$\rho_i$ ignores the continuation values arising from the dynamic sequential nature of the schooling decision where information is updated and schooling at one stage opens up options for schooling at later stages.\textsuperscript{4} The distribution of $\rho_i$ and its correlation with $S_i$ have become central targets of empirical studies of the causal effects of education. A positive correlation is consistent with a meritocratic society. People who benefit from schooling get more of it. A negative correlation indicates problems with access to schooling.

Two approaches have been developed to address challenge (b) and estimate rates of return in the general case where $\rho_i$ is correlated with $S_i$ (sorting bias) and $S_i$ is correlated with $\alpha_i$ (ability bias). They are: (I) structural models that jointly analyze outcomes and schooling choices,\textsuperscript{5} and (II) treatment effect models that use instrumental variables methods (including randomization and regression discontinuity methods) as well as matching on observed variables to identify “causal parameters.”\textsuperscript{6}

The structural approach explicitly models agent decision rules. It uses a variety of sources of identification including exclusion restrictions (instrumental variables), conditional independence assumptions on unobservables and functional form assumptions (see, e.g., Blevins, 2014). The final two sources of identification are often controversial. The structural approach identifies the margins of choice identified by instruments and can evaluate the impacts of different policies never previously implemented.

The instrumental variable approach is agnostic about agent decision rules and relies solely

\textsuperscript{2}See, e.g., Cutler and Lleras-Muney (2010) who apply model (1) to estimate the causal effect of education on health.

\textsuperscript{3}The stringent conditions under which $\rho_i$ is an internal rate of return and evidence that they are not satisfied in many commonly used samples are presented in Heckman, Lochner, and Todd (2006).

\textsuperscript{4}Weisbrod (1962) first raised this point. There is later work by e.g., Comay, Melnik, and Pollatschek (1973); Altonji (1993); Cameron and Heckman (1993), and Eisenhauer, Heckman, and Mosso (2015).

\textsuperscript{5}See e.g., Willis and Rosen (1979); Keane and Wolpin (1997); Eisenhauer, Heckman, and Mosso (2015).

on exclusion restrictions to identify its estimands. This approach is often more transparent in securing identification than is the structural approach. However, the economic interpretation of its estimands is often obscure. In a model with multiple levels of schooling, LATE often does not identify the separate margins of choice traced out by instruments or the subpopulations affected by them. Its estimands are irrelevant for addressing policy questions except when the variation induced by the instruments corresponds closely to variation induced by the policies of interest.\textsuperscript{7}

This paper develops and applies a methodology that offers a middle ground between the reduced form treatment approach and the fully structural dynamic discrete choice approach. Like the structural literature, we estimate causal effects at clearly identified margins of choice. Our methodology identifies which agents are affected by instruments as well as which agents would be affected by alternative policies never previously implemented. Like the treatment effect literature, we are agnostic about the precise rules used by agents to make decisions. Unlike that literature, we recognize the possibility that people somehow make decisions and account for the consequences of their choices. We approximate agent decision rules and do not impose the cross-equation restrictions that are the hallmark of the structural approach.\textsuperscript{8}

Using a generalized Roy framework, we estimate a multistage sequential model of educational choices and their consequences.\textsuperscript{9} An important feature of our model is that educational choices at one stage open up educational options at later stages. Each educational decision is approximated using a reduced form discrete choice model. The anticipated consequences of future choices and their costs are implicitly valued by individuals when deciding whether or not to continue their schooling. Our model approximates a dynamic discrete choice model

\textsuperscript{7}See Heckman (2010).
\textsuperscript{8}Such approximations are discussed in Heckman (1981), Eckstein and Wolpin (1989), Cameron and Heckman (2001), and Geweke and Keane (2001).
\textsuperscript{9}Our approach is related to the analyses of Heckman and Vytlacil (1999, 2005, 2007a,b), Carneiro, Heckman, and Vytlacil (2010, 2011), and Eisenhauer, Heckman, and Vytlacil (2015), who introduce choice theory into the instrumental variables literature. They focus their analysis on binary choice models but also analyze ordered and unordered choice models with multiple outcomes to estimate economically interpretable treatment effects. Expanding on that body of research, we consider multiple sources of identification besides instrumental variables, and link our analysis more closely than they do to the dynamic discrete choice literature.
without taking a stance on exactly what agents are maximizing or how their information sets are being updated.

Like structural models, our model is identified though multiple sources of variation. Drawing from the matching literature, we identify the causal effects of schooling at different stages of the life cycle by using a rich set of observed variables and by proxying unobserved endowments. Unlike previous work on matching, we correct our proxies for measurement error and the bias introduced into the measurements by family background. We can also use exclusion restrictions to identify our model as in the IV and control function literatures. Unlike many structural models, we provide explicit proofs of model identification.

Our framework allows for *ex-ante* valuations as in dynamic discrete choice models but does not explicitly identify them.\(^{10}\) However, we can estimate *ex-post* returns to schooling, and model how they depend on both observed and unobserved variables. We decompose the *ex-post* treatment effects into (a) the direct benefits of going from one level of schooling to the next\(^{11}\) and (b) continuation values arising from access to additional education beyond the immediate next step.

Estimating our model on NLSY79 data, we investigate foundational issues in human capital theory. We report the following findings. (1) Ability bias accounts for a substantial portion (ranging between a third and two-thirds) of the raw differences in outcomes classified by education. At the same time, there are substantial causal effects of education on earnings and health.\(^{12}\) (2) Estimated causal effects differ by schooling level and depend on observed and unobserved characteristics of individuals. While the returns to high school are roughly the same across endowment levels, only high-endowment individuals benefit from college graduation. There is positive sorting on gains ("sorting bias" or "pursuit of comparative advantage") only at higher educational levels, but there is sorting into schooling based on

\(^{10}\)See, e.g., Eisenhauer, Heckman, and Mosso (2015), where this is done.

\(^{11}\)The human capital literature traditionally focused on the direct causal benefits of one final schooling level compared to another, but makes sequential comparisons from the lowest levels of schooling to the highest (Becker, 1964)

\(^{12}\)This finding runs counter to a common interpretation in the literature based on comparing IV and OLS estimates of Equation (1). See, e.g., Griliches (1977) and Card (1999, 2001).
observed and unobserved variables in earnings equations across all schooling levels ("ability bias").

(3) The early literature ignored the dynamics of schooling decisions. We find that continuation values arising from sequential choices are empirically important. Continuation values depend on cognitive and noncognitive endowments. Low endowment individuals gain mostly from the direct effect of high school graduation while high endowment individuals gain mostly in terms of continuation values. Low endowment individuals do not benefit from graduating college.

(4) Our schooling choice model is consistent with a variety of decision rules and allows for time inconsistency, regret and systematic mistakes due to cognitive failures. We use model estimates to test the assumptions of forward looking behavior and selection on gains often assumed in estimating dynamic discrete choice models. We find that agents do not know, or act on, publicly available information on college tuition costs in making decisions about graduating high school. Nonetheless, agents sort into schooling on ex-post gains, especially at higher schooling levels. A core tenet of human capital theory is thus confirmed.

(5) Our evidence of sorting into schooling on gains that varies across schooling levels does not support the Mincer model (1). The distributions of annualized gains vary greatly across schooling levels.

(6) Our paper contributes to an emerging literature on the importance of both cognitive and noncognitive endowments in shaping life outcomes. Consistent with the recent literature, we find that both cognitive and noncognitive endowments are important predictors of educational attainment. Within schooling levels, cognitive and noncognitive endowments have additional impacts on most outcomes.

(7) We meet challenge (c) and estimate substantial causal effects of education on health

---

13See e.g., Rust (1994); Keane and Wolpin (1997); Blevins (2014).
15Our estimates of the causal effects of education do not require that we separately isolate the effects of individual cognitive and noncognitive endowments on outcomes, just that we control for them as a set.
and healthy behaviors in addition to its large effects on wages.¹⁶

Using our estimated model, we conduct three policy experiments. In the first, we examine the impact of a tuition subsidy on college enrollment. We identify who is impacted by the policy, how their decisions change, and how much they benefit. Those induced to enroll benefit from the policy, and many go on to graduate from college. In a second experiment, we exploit the structural properties of our model. We analyze a policy that improves the endowments of those at the bottom of the distribution to see how this impacts educational choices and outcomes. Such improvements are produced by early intervention programs.¹⁷ Increasing cognitive endowments has a positive impact on all outcomes, while increasing noncognitive endowments mostly impacts health outcomes. In a third policy experiment, we use our estimated model to interpret what instrumental variables associated with a policy of reducing distance to the nearest college identify and the choice margins and characteristics of the people affected by this policy.

This paper proceeds in the following way. Section 2 presents our model. Section 3 presents the economically interpretable treatment effects that can be derived from it. Section 4 discusses identification. Section 5 discusses the data analyzed and presents unadjusted associations and regression adjusted associations between different levels of education and the outcomes analyzed in this paper. Section 6 reports our estimated treatment effects and their implications. Section 7 uses the estimated model to address three policy-relevant questions and to interpret what IV estimates. Section 8 considers the robustness of our estimates to alternative methodological approaches. It also shows how our estimates differ from estimates from OLS matching and IV procedures. Section 9 concludes.

¹⁶There is a small, but growing literature on this topic. See Grossman (2000); McMahon (2000); Lochner (2011); Oreopoulos and Salvanes (2011); Cutler and Lleras-Muney (2010). For a review of this literature see Web Appendix A.1.

¹⁷Heckman, Pinto, and Savelyev (2013).
The Causal Effects of Education on Earnings and Health

2 Model

This paper estimates a multistage sequential model of educational choices and their consequences. Let $J$ denote an ordered set of possible terminal states (see Figure 1). At each node there are only two possible choices: remain at $j$ or transit to $j + 1$. $D_j = 0$ if a person at $j$ does not stop there and goes on to $j + 1$. $D_j = 1$ if the person stops at $j$. $D_j \in D$ is the set of possible transition decisions that can be taken by the individual over the decision period. Let $S = \{0, \ldots, s\}$ denote the finite and bounded set of stopping states with $S = s$ if the agent stops at $s \in S$ and $D_s = 1$. Define $\overline{s}$ as the highest attainable element in $S$. We assume that the environment is time-stationary and decisions are irreversible.

$Q_j = 1$ denotes that an agent gets to decision node $j$. $Q_j = 0$ if the person never gets there. The history of nodes visited by an agent can be described by the collection of the $Q_j$ such that $Q_j = 1$.

![Figure 1: A Multistage Dynamic Decision Model](image)

2.1 A Sequential Decision Model

The decision process at each node is characterized by an index threshold-crossing property:

$$D_j = \begin{cases} 
0 & \text{if } I_j \geq 0, \quad j \in J = \{0, \ldots, \overline{s} - 1\} \\
1 & \text{otherwise,} 
\end{cases} \quad \text{for } Q_j = 1, \quad j \in \{0, \ldots, \overline{s} - 1\} \quad (2)$$

where $I_j$ is the perceived value at node $j$ of going on to $j + 1$ for a person at node $j$. The requirement $Q_j = 1$ ensures that agents are able to make the transition from $j$ to $j + 1$.

Associated with each final state $s \in S = \{0, \ldots, \overline{s}\}$ is a set of $K_s$ potential outcomes for

---

18The restriction that $J$ is an ordered set can readily be relaxed at the cost of greater notational burden. For example, we can include the GED as a choice after students drop out of high school. See Web Appendix Section A.2 for details.
each agent with indices $k \in \mathcal{K}_s$. We define $\tilde{Y}_s^k$ as latent variables that map into potential outcomes $Y_s^k$:

$$Y_s^k = \begin{cases} 
\tilde{Y}_s^k & \text{if } Y_s^k \text{ is continuous}, \\
1(\tilde{Y}_s^k \geq 0) & \text{if } Y_s^k \text{ is a binary outcome}, 
\end{cases} \quad k \in \mathcal{K}_s, \quad s \in \mathcal{S}. \quad (3)$$

Using the switching regression framework of Quandt (1958, 1972), the observed outcome $Y^k$ for a $k$ common across transitions is

$$Y^k = \sum_{s \in \mathcal{S}} D_s Y_s^k, \quad k \in \mathcal{K}_s. \quad (4)$$

### 2.2 Parameterizations of the Decision Rules and Potential Outcomes for Final States

Following a well-established tradition in the treatment effect and structural literatures, we approximate $I_j$ using a separable model:

$$I_j = \phi_j(Z) - \eta_j, \quad j \in \{0, \ldots, \bar{s} - 1\} \quad (5)$$

where $Z$ is a vector of variables observed by the analyst, components of which determine the transition decisions of the agent at different stages and $\eta_j$ is unobserved by the analyst. A separable representation of the choice rule is an essential feature of LATE (Vytlacil, 2002) and dynamic discrete choice models (Blevins, 2014).

Outcomes are also separable:

$$\tilde{Y}_s^k = \tau_s^k(X) + U_s^k, \quad k \in \mathcal{K}_s, \quad s \in \mathcal{S}, \quad (6)$$

where $X$ is a vector of observed determinants of outcomes and $U_s^k$ is unobserved by the
analyst. Separability of the unobserved variables in the outcome equations is often invoked in the structural literature but is not strictly required (see Blevins, 2014). It is not required in the IV literature.

2.3 Structure of the Unobservables

Central to our main empirical strategy is the existence of a finite dimensional vector $\theta$ of unobserved (by the economist) endowments that generate all of the dependence across the $\eta_j$ and the $U^k_s$. We assume that

$$\eta_j = -(\theta' \alpha_j - \nu_j), \ j \in \{0, \ldots, s - 1\} \quad (7)$$

and

$$U^k_s = \theta' \alpha^k_s + \omega^k_s, \ k \in K_s, s \in S, \quad (8)$$

where $\nu_j$ is an idiosyncratic error term for transition $j$.

Conditional on $\theta, X, Z$, choices and outcomes are statistically independent. Thus controlling for this set of variables eliminates selection effects. If the analyst knew $\theta$, he/she could use matching to identify the model.$^{19}$

Standard “random effects” approaches in the structural literature integrate out $\theta$ and do not interpret it. Our approach is different. We proxy $\theta$ using multiple measurements of it and we identify, and correct for, errors in the proxy variables. The measurements facilitate the interpretation of $\theta$. We develop this intuition further in Section 4, after presenting the rest of our model.

We array the $\nu_j, j \in J$, into a vector $\nu = (\nu_0, \nu_1, \ldots, \nu_{s-1})$ and the $\eta_j$ into $\eta = (\eta_0, \ldots, \eta_{s-1})$. $\omega^k_s$ represents an idiosyncratic error term for outcome $k$ in state $s$. Array the $\omega^k_s$ into a vector $\omega_s = (\omega^1_s, \ldots, \omega^K_s)$. Array the $U^k_s$ into vector $U^s = (U^1_s, \ldots, U^K_s)$ and array the $U_s$ into $U = (U^0_s, \ldots, U^s)$.

Letting “⊥” denote statistical independence, we assume that

\[ \nu_j \perp \nu_l \quad \forall \ l \neq j \quad l, j \in \{0, \ldots, s - 1\} \]  
\( \text{(A-1a)} \)

\[ \omega_s^k \perp \omega_{s'}^{k'} \quad \forall \ k \neq k', \quad s, s' \in S \]  
\( \text{(A-1bi)} \)

\[ \omega_s^k \perp \omega_{s'}^{k'}, \quad \forall \ k, k', \quad s \neq s', \quad s, s' \in S \]  
\( \text{(A-1bii)} \)

\[ \omega_s \perp \nu, \quad \forall \ s \in S \]  
\( \text{(A-1c)} \)

\[ \theta \perp (X, Z) \]  
\( \text{(A-1d)} \)

\[ (\omega_s, \nu) \perp (\theta, X, Z) \quad \forall \ s \in S. \]  
\( \text{(A-1e)} \)

Assumption (A-1a) maintains independence of the shocks affecting schooling transitions; assumption (A-1bi) and (A-1bii) maintain independence of shocks across outcomes within states and across states; assumption (A-1c) maintains the independence of the shocks to choice equations with the shocks to outcomes; assumption (A-1d) maintains independence of \( \theta \) with the observables; and assumption (A-1e) maintains independence of the shocks and \( \theta \) with the observed variables. Representations (7) and (8) and versions of assumptions (A-1d) and (A-1e) play fundamental roles in the structural dynamic discrete choice literature.\(^{21}\)

### 2.4 Measurement System for Unobserved Factors \( \theta \)

We allow for the possibility that \( \theta \) cannot be measured precisely, but that it can be proxied with multiple measurements. We correct for the effects of measurement error in the proxy.

\(^{20}\)Conditioning on \( X \), we can weaken (A-1d) to

\[ \theta \perp Z|X \]  
\( \text{(A-1d')} \)

and (A-1e) to

\[ (\omega_s, \nu) \perp \theta, Z|X \]  
\( \text{(A-1e')} \)

All of the assumptions (A-1a)-(A-1e) can be reformulated to be conditional on \( X \).

\(^{21}\)The Keane and Wolpin (1997) “types” can be interpreted as versions of \( \theta \) that arise from the initial conditions of their model. (7) and (8) capture the random effects specifications widely used in the discrete choice literature. (See Aguirregabiria, 2010 and Adda and Cooper, 2003.) Our model does not impose any particular information updating structure (e.g., iid shocks), the risk-neutrality of decision makers or the Bellman equation decision structure widely used in the structural literature.
The structural literature treats the $\theta$ as nuisance variables, invokes conditional independence assumptions, and integrates $\theta$ out using random effect procedures. Instead, we link $\theta$ to measurements, and adjoin measurement equations to choice and outcome equations, rendering $\theta$ interpretable.

Let $T$ be a vector of $M$ measurements on $\theta$. They may consist of lagged or future values of the outcome variables or additional measurements. The system of equations determining $T$ is:

$$T = \Phi(X, \theta, e),$$

where $X$ are observed variables, $\theta$ are the factors and

$$T = \begin{pmatrix} T_1 \\ \vdots \\ T_M \end{pmatrix} = \begin{pmatrix} \Phi_1(X, \theta, e_1) \\ \vdots \\ \Phi_M(X, \theta, e_M) \end{pmatrix},$$

where we array the $e_j$ into $e = (e_1, \ldots, e_M)$. We make the additional assumptions that:

$$e_j \perp \perp e_l, \quad j \neq l, \quad j, l \in \{1, \ldots, M\}$$

and

$$e \perp \perp (X, Z, \theta, \nu, \omega).$$

For the purpose of identifying treatment effects, we do not need to identify each equation of system (9). We just have to identify the span of $\theta$ that preserves the information on $\theta$ in (9), and that is sufficient to produce conditional independence between choices and outcomes. However, in this paper we estimate equation system (9).

---

See e.g., Keane and Wolpin (1997); Rust (1994); Adda and Cooper (2003); Blevins (2014).

See, e.g., Abbring and Heckman (2007); Schennach, White, and Chalak (2012).

See e.g., Heckman, Schennach, and Williams (2013).
3 Defining Treatment Effects

A variety of \textit{ex post} counterfactual outcomes and associated treatment effects can be generated from our model. They can be used to predict the effects of manipulating education levels through different policies for people of different backgrounds and abilities. They allow us to understand the effectiveness of policies for different identifiable segments of the population, and the benefits to people at different margins of choice.

In principle, we can define and estimate a variety of treatment effects, many of which are implausible. For example, many empirical economists would not find estimates of the effect of fixing (manipulating) \(D_j = 0\) if \(Q_j = 0\) to be credible (i.e., the person for whom we fix \(D_j = 0\) is not at the decision node to take the transition).\(^{25}\) In the spirit of credible econometrics, we define treatment effects associated with fixing \(D_j = 0\) conditioning on \(Q_j = 1\). This approach blends structural and treatment effect approaches. Our causal parameters recognize agent heterogeneity and are allowed to differ across populations, contrary to standard approaches in structural econometrics.\(^{26}\)

The person-specific treatment effect \(T^k_j\) for outcome \(k\) for an individual selected from the population \(Q_j = 1\) with characteristics \(X = x, Z = z, \theta = \bar{\theta}\), making a decision at node \(j\) between going on to \(j + 1\) or stopping at \(j\) is the difference between the individual’s outcomes under the two actions. This can be written as

\[
T^k_j[Y^k|X = x, Z = z, \theta = \bar{\theta}] := (Y^k|X = x, Z = z, \theta = \bar{\theta}, Fix D_j = 0, Q_j = 1) - (Y^k|X = x, Z = z, \theta = \bar{\theta}, Fix D_j = 1, Q_j = 1). \tag{10}
\]

The random variable \((Y^k|X = x, Z = z, \theta = \bar{\theta}, Fix D_j = 0, Q_j = 1)\) is the outcome variable \(Y^k\) at node \(j\) for a person with characteristics \(X = x, Z = z, \theta = \bar{\theta}\) from the population who attain node \(j\) (or higher), \(Q_j = 1\), and for whom we fix \(D_j = 0\) so they go on to the next

\(^{25}\)The distinction between \textit{fixing} and \textit{conditioning} traces back to Haavelmo (1943). White and Chalak (2009) use the terminology “setting” for the same notion. For a recent analysis of this crucial distinction, see Heckman and Pinto (2015).

\(^{26}\)See, e.g., Hansen and Sargent (1980).
node. Random variable \((Y^k|X = x, Z = z, \theta = \bar{\theta}, Fix \ D_j = 1, Q_j = 1)\) is defined for the same individual but forces the person with these characteristics not to transit to the next node.

We next present population level treatment effects based on (10). We focus our discussion on means but we can also formulate distributional counterparts for all of the treatment effects considered in this paper.

### 3.1 Dynamic Treatment Effects

A main contribution of this paper is to define and estimate treatment effects that take into account the direct effect of moving to the next node of a decision tree, plus the benefits associated with the further schooling that such movement opens up. This treatment effect is the difference in expected outcomes arising from changing a single educational decision in a sequential schooling model and tracing through its consequences, accounting for the dynamic sequential nature of schooling.

The person-specific treatment effect can be decomposed into two components: the Direct Effect of going from \(j\) to \(j + 1\): \(DE^k_j = Y^k_{j+1} - Y^k_j\), the effect often featured in the literature on the returns to schooling, and the Continuation Value of going beyond \(j + 1\):

\[
C^k_{j+1} = \sum_{r=1}^{\pi-(j+1)} \left[ \prod_{l=1}^{r} (1 - D_{j+l}) \right] (Y^k_{j+r+1} - Y^k_{j+r}).
\]

Thus, at the individual level, the Total Effect of fixing \(D_j = 0\) on \(Y^k\) is decomposed into

\[
T^k_j = DE^k_j + C^k_{j+1}.
\]
The associated population level average treatment effect conditional on \( Q_j = 1 \) is

\[
ATE_j^k := \iint \int E[T_j^k(Y^k|X = x, Z = z, \theta = \overline{\theta})] \, dF_{X,Z,\theta}(x, z, \theta|Q_j = 1) \tag{12}
\]

which can be decomposed into direct and continuation value components.

Because we do not specify or attempt to identify choice-node-specific agent information sets, we can only identify \textit{ex-post} treatment effects. Hence, we can identify continuation values associated with choices, but cannot identify option values. However, a benefit of this more agnostic approach is that it does not impose specific decision rules. Our model allows for irrationality, regret, and mistakes in decision-making associated with agent maturation and information acquisition.

### 3.2 Mean Differences Across Final Schooling Levels

Becker’s original approach (1964) can be interpreted to define returns to education as the gains from choosing between a base and a terminal schooling level. Let \( Y_{s}^k \) be outcome \( k \) at schooling level \( s' \) and \( Y_{s}^k \) be outcome \( k \) at schooling level \( s \). Conditioning on \( X = x \) and \( \theta = \overline{\theta} \), the average treatment effect of \( s \) compared to \( s' \) is \( E(Y_{s}^k - Y_{s'}^k|X = x, Z = z, \theta = \overline{\theta}) \).

Integrating out \( X, Z, \theta \) produces a pairwise ATE parameter over the available supports of these variables.

A more empirically credible version, and the one we report here, calculates the mean gain for the subset of the population that completes one of the two final schooling levels:

\[
ATE_{s,s'}^k \equiv \iint \int E(Y_{s}^k - Y_{s'}^k|X = x, Z = z, \theta = \overline{\theta}) \, dF_{X,Z,\theta}(x, z, \theta|D_s + D_{s'} = 1). \tag{13}
\]

Conditioning in this fashion recognizes that the characteristics of people not making either final choice could be far away from the population making one of those choices and hence might be far away from having any empirical or policy relevance.\(^{28}\)

\(^{28}\)The estimated differences in treatment effects for the conditional and unconditional population are
3.3 Average Marginal Treatment Effects

In order to understand treatment effects for persons at the margin of indifference at each node of the decision tree of Figure 1, we estimate the Average Marginal Treatment Effect (AMTE).\textsuperscript{29} It is the average effect of transiting to the next node for individuals at the margin of indifference between the two nodes:

\[
AMTE^k_j := \int\int\int E[T^k_j(Y^k|X = x, Z = z, \theta = \bar{\theta}, |I_j| < \varepsilon)] \, dF_{X, Z, \theta}(x, z, \bar{\theta} | Q_j = 1, |I_j| \leq \varepsilon),
\]

where \(\varepsilon\) is an arbitrarily small neighborhood around the margin of indifference. These effects are inclusive of all consequences of taking the transition at \(j\), including the possibility of attaining final schooling levels well beyond \(j\). AMTE defines causal effects at well-defined and empirically identified margins of choice. It is the proper measure of the marginal gross benefit for evaluating the gains from moving from one stage of the decision tree to the next for those at that margin of choice. In general it is distinct from LATE, which is not defined for any specific margin of choice.\textsuperscript{30} Since we identify the distribution of \(I_j\), we can identify the characteristics of agents in the indifference set, something not possible using IV or matching.

The population distribution counterpart of AMTE is defined over the set of agents for whom \(|I_j| \leq \varepsilon\), which can be generated from our model: \(Pr(T^k_j < t^k_j|Q_j = 1, |I_j| \leq \varepsilon)\). Distributional versions can be defined for all of the treatment effects considered in this section.

3.4 Policy Relevant Treatment Effects

The policy relevant treatment effect (PRTE) is the average treatment effect for those induced to change their choices in response to a particular policy intervention. Let \(Y^k(p)\) be the aggregate outcome under policy \(p\) for outcome \(k\). Let \(S(p)\) be the final state selected by an agent under policy \(p\). The policy relevant treatment effect from implementing policy \(p\)

not large for outcomes associated with the decision to enroll in college, but is substantial for the choice to graduate from college. See Tables A55, A57, A59, and A61 in the Web Appendix.

\textsuperscript{29}See Carneiro, Heckman, and Vytlacil (2010, 2011).

\textsuperscript{30}See Heckman and Vytlacil (2007a) and Carneiro, Heckman, and Vytlacil (2010). The LATE can correspond to people at multiple margins. See Angrist and Imbens (1995).
compared to policy $p'$ for outcome $k$ is:

$$PRTE_{p,p'}^k := \frac{1}{\mathbb{P}(S(p) \neq S(p'))} \int \frac{1}{\mathbb{P}(S(p) \neq S(p'))} \left( \int \frac{1}{\mathbb{P}(S(p) \neq S(p'))} \left( \int \frac{1}{\mathbb{P}(S(p) \neq S(p'))} \right) \right) \right),$$

where $S(p) \neq S(p')$ denotes the set of the characteristics of people for whom attained states differ under the two policies. In general, it is different from AMTE because the agents affected by a policy can be at multiple margins of choice. PRTE is often confused with LATE. In general, they are different unless the proposed policy change coincides with the instrument used to define LATE.\textsuperscript{31}

### 4 Identification and Model Likelihood

The treatment effects defined in Section 3 can be identified using alternative empirical approaches. The main approach used in this paper exploits the fact that conditional on $\theta, X, Z$, outcomes and choices are statistically independent. $X$ and $Z$ are observed. $\theta$ is not. If $\theta$ were observed, one could condition on $\theta, X, Z$ and identify the model of Equations (2) - (8) and the treatment effects that can be generated from it. We use nonlinear factor model (9) to proxy $\theta$.

Under the conditions presented in Web Appendix A.4, we can nonparametrically identify the model of Equations (2) - (8) including the distribution of $\theta$, as well as the $\Phi$ functions and the distribution of $e$ (which can be interpreted as measurement errors). Effectively, we match on proxies for $\theta$ and correct for the effects of measurement error ($e$) in creating the proxies. Such corrections are possible because with multiple measures on $\theta$ we can identify the distribution of $e$.

Under full linearity assumptions, one can directly estimate the $\theta$ and use factor regression

\textsuperscript{31}See Carneiro, Heckman, and Vytlacil (2011) for an empirical example. The differences between the two parameters can be substantial as we show in Web Appendix A.5.2.
methods.\textsuperscript{32} Full details of this approach are spelled out in Web Appendix A.4.\textsuperscript{33} Another approach to identification uses instrumental variables which, if available, under the conditions presented in Web Appendix A.4 can be used to identify the structural model (2) - (8) without factor structure (7) and (8).

The precise parameterization and the likelihood function for the model we estimate is presented in Web Appendix A.6. While in principle it is possible to identify the model semi-parametrically, in this paper we make parametric assumptions in order facilitate computation. We subject the estimated model to rigorous goodness of fit tests which we pass.\textsuperscript{34}

5 Our Data, A Benchmark OLS Analysis of the Outcomes We Study and Our Exclusion Restrictions

We estimate our model on a sample of males extracted from the widely used National Longitudinal Sample of Youth (NLSY 79).\textsuperscript{35} Before discussing estimates from our model, it is informative to set the stage and present adjusted and unadjusted associations between the outcomes we study and schooling. Figure 2 presents estimated regression relationships between different levels of schooling (relative to high school dropouts) and the four outcomes analyzed in this paper: wages, log present value of wages, health limitations, and smoking.\textsuperscript{36}

The black bars in each panel show the unadjusted mean differences in outcomes for persons at the indicated levels of educational attainment compared to those for high school dropouts. Higher ability is associated with higher earnings and more schooling. However, as shown by the grey bars in Figure 2, adjusting for family background and adolescent measures of ability attenuates, but does not eliminate, the estimated effects of education.

\textsuperscript{32}See, e.g., Heckman, Pinto, and Savelyev (2013) and the references cited therein.

\textsuperscript{33}As noted in Web Appendix A.4.1, and Heckman, Schemach, and Williams (2011), we do not need to solve classical identification problems associated with estimating equation system (9) in order to extract measure-preserving transformations of \( \theta \) on which we can condition in order to identify treatment effects. In the linear factor analysis literature these are rotation and normalization problems.

\textsuperscript{34}See Web Appendix A.7.

\textsuperscript{35}Web Appendix A.8 presents a detailed discussion of the data we analyze and our exclusion restrictions.

\textsuperscript{36}Adjustments are made through linear regression.
Figure 2 shows that controlling for proxies for ability substantially reduces the observed differences in earnings across educational groups. Nonetheless, there are still strong causal effects of education. It has been claimed that a model that is linear in years of schooling fits the data well.\textsuperscript{37} The white bar in Figure 2 displays the estimated adjusted effect of schooling controlling for years of completed schooling as in Equation (1).\textsuperscript{38} The white bars in all figures suggest that the linear-in-years-of-schooling Mincer specification (1) does not describe our data. There are effects of schooling beyond those captured by a linear years of schooling specification.

5.1 Exclusion Restrictions

As noted in Section 4, identification does not depend exclusively on conditional independence assumptions associated with our factor model although they alone justify the identification of our model using matching on mismeasured variables.\textsuperscript{39} Node-specific instruments can nonparametrically identify treatment effects without invoking the full set of conditional independence assumptions.\textsuperscript{40} We have a variety of exclusion restrictions that affect choices but not outcomes. Table 1 documents the $X$ and $Z$ used in this paper. Our instruments are traditional in the literature that estimates the causal effects of education.\textsuperscript{41}

\textsuperscript{38}Mis-measurement of schooling is less of a concern in our data as the survey asks numerous educational questions every year which we use to determine an individual’s final schooling state.
\textsuperscript{39}See Carneiro, Hansen, and Heckman (2003).
\textsuperscript{40}See Web Appendix A.4.
\textsuperscript{41}For example, presence of a nearby college or distance to college is used by Card (2012); Cameron and Taber (2004); Kling (2001); Carneiro, Meghir, and Parey (2013); Cawley, Conneely, Heckman, and Vyltacil (1997); Heckman, Carneiro, and Vyltacil (2011); and Eisenhauer, Heckman, and Vyltacil (2015). Local tuition at two or four year colleges is used as an instrument by Kane and Rouse (1993); Heckman, Carneiro, and Vyltacil (2011); Eisenhauer, Heckman, and Vyltacil (2015); and Cameron and Taber (2004). Local labor market shocks are used by Heckman, Carneiro, and Vyltacil (2011) and Eisenhauer, Heckman, and Vyltacil (2015).
Figure 2: Raw and Adjusted Benefits from Education

![Graphs showing the benefits from education on log wages, log PV of wages, health limits work, and daily smoking.](image)

Notes: The bars represent the coefficients from a regression of the designated outcome on dummy variables for educational attainment, where the omitted category is high school dropout. Regressions are run adding successive controls for background and proxies for ability. Background controls include race, region of residence in 1979, urban status in 1979, broken home status, number of siblings, mother’s education, father’s education, and family income in 1979. Proxies for ability are average score on the ASVAB tests and ninth grade GPA in core subjects (language, math, science, and social science). “Some College” includes anyone who enrolled in college, but did not receive a four-year college degree. The white bar additionally controls for highest grade completed (HGC). Source: NLSY79 data.

6 Estimated Causal Effects

We next present the main treatment effects estimated from our model. Since our model is nonlinear and multidimensional, in the main body of the paper we report interpretable functions derived from it.\textsuperscript{42} We randomly draw sets of regressors from our sample and a

\textsuperscript{42}Parameter estimates for individual equations are reported in Web Appendix A.9.
Table 1: Control Variables and Instruments Used in the Analysis

<table>
<thead>
<tr>
<th>Variables</th>
<th>Measurement Equations</th>
<th>Choice</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Broken Home</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Number of Siblings</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Parents’ Education</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Family Income (1979)</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Region of Residence&lt;sup&gt;a&lt;/sup&gt;</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Urban Status&lt;sup&gt;a&lt;/sup&gt;</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Age&lt;sup&gt;b&lt;/sup&gt;</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Local Unemployment&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Local Long-Run Unemployment</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Instruments**

<table>
<thead>
<tr>
<th>Instruments</th>
<th>Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local Unemployment at Age 17&lt;sup&gt;d&lt;/sup&gt;</td>
<td>x</td>
</tr>
<tr>
<td>Local Unemployment at Age 22&lt;sup&gt;e&lt;/sup&gt;</td>
<td>x</td>
</tr>
<tr>
<td>College Present in County 1977&lt;sup&gt;f&lt;/sup&gt;</td>
<td>x</td>
</tr>
<tr>
<td>Local College Tuition at Age 17&lt;sup&gt;g&lt;/sup&gt;</td>
<td>x</td>
</tr>
<tr>
<td>Local College Tuition at Age 22&lt;sup&gt;h&lt;/sup&gt;</td>
<td>x</td>
</tr>
</tbody>
</table>

**Notes:**

<sup>a</sup> Region and urban dummies are specific to the age that the measurement, educational choice, or outcome occurred.

<sup>b</sup> Age in 1979 is included as a cohort control. We also included individual cohort dummies which did not change the results.

<sup>c</sup> For economic outcomes, local unemployment at the time the outcome is measured.

<sup>d</sup> This is an instrument for choices at nodes 1 and 2. It represents opportunity costs at the time schooling decisions are made.

<sup>e</sup> This is an instrument for the choice at node 3.

<sup>f</sup> Presence of a 4-year college in the county in 1977 is constructed from Kling (2001) and enters the choice to enroll and the choice to graduate from college.

<sup>g</sup> Local college tuition at age 17 only enters the college enrollment graduation decisions.

<sup>h</sup> Local college tuition at age 22 only enters the college completion equation.

We first present (Section 6.1) the main treatment effects across final schooling levels, by node, and their decomposition into direct and indirect effects. We discuss how endowments affect the treatment effects. We next (Section 6.2) discuss distributions of treatment effects. In Section 6.3 we interpret these estimates for each of the four outcomes studied.

Educational decisions at each node depend on both endowments. In addition we find sorting on gains (comparative advantage) for the college enrollment and college graduation decisions, but not the high school graduation decision. Sorting on levels (ability bias) is found across all educational levels. The Mincer model (Equation (1)) does not capture these

---

<sup>43</sup>We randomly draw an individual and use their full set of regressors.

<sup>44</sup>This finding generalizes the analysis of Willis and Rosen (1979) to multiple schooling levels.
types of sorting patterns. It overlooks the differences in the distributions of returns across schooling levels.

6.1 The Estimated Causal Effects of Educational Choices

We first compare the outcomes from final schooling level $s$ with those from $s - 1$. The estimated treatment effects of education on log wages, log PV wage income, smoking, and health limits work are shown in Figure 3. For each outcome, the bars labeled “Observed” display the unadjusted raw differences in the data. The bars labeled “Causal Component” display the average treatment effect obtained from comparing the outcomes associated with a particular schooling level $s$ relative to $s - 1$. These are defined for individuals at $s$ or $s - 1$. There are substantial causal effects on earnings and health at each level of schooling. But at most levels there is also considerable ability bias.

---

45 See expression (13) for the case $s' = s - 1$.
46 These are calculated by simulating the mean outcomes for the designated state and comparing it with the mean-simulated outcome for the state directly below it for the subpopulation of persons who are in either of the states.
Figure 3: Causal Versus Observed Differences by final schooling level (compared to next lowest level)

3A. Decomposition of Schooling Effects
Log Wages

3B. Decomposition of Schooling Effects
Log PV Wages

3C. Decomposition of Schooling Effects
Daily Smoking

3D. Decomposition of Schooling Effects
Health Limits Work

Notes: Each bar compares the outcomes from a particular schooling level $j$ and the next lowest level $j - 1$. The “Observed” bar displays the observed differences in the data. The “Causal Component” bar displays the estimated average treatment effects (ATE). The difference between the observed and causal treatment effect is attributed to the effect of selection and ability. The error bars and significance levels for the estimated ATE are calculated using 200 bootstrap samples. Error bars show one standard deviation and correspond to the 15.87th and 84.13th percentiles of the bootstrapped estimates, allowing for asymmetry. Significance at the 5% and 1% levels is shown by open and filled circles on the plots, respectively.

6.1.1 Dynamic Treatment Effects

We next report treatment effects by decision node (see Figure 4). We compute the gains to achieving (and possibly exceeding) the designated level of schooling (including continuation values) and compare them to the outcomes associated with not achieving that level. The Average Marginal Treatment Effect, AMTE, is the average treatment effect for those indifferent
to the two options of the choice studied.47

Each box of Figure 4 presents the average effects of educational choices on the specified outcome. The effects are presented as the height of different bars in each figure. They are defined as the differences in the outcomes associated with being at the designated level, compared to the one preceding it (not necessarily final or terminal schooling levels), for those for whom $Q_j = 1$. The ATE is calculated for the population that reaches the decision node. At each node $j$, the treatment effect is $E(Y_k|\text{fix } D_j = 0, Q_j = 1) - E(Y_k|\text{fix } D_j = 1, Q_j = 1)$. ATE (high) and ATE (low) are the ATEs for different ability groups. The high- (low-) ability group is defined for individuals with both cognitive and socioemotional endowment above (below) the median of the full population. The table below the figure displays the fraction of individuals at each educational choice who are in the high- or low-ability group.
Figure 4: Treatment Effects of Outcomes by Decision Node
\[ E(Y^k|\text{Fix } D_j = 0, Q_j = 1) - E(Y^k|\text{Fix } D_j = 1, Q_j = 1) \]

Notes: Each schooling level might provide the option to pursuing higher schooling levels. Only final schooling levels do not provide an option value. The error bars and significance levels for the estimated ATE are calculated using 200 bootstrap samples. Error bars show one standard deviation and correspond to the 15.87th and 84.13th percentiles of the bootstrapped estimates, allowing for asymmetry. Significance at the 5% and 1% level are shown by hollow and black circles on the plots respectively. The figure reports various treatment effects for those who reach the decision node, including the estimated ATE conditional on endowment levels. The high- (low-) ability group is defined as those individuals with cognitive and socioemotional endowments above (below) the median in the overall population. The table below the figure shows the proportion of individuals at each decision \((Q_j = 1)\) that are high and low ability. The larger proportion of the individuals are high ability and a smaller proportion are low ability in later educational decisions. In this table, final schooling levels are highlighted using bold letters.
6.1.2 Continuation Values

We next decompose the node-specific treatment effects reported in Table 4 into the total effect and the continuation value components. Figure 5 presents graphs of each causal effect in Figure 4 and shows the continuation value component (in white). Continuation values are important components of the dynamic treatment effects for all outcomes except health limits work.

Figure 5: Dynamic Treatment Effects:
Continuation Values and Total Treatment Effects by Node

Notes: High-ability individuals are those in the top 50% of the distributions of both cognitive and socioemotional endowments. Low-ability individuals are those in the bottom 50% of the distributions of both cognitive and socioemotional endowments. The error bars and significance levels for the estimated ATE are calculated using 200 bootstrap samples. Error bars show one standard deviation and correspond to the 15.87th and 84.13th percentiles of the bootstrapped estimates, allowing for asymmetry. Significance at the 5% and 1% level are shown by hollow and black circles on the plots respectively. Statistical significance for continuation values at the 5% level are shown by x. Section 3 provides details on how the continuation values and treatment effects are calculated.
6.1.3 The Effects on Cognitive and Noncognitive Endowments on Treatment Effects

While we disaggregate the treatment effects for “high” and “low” endowment individuals in Figure 4, this division is coarse. A byproduct of our approach is that we can determine the contribution of cognitive and noncognitive endowments (θ) to explaining estimated treatment effects. We can decompose the overall effects of θ into their contribution to the causal effects at each node and the contribution of endowments to attaining that node. We find substantial contributions of θ to each component at each node.

To illustrate, the panels in Figure 6 display the estimated average treatment effect of getting a four-year degree (compared to stopping with some college) for each decile pair of cognitive and noncognitive endowments. Treatment effects in general depend on both measures of ability. Moreover, different outcomes depend in different ways on the two dimensions of ability. For example, the treatment effect of graduating college is increasing in both dimensions for present value of wages, but the reductions in health limitations with education depend mostly on cognitive endowments.

48Web Appendix A.10 reports a full set of results.
49They show average benefits by decile over the full population, rather than for the population that reaches each node.
Figure 6: Average Treatment Effect of Graduating from a Four-Year College by Outcome

A. (log)Wages

B. PV Wages

C. Smoking

D. Health Limits Work

Notes: Each panel in this figure studies the average effects of graduating with a four-year college degree on the outcome of interest. The effect is defined as the differences in the outcome between those with a four-year college degree and those with some college. For each panel, let $Y_{somecoll}$ and $Y_{4\text{-yr degree}}$ denote the outcomes associated with attaining some college and graduating with a four-year degree, respectively. For each outcome, the first figure (top) presents $E(Y_{4\text{-yr degree}} - Y_{somecoll}|d^C, d^{SE})$ where $d^C$ and $d^{SE}$ denote the cognitive and socioemotional deciles computed from the marginal distributions of cognitive and socioemotional endowments. The second figure (bottom left) presents $E(Y_{4\text{-yr degree}} - Y_{somecoll}|d^C)$ so that the socioemotional factor is integrated out. The bars in this figure display, for a given decile of cognitive endowment, the fraction of individuals visiting the node leading to the educational decision involving graduating from a four-year college. The last figure (bottom right) presents $E(Y_{4\text{-yr degree}} - Y_{somecoll}|d^{SE})$ and the fraction of individuals visiting the node leading to the educational decision involving graduating from a four-year college for a given decile of socioemotional endowment.

6.2 Distributions of Treatment Effects

One benefit of our approach over the standard IV approach is that we can identify the distributions of expected treatment effects. This feature is missing from the standard
treatment effect literature. Figure 7 plots the distribution of gains for persons who graduate from college (compared to attending college but not attaining a four-year degree) along with the mean treatment effects.\textsuperscript{50} Expectations are computed over the idiosyncratic error terms ($\omega_k^s$).\textsuperscript{51} Variation in the expected treatment effect comes from the variation in observed variables ($X$) and the unobserved endowments ($\theta$).

\textbf{Figure 7: Distributions of Expected Treatment Effects: College Graduation}

\textit{Notes:} Distributions of treatment effects are for those who reach the educational choice.

6.3 Interpreting the Estimated Treatment Effects

\textbf{Treatment Effects on Log Wages} Comparing final educational levels, the average treatment effect is statistically significant for graduating from high school, attending college,
and attaining a four-year college degree. About half of the observed difference in wages at age 30 is explained by the $X$, $Z$, and $\theta$.

Estimates for node-specific treatment effects show that more schooling causally boosts wages although low-endowment individuals gain very little from getting a four-year college degree. Figure 4 shows that individuals with high cognitive ability capture most of the gains from a four-year degree. In fact, our estimates suggest those with very low cognitive and socioemotional endowments lose wage income at age 30 by graduating from college. Figure 5 shows that continuation values are an important component of average treatment effects for high ability individuals. Figure 6 shows that most of the effect of abilities on the average treatment effect of college graduation comes from cognitive channels. Figure 7 shows the sorting pattern for college graduation. Even though it is not imposed by our estimation procedure, we find sorting on gains.

**Treatment Effects on Present Value of Wage Income** The pattern for the present value of wages is similar to that for wages with some interesting exceptions. Figures 4 and 5 show that low ability students appear to benefit substantially from graduating from high school, while only high ability individuals benefit from enrolling in and graduating from college. The treatment effect of college graduation is especially strong for high ability students. The benefits to low ability people and people at the margin of graduating high school come primarily from direct effects. The larger effects for present values than for wages comes from labor supply responses of high school graduates. We find sorting on gains for the higher educational nodes. Figure 6 shows that noncognitive endowments play a much stronger role in generating the average treatment effect of college graduation on the PV of wages than they do for wages.

---

52See Web Appendix Section A.1.2 for a brief overview of the literature on the outcomes considered in this paper.

Treatment Effects on Smoking  Controlling for unobserved endowments, education causally reduces smoking. The endowments and observables account for about one-third of the observed effect of education. The effects are especially strong for high school graduation. Looking at the node-specific treatment effects, each level of education has a substantial causal effect in reducing smoking. For high-endowment individuals, more than half of the average treatment effect of graduating high school and enrolling in college is derived from continuation values. Almost all of the treatment effect comes from the direct effect for low-endowment individuals.

Treatment Effects on Health Limits Work  There are strong treatment effects for graduating high school but much weaker, and less precisely determined, treatment effects at higher levels of education. Continuation values are small and generally statistically insignificant. As in the case of smoking, the treatment effects are especially strong for high ability individuals except in this case noncognitive endowments play a small role.

6.4 Heterogeneity in Returns Across Schooling Levels

Mincer equation (1) is often invoked in the empirical literature.\textsuperscript{54} Its linear specification avoids the need for stage-specific instruments. It implies that the return to an additional year of schooling is the same across schooling levels (although it may differ across persons). Returns across different schooling levels are summarized by a single number, $\rho_i$.

Most of the literature estimates linear-in-years-of-schooling models using OLS. Figure 2 shows that adding dummy variables to a linear-in-schooling specification substantially improves the fit of the model.\textsuperscript{55} We have already noted that sorting by gains differs by educational level. Using this methodology, linearity in years of schooling for wage equations is decisively rejected in many data sets and it is rejected in our data.\textsuperscript{56}

\textsuperscript{55}Card’s (1999, 2001) claims about linearity are based on OLS, as are most claims in the literature. (Heckman, Lochner, and Todd (2006)). We formally test this proposition in Web Appendix A.12.
\textsuperscript{56}See the evidence discussed in Heckman, Lochner, and Todd (2006) and in Web Appendix A.12.
Figure 8 plots the distributions of smoothed annualized returns $E_{\omega}(\rho_i) = E_{\omega}\left(\frac{Y_i - Y_{i-1}}{q_j - q_{j-1}}\right)$ for the full population for each node of schooling choices where $q_j$ is the years of schooling associated with node $j$.\(^{57}\) While mean returns are similar across schooling levels, the distributions of returns are quite different.\(^{58}\)

**Figure 8: Population Distribution of Annualized Distribution of Returns**

![Distribution of Expected TEs on Log Wage](image1)

![Distribution of Expected TEs on PV-Wage](image2)

![Distribution of Expected TEs on Smoking](image3)

![Distribution of Expected TEs on Health Limits Work](image4)

**Notes:** Distributions of treatment effects are for the whole population. Distributions have been normalized by the difference in the average number of years of completed schooling between the two educational choices. To smooth the plots, we display the distributions of treatment effects $E_{\omega}(\rho_i) = E_{\omega}\left(\frac{Y_i - Y_{i-1}}{q_j - q_{j-1}}\right)$, where we are taking the expectation over the idiosyncratic shocks to the outcomes the agent does not know or act on. The vertical lines indicate the means for each distribution.

Not only are the mean annualized returns somewhat different across schooling levels but the correlations of returns across schooling levels are different. Table 2 below shows the Pearson correlations for wages. The correlations are far from 1—what is predicted from the

---

\(^{57}\)We smooth the estimates by integrating out the idiosyncratic components ($\omega$).

\(^{58}\)See Web Appendix A.13 for distributions conditional on $Q_j = 1$ and for distributions of returns inclusive of continuation values.
Mincer model—and in fact for some transitions are negative. The negative correlations are consistent with the evidence that high ability individuals have low or negative direct benefits from early transitions but benefit most in later transitions. They are also consistent with the pattern of negative sorting reported in Willis and Rosen (1979) for a single educational transition (high school to college). Similarly, we know that low ability individuals have high direct returns from graduating from high school, but low or even negative returns from college. Spearman rank correlations agree with Pearson correlations. This is documented in Web Appendix A.13, which also shows the patterns for other outcomes. It also shows correlations far from 1 and sometimes negative.\(^{59}\)

**Table 2: Pearson Correlation in Treatment Effects: Wages**

<table>
<thead>
<tr>
<th>Pearson:</th>
<th>HS-DO</th>
<th>SC-HS</th>
<th>COLL-SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>High School vs Dropout</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some College vs High School</td>
<td>-0.6622</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>College Grad. vs Some College</td>
<td>0.2946</td>
<td>-0.5375</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Notes: Table shows the correlation between \(\rho_i\) and \((Y_j - Y_{j-1})\) for \(j \in \{1, 2, 3\}\). Estimates use the full population from simulations of our model.

The estimates from our model generally do not support Mincer specification (1). Sorting by gains differs across education levels. After adjusting for ability and sorting bias, there is no strong evidence supporting linearity of annualized ATEs across schooling levels.\(^{60}\)

### 6.5 Decomposing the Correlation Between \(\rho\) and \(S\): Are Those Who Go to School the Ones Who Benefit from It?

The correlation between \(\rho_i\) and \(S_i\) is a measure of the quality of sorting of people into schooling by their gain from it—a topic Becker investigated in depth in his Woytinsky lecture.

---

\(^{59}\)Web Appendix A.13 shows comparable patterns for dynamic treatment effects.

\(^{60}\)See Web Appendix A.12 especially Table A.12 and Figure A13. The tests and plot reported in that Appendix use the strong assumption that \(q_j - q_{j-1}\) is the same for everyone moving from \(j-1\) to \(j\) (imputing sample means to all) when in truth, it varies greatly among individuals. The regressions underlying Figure 2 use actual values of years of schooling completed and is more accurate as a measure of actual schooling attainment.
We have already established that the distributions of returns differ across schooling levels and the returns across schooling levels are far from perfectly correlated. It is thus of interest to push our analysis a bit further and investigate the correlation of annualized returns with attained schooling levels. Table 3 shows the correlations between educational choices and the node-specific annualized terminal gains \((\frac{Y_j - Y_{j-1}}{q_j - q_{j-1}})\) as well as the overall correlation.\(^{61}\)

The correlations between \(\rho\) and \(S\) are shown in column 1. Columns 2 through 4 show the correlations between the individual treatment effects \(\rho_j\) and choices at node \(D_j\). For columns 2 through 4, each correlation is estimated conditional on the population that makes it to the specific decision \((Q_j = 1)\).

**Table 3: Correlation Between Annualized Returns and Educational Choices**

<table>
<thead>
<tr>
<th></th>
<th>(Corr(\rho, S))</th>
<th>(Corr(\rho_1, (1 - D_1)))</th>
<th>(Corr(\rho_2, 1 - D_2))</th>
<th>(Corr(\rho_3, (1 - D_3)))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Wage</strong></td>
<td>0.069</td>
<td>0.011</td>
<td>-0.041</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.061)</td>
<td>(0.002)</td>
<td>(0.018)</td>
</tr>
<tr>
<td><strong>PVwage</strong></td>
<td>-0.080</td>
<td>-0.193</td>
<td>-0.068</td>
<td>0.084</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.034)</td>
<td>(0.023)</td>
<td>(0.021)</td>
</tr>
<tr>
<td><strong>Smoking</strong></td>
<td>-0.082</td>
<td>0.202</td>
<td>-0.110</td>
<td>-0.034</td>
</tr>
<tr>
<td></td>
<td>(0.069)</td>
<td>(0.171)</td>
<td>(0.030)</td>
<td>(0.063)</td>
</tr>
<tr>
<td><strong>Health Limits Work</strong></td>
<td>-0.102</td>
<td>-0.225</td>
<td>0.227</td>
<td>-0.065</td>
</tr>
<tr>
<td></td>
<td>(0.064)</td>
<td>(0.029)</td>
<td>(0.022)</td>
<td>(0.056)</td>
</tr>
</tbody>
</table>

**Notes:** Let \(q_j\) be the years of schooling associated with node \(j\). The annualized terminal node \(j\) return is \(\rho_j := \frac{Y_j - Y_{j-1}}{q_j - q_{j-1}}\) and we define \(\rho := \frac{Y_j - Y_0}{q_j - q_0}\). Total years of schooling is \(S = \sum_{j=1}^{T} q_j D_j\). Note \(D_1 = 1\) if individuals stop their education as a high school graduate. \(D_2 = 1\) and \(D_3 = 1\) denote stopping at some college and college respectively. Standard errors are estimated using 200 bootstrap samples and show the standard deviation of the estimate across the samples.

The overall correlation and the correlation by node differ substantially. The general pattern is that for wages, people sort on terminal gains although the effect is only strong for graduating college, and for most outcomes it is perverse for some college. The sorting

---

\(^{61}\)Precise definitions are given at the base of Table 3.
is negative for PV wages, except for college graduation. For smoking, the overall effect is negative, but is positive for high school graduation. For health limits work, the correlations differ but are negative except for the anomalous correlation for some college.

Table A52 in the Web Appendix reports a version of Table 3 using annualized dynamic treatment effects ($T_j$) inclusive of continuation values. It is more likely that these correspond to the gains that agents act on in making their transition decisions. Accounting for standard errors, the patterns are similar across the two tables. The correlations are consistently negative for smoking across all transitions. The strongest negative correlation for health limits work is for high school graduation. The correlation with some college is anomalous. Using either terminal level treatment effects or dynamic treatment effects sorting is generally positive, broadly consistent with a meritocratic society.

7 Policy Simulations from our Model: Interpreting what IV Estimates

Using our model, it is possible to conduct a variety of counterfactual policy simulations, a feature not shared by standard treatment effect models. We achieve these results without imposing strong assumptions on the choice model. In addition, we can interpret what IV estimates — the gains and the characteristics of the subpopulations affected. We consider three policy experiments: (1) a tuition subsidy; (2) an increase in the cognitive or non-cognitive endowments of those at the bottom of the endowment distribution; and (3) a policy experiment that simulates the effect of opening colleges in the county on college attendance. The first policy experiment is similar to what is estimated by LATE only in the special case where the instrument corresponds to the exact policy experiment. The second policy experiment is of interest because early childhood programs boost these endowments (Heckman, Pinto, and Savelyev, 2013). The counterfactuals generated cannot be estimated

\footnote{The category is a catch all for those who attend college for remedial education, those who seek certificates, those getting associate’s degrees and dropouts from 4 year college programs.}
by instrumental variable methods. The third policy experiment uses our model to interpret the components of what IV estimates. We ignore general equilibrium effects in all of these simulations.

### 7.1 Policy Relevant Treatment Effects

Unless the instruments correspond to policies, IV does not identify policy relevant treatment effects. The PRTE allows us to identify who would be induced to change educational choices under specific policy changes, and how these individuals would benefit on average. As an example, we simulate the response to a policy intervention that provides a one standard deviation subsidy to early college tuition (approximately $850 dollars per year). Column 1 of Table 4 presents the average treatment effect (including continuation values) in our estimated model for those who are induced to change education levels by the tuition subsidy. Since tuition at age 17 only enters the choice to enroll in college, the subsidy only induces high school graduates to change their college enrollment decisions. Those induced to enroll may then go on to graduate with a four year degree. Columns 2 and 3 of Table 4 decompose the PRTE into the average gains for those induced to enroll and then go on to earn 4 year degrees and the average gains for those who do not. For the most part, the PRTE is larger for those who go on to earn 4 year degrees.

Figure 9 shows which individuals are induced to enroll in college within the deciles of the distribution of the unobservable in the choice equation for node 2, conditional on $Q_2 = 1$ (the node determining college enrollment). These are the unobserved components of heterogeneity acted upon by the agent but unobserved by the economist.

The policy induces some individuals at every decile to switch, but places more weight on those in the middle deciles of the distribution. The figure further decomposes the effect of

---

63 Models were estimated that include tuition as a determinant of the high school graduation decision. However, estimated effects of tuition on high school graduation are small and statistically insignificant. We do not impose the requirement that future values of costs affect current educational choices. This highlights the benefits of our more robust approach.

64 $(\eta_2 = \theta' \alpha_2 - \nu_2)$
those induced to switch into the effect for those who go on to graduate with four year degrees and the effect for those who do not. Those induced to switch in the top deciles are more likely to go on to graduate.

Table 4: PRTE: Standard Deviation Decrease in Tuition

<table>
<thead>
<tr>
<th></th>
<th>PRTE</th>
<th>4-year degree</th>
<th>no 4-year degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log Wages</td>
<td>0.125 (0.023)</td>
<td>0.143 (0.027)</td>
<td>0.114 (0.027)</td>
</tr>
<tr>
<td>PV Log Wages</td>
<td>0.129 (0.03)</td>
<td>0.138 (0.033)</td>
<td>0.123 (0.028)</td>
</tr>
<tr>
<td>Health Limits Work</td>
<td>-0.036 (0.022)</td>
<td>-0.025 (0.021)</td>
<td>-0.043 (0.023)</td>
</tr>
<tr>
<td>Smoking</td>
<td>-0.131 (0.029)</td>
<td>-0.166 (0.030)</td>
<td>-0.108 (0.030)</td>
</tr>
</tbody>
</table>

Notes: Table shows the policy relevant treatment effect (PRTE) of reducing tuition for the first two years of college by a standard deviation (approx. $850). The PRTE is the average treatment effect of those induced to change educational choices as a result of the policy:

\[ PRTE_{k_{p',p}} := \int \int \int E(T^k(p) - T^k(p')|X = x, Z = z, \theta = \theta, S(p) \neq S(p')) dF_{X, Z, \theta}(x, z, \theta|S(p) \neq S(p')) \]

Column 1 shows the overall PRTE. Column 2 shows the PRTE for those induced to enroll by the policy who then go on to complete 4-year college degrees. Column 3 shows the PRTE for individuals induced to enroll but who do not complete 4-year degrees.

Figure 9: PRTE: Who is induced to switch

Notes: The figure plots the proportion of individuals induced to switch from the policy that lay in each decile of \( \eta_2 \), where \( \eta_2 = \theta_1 a_2 - \nu_2 \). \( \eta_2 \) is the unobserved component of the educational choice model. The deciles are conditional on \( Q_2 = 1 \), so \( \eta_2 \) for individuals who reach the college enrollment decision. The bars are further decomposed into those that are induced to switch that then go on to earn 4-year degrees and those that are induced to switch but do not go on to graduate.

The $850 subsidy induces 12.8% of high school graduates who previously did not attend college to enroll in college. Of those induced to enroll, more than a third go on to graduate.
with a 4 year degree. For outcomes such as smoking, the benefits are larger for those who
graduate with a 4-year degree. The large gains for marginal individuals induced to enroll is
consistent with the literature that finds large psychic costs are necessary to justify why more
individuals do not attend college.

Using the estimated benefits, we can determine if the monetary gains in the present value
of wages at age 18 is greater than the $850 subsidy.\footnote{However, a limitation of our model is that we can only estimate the monetary costs and do not estimate psychic costs.} Given a PRTE of 0.13 for log present
value of wage income, the average gains for those induced to enroll is $36,401. If the subsidy
is given for the first two years of college, then the policy clearly leads to monetary gains for
those induced to enroll. If the subsidy is also offered to those already enrolled, the overall
monetary costs of the subsidy is much larger because it is given to more than 8 students
previously enrolled for each new student induced to enroll (dead weight).

### 7.2 Boosting Cognitive and Noncognitive Endowments

Using simulation, it is possible to conduct counterfactual policy simulations unrelated to
any particular set of instruments. For example, some early childhood programs have been
shown to have lasting impacts on the cognitive or non-cognitive endowments of low ability
children (see \textit{Heckman, Pinto, and Savelyev, 2013}). We simulate two policy experiments:
(1) increasing the cognitive endowment of those in the lowest decile and (2) increasing the
non-cognitive endowment of those in the lowest decile.\footnote{The details of how these simulations were conducted are presented in Web Appendix A.14. Our model
does not address general equilibrium effects of such a change in the endowment distribution.}

The panels of Figure 10 show the average gains for increasing the cognitive or non-cognitive
endowment of those in the lowest decile of each ability. Increased cognition helps individuals
across the board. Increasing socio-emotional endowments has a smaller effect on labor market
outcomes but substantial effects on health.
Figure 10: Policy Experiments

9A. Policy Experiment: Log Wages (improving bottom decile of skills)

9B. Policy Experiment: Log PV Wages (improving bottom decile of skills)

9C. Policy Experiment: Health Limits Work (improving bottom decile of skills)

9D. Policy Experiment: Daily Smoking (improving bottom decile of skills)

Notes: This plot shows the average gains for those in the bottom deciles of cognitive ability (left) and socioemotional ability (right), from an increase in the endowment.

7.3 Introducing Colleges into Localities: Interpreting What IV Estimates

We use simulations from our model to interpret what LATE estimates in a model with multiple schooling transitions when agents are subject to changes in the constraints facing them. Let $S(z)$ denote the final level of schooling selected when $Z = z$. As before, $q_j$ is the years of schooling associated with attainment at node $j$. Angrist and Imbens (1995) show
that for a binary valued instrument \((Z \in \{z_1, z_2\})\) assuming that \(S(z_2) > S(z_1)\),

\[
\text{LATE}(z_1, z_2) = \sum_{j=2}^{3} E(\rho_j | S(z_2) \geq q_j > S(z_1)) W_j
\]

where \(W_j = \frac{Pr(S(z_2) \geq q_j > S(z_1))}{\sum_{k=2}^{3} (q_k - q_{k-1}) Pr(S(z_2) \geq q_k > S(z_1))}\) and \(E(\rho_j | S(z_2) \geq q_j > S(z_1))\) is an annualized “causal effect” of transiting from \(j - 1\) to node \(j\) for people who transit from whatever node they start at when \(Z = z_1\) to wherever node the instrument \(Z = z_2\) shifts them to.

\(\text{LATE}(z_1, z_2)\) includes the mean gain of going from \(j - 1\) to \(j\) for people who pass through \(j\) even though they may end up far above \(j\) and start far below it. It may differ across \(j\) because the distributions of the returns \((\rho_j)\) are different across conditioning sets and because of the intrinsic nonlinearity in the returns.

Consider the college-in-county indicator used in the IV analysis of Kling (2001). Let \(Z = z_1\) when no college is present in the county and \(Z = z_2\) when a college is present.\(^{67}\) This instrument affects college enrollment and college graduation decisions. For a binary instrument that affects only the final two educational choices, there are three different affected sub-populations: (1) those who are induced to enroll in college but do not graduate, (2) those induced to enroll in college and who go on to graduate, and (3) those who previously enrolled in college and who would not have graduated who are induced to graduate by the policy.

Simulating our model, we decompose the Wald estimator for the effect of schooling on log PV wages into the expected gains at each transition for each sub-population as well as the weights \(W_j, j = 2, \ldots, \bar{s}\). LATE is not the return to any particular population at any particular margin, but rather a weighted average of returns to a year of schooling for each of the affected sub-populations across different transitions. It weights the expected gains differentially. LATE depends on which sub-populations are induced to change schooling levels. Given that returns to schooling vary by educational stage and individual characteristics, the margins and sub-populations a specific instrument affects directly impacts the estimated

\(^{67}\)Oreopoulos and Salvanes (2011) present a summary of papers using presence of a local college in a county as an instrument.
returns to a year of schooling produced by LATE.

**Table 5: Decomposing the LATE Estimator of the Effect of Education on Log PV Wages**

<table>
<thead>
<tr>
<th>Group</th>
<th>Return</th>
<th>Weight</th>
<th>Δq_j</th>
<th>Avg Yearly Return</th>
</tr>
</thead>
<tbody>
<tr>
<td>S(z_1) = 1, S(z_2) = 2</td>
<td>E[Y_2 - Y_1] = 0.087</td>
<td>W_{1,2} = 0.289</td>
<td>1.790</td>
<td>0.049</td>
</tr>
<tr>
<td>S(z_1) = 1, S(z_2) = 3</td>
<td>E[Y_2 - Y_1] = 0.039</td>
<td>W_{1,3} = 0.204</td>
<td>1.790</td>
<td>0.022</td>
</tr>
<tr>
<td>S(z_1) = 1, S(z_2) = 3</td>
<td>E[Y_3 - Y_2] = 0.150</td>
<td>W_{1,3} = 0.341</td>
<td>2.988</td>
<td>0.050</td>
</tr>
<tr>
<td>S(z_1) = 2, S(z_2) = 3</td>
<td>E[Y_3 - Y_2] = 0.169</td>
<td>W_{2,3} = 0.166</td>
<td>2.988</td>
<td>0.057</td>
</tr>
</tbody>
</table>

Total=1 \[ \rho_{LATE} = 0.045 \]

*Notes:* The LATE estimator is the weighted sum of returns. The “Avg Yearly Return”, is the average return per year of schooling for that specific sub population \( E_\omega(\rho_j) = \frac{E_\omega(Y_j - Y_{j-1})}{(q_j - q_{j-1})} \). The Return column shows the expected return for the specific group defined in the Group column. \( \Delta q_j \) is the difference in the average number of years of schooling for individuals transiting from \( j - 1 \) to \( j \). Of those induced to change schooling levels by the instrument, 48.7% moved from high school graduate to some college, 34.5% moved from high school graduate to college graduate, and 16.8% moved from some college to college graduate.

The subpopulations induced to switch by the change in the instrument are very different. Table 6 shows how the average levels of cognitive and socio-emotional endowments and other background characteristics differ across the sub-populations induced to change final schooling levels by the college introduction instrument.\(^{68}\) The columns show the three different populations induced to switch by changing \( Z \). These subpopulations differ substantially in their average levels of cognitive and non-cognitive endowments. The marginal gains differ across transitions. The LATE estimator does not identify the gains to improving access for enrollment in college \((1 \rightarrow 2)\), but a composite effect across multiple margins.

\(^{68}\) For a full table, see Web Appendix Table A6.
Table 6: Means of Observables and Endowments for Groups that Constitute the LATE Decomposition

<table>
<thead>
<tr>
<th></th>
<th>Full</th>
<th>HS to Some Coll (1 → 2)</th>
<th>HS to Coll Grad (1 → 3)</th>
<th>Some Coll to Coll Grad (2 → 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive</td>
<td>0.001</td>
<td>-0.006</td>
<td>0.286</td>
<td>0.321</td>
</tr>
<tr>
<td></td>
<td>(0.679)</td>
<td>(0.605)</td>
<td>(0.579)</td>
<td>(0.570)</td>
</tr>
<tr>
<td>Non-Cognitive</td>
<td>0.001</td>
<td>0.032</td>
<td>0.249</td>
<td>0.229</td>
</tr>
<tr>
<td></td>
<td>(0.652)</td>
<td>(0.596)</td>
<td>(0.566)</td>
<td>(0.567)</td>
</tr>
<tr>
<td>Black</td>
<td>0.117</td>
<td>0.116</td>
<td>0.083</td>
<td>0.096</td>
</tr>
<tr>
<td></td>
<td>(0.322)</td>
<td>(0.320)</td>
<td>(0.275)</td>
<td>(0.295)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.067</td>
<td>0.067</td>
<td>0.044</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>(0.250)</td>
<td>(0.250)</td>
<td>(0.204)</td>
<td>(0.231)</td>
</tr>
<tr>
<td>Broken Home</td>
<td>0.241</td>
<td>0.235</td>
<td>0.157</td>
<td>0.197</td>
</tr>
<tr>
<td></td>
<td>(0.428)</td>
<td>(0.424)</td>
<td>(0.364)</td>
<td>(0.397)</td>
</tr>
</tbody>
</table>

Notes: Mean coefficients; sd in parentheses

* p < 0.10, ** p < 0.05, *** p < 0.01

8 Robustness and Comparison of Our Estimates with those Obtained from Other Methods

This section examines the robustness of some of the key assumptions maintained in this paper. It also examines whether simpler methods can be used to obtain average treatment effects. We first test the robustness of our model to relaxing key assumptions. We then consider whether it is possible to obtain reliable estimates using conventional methods in the treatment effect literature. We go on to compare our estimates of AMTE and PRTE with those obtained from the standard set of instruments used in this paper.

8.1 Testing the Two Factor Assumption

Throughout this paper, we have assumed that selection of outcomes occurs on the basis of a two component vector \( \theta \), where the components can be proxied by our measures of cognitive and noncognitive endowments. An obvious objection is that there may be unproxied
endowments that affect both choices and outcomes that we do not measure.

Cunha and Heckman (2015) estimate a related model using the same data source. They find that a three factor model explains wages and present value of wages. Two of their factors correspond to the factors used in this paper. Their third factor improves the fit of the wage outcome data but does *not* enter agent decision equations or affect selection or sorting bias. Our results are consistent with these findings.

In order to test for the presence of a third factor that influences both choices and outcomes, we test whether the simulated model fits the sample covariances between $Y^k$ and $D_j, j = 1, \ldots, \bar{s}, k = 1, \ldots, 4$. If an important third factor common to both outcome and choice equations has been omitted, the agreement should be poor. In fact, we find close agreement.\footnote{See Web Appendix A.15 and Table A38.} Like Cunha and Heckman (2015), we find that adding a third factor that appears in outcome equations but not choice equations improves goodness of fit, but has no effect on our estimated treatment effects.\footnote{See Web Appendix A.15.}

### 8.2 Comparisons with Alternative Treatment Effect Estimators

Throughout this paper we have exploited the assumption of conditional independence of outcomes and choices given $X, Z, \theta$. This raises the question of how similar our results would be if we had used simple matching methods that also control for $\theta, X, Z$.\footnote{OLS is a version of matching.} We estimate Bartlett factor scores based on our measures using standard statistical software.\footnote{See Bartlett (1937, 1938).} Using these extracted factors, we estimate average treatment effects using (a) Linear regression (with and without factors) and (b) Matching. Table 8 presents our estimates. All models are estimated for individuals who attain each decision node ($Q_j = 1$) and include those who may go on to attain further education in order to make the alternative models comparable to our ATE estimate that includes continuation values.

The first three columns show estimates from linear models. The first two columns
introduce schooling by using dummy shifts in intercepts. The first column uses measures of cognitive and noncognitive endowments directly, while the second column includes the extracted factor scores. The third column allows loadings on covariates and factor scores to vary by schooling level.\textsuperscript{73}

The fourth and fifth columns show estimates from matching using the Bartlett factor scores previously described as well as an index of covariates. The fourth column shows results from nearest neighbor matching using the 3 nearest neighbors. The fifth column shows results from propensity score matching using the estimated probability of the educational decision as the propensity score.\textsuperscript{74}

The estimates differ greatly from the OLS estimates obtained without any adjustment for $\theta$. Controlling for ability has substantial effects on the estimated average treatment effects. Across schooling nodes, all of the other estimates are roughly “within the ball park” of the estimates produced from our model, provided that we control for $\theta$. This is good news for applied economists mainly interested in using simple methods to estimate average treatment effects. However, these simple methods are powerless in estimating AMTE and PRTE or answering many of the other questions addressed in this paper.\textsuperscript{75}

8.3 Comparison of Estimates from Our Model with Those from IV

Table 8 compares the AMTE and PRTE parameters generated from our model with standard estimates based on the widely used instruments of Table 1. The IV estimates are not close to our estimates of the interpretable policy relevant treatment effects or average marginal treatment effects. The IV estimates vary substantially based on which instruments we use and

\textsuperscript{73}See Section A.16.3 for details on these estimators.

\textsuperscript{74}Precise specifications of the estimating equations are presented in Web Appendix A.16.

\textsuperscript{75}Table A73 of the Web Appendix compares OLS estimates of direct effects and continuation values with our model estimates. The OLS estimates are “within ballpark” for smoking and health limits work, but they are wide of the mark for wages and PV wages.
Table 7: Average Treatment Effects - Comparison of Estimates from Our Model to Those from Simpler Methods

<table>
<thead>
<tr>
<th></th>
<th>Linear</th>
<th>Matching</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>OLS-P</td>
<td>OLS-F</td>
<td>RA-F</td>
<td>NNM(3)-F</td>
<td>PSM-F</td>
<td>ATE</td>
</tr>
<tr>
<td>HS Grad.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wages</td>
<td>0.205</td>
<td>0.073</td>
<td>0.155</td>
<td>0.159</td>
<td>0.106</td>
<td>0.127</td>
<td>0.094</td>
</tr>
<tr>
<td>SE</td>
<td>(0.025)</td>
<td>(0.026)</td>
<td>(0.025)</td>
<td>(0.035)</td>
<td>(0.039)</td>
<td>(0.049)</td>
<td>(0.056)</td>
</tr>
<tr>
<td>PV-Wage</td>
<td>0.380</td>
<td>0.213</td>
<td>0.318</td>
<td>0.277</td>
<td>0.203</td>
<td>0.227</td>
<td>0.173</td>
</tr>
<tr>
<td>SE</td>
<td>(0.030)</td>
<td>(0.031)</td>
<td>(0.030)</td>
<td>(0.041)</td>
<td>(0.053)</td>
<td>(0.055)</td>
<td>(0.059)</td>
</tr>
<tr>
<td>Smoking</td>
<td>-0.327</td>
<td>-0.246</td>
<td>-0.281</td>
<td>-0.301</td>
<td>-0.287</td>
<td>-0.286</td>
<td>-0.263</td>
</tr>
<tr>
<td>SE</td>
<td>(0.028)</td>
<td>(0.029)</td>
<td>0.028</td>
<td>0.041</td>
<td>0.056</td>
<td>0.068</td>
<td>(0.056)</td>
</tr>
<tr>
<td>Health-Limits-Work</td>
<td>-0.178</td>
<td>-0.115</td>
<td>-0.151</td>
<td>-0.150</td>
<td>-0.037</td>
<td>-0.124</td>
<td>-0.108</td>
</tr>
<tr>
<td>SE</td>
<td>(0.022)</td>
<td>(0.024)</td>
<td>(0.023)</td>
<td>(0.033)</td>
<td>(0.031)</td>
<td>(0.044)</td>
<td>(0.042)</td>
</tr>
<tr>
<td>Coll. Enroll</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wages</td>
<td>0.223</td>
<td>0.121</td>
<td>0.186</td>
<td>0.190</td>
<td>0.185</td>
<td>0.225</td>
<td>0.134</td>
</tr>
<tr>
<td>SE</td>
<td>(0.023)</td>
<td>(0.024)</td>
<td>(0.024)</td>
<td>(0.023)</td>
<td>(0.031)</td>
<td>(0.032)</td>
<td>(0.025)</td>
</tr>
<tr>
<td>PV-Wage</td>
<td>0.221</td>
<td>0.109</td>
<td>0.176</td>
<td>0.171</td>
<td>0.198</td>
<td>0.238</td>
<td>0.137</td>
</tr>
<tr>
<td>SE</td>
<td>(0.027)</td>
<td>(0.029)</td>
<td>(0.028)</td>
<td>(0.027)</td>
<td>(0.029)</td>
<td>(0.031)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>Smoking</td>
<td>-0.177</td>
<td>-0.138</td>
<td>-0.165</td>
<td>-0.170</td>
<td>-0.174</td>
<td>-0.175</td>
<td>-0.139</td>
</tr>
<tr>
<td>SE</td>
<td>(0.026)</td>
<td>(0.028)</td>
<td>(0.027)</td>
<td>(0.028)</td>
<td>(0.032)</td>
<td>(0.033)</td>
<td>(0.028)</td>
</tr>
<tr>
<td>Health-Limits-Work</td>
<td>-0.085</td>
<td>-0.037</td>
<td>-0.066</td>
<td>-0.057</td>
<td>-0.055</td>
<td>-0.056</td>
<td>-0.037</td>
</tr>
<tr>
<td>SE</td>
<td>(0.020)</td>
<td>(0.022)</td>
<td>(0.021)</td>
<td>(0.021)</td>
<td>(0.022)</td>
<td>(0.037)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>Coll. Grad</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wages</td>
<td>0.210</td>
<td>0.146</td>
<td>0.184</td>
<td>0.185</td>
<td>0.176</td>
<td>0.108</td>
<td>0.114</td>
</tr>
<tr>
<td>SE</td>
<td>(0.032)</td>
<td>(0.034)</td>
<td>(0.033)</td>
<td>(0.035)</td>
<td>(0.039)</td>
<td>(0.061)</td>
<td>(0.037)</td>
</tr>
<tr>
<td>PV-Wage</td>
<td>0.243</td>
<td>0.163</td>
<td>0.208</td>
<td>0.228</td>
<td>0.207</td>
<td>0.251</td>
<td>0.171</td>
</tr>
<tr>
<td>SE</td>
<td>(0.037)</td>
<td>(0.040)</td>
<td>(0.038)</td>
<td>(0.037)</td>
<td>(0.040)</td>
<td>(0.042)</td>
<td>(0.040)</td>
</tr>
<tr>
<td>Smoking</td>
<td>-0.209</td>
<td>-0.171</td>
<td>-0.195</td>
<td>-0.192</td>
<td>-0.145</td>
<td>-0.157</td>
<td>-0.172</td>
</tr>
<tr>
<td>SE</td>
<td>(0.032)</td>
<td>(0.035)</td>
<td>(0.033)</td>
<td>(0.035)</td>
<td>(0.040)</td>
<td>(0.042)</td>
<td>(0.043)</td>
</tr>
<tr>
<td>Health-Limits-Work</td>
<td>-0.085</td>
<td>-0.069</td>
<td>-0.078</td>
<td>-0.077</td>
<td>-0.047</td>
<td>-0.059</td>
<td>-0.064</td>
</tr>
<tr>
<td>SE</td>
<td>(0.024)</td>
<td>(0.026)</td>
<td>(0.025)</td>
<td>(0.026)</td>
<td>(0.030)</td>
<td>(0.028)</td>
<td>(0.031)</td>
</tr>
</tbody>
</table>

Notes: We estimate the ATE for each outcome and educational choice using a number of methods. All models are estimated for the population who reaches the choice being considered \( (Q_j = 1) \), inclusive of those who then may go on to further schooling in order to make them comparable to the ATE from our model that includes continuation value. All models use the full set of controls listed in Table 1. “OLS” estimates a linear model using a schooling dummy, and controls \( (Y = D_j + X\beta + \epsilon) \). “OLS-P” estimates a linear model using a schooling dummy, controls, the sum of the ASVAB scores used, gpa, and an indicator of risky behavior \( (Y = D_j + X\beta + A\gamma + \epsilon) \), where \( A \) are the proxies for cognitive and socio-emotional endowments. All models ending in “-F” are estimated using Bartlett factor scores \( (Bartlett (1937, 1938)) \) estimated using our measurement system, but assuming a bivariate normal distribution and not accounting for schooling at the time of the test. “OLS-F” estimates the model \( (Y = D_j + X\beta + \hat{\theta}\gamma + \epsilon) \) where \( \hat{\theta} \) are the Bartlett factor scores described above. “RA-F” extends OLS-F by letting the loadings on the covariates and factors vary by schooling level as described in Web Appendix A.16.3. “NNM(3)-F” presents the estimated treatment effect of nearest-neighbor matching with 3 neighbors. Neighbors are matched on their Bartlett cognitive factor, Bartlett non-cognitive factor, and an index constructed from their observable characteristics as described in Web Appendix A.16.1. “PSM-F” presents the estimated average treatment effect from propensity score matching, using the Bartlett cognitive factor, Bartlett non-cognitive factor, and control variables as described in Web Appendix A.16.2. “ATE” presents the estimated average treatment effect from the model presented in this paper. See Table A54 for additional comparisons.
which covariates are controlled for. Overall, IV estimates do not produce the population average treatment effects, or the treatment effect of those at the margin of indifference for a particular educational choice.

### Table 8: Comparing Treatment Effects

<table>
<thead>
<tr>
<th></th>
<th>Graduate High School</th>
<th>Enroll in College</th>
<th>Graduate College</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Wages</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMTE</td>
<td>0.093** (0.028)</td>
<td>0.101** (0.023)</td>
<td>0.110** (0.034)</td>
</tr>
<tr>
<td>PRTE</td>
<td>na</td>
<td>0.114** (0.027)</td>
<td>0.143** (0.027)</td>
</tr>
<tr>
<td>IV</td>
<td>-0.761 (0.678)</td>
<td>0.511* (0.227)</td>
<td>0.025 (0.546)</td>
</tr>
<tr>
<td><strong>PV Wages</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMTE</td>
<td>0.282** (0.041)</td>
<td>0.112** (0.031)</td>
<td>0.146** (0.042)</td>
</tr>
<tr>
<td>PRTE</td>
<td>na</td>
<td>0.123** (0.028)</td>
<td>0.138** (0.033)</td>
</tr>
<tr>
<td>IV</td>
<td>0.542 (0.659)</td>
<td>0.373 (0.233)</td>
<td>0.72 (0.640)</td>
</tr>
<tr>
<td><strong>Smoking</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMTE</td>
<td>-0.242** (0.033)</td>
<td>-0.131** (0.027)</td>
<td>-0.173** (0.038)</td>
</tr>
<tr>
<td>PRTE</td>
<td>na</td>
<td>-0.108 (0.030)</td>
<td>-0.166 (0.030)</td>
</tr>
<tr>
<td>IV</td>
<td>-1.311 (0.744)</td>
<td>-0.345 (0.215)</td>
<td>-0.377 (0.693)</td>
</tr>
<tr>
<td><strong>Health Limits Work</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMTE</td>
<td>-0.110** (0.031)</td>
<td>-0.029 (0.022)</td>
<td>-0.067* (0.030)</td>
</tr>
<tr>
<td>PRTE</td>
<td>na</td>
<td>-0.043 (0.023)</td>
<td>-0.025 (0.021)</td>
</tr>
<tr>
<td>IV</td>
<td>-0.560 (0.577)</td>
<td>0.081 (0.178)</td>
<td>-0.763 (0.664)</td>
</tr>
</tbody>
</table>

**Notes:** Standard errors (in parenthesis) and significance levels (*$p < 0.05$, **$p < 0.01$). No PRTE is estimated for the high school graduation decision because as the policy experiment considered is a tuition subsidy and college tuition does not determine the high school graduation decision in our model. “na” in the table above stands for not applicable. Each column presents the average effect of an educational decision. AMTE presents the average effect for those who are indifferent between choosing a higher level of schooling or not. PRTE presents the average gains from those induced to enroll in college or enroll and then graduate from college from a two-year one standard deviation decrease in tuition. The table also presents the estimated treatment effect from two stage least squares controlling for family background variables, factors, and decision-specific instruments by educational choice (“IV”). For graduating from high school the instruments are long run local unemployment rate and current local unemployment rate at 17. The choice to enroll in college includes long run local unemployment rate, current local unemployment rate at 17, presence of a 4-year college in the county, and local college tuition. The choice to graduate from college includes local long-run unemployment at age 22, current local unemployment at age 22, and local college tuition at age 22. The IV models include background controls and factors. The IV model uses two stage least squares where the endogenous covariate is the single schooling decision dummy and the regression is restricted to those who reach the choice. The first stage and second stage control for the standard set of covariates used in the paper (see Section A.8.7.2) as well as cognitive and socio-emotional factor scores. The first stage includes the decision specific instruments discussed in Section A.8.7.4. See Web Appendix Section A.16.4 for details on the construction of the IV estimator which is two stage least squares.

---

Heckman and Vytlacil (2007b) present a formal analysis of what different instruments identify for the binary case, an ordered choice model and a multinomial discrete choice model.
9 Summary and Conclusion

Gary Becker’s seminal research on human capital launched a large and active industry on estimating causal effects and rates of return to schooling. Multiple methodological approaches have been used to secure estimates ranging from reduced form treatment effect methods to fully structural methods. Each methodology has its benefits and limitations.

This paper develops and estimates a robust dynamic causal model of schooling and its consequences for earnings, health and healthy behaviors. The model recognizes the sequential dynamic nature of educational decisions. We borrow features from both the reduced form treatment effect literature and the structural literature. Our estimated model passes a variety of goodness of fit and model specification tests.

We allow agents to be irrational and myopic in making schooling decisions. Hence we can use our model to test some of the maintained assumptions in the dynamic discrete choice literature on schooling.

We use our dynamic choice model to interpret what IV estimates and to identify the sets of agents affected by various instruments. We estimate causal effects from multiple levels of schooling rather than the binary comparisons typically featured in the literature on treatment effects and in many structural papers.\textsuperscript{77}

By estimating a sequential model of schooling in a unified framework, we are able analyze the \textit{ex post} returns to education for people at different margins of choice and analyze a variety of interesting policy counterfactuals. We are able to characterize who benefits from education across a variety of market and nonmarket outcomes.

The early literature on human capital ignored the dynamics of schooling choices. We decompose these benefits into direct components and indirect components arising from continuation values. We estimate substantial continuation value components especially for high ability individuals. For them, schooling opens up valuable options for future schooling. Standard estimates of the benefits of education based only on direct components underestimate

\footnotesize\textsuperscript{77}See, e.g., Willis and Rosen (1979).
the full benefits of education.

Without imposing rationality, we nonetheless find evidence consistent with it. We find positive sorting into schooling based on gains, especially for higher schooling levels. Schooling has strong causal effects on earnings, health and healthy behaviors even though we also find strong evidence of ability bias at all levels of schooling. Both cognitive and noncognitive endowments affect schooling choices and outcomes for each level of schooling.

We test numerous implications of the Becker - Chiswick - Mincer model (1) and find substantial evidence against it. The distributions of annualized returns differ across schooling levels. The returns are also far from perfectly correlated across schooling transitions and some are even negatively correlated. The correlation of returns and schooling attained differs greatly across schooling levels. There is heterogeneity in the effects of schooling across levels of schooling for the same person and heterogeneity in the effects of extra schooling across persons.

We link the structural and matching literatures using conditional independence assumptions. We investigate how simple methods used in the treatment effect literature perform in estimating average treatment effects. They roughly approximate our model estimates provided analysts condition on endowments of cognitive and noncognitive skills, and correct for measurement error in the proxies. However, these simple methods do not identify the treatment effects for persons at the margins of different choices, the policy relevant treatment effects, or the continuation values analyzed in this paper. Estimates from IV model are very different from the economically interpretable and policy relevant estimates produced from our model.

We deepen understanding of LATE by identifying the groups affected by variations in instruments and policies. We compare our estimates to those estimated by LATE and use our model to decompose the LATE estimator into economically interpretable components.

Our analysis is broadly consistent with the pioneering analysis of Becker (1964) but enriches it. The early research on human capital was casual about agent heterogeneity. It
ignored selection bias and comparative advantage in schooling. We quantify the magnitude and sources of selection bias. We find evidence of both ability bias and sorting bias (comparative advantage). Nonetheless, we find strong causal effects of education at most margins for most outcomes.

Our findings thus support the basic insights of Becker (1964). Schooling has strong causal effects on market and nonmarket outcomes. Both cognitive and noncognitive endowments affect schooling choices and outcomes. People tend to sort into schooling based on gains.
References


