Human Capital and Economic Opportunity Global Working Group

Working Paper Series

Working Paper 2016-014

Returns to Education: The Causal Effects of Education on Earnings, Health and Smoking

James J. Heckman
John Eric Humphries
Gregory Veramendi

May, 2016

Human Capital and Economic Opportunity Global Working Group
Economics Research Center
University of Chicago
1126 E. 59th Street
Chicago IL 60637
www.hceconomics.org

Returns to Education: The Causal Effects of Education on Earnings, Health and Smoking*

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

John Eric Humphries
University of Chicago

Gregory Veramendi
Arizona State University

May 19, 2016

*James J. 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 on an early draft. We also thank Ariel Pakes and other participants at a Harvard Labor Economics Workshop in April, 2014, for helpful comments on a previous draft. We thank Eleanor Dillon and Matthew Wiswall for comments received at a seminar at Arizona State University, February, 2015. We thank the special editor, Ed Lazear, and an anonymous referee for helpful comments. We also thank Jessica Yu Kyung Koh, Joshua Shea, Jennifer Pachon, and Anna Ziff for comments on this draft. 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, NICHD R01HD054702, and NIA R24AG048081; 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 returns to education using a dynamic model of educational choice that synthesizes approaches in the structural dynamic discrete choice literature with approaches used in the reduced form treatment effect literature. It is an empirically robust middle ground between the two approaches which estimates economically interpretable and policy-relevant dynamic treatment effects that account for heterogeneity in cognitive and non-cognitive skills and the continuation values of educational choices. Graduating college is not a wise choice for all. Ability bias is a major component of observed educational differentials. For some, there are substantial causal effects of education at all stages of schooling.

Keywords: education, earnings, health, rates of return, causal effects of education, cognitive skills, non-cognitive skills

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

James J. 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 analysis of human capital, Gary Becker (1962; 1964) emphasized the importance of the rate of return for evaluating the effectiveness of human capital investments. He launched an active industry estimating returns to schooling.[1]

At the time Becker crafted his analysis, modern economic dynamics was in its infancy, as was research on the economics of uncertainty in dynamic sequential models. In an early contribution, Burton Weisbrod (1962) noted that each year of schooling attained opened up options for additional schooling and training and provided opportunities for learning about personal abilities and life opportunities.[2]

A parallel development in empirical economics was the growing awareness of heterogeneity and diversity among individual cognitive and non-cognitive abilities.[3] Agents differ in their returns to schooling. Failure to account for this heterogeneity leads to confusion in interpreting estimated effects of schooling.

Becker's early work focused on internal rates of return that equated *ex post* discounted values of earnings streams net of monetary and psychic costs at different levels of education. He noted that the full return to schooling includes non-market benefits and non-pecuniary costs. In modern parlance, individuals should continue their schooling as long as their *ex ante* marginal return exceeds their *ex ante* marginal opportunity cost of funds.

Formidable empirical challenges arise in estimating *ex post* internal rates of return: lifetime earnings profiles are required; observed earnings profiles are subject to the selection bias that arises from the fact that earnings are observed only at schooling levels selected by agents; and quantifying non-market benefits and non-pecuniary costs is a difficult task. For estimating

[1] Becker (1964) also estimated rates of return. For surveys of this literature, see, e.g., Card (1999, 2001); Heckman et al. (2006a); Oreopoulos and Salvanes (2011); McMahon (2009); Oreopoulos and Petronijevic (2013).

[2] Weisbrod's paper stimulated research on the option value of schooling. See, e.g., Comay et al. (1973); Dothan and Williams (1981); Bamberger (1987); Altonji (1993); Cameron and Heckman (1993); Keane and Wolpin (1997); Arcidiacono and Miller (2011); Heckman et al. (2008); Stange (2012); Eisenhauer et al. (2015a).

[3] See Heckman (2001).

ex ante returns, information on how agents forecast future events is also required.

In a neglected paper, Becker and Chiswick (1966) developed a tractable framework for measuring *ex post* returns to schooling that utilizes cross-section synthetic cohort data on earnings to approximate life cycle earnings data.[4] Mincer (1974) improved on this model by adding work experience. The "Mincer Equation" has become the workhorse of the empirical literature on estimating *ex post* rates of return:

$$\ln Y(S_i, \boldsymbol{X}_i) = \gamma_i + \rho_i \underbrace{S_i}_{\substack{\text{years of} \\ \text{schooling}}} + \phi(\underbrace{\boldsymbol{X}_i}_{\substack{\text{other} \\ \text{determinants}}}) \tag{1}$$

where $Y(S_i, \boldsymbol{X}_i)$ is the earnings of individual i with S_i years of schooling and a vector of other determinants $\boldsymbol{X}_i$.

This equation is interpreted as a causal relationship generated by hypothetical variations of each of γ_i, ρ_i, and $\phi(\boldsymbol{X}_i)$, holding other components on the right-hand side of (1) fixed.[5] γ_i is what person i would earn independent of any influence of schooling $\boldsymbol{X}_i$. Correlation between γ_i and S_i is the source of "ability bias" (Griliches, 1977). Strictly speaking, γ_i may or may not be related to ability. It is a determinant of earnings that may also be correlated with S_i. ρ_i is the "return to a unit of schooling" for person i and is allowed to vary among individuals. It is a causal parameter realized by acquiring one more unit of schooling. There are both *ex ante* and *ex post* definitions of γ_i and ρ_i. The early literature and most of the empirical literature today focuses on estimating *ex post* returns.

This paper examines the economic foundations of Equation (1) and its generalizations accounting for the dynamics of educational decision-making and multidimensional heterogeneity in abilities among agents. We develop and estimate an empirically robust dynamic discrete choice model that allows for agent fallibility arising from imperfect information and learning, as well as time inconsistency. We allow agents to make schooling decisions based

[4] It is based on the assumption that the earnings of a person age a in a given cross section when that person turns $a'(> a)$ is well-approximated by the earnings of agents a' in that same cross section. This synthetic cohort assumption is standard in the literature.

[5] See Heckman (2008) and Heckman and Pinto (2015) for a discussion of causality and the role of fixing.

on expected future values. We test and reject strong forms of forward-looking behavior, but nonetheless find that agents sort on *ex post* gains.

We develop and estimate a variety of economically motivated and policy-relevant treatment effects. For most of the outcomes studied in this paper, we find strong evidence of ability bias at all levels of education, where ability includes both cognitive and non-cognitive skills, but only find sorting on gains (a relationship between ρ_i and S_i) at higher levels of schooling.

1.1 Interpreting Returns to Education

The Becker-Chiswick-Mincer Equation (1), and variants of it, have become the standard framework for estimating *ex post* returns to schooling for a variety of outcomes.[6] While ρ_i is not, in general, an internal rate of return for individual i, it is the *ex post* causal effect of increasing final schooling by exactly one year from any base state of schooling, holding γ_i and $\boldsymbol{X}_i$ fixed.[7] It is the slope of an hedonic wage function—the derivative of the aggregate production function evaluated at $S = s$ for a fixed γ_i and $\boldsymbol{X}_i$.

ρ_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. More generally, for a person at $s-1$, the perceived *ex ante* gain in log earnings of moving to schooling level s is the anticipated direct effect ρ_i and the (undiscounted) perceived continuation value of schooling for person i:

$$R_{s,i} = \rho_i + \underbrace{\rho_i \sum_{l=s+1}^{\overline{s}} P_{s,l,i}}_{\text{Continuation Value}}. \tag{2}$$

Under an *ex ante* interpretation, $P_{s,l,i}$ is the agent's perceived probability of attaining (at least) schooling level $S = l$ for a person starting at schooling level s, including any relevant

[6] See, e.g., Cutler and Lleras-Muney (2010) who apply model (1) to estimate the causal effect of education on health.

[7] The stringent conditions under which ρ_i is an internal rate of return, and evidence that they are not satisfied in many commonly used samples, are presented in Heckman et al. (2006a).

discounting of future benefits; $\overline{s}$ is the highest attainable value of S. $R_{s,i}$ captures Weisbrod's notion of valuing the future options that attaining schooling level s opens up.[8] It is the individual causal effect of an extra year of schooling inclusive of continuation values. As long as $\rho_i \neq 0$, it is distinct from $R_{s,i}$. One can define different versions of $R_{s,i}$ depending on how $P_{s,l,i}$ and ρ_i are specified.

Determining (2) poses major empirical challenges. There are multiple sources of heterogeneity in $R_{s,i}$. Individuals may differ in their values of ρ_i. Even if all people have the same ρ_i, they may differ in their expected anticipated probabilities of attaining schooling level s' $(P_{s,s',i},\ s' > s)$.[9]

The causal effects ρ_i and $R_{s,i}$ are formulated at the individual level. The modern treatment effect literature defines versions of these parameters for different groups and typically estimates *ex post* effects.[10] Thus, one can define the mean causal effect for the whole population $E(\rho)$.[11] Another possible causal effect is $E(R_s)$ defined for schooling level s for the entire population. One could also define the direct return to schooling for those who *choose* to be at a given level of schooling $E(\rho|S = s)$. This is the causal effect of one more unit of schooling for those who *stop* at $S = s$. One can define causal parameters for samples defined by other choices (e.g., for those indifferent between s and s'; for those who would stop at $s - 1$; etc.), and for different notions of returns, e.g., $E(R|S = s)$.

$E(\gamma|S = s)$ is the population mean γ arising solely from statistical dependence between γ and S. It has no causal basis and is the source of ability bias. Since dependence between γ and S may arise from multiple sources, we refer to "ability bias" as selection bias throughout much of this paper.

The early literature adopted a simple approach to identifying returns. It assumed that ρ_i

[8]Note, however, that the continuation value is different from the option value. See, e.g., Stange (2012) and Eisenhauer et al. (2015a).

[9]Rational expectations models assume that objectively measured probabilities are subjective probabilities. We do not impose this assumption in our analysis. For a survey of the expectation elicitation literature, see, e.g., Manski (2004).

[10]See Heckman (2008).

[11]See, e.g., Card (1999; 2001).

is identical for persons with the same observed characteristics. In this case, the only source of bias in estimating (1) is the statistical dependence between γ_i and S_i (selection bias). The recent literature recognizes heterogeneity in both γ_i and ρ_i. Both may be statistically dependent on S_i, giving rise to both selection bias and sorting on gains. The latter arises because the causal effect of S may be moderated by other variables. Whether or not sorting gains are a source of bias depends on the question being addressed.

To illustrate the importance of accounting for continuation values, consider a compulsory schooling policy that forces all persons to take a minimum level of schooling ($S \geq \underline{s}$). What causal effect is identified by this "natural experiment?" Abstracting from general equilibrium effects, any estimated treatment effect is defined conditional on the set of people who change their schooling from below $\underline{s}$ to at or above $\underline{s}$. However, there is no presumption that such agents will stop at $\underline{s}$ if they are forced to attain it. They may learn things about themselves and their possibilities, so they continue beyond $\underline{s}$ and thereby generate continuation values.[12] Thus, an experiment that evaluates the effects of this policy does not, in general, estimate $E(\rho)$ or even $E(\rho|S = \underline{s})$. It does not, in general, estimate the marginal effect of a change in S on the log marginal price of schooling.

The analysis just presented can be generalized to incorporate non-linear structural (causal) returns to schooling by allowing the ρ_i to depend on the origin and destination schooling states ($\rho_{s,s',i}$) for $s' > s$. Non-linearities associated with sheepskin effects associated with graduation are a potentially important source of continuation values.

1.2 Approaches to Identifying Causal Effects and Causal Rates of Return

Two general approaches have been developed to estimate returns to schooling in the general case. They are: (i) structural models that jointly analyze outcomes and schooling choices; and

[12] This is recognized in the LATE literature. See Angrist and Imbens (1995). What is not recognized in that literature is that LATE estimates the returns expected by agents only under a rational expectations assumption.

(ii) treatment effect models that use instrumental variables methods (including randomization and regression discontinuities as instruments) as well as matching on observed variables to identify causal parameters.[13]

The structural approach explicitly models agent decision rules that generate $P_{s,l,i}$ and the dependence between ρ_i and S_i. The modern version explicitly models agent expectations and distinguishes *ex ante* from *ex post* returns.[14] It uses a variety of sources of identification, including exclusion restrictions (instrumental variables), conditional independence assumptions about unobservables, and functional form assumptions (see, e.g., Blevins, 2014). Among other features, the structural approach identifies causal effects at well-defined margins of choice and can evaluate the impacts of different policies never previously implemented.[15]

The treatment effect approach is typically agnostic about agent decision rules and relies on exclusion restrictions to identify its estimands. It rarely distinguishes *ex ante* from *ex post* returns.[16] This approach is more transparent in securing identification than the structural approach.[17] However, the economic interpretation of its estimated parameters is often quite obscure. In a model with multiple levels of schooling, LATE typically does not identify returns at the various margins of choice that generate outcomes or the sub-populations (defined in terms of observables and unobservables) affected by the instruments used.[18] Its estimands do not identify a variety of well-posed policy questions except when the variation induced by the instruments corresponds closely to the variations induced by the policies of interest.[19]

We build on the analyses of Heckman and Vytlacil (1999, 2005, 2007a,b), Carneiro et al. (2010, 2011), and Eisenhauer et al. (2015b), who introduce choice theory into the modern

[13] See, e.g., Angrist and Imbens (1995) and Angrist and Pischke (2009) for IV, and Heckman et al. (1998) for matching.

[14] See, e.g., Keane and Wolpin (1997); Eisenhauer et al. (2015a).

[15] See Heckman (2010) and Heckman and Urzúa (2010).

[16] Eisenhauer et al. (2015b) distinguish and estimate *ex ante* and *ex post* returns in an instrumental variable model.

[17] The modern instrumental variables case requires assumptions about the validity of the instruments. If there are heterogeneous treatment effects, additional assumptions such as "monotonicity" (better termed uniformity) are required to interpret IV estimates. See Imbens and Angrist (1994); Heckman and Vytlacil (2005); Angrist and Pischke (2009) for details.

[18] See Heckman et al. (2006c) and Heckman et al. (2016) for a discussion.

[19] See Heckman (2010).

analysis of instrumental variables. They focus 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. We do not rely on continuous instruments. In addition, we link our analysis to the dynamic discrete choice literature.

1.3 Our Approach

This paper develops a methodological middle ground between the reduced form treatment approach and the fully structural dynamic discrete choice approach. As in 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 persons would be affected by alternative policies not previously implemented. As in 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 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, nor do we explicitly model agent expectations about costs and returns.[20]

Using a generalized Roy framework, we estimate a multistage sequential model of educational choices and their consequences. An important feature of our model is that educational choices at one stage open up educational options at later stages. Each educational decision is characterized using a flexible discrete choice model. The anticipated consequences of future choices and their costs can be assessed in a variety of ways by individuals in deciding whether or not to continue their schooling. Our model approximates a dynamic discrete choice model without taking a stance on exactly what agents are maximizing or their information sets.

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

[20]Such approximations are discussed in Heckman (1981), Eckstein and Wolpin (1989), Cameron and Heckman (2001), and Geweke and Keane (2001).

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 the match variables for measurement error and the bias introduced into the measurements by family background. We also use exclusion restrictions to identify our model as in the IV and control function literatures. Unlike many structural papers, we provide explicit proofs of model identification.[21]

Our framework allows agents to make *ex ante* valuations as in dynamic discrete choice models but does not explicitly identify them.[22] However, we estimate a variety of *ex post* returns to schooling, and model how they depend on both observed and unobserved variables. We decompose *ex post* treatment effects into (i) the direct benefits of going from one level of schooling to the next;[23] and (ii) continuation values arising from access to additional education beyond the next step.

Estimating our model on NLSY79 data, we investigate foundational issues in human capital theory. We report the following findings.

(1) There are substantial returns/causal effects of education on wages, the present value of wages, health, and smoking.[24]

(2) The continuation values arising from sequential choices are empirically important components of returns to education. Low-ability individuals gain mostly from graduating high school and stopping there. High-ability individuals have substantial post-high school continuation values.

(3) Estimated returns (causal effects) differ by schooling level and depend on observed and unobserved characteristics of individuals. Graduating high school benefits all—and especially low-ability persons. Only high-ability individuals receive substantial benefits from college graduation. There is positive sorting on gains only at higher educational levels.

[21] Heckman and Navarro (2007) and Blevins (2014) also proof identifiability of structural models.

[22] See, e.g., Eisenhauer et al. (2015a).

[23] 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).

[24] There is a small, but growing literature on the effects of education on health and healthy behaviors. 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.

(4) People sort on *ex post* gains, especially more able people at higher schooling levels, confirming a core tenet of human capital theory. Yet, at the same time, people do not know or act on publicly available information when making decisions about high school graduation.

(5) This paper contributes to an emerging literature on the importance of both cognitive and non-cognitive abilities in shaping life outcomes.[25] Consistent with the recent literature, we find that both types of abilities are important predictors of educational attainment. Within schooling levels, cognitive and non-cognitive abilities have impacts on most outcomes.[26]

(6) Selection bias arising from both observed and unobserved variables accounts for a substantial portion (typically over one half) of the observed differences in wage outcomes classified by education. This finding runs counter to a common interpretation in the literature based on comparing IV and OLS estimates of Equation (1).[27]

Using our estimated model, we conduct two policy experiments. In the first, we examine the impact of a tuition subsidy on college enrollment. We identify who is affected 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 analyze a policy that improves the ability 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.[28] Increasing cognitive endowments positively impacts all outcomes, while increasing non-cognitive endowments mostly impacts smoking and health outcomes.

Our paper proceeds in the following way. Section 2 presents our model. Section 3 presents economically interpretable treatment effects (rates of return) 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

[25] See, e.g., Borghans et al. (2008); Heckman et al. (2006b); Almlund et al. (2011).

[26] Our estimates of the causal effects of education do *not* require that we separately isolate the effects of individual cognitive and non-cognitive endowments on outcomes, just that we control for them as a set.

[27] See, e.g., Griliches (1977) and Card (1999, 2001).

[28] Heckman et al. (2013a).

of education and the outcomes analyzed in this paper. Section 6 reports our estimated treatment effects and interprets them. Section 7 uses the estimated model to address two policy-relevant questions. Section 8 tests a key identifying assumption. Section 9 compares our estimates to those derived from alternative methodological approaches such as OLS and matching. Section 10 concludes.

2 Model

This paper estimates a multistage sequential model of educational choices with transitions and decision nodes shown in Figure 1. Let $\mathcal{J}$ denote a set of possible terminal states. At each node there are only two possible choices: remain at j or transit to the next node ($j+1$ if $j \in \{1, \ldots, \overline{s}-1\}$). $D_j = 0$ if a person at j does not stop there and goes on to the next node. $D_j = 1$ if the person stops at j for $j \neq 0$. $D_0 = 1$ opens an additional branch of the decision tree. A person may remain a dropout or get the GED.[29] For $D_0 = 1$, we define the attainable set as $\{0, G\}$. Thus, in the lower branch ($D_0 = 1$), agents can terminate as a dropout ($D_0 = 1, D_G = 1$) or as a dropout who gets a GED certificate ($D_0 = 1, D_G = 0$). $D_j \in \mathcal{D}$ is the set of possible transition decisions that can be taken by the individual over the decision horizon. Let $\mathcal{S} = \{G, 0, \ldots, \overline{s}\}$ denote the set of stopping states with $S = s$ if the agent stops at $s \in \mathcal{S}(D_s = 1$ for $s \in \mathcal{S} \backslash \{0, G\})$. Define $\overline{s}$ as the highest attainable element in $\mathcal{S}$ in the ordered subset $\{0, \ldots, \bar{s}\}$. We assume that the environment is time-stationary and decisions are irreversible.[30]

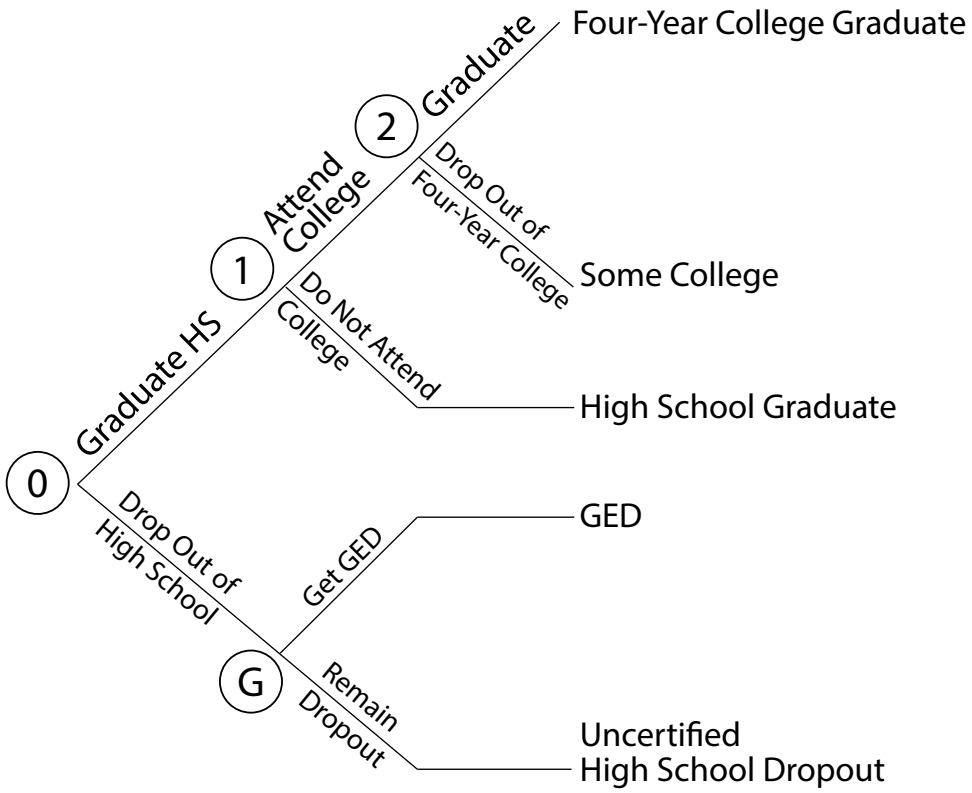
$Q_j = 1$ indicates that an agent *gets to* decision node j and acquires at least the education associated with j. $Q_j = 0$ if the person never gets there. $Q_G = 1$ if the agent drops out of high school and faces the GED option. The history of nodes visited by an agent can be described by the collection of the Q_j such that $Q_j = 1$. Observe that $D_s = 1$ is equivalent to

[29] The GED is a test high school dropouts can take to earn state-issued high school equivalency credentials. For strong evidence on the nonequivalence of GEDs to high school dropouts, see Heckman et al. (2014).

[30] Versions of this model are also analyzed in Cunha et al. (2007), Heckman and Navarro (2007), and Heckman et al. (2016).

$S = s$ for $s \in \{1, \ldots, \bar{s}\}$ and $D_{\bar{s}} = 1$ if $D_j = 0, \forall j \in \mathcal{S} \backslash \{\bar{s}\}$.[31] Finally, $D_0 = 1$ and $D_G = 0$ is equivalent to $S = G$.

Figure 1: A Multistage Dynamic Decision Model



2.1 A Sequential Decision Model

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

$$D_j = \left\{ \begin{array}{ll} 0 & \text{if } I_j \geq 0 \\ 1 & \text{otherwise} \end{array} \right\} \text{for} \quad Q_j = 1, \quad j \in \mathcal{J} = \{G, 0, \ldots, \bar{s} - 1\} \tag{3}$$

[31] For notational convenience, we assign $D_j = 0$ for all $j > s$.

where I_j is the agent's *perceived* value at node j of going on to the next node. The requirement $Q_j = 1$ ensures that agents are able to make the transition at j by conditioning on the population eligible to make the transition.

Associated with each final state $s \in \mathcal{S}$ is a set of K_s potential outcomes for each agent with indices $k \in \mathcal{K}_s$. We define the $\tilde{Y}_s^k$ as latent variables that map into potential outcomes Y_s^k:

$$Y_s^k = \left\{ \begin{array}{ll} \tilde{Y}_s^k & \text{if } Y_s^k \text{ is continuous} \\ \boldsymbol{1}(\tilde{Y}_s^k \geq 0) & \text{if } Y_s^k \text{ is a binary outcome} \end{array} \right\} \text{ for } k \in \mathcal{K}_s, \quad s \in \mathcal{S}. \tag{4}$$

The outcome variables may be in levels, logs, or other transformations. Using the switching regression framework of Quandt (1958, 1972), the observed outcome Y^k for a k common across all decision nodes is

$$Y^k = \left(\sum_{\mathcal{S} \setminus \{0,G\}} D_s Y_s^k \right) (1 - D_0) + \left(Y_0^k D_G + Y_G^k (1 - D_G) \right) D_0. \tag{5}$$

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:

$$I_j = \phi_j \underbrace{(\boldsymbol{Z})}_{\substack{\text{Observed} \\ \text{by analyst}}} - \underbrace{\eta_j}_{\substack{\text{Unobserved} \\ \text{by analyst}}}, \quad j \in \mathcal{J}, \tag{6}$$

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

This specification of agent decision-making is quite agnostic. It does not impose forward-looking behavior. Agents may be myopic or time-inconsistent and may be confronted by surprises. Because we do not impose particular expectation formation assumptions, we are not tied to a particular set of assumptions about agent rationality. A drawback of this approach is that we cannot identify *ex ante* versions of the economic parameters we estimate.

Outcomes are also assumed to be separable:

$$\tilde{Y}_s^k = \tau_s^k(\underbrace{\boldsymbol{X}}_{\substack{\text{Observed} \\ \text{by analyst}}}) + \underbrace{U_s^k}_{\substack{\text{Unobserved} \\ \text{by analyst}}}, \quad k \in \mathcal{K}_s, \quad s \in \mathcal{S}, \tag{7}$$

where $\boldsymbol{X}$ is a vector of observed determinants of outcomes and U_s^k is unobserved by the analyst.[32] Separability of the unobserved variables in the outcome equations is often invoked in the structural literature but is not strictly required in the structural or discrete choice literatures.[33]

2.3 Assumptions about the Unobservables

Central to our main empirical strategy is the existence of a finite dimensional vector $\boldsymbol{\theta}$ of unobserved (by the economist) endowments that generate all of the dependence across the η_j and the U_s^k. We assume that

$$\eta_j = -(\boldsymbol{\theta}'\boldsymbol{\lambda}_j - \nu_j), \quad j \in \mathcal{J} \tag{8}$$

and

$$U_s^k = \boldsymbol{\theta}'\boldsymbol{\alpha}_s^k + \omega_s^k, \quad k \in \mathcal{K}_s, s \in \mathcal{S}, \tag{9}$$

where ν_j is an idiosyncratic error term for transition j. ω_s^k represents an idiosyncratic error term for outcome k in state s.

[32] In our model, $\boldsymbol{X}$ and $\boldsymbol{Z}$ can vary by decision or outcome depending on the specification of functions $\tau_s^k(\boldsymbol{X})$ and $\phi_j(\boldsymbol{Z})$. See Table 1 for details.

[33] Moreover, we can condition on observable covariates $\boldsymbol{X}$.

Conditional on $\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}$, choices and outcomes are statistically independent. Controlling for this set of variables eliminates selection effects. If the analyst knew $\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}$, he/she could use matching to identify the model.[34]

The standard "random effects" approach in the structural literature treats $\boldsymbol{\theta}$ as a nuisance variable and does not interpret it.[35] Our approach is to proxy $\boldsymbol{\theta}$ using multiple interpretable measurements of it. We correct for errors in the proxy variables. The measurements facilitate the interpretation of $\boldsymbol{\theta}$. We develop this intuition further in Section 4, after presenting the rest of our model.

We array the $\nu_j, j \in \mathcal{J}$ into a vector $\boldsymbol{\nu} = (\nu_G, \nu_0, \nu_1, \ldots, \nu_{\bar{s}-1})$, and the η_j into $\boldsymbol{\eta} = (\eta_G, \eta_0, \ldots, \eta_{\bar{s}-1})$. Array the ω_s^k into a vector $\boldsymbol{\omega}_s = (\omega_s^1, \ldots, \omega_s^{Ks})$. Array the U_s^k into vector $\boldsymbol{U}_s = (U_s^1, \ldots, U_s^{Ks})$, and array the $\boldsymbol{U}_s$ into $\boldsymbol{U} = (\boldsymbol{U}_G, \boldsymbol{U}_0, \ldots, \boldsymbol{U}_{\bar{s}})$.

Letting "$\perp\!\!\!\perp$" denote statistical independence, we assume that, conditional on $\boldsymbol{X}$

$$\nu_j \perp\!\!\!\perp \nu_l, \quad \forall l \neq j \quad l, j \in \mathcal{J} \tag{A-1a}$$

$$\omega_s^k \perp\!\!\!\perp \omega_{s'}^k, \quad \forall s \neq s' \quad \forall k \tag{A-1b}$$

$$\boldsymbol{\omega}_s \perp\!\!\!\perp \boldsymbol{\nu}, \quad \forall s \in \mathcal{S} \tag{A-1c}$$

$$\boldsymbol{\theta} \perp\!\!\!\perp \boldsymbol{Z} \tag{A-1d}$$

$$(\boldsymbol{\omega}_s, \boldsymbol{\nu}) \perp\!\!\!\perp (\boldsymbol{\theta}, \boldsymbol{Z}), \quad \forall s \in \mathcal{S}. \tag{A-1e}$$

Assumption (A-1a) maintains independence of the shocks affecting transitions; (A-1b) assumes independence of shocks across all states; (A-1c) assumes independence of the shocks to transitions and the outcomes; (A-1d) assumes independence of $\boldsymbol{\theta}$ with respect to the observables; and (A-1e) assumes independence of the shocks with the factors $\boldsymbol{\theta}$ and $\boldsymbol{Z}$. Versions of assumptions (A-1d) and (A-1e) play fundamental roles in the structural dynamic

[34] See Carneiro et al. (2003).

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

discrete choice literature.[36] Any dependence postulated across the $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$ can be captured by introducing factors in $\boldsymbol{\theta}$.

2.4 Measurement System for Unobserved Factors $\boldsymbol{\theta}$

We allow for the possibility that $\boldsymbol{\theta}$ cannot be measured precisely, but that it can be proxied with multiple measurements. We correct for the effects of measurement error in the proxy. We link $\boldsymbol{\theta}$ to measurements, and adjoin measurement equations to choice and outcome equations, making $\boldsymbol{\theta}$ interpretable.

Let $\boldsymbol{M}$ be a vector of N_M measurements on $\boldsymbol{\theta}$. They may consist of lagged or future values of the outcome variables or additional measurements.[37] The system of equations determining $\boldsymbol{M}$ is

$$w\boldsymbol{M} = \Phi(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{e}), \tag{10}$$

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

$$\boldsymbol{M} = \begin{pmatrix} M_1 \\ \vdots \\ M_{N_M} \end{pmatrix} = \begin{pmatrix} \Phi_1(\boldsymbol{X}, \boldsymbol{\theta}, e_1) \\ \vdots \\ \Phi_{N_M}(\boldsymbol{X}, \boldsymbol{\theta}, e_{N_M}) \end{pmatrix},$$

where we array the e_j into $\boldsymbol{e} = (e_1, \ldots, e_{N_M})$. We assume, in addition to the previous assumptions that, conditional on $\boldsymbol{X}$,

$$e_j \perp\!\!\!\perp e_l, \quad j \neq l, \quad j, l \in \{1, \ldots, N_M\} \tag{A-1f}$$

$$\text{and} \qquad \boldsymbol{e} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{\omega}). \tag{A-1g}$$

For the purpose of identifying treatment effects, we do not need to identify each equation of system (10). We just need to identify the span of $\boldsymbol{\theta}$ that preserves the information on $\boldsymbol{\theta}$ in

[36] For example, the widely-used "types" assumption of Keane and Wolpin (1997) postulates conditional independence between choices and outcomes conditional on types ($\boldsymbol{\theta}$) that operate through the initial conditions of their model.

[37] See, e.g., Abbring and Heckman (2007); Schennach et al. (2012).

(10). That is sufficient to produce conditional independence between choices and outcomes.[38] However, in this paper we estimate equation system (10) to enhance interpretability.

3 Defining Returns/Causal Effects of Education

A variety of *ex post* counterfactual outcomes and associated treatment effects can be generated from our model. There is no single "causal effect" of education. The causal effects we analyze can be used to predict the effects of changing education levels through different policies for people of different backgrounds and abilities. They allow us to improve on the "effects" reported in the literature on instrumental variables to understand the effectiveness of policies for different identifiable segments of the population, and the benefits to people at different margins of choice. These effects are defined for different conditioning sets and thought experiments. Our dynamic model suggests a new range of treatment parameters that do not arise in models with binary treatments. This section makes precise the notions of returns to education discussed in Section 1.

In principle, we could define and estimate a variety of causal effects, many of which are not plausible. 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).[39] In the spirit of credible econometrics, we define such treatment effects conditional on $Q_j = 1$. This approach blends structural and treatment effect approaches. Our causal parameters recognize agent heterogeneity and are allowed to differ across different subsets of the population.

The person-specific treatment effect T_j^k for outcome k for an individual selected from the population $Q_j = 1$ with characteristics $\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{\theta} = \overline{\boldsymbol{\theta}}$, making a decision at node j between going on to the next node or stopping at j, is the difference between the individual's

[38] See, e.g., Heckman et al. (2013b).

[39] The distinction between *fixing* and *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).

outcomes under the two actions:

$$T_j^k[Y^k|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}] := (Y^k|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{\theta}=\overline{\boldsymbol{\theta}},Q_j=1,Fix\ D_j=0) - (Y^k|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{\theta}=\overline{\boldsymbol{\theta}},Q_j=1,Fix\ D_j=1). \quad (11)$$

The random variable $(Y^k|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{\theta}=\overline{\boldsymbol{\theta}},Q_j=1,Fix\ D_j=0)$ is the outcome at node j for a person with characteristics $\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}$ from the population that attains node j (or higher), $Q_j=1$, and for whom we fix $D_j=0$ so they go on to the next node. They may choose to go even further. Random variable $(Y^k|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{\theta}=\overline{\boldsymbol{\theta}},Q_j=1,Fix\ D_j=1)$ is defined for the same population but forces persons with those characteristics not to transit to the next node.

We present population-level treatment effects based on (11). We focus our discussion on means, but we also discuss distributional counterparts for all of the treatment effects considered in this paper.

3.1 Direct Effects and Continuation Values

A principal contribution of this paper is the definition and estimation of 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. The associated mean 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.

Person-specific treatment effects at node j can be decomposed into two components. The first component is the *direct effect* of going from j to $j+1$: $DE_j^k = Y_{j+1}^k - Y_j^k$, the effect often featured in the literature on the returns to schooling when comparing schooling levels $j+1$ and j (Becker, 1964). The second component is the *continuation value* of going beyond

$j+1$ for persons with $D_0 = 0$ (the upper branch of Figure 1), which is

$$C^k_{j+1} := \sum_{r=1}^{\overline{s}-(j+1)} \left[\prod_{l=1}^{r}(1-D_{j+l})\right](Y^k_{j+r+1} - Y^k_{j+r}).^{40}$$

The continuation value for the lower branch of Figure 1 ($D_0 = 1$) is defined for the attainable set $\{0, G\}$. G is the only option available to a high school dropout in that branch. In the following, we analyze the upper branch of Figure 1. The analysis for the lower branch is similar.

At the individual level, the *total effect* of fixing $D_j = 0$ on Y^k is decomposed into

$$T^k_j = DE^k_j + C^k_{j+1}. \tag{12}$$

The associated population level average treatment effect at node j inclusive of continuation values, conditional on $Q_j = 1$, is

$$ATE^k_j := \int \dots \int E(T^k_j[Y^k|\boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Z}=\boldsymbol{z}, \boldsymbol{\theta}=\overline{\boldsymbol{\theta}}])\, dF_{\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}, \overline{\boldsymbol{\theta}} \,|\, Q_j = 1), \tag{13}$$

which can be decomposed into direct and continuation value components.

Integrating over the $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}$, conditioning on $Q_j = 1$, the component of (13) due to the population continuation value at $j+1$ is

$$E_{\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta}}(C^k_{j+1}) = E_{\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta}}\Bigg[\sum_{l=j+1}^{\overline{s}-1}\Big\{E(Y^k_{l+1} - Y^k_l|\boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Z}=\boldsymbol{z}, \boldsymbol{\theta}=\overline{\boldsymbol{\theta}}, Q_{l+1}=1,\ Fix\ Q_{j+1}=1)$$
$$\cdot Pr(Q_{l+1}=1|\boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Z}=\boldsymbol{z}, \boldsymbol{\theta}=\overline{\boldsymbol{\theta}}, Q_j=1, Fix\ Q_{j+1}=1)\Big\}|Q_j=1\Bigg], \tag{14}$$

where $Q_{\overline{s}} = 1$ if $S = \overline{s}$.

We can also define conditional (on $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}$) population distributions of total effects as in

[40] The relationship between this notion of continuation values and the definition used in the dynamic discrete choice literature is explored in Web Appendix A.3.

Heckman et al. (1997):[41]

$$Pr(T_j^k < t_j^k | \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{\theta} = \overline{\boldsymbol{\theta}}, Q_j = 1) \tag{15}$$

and the population counterpart, integrating over $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}$, which can be further decomposed into the distributions of direct effects and of continuation values.[42]

Because we do not specify or attempt to identify choice-node-specific agent information sets, we can only identify *ex post* treatment effects. Hence, we can identify continuation values associated with choices, but cannot identify option values. A benefit of this more agnostic approach is that it does not impose specific decision rules or assumptions about agent expectations. Our model allows for irrationality, regret, and mistakes in agent decision-making associated with maturation and information acquisition and allows us to test the validity of certain assumptions commonly made about agent expectations.

3.2 Average Marginal Treatment Effects

In order to understand the economic returns to an additional unit of schooling 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).[43] It is the average effect of transiting to the next node for individuals at or near the margin of indifference between the two nodes:

$$AMTE_j^k := \iiint E\left[T_j^k\left(Y^k | \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{\theta} = \overline{\boldsymbol{\theta}}\right)\right] dF_{\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}, \bar{\boldsymbol{\theta}} \mid Q_j = 1, |I_j| \leq \varepsilon), \tag{16}$$

where ε is an arbitrarily small neighborhood around the margin of indifference.[44] These effects are inclusive of all consequences of taking the transition at j, including the possibility

[41] See Abbring and Heckman (2007) for a review of the literature.

[42] The modifications for the unordered case require that we define these terms over the admissible options available for $D_0 = 1$ or $D_0 = 0$.

[43] See Carneiro et al. (2010, 2011).

[44] Note that the limit of (16) as $\varepsilon \to 0$ is not well-defined without further assumptions. This is the so-called "Borel paradox" discussed in this context in Carneiro et al. (2010). We avoid this problem by assuming a functional form for the distribution of ε.

of attaining final schooling levels well beyond j.[45] AMTE defines causal effects at well-defined and empirically identified margins of choice. It is the proper measure of the *ex post* 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, and generally does not estimate $E(\rho)$ or $E(\rho|S=s)$, and includes the effects on outcomes for transitions induced by instruments beyond any schooling level at which the instrument operates.[46] Since we identify the distribution of I_j, we can identify the characteristics of agents in the indifference set, something not possible using LATE.[47]

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

3.3 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 compared to policy p' for outcome k is:

$$PRTE^k_{p,p'} := \iiint E(Y^k(p') - Y^k(p) | \boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Z}=\boldsymbol{z}, \boldsymbol{\theta}=\overline{\boldsymbol{\theta}}) dF_{\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z},\overline{\boldsymbol{\theta}} | S(p) \neq S(p')), \quad (17)$$

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

[45] One might also define a version of this treatment effect for two adjacent states ignoring continuation values.

[46] See Heckman and Vytlacil (2007a) and Carneiro et al. (2010). The LATE can correspond to people at multiple margins. See Angrist and Imbens (1995) and Heckman et al. (2016).

[47] Note that the indifference set may contain multiple margins, as in Heckman and Vytlacil (2007b) and Heckman and Urzúa (2010).

general, they are different unless the proposed policy change coincides with the instrument used to define LATE.[48]

3.4 Differences Across Final Schooling Levels

Becker's original approach to estimating returns to schooling (1964) focused on the upper branch of Figure 1 and reported estimates from pairwise comparisons of returns at final schooling levels. He defines returns to education as the gains from choosing between a terminal base state and a terminal final schooling level, implicitly assuming that the probabilities of all intervening transitions in Equation (2) are 1. Following Becker, but controlling for $\boldsymbol{\theta}$, $\boldsymbol{Z}$, and $\boldsymbol{X}$, the mean gain for the subset of the population that completes one of the two adjacent schooling level $S \in \{s, s'\}$ is:

$$ATE^k_{s,s'} := \iiint E(Y^k_{s'} - Y^k_s | \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{\theta} = \overline{\boldsymbol{\theta}})\, dF_{\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}, \overline{\boldsymbol{\theta}} \,|\, S \in \{s, s'\}). \quad (18)$$

Unlike (13), this parameter ignores continuation values.

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 two choices, and hence, might be far away from having any empirical or policy relevance.[49] One can also compute parameters of $ATE_{s,s'}$ for other conditioning sets, such as $S = s'$ (treatment on the treated). We report estimates of different versions of these treatment effects in the Web Appendix A.14.1.

[48] See Carneiro et al. (2011) for an empirical example. The differences between the two parameters can be substantial as we show in Heckman et al. (2016).

[49] The estimated differences in treatment effects for the conditional and total populations are not large for outcomes associated with the decision to enroll in college, but are substantial for the choice to graduate from college. See Web Appendix A.14.2.

3.5 Decomposing Observed Differences in Outcomes into Selection Bias, Sorting Gains, and Average Treatment Effects

Using our model, we interpret "ability bias" (really selection bias) and sorting on gains using the traditional Becker-Chiswick-Mincer model (1) and its extensions as a benchmark. To simplify the exposition, we focus on the upper branch of Figure 1 ($D_0 = 0$) and analyze continuous outcomes.[50]

There are two basic models used in the empirical literature estimating returns to schooling. One version studies outcomes and selection bias in terms of pairwise final schooling levels (s_0, s) attained by agents $(D_{s_0} + D_s = 1)$, $s_0 \neq s$. It is defined for the population at one of these two terminal schooling states. It does not include terminal values beyond s. Another version studies gains and ability bias in terms of benefits associated with attaining (and possibly exceeding) given schooling levels $(Q_j = 1)$. This includes continuation values. In the text, we develop both widely-used versions.

The effect of additional schooling starting at s_0 and stopping at s is captured by $Y_s^k - Y_{s_0}^k = \rho_{s_0,s}^k$.[51] This is the direct gain of going from s_0 to s. It does not include any gains from transitions beyond s:

$$\rho_{s_0,s}^k = Y_s^k - Y_{s_0}^k = \tau_s^k(\boldsymbol{X}) - \tau_{s_0}^k(\boldsymbol{X}) + \boldsymbol{\theta}'(\boldsymbol{\alpha}_s^k - \boldsymbol{\alpha}_{s_0}^k) + \omega_s^k - \omega_{s_0}^k.$$

In this notation we may write the outcome Y^k relative to base state Y_0^k as

$$Y^k = Y_{s_0}^k + \sum_{s \in \mathcal{S}} \rho_{s_0,s}^k D_s. \tag{19}$$

This is a version of (1) where schooling is discretized at final schooling attainment levels: $S = s$ if $D_s = 1$. $E(\rho_{s_0,s}^k)$ is one version of the returns to schooling compared to benchmark

[50] The analysis for discrete outcomes is straightforward.

[51] Note that Y_s^k can be log outcomes as in (1). We can also formulate the outcomes in terms of latent variables.

s_0 defined for the entire population.

Except for knife-edge cases, if $\boldsymbol{\lambda}_s \neq 0$, dependence between D_s and $\rho^k_{s_0,s}$ is generated if either $\tau^k_{s_0}(\boldsymbol{X}) \neq \tau^k_s(\boldsymbol{X})$, or $\boldsymbol{\alpha}^k_{s_0} \neq \boldsymbol{\alpha}^k_s$, or both.[52] Sorting on gains (correlation between $\rho^k_{s_0,s}$ and D_s) may not appear in empirical estimates if agents are sorting on gains beyond s and not on direct effects (i.e., sorting on components of $R_{s,i}$ as defined in (2)). Only in the case where there is no continuation value can we conclude from empirical estimates that absence of sorting effects defined in this fashion implies absence of sorting on potential future gains.

The traditional Griliches (1977) analysis of returns to schooling ignores sorting on gains and only considers ability bias. Assuming analysts condition on $\boldsymbol{X}$ (in levels and in interactions with D_s), sorting gains arise only if $\boldsymbol{\alpha}^k_s - \boldsymbol{\alpha}^k_{s_0} \neq 0$ and $\boldsymbol{\lambda}_s \neq 0$. Even if $\boldsymbol{\alpha}^k_s - \boldsymbol{\alpha}^k_{s_0} = 0$, as long as $\boldsymbol{\alpha}^k_{s_0} \neq 0$, ability bias will arise in estimating the mean of the gains $\rho^k_{s_0,s}$ in (19), provided $\boldsymbol{\lambda}_s \neq 0$.[53]

Note that the choice of a base state matters for estimating sorting gains in the general case where the magnitude of $\boldsymbol{\alpha}^k_s - \boldsymbol{\alpha}^k_{s_0}$ changes depending on the base state selected. Some representations may generate sorting gains that are absent from other representations with different base states.[54]

Within this framework, there are several meaningful ways to decompose the observed difference in outcomes between those at j who go on to $S = j + 1$. The observed difference

[52] Notice that even if there is no such dependence, some agents may still choose to go beyond s because of later gains in outcomes.

[53] For a given level of s, selection bias is defined as $E(Y^k_{s_0}|D_s = 1) - E(Y^k_{s_0}|D_{s_0} = 1)$, the mean difference in baseline outcomes for persons who stop at $S = s$ compared to those who stop at $S = s_0$.

[54] The extension of this analysis to more general model (5) with the GED is straightforward.

can be decomposed as follows:

$$
\begin{aligned}
&\underbrace{E[Y_{j+1}^k|S=j+1]-E[Y_j^k|S=j]}_{\text{Observed difference}} \\
&= \underbrace{E[Y_{j+1}^k-Y_j^k|S=j+1]}_{\text{Treatment on the treated } TT_{j,j+1}} + \underbrace{E[Y_j^k|S=j+1]-E[Y_j^k|S=j]}_{\substack{\text{Selection bias } SB_{j,j+1} \\ \text{from base state } j}} \\
&= \underbrace{E[Y_{j+1}^k-Y_j^k|S\in\{j,j+1\}]}_{\substack{\text{Pairwise average treatment effect } ATE_{j,j+1} \\ \text{for people in conditioning set } \{j,j+1\}}} + \\
&\underbrace{E[Y_{j+1}^k-Y_j^k|S=j+1]-E[Y_{j+1}^k-Y_j^k|S\in\{j,j+1\}]}_{\text{Sorting gains } SG_{j,j+1}} + \underbrace{E[Y_j^k|S=j+1]-E[Y_j^k|S=j]}_{\text{Selection bias } SB_{j,j+1}}.^{55}
\end{aligned}
\tag{20}
$$

Note that the ATE parameter depends on the distributions of characteristics of $\boldsymbol{X}$ and $\boldsymbol{\theta}$ for persons at node j, as do the sorting on gains and selection bias parameters. These components can be further decomposed into selection on observed variables and selection on unobserved ability components $\boldsymbol{\theta}$, and the ability components can be further decomposed into cognitive and non-cognitive components.

These decompositions focus on gains *up to* final schooling states. They compare observed differences across pairs of final schooling levels. The empirical literature on the returns to schooling also compares the observed differences in outcomes between persons at a given node ($Q_j = 1$) who make a particular schooling transition with those who do not make that transition.

Thus, we can decompose the observed gain from going to $j+1$ from j for those at j ($Q_j = 1$) into a gain for those who take the transition ($D_j = 0$) and a selection bias term (the difference in the mean outcomes between those who would have gone on ($D_j = 0$), but are stopped at j (Fix $D_j = 1$), and those who chose not to go on). We can further decompose the treatment on the treated parameter into a node-specific ATE (the mean difference between

[55] Appendix A.15.1 gives the exact decomposition for our specific functional forms.

those for whom $Q_j = 1$ where we fix $D_j = 0$ and we fix $D_j = 1$, respectively), and a "sorting gains" term which is the difference between the node-specific treatment on the treated term and the node-specific ATE.

In Web Appendix A.15.3, we decompose the values of being at j into components associated with stopping at j and continuing beyond j where, for the upper branch of Figure 1 ($D_0 = 0$),

$$Y^k = Y_0^k + \sum_{j \geq 1}^{\bar{s}} \rho_{j-1,j}^k Q_j, \tag{21}$$

where $\rho_{j-1,j}^k = Y_j^k - Y_{j-1}^k$. The expected future gain for a person at j (≥ 1) is

$$\begin{aligned} & E_j \left(\sum_{l>j}^{\bar{s}} \rho_{l-1,l}^k Q_l | Q_j = 1 \right) \\ & = \sum_{l>j}^{\bar{s}} \left[E_j(\rho_{l-1,l}^k | Q_l = 1) P(Q_l = 1 | Q_j = 1) \right], \quad j \geq 1, \end{aligned}$$

where the conditioning $D_0 = 0$ is kept implicit.[56]

Analogous to decomposition (20), we can decompose the observed difference between those with $D_j = 0$ and those with $D_j = 1$, i.e., the observed difference between those that do and do not make a particular transition conditional on making that transition. $E(\rho_{j,j+1}^k)$ is the expected incremental gain of proceeding to the next stage. For the upper branch

[56] The more general expression incorporating D_0 is presented as Equation (A.10) in the Web Appendix.

$(D_0 = 0)$, we may write for the kth outcome at node j:

$$
\begin{aligned}
&\underbrace{E[Y^k|D_j=0,Q_j=1]-E[Y^k|D_j=1,Q_j=1]}_{\text{Observed difference}}\\
=&\underbrace{E[Y^k|D_j=0,Q_j=1]-E[Y^k|D_j=0,Q_j=1,Fix\ D_j=1]}_{\text{Dynamic treatment on the treated for those at } j}\\
+&\underbrace{E[Y^k|D_j=0,Q_j=1,Fix\ D_j=1]-E[Y^k|D_j=1,Q_j=1]}_{\text{Selection bias for those at } j}\\
=&\underbrace{E[Y^k|Q_j=1,Fix\ D_j=0]-E[Y^k|Q_j=1,Fix\ D_j=1]}_{\text{ATE for those at } j}\\
+&\underbrace{\left\{\begin{array}{l}\big(E[Y^k|D_j=0,Q_j=1]-E[Y^k|D_j=0,Q_j=1,Fix\ D_j=1]\big)\\ -\big(E[Y^k|Q_j=1,Fix\ D_j=0]-E[Y^k|Q_j=1,Fix\ D_j=1]\big)\end{array}\right\}}_{\text{TT - ATE: Sorting gain at } j \text{ for those who transit to } j+1}\\
+&\underbrace{E[Y^k|D_j=0,Q_j=1,Fix\ D_j=1]-E[Y^k|D_j=1,Q_j=1]}_{\text{Selection bias}}. \qquad (22)
\end{aligned}
$$

The node-specific ATE is defined for the population at $Q_j = 1$ and considers either forcing population members to stay at j, or moving the entire group from j to $j+1$ (i.e, $Fix\ D_j = 1$ and $Fix\ D_j = 0$, respectively). The sorting gain is the average net gain beyond ATE to those who actually take the transition $(D_j = 0)$.

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 $\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}$, outcomes and choices are statistically independent where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are observed and $\boldsymbol{\theta}$ is not. If $\boldsymbol{\theta}$ were observed, one could condition on $\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}$ and identify the model of Equations (3)–(9) and the treatment effects that can be generated from it. We use factor model (10) to proxy $\boldsymbol{\theta}$ using measurements $\boldsymbol{M}$.

Under the conditions presented in Heckman et al. (2016), we can non-parametrically identify the model of Equations (3)–(7) including the distribution of $\boldsymbol{\theta}$, as well as the Φ functions and the distribution of $\boldsymbol{e}$ (which can be interpreted as measurement errors). Effectively, we match on proxies for $\boldsymbol{\theta}$ and correct for the effects of measurement error ($\boldsymbol{e}$) in creating the proxies. Such corrections are possible because with multiple measures on $\boldsymbol{\theta}$ we can identify the distribution of $\boldsymbol{e}$.[57] We can identify treatment effects even though we do not isolate individual factors. We only need that the factors $\boldsymbol{\theta}$ are spanned by $\boldsymbol{M}$, not that Equations (10) are separately identified.[58]

Another approach to identification uses instrumental variables which, if available, under the conditions presented in Heckman et al. (2016) can be used to identify the structural model (3)–(9) without invoking the factor structure (8) and (9) or the postulated conditional independence assumptions.

The precise parameterization and the likelihood function for the model we estimate is presented in Web Appendix A.4. While, in principle, it is possible to identify the model non-parametrically, in this paper we make parametric assumptions in order facilitate computation. We subject the estimated model to rigorous goodness-of-fit tests which the model passes.[59]

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).[60] Before discussing estimates from our model, it is informative to set the stage for what follows and present adjusted and unadjusted associations

[57] Under linear specifications for (10), we can directly estimate the $\boldsymbol{\theta}$ and use factor regression methods. See, e.g., Heckman et al. (2013a), Heckman et al. (2016), and the references cited therein.

[58] As noted in Heckman et al. (2011b), we do not need to solve classical identification problems associated with estimating equation system (10) in order to extract measure-preserving transformations of $\boldsymbol{\theta}$ on which we can condition in order to identify treatment effects. In the linear factor analysis literature these are the classical rotation and normalization problems.

[59] See Web Appendix A.5.

[60] Web Appendix A.2 presents a detailed discussion of the data we analyze and our exclusion restrictions.

between the outcomes we study and schooling. Figure 2 presents estimated linear 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 (or PV of wages), health limitations, and smoking.[61] These are least squares regressions using the regressors indicated at the base of the figure, including our proxies for ability. They do not separate out the roles of $\boldsymbol{X}$ and $\boldsymbol{\theta}$ in contributing to the causal and selection bias components of the observed differences. Least squares estimates of this form are commonly reported in the literature that investigates the effects of schooling controlling for $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{\theta}$.

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 least squares estimates of the effects of education.

Figure 2 shows that controlling for proxied ability substantially reduces the observed differences in earnings across educational groups. These regression estimates suggest, but do not identify, substantial causal effects which we report below.

Entering $\boldsymbol{\theta}$ as a regressor is a traditional way to control for ability bias. It eliminates the ability bias emphasized by Griliches (1977). If there is sorting on gains that depend on $\boldsymbol{X}$ and $\boldsymbol{\theta}$, this approach over-controls for those variables that are components of the causal effect of treatment on the treated as defined in (20).[62] Figure 2 reports traditional measures of regression-adjusted causal effects of schooling. At the same time, such regressions do not discriminate among the components of (20) which have different causal interpretations. In Web Appendix A.17.2, we compare the OLS pairwise causal effects implicit in the estimates reported in Figure 2 with the estimates from our version of a structural model discussed

[61] Adjustments are made through linear regression. This decomposition uses high school dropout as the base category (s_0).

[62] See also the decompositions in (A.7).

below.[63]

It is sometimes claimed that a linear-in-years-of-schooling model fits the data well.[64] The white bar in Figure 2 displays the OLS-adjusted effect of schooling controlling for years of completed schooling as in Equation (1).[65] The white bars in all figures show that, even after controlling for years of schooling, the educational indicators still play an important role. OLS estimates of Mincer specification (1) do not precisely describe the data. There are effects of schooling beyond those captured by a linear years-of-schooling specification.[66]

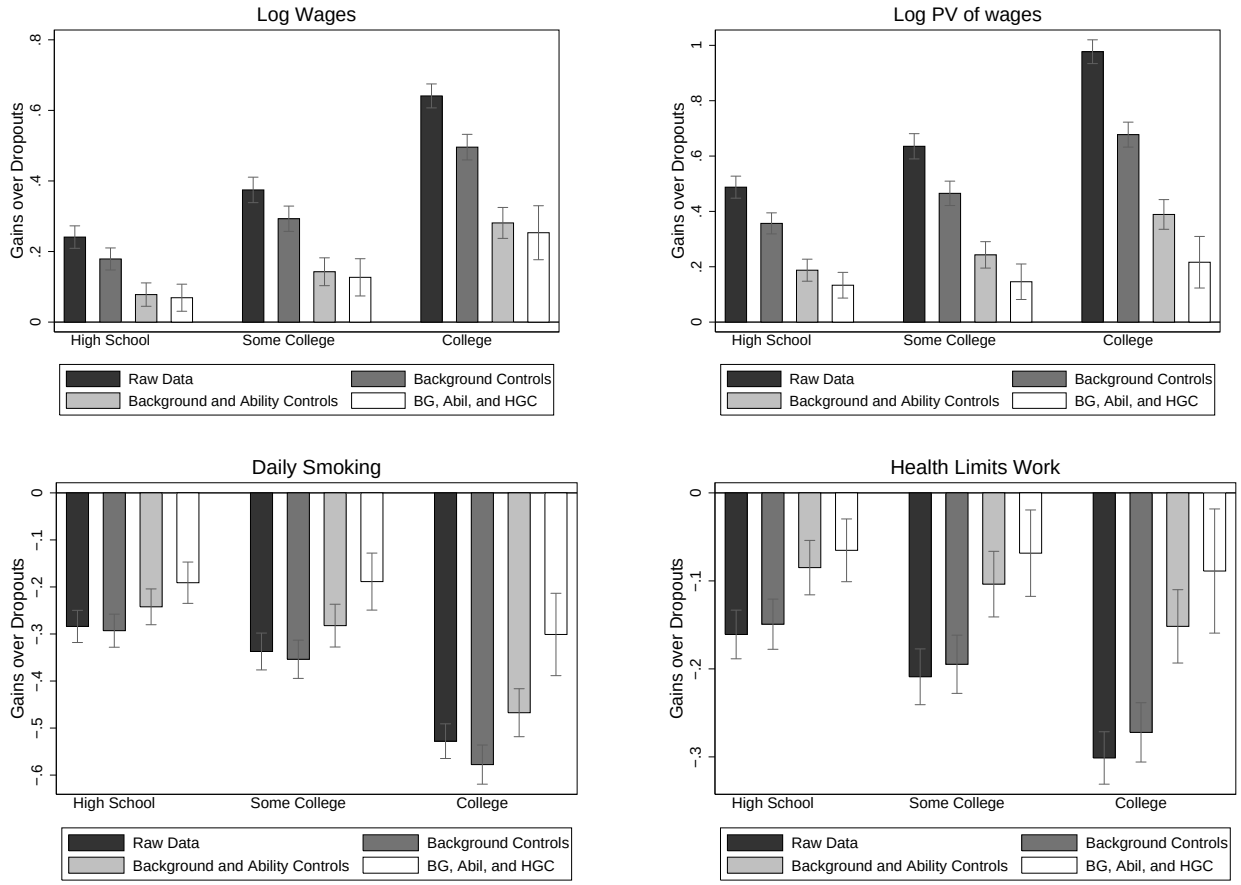
[63]The OLS estimates do not identify treatment on the treated parameters. They are in rough agreement with ATEs except for the log present value of earnings.

[64]See, e.g., Card (1999, 2001). Heckman et al. (2006a) dispute this claim.

[65]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.

[66]Using our estimated model, we find, however, that population ATEs are well described by a linear-in-schooling specification. See Web Appendix A.8.

Figure 2: Observed and Adjusted Benefits from Education



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, age in 1979, 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 Armed Services Vocational Aptitude Battery (ASVAB) tests and ninth grade GPA in core subjects (language, math, science, and social science). See the discussion surrounding Table 1 (below) and Web Appendix A.2 for additional details. "Some College" includes anyone who enrolled in college, but did not receive a four-year college degree. The white bars additionally control for highest grade completed (HGC). *Source:* NLSY79 data.

5.1 Control Variables and Exclusion Restrictions

As previously noted, identification of our model and the associated treatment effects does not depend exclusively on conditional independence assumptions associated with our factor model.[67] Node-specific instruments can non-parametrically identify treatment effects without invoking the full set of conditional independence assumptions.[68] We have a variety of exclusion restrictions that affect choices but not outcomes. Table 1 documents the control variables

[67] See Carneiro et al. (2003).
[68] See Heckman et al. (2016).

($\boldsymbol{X}$) and the exclusion restrictions (components of $\boldsymbol{Z}$ not in $\boldsymbol{X}$) used in this paper. Our instruments are traditional in the literature that estimates the causal effects of education.[69]

Table 1: Control Variables and Instruments Used in the Analysis

Control Variables	Measurement Equations	Choice	Outcomes
Race	x	x	x
Broken Home	x	x	x
Number of Siblings	x	x	x
Parents' Education	x	x	x
Family Income (1979)	x	x	x
Region of Residence[a]	x	x	x
Urban Status[a]	x	x	x
Age[b]	x	x	x
Local Unemployment[c]			x
Local Long-Run Unemployment		x	
Instruments (Exclusion Restrictions)			
Local Unemployment at Age 17[d]		x	
Local Unemployment at Age 22[e]		x	
College Present in County 1977[f]		x	
Local College Tuition at Age 17[g]		x	
Local College Tuition at Age 22[h]		x	

Notes: [a]Region and urban dummies are specific to the age that the measurement, educational choice, or outcome occurred. [b]Age in 1979 is included as a cohort control. We also included individual cohort dummies which did not change the results. [c]For economic outcomes, local unemployment at the time the outcome is measured. [d]This is an instrument for choices at nodes 0 and 1. It represents opportunity costs at the time schooling decisions are made. [e]This is an instrument for the choice at node 2. [f]Presence of a four-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. [g]Local college tuition at age 17 only enters the college enrollment graduation decisions. [h]Local college tuition at age 22 only enters the college completion equation. The measurement system includes the arithmetic reasoning, coding speed, paragraph comprehension, word knowledge, mathematical knowledge, and numerical operations sub-tests of the ASVAB, 9th grade GPA in math, English, science, and social studies, and early risky and reckless behavior. We assume ASVAB only loads on the cognitive factor. See Web Appendix Section A.2 for details.

[69]For example, presence of a nearby college or distance to college is used by Cameron and Taber (2004), Kling (2001), Carneiro et al. (2013), Cawley et al. (1997), Heckman et al. (2011a), and Eisenhauer et al. (2015b). Local tuition at two- or four-year colleges is used as an instrument by Kane and Rouse (1993), Heckman et al. (2011a), Eisenhauer et al. (2015b), and Cameron and Taber (2004). Local labor market shocks are used by Heckman et al. (2011a) and Eisenhauer et al. (2015b).

6 Estimated Causal Effects

In this section of the paper, we move beyond OLS analyses of causal effects of schooling and present the estimated causal effects of schooling from our model. Since the model is non-linear and multidimensional, in the main body of the paper we only report the treatment effects derived from it.[70] We randomly draw sets of regressors from our sample and a vector of factors from the estimated factor distribution to simulate the reported treatment effects.[71]

Section 6.1 presents estimated treatment effects across final schooling levels. These are based on Equation (18) and extend Becker (1964) by controlling for observed and proxied unobserved variables. Section 6.2 presents the main empirical analysis of this paper. We estimate dynamic treatment effects, inclusive of continuation values. We analyze the contribution of continuation values, sorting on gains, and selection bias to measured differences in education across levels. Section 6.3 analyzes the effects of cognitive and non-cognitive endowments on estimated treatment effects. Section 6.4 presents estimates of distributions of treatment effects. Section 6.5 examines the implications of our analysis for the validity of the Becker-Chiswick-Mincer model. Section 6.6 summarizes our analysis. In the text of our paper, we focus on the transitions in the upper branch of Figure 1, although our model is estimated over both branches.

6.1 The Estimated Average Causal Effect of Educational Choices by Pairwise Final Schooling Levels

We first present estimates of average treatment effects $ATE_{s-1,s}$ (18) for the four outcomes studied in this paper at final schooling level s compared to final schooling level $s-1$.[72] They ignore continuation values.

The shaded regions labeled "Observed" in Figure 3 are the raw differences found in our

[70] Parameter estimates for individual equations are reported in Web Appendix A.6.
[71] We randomly draw an individual and use their full set of regressors.
[72] This is Expression (18) for the case $s' = s + 1$.

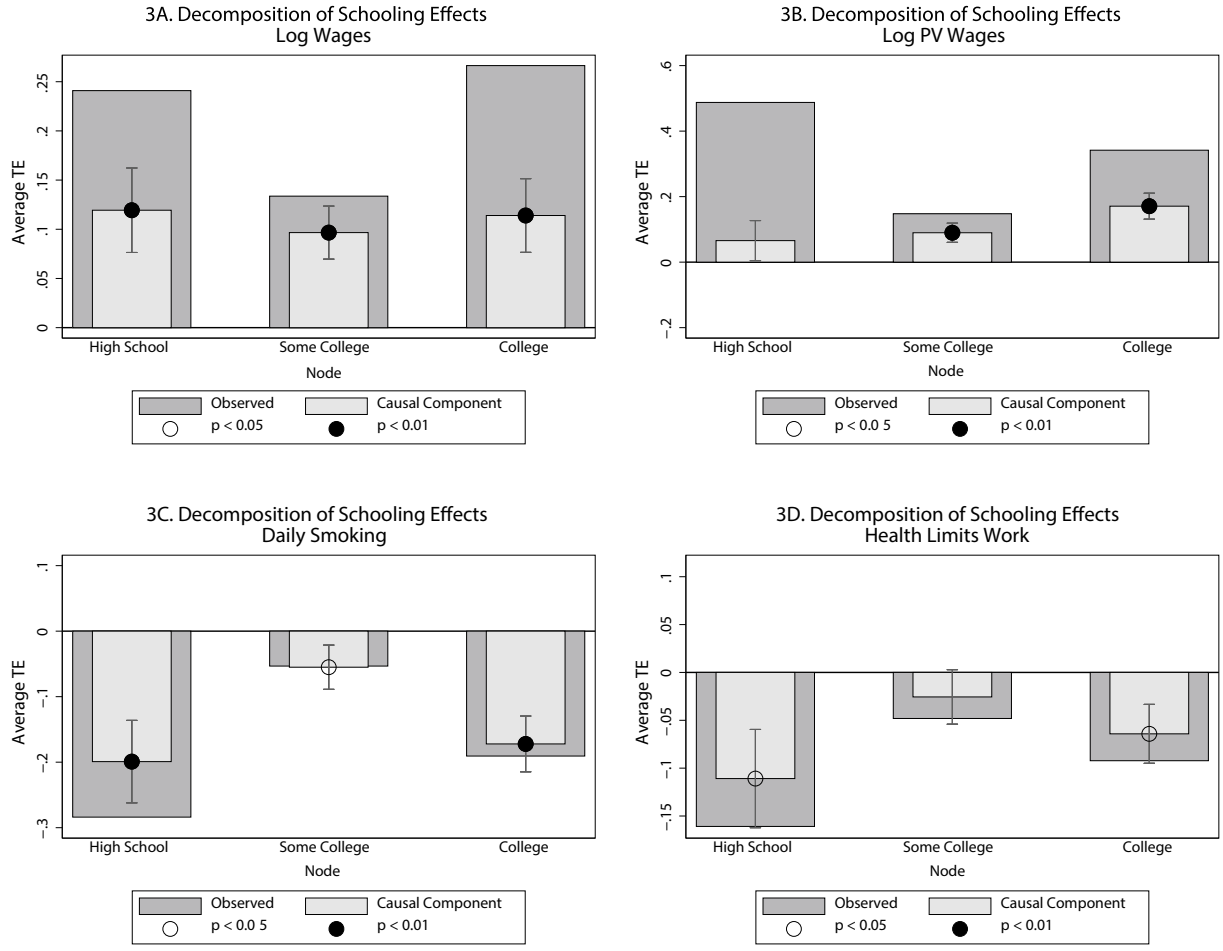
data. The estimated average causal effects (displayed in the light blocks) are large and statistically significant for all outcomes except for the log PV wages for graduating high school (compared to dropping out).[73] For example, the leftmost bar in panel 3a can be interpreted as follows: while high school graduates make on average 24 log points higher wages than high school dropouts, we find that the average causal effect of graduating high school is on average 12 log points for the same population.

Web Appendix A.14.1 reports traditional treatment effects (treatment on the treated, treatment on the untreated, as well as the ATEs displayed in Figure 3). Web Appendix A.15.2 presents estimates of decomposition (20) for all four outcomes. The decompositions show substantial gains for high-ability persons who graduate college. A large component of the observed difference is properly attributed to selection bias for most outcomes.[74]

[73]For that group, the delay in receiving high school wage rates is not sufficiently compensated by higher wage rates.

[74]For a comparison of the treatment effects implicit in Figure 2 with those implicit in Figure 3, see Web Appendix A.17.2, Tables A71–A74. For most outcomes, the agreement is rather close, except for the ln PV of wages.

Figure 3: Causal Versus Observed Differences by Final Schooling Level (compared to next lowest level)



Notes: These figures report pairwise treatment effect (18) for the indicated schooling nodes. Each bar compares the mean outcomes from a particular schooling level j and the next lowest level $j-1$ defined for the set of persons who complete schooling at $j-1$ or j. The "Observed" bar displays the observed differences in the data. The "Causal Component" bar displays the estimated average treatment effect to those who get treated (ATE) for the indicated group. The difference between the observed and causal treatment effect is attributed to the effect of selection and ability. Selection includes sorting on gains. 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.

6.2 Dynamic Treatment Effects

A major contribution of this paper is the estimation of dynamic treatment effects that include continuation values. These are defined for populations that achieve a node ($Q_j = 1$) which includes people who might go beyond j and even $j+1$. Specifically, we calculate the average

gains to fixing $D_j = 0$ (and possibly going beyond $j+1$) compared to those at j ($Q_j = 1$) who stop at j ($D_j = 1$). See Equation (11) for the precise expression. Figure 4 plots these treatment effects by the level of educational decision faced by the agent. These treatment effects are also broken down into those for low-ability and high-ability populations using the ability categories defined at the base of the figure. The figure also reports AMTE for individuals at the margin of indifference at each transition.[75]

There are large and statistically significant average causal effects of education for all wage outcomes.[76] Disaggregating by ability, the effects are strong for high-ability people who enroll in college. They are especially strong for those who graduate college. We find little to no evidence of any benefit of graduating college for low-ability individuals.[77] In fact, the point estimates are negative, albeit imprecisely estimated. Although there are wage rate benefits to low-ability people for enrolling in college (Figure 4A), the benefits in terms of the log present value of wages are minimal. For these people, the wage benefits of attending college barely offset the lost work experience and earnings from attending school.

At all levels of education, the estimated AMTE is substantial: there are marginal benefits to additional education at every transition node for individuals at or near the margin of indifference for that transition. The marginal benefits are close to (but generally somewhat below) the average benefits. This is consistent with diminishing benefits of educational expansion.[78] For people at all margins, there are benefits to taking the next transition that are especially pronounced for high school graduation. There are unrealized potential gains in the current system.

We probe more deeply in the Web Appendices. In A.14.2, we present a variety of treatment effects, including treatment on the treated (TT), treatment on the untreated (TUT), and the average treatment effect defined for the entire population, and not just for those at a

[75] We define the margin of indifference to be $|| I_j/\sigma_j || \leq 0.01$, where σ_j is the standard deviation of I_j.

[76] Across all outcomes, the GED has no benefit.

[77] These estimates are imprecisely determined, in part, because there are few low-ability persons in this category.

[78] A notable exception is for the AMTE for log PV of wages for high school graduation, for which the marginal benefits greatly exceed the average benefits.

particular node in our decision tree. This enables us to examine the extent of sorting on gains. In A.15.3, we go further and decompose observed differences in the data into average treatment effects, sorting gains, and selection bias (Equation (22)).

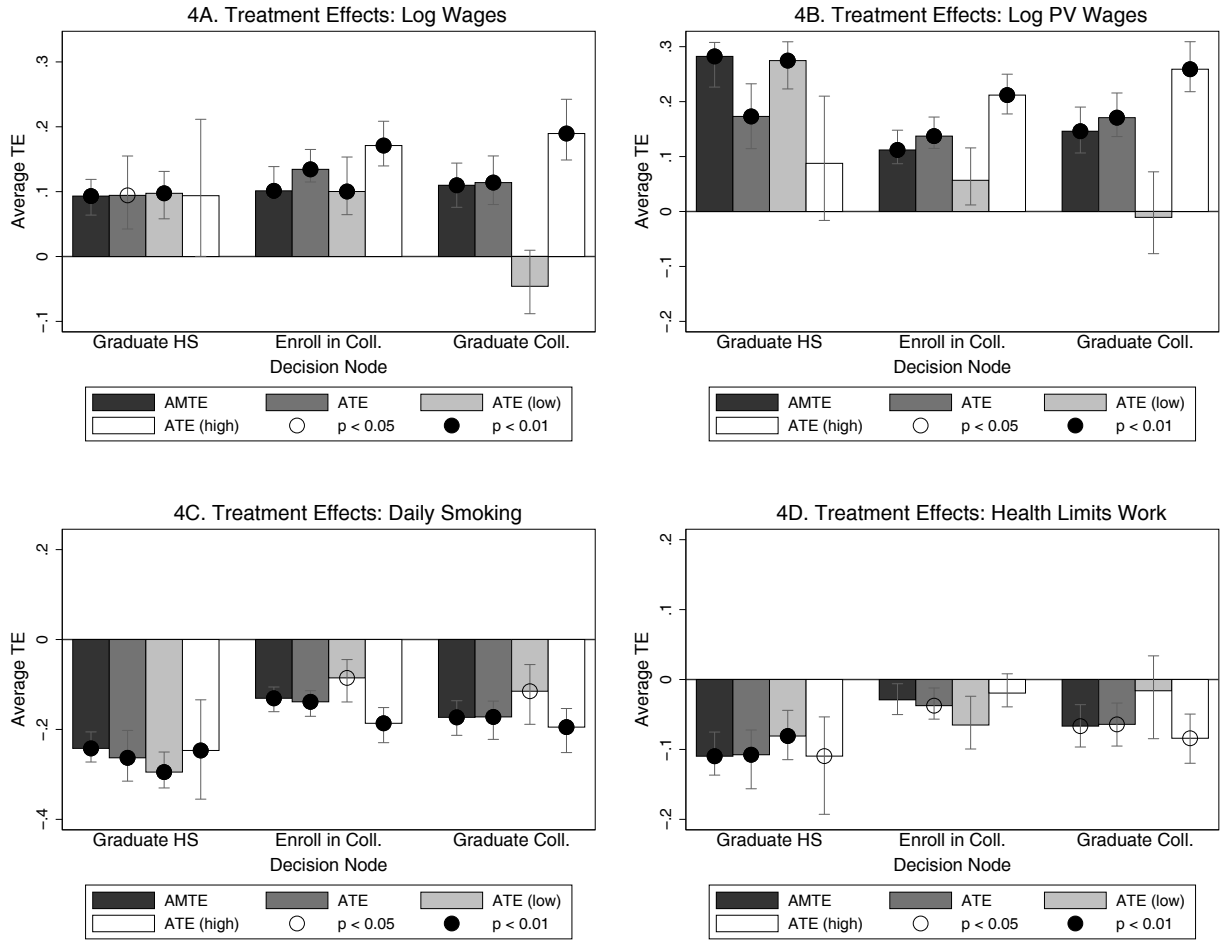
Broadly speaking, for wage outcomes we find sorting on gains for college graduation. This arises primarily from the gains to high-ability people documented in Figure 4. We find the reverse pattern for high school graduation: *negative* sorting for graduating high school that is especially pronounced for log present value of our earnings.[79] Consistent with our analysis of AMTE, there are unrealized gains available in the system for a policy of promoting high school graduation for low-ability individuals. For all wage outcomes, there are substantial selection effects, ranging from 50–70% of the observed differences.

The story is different for the estimated educational causal effects on non-market outcomes. We first discuss smoking. There are strong average causal effects on reducing smoking. They are particularly strong for graduating high school for low-ability individuals. Nonetheless, there are substantial negative average causal effects from education on smoking for all nodes. There is little evidence of sorting on gains. Unlike the evidence for wages, there is substantially less evidence of selection bias at any transition.

The evidence for "Health Limits Work" indicates strongly beneficial causal effects for high school graduation, and weak—but generally precisely determined—causal effects for college graduation, which essentially vanish for low-ability persons. There is little evidence of causal effects of attending some college. There is no evidence of sorting on gains. Selection bias is a strong component of observed differences.

[79] TUT > ATE > TT.

Figure 4: Treatment Effects of Outcomes by Decision Node
$E(Y^k|\textbf{Fix } D_j = 0, Q_j = 1) - E(Y^k|\textbf{Fix } D_j = 1, Q_j = 1)$



Sorting on Ability		
	Low Ability	High Ability
D_1: Dropping from HS vs. Graduating from HS	0.31	0.31
D_2: **HS Graduate** vs. College Enrollment	0.22	0.38
D_3: **Some College** vs. **Four-Year College Degree**	0.13	0.51

Notes: The nodes in the table correspond to the next stage of the transition analyzed. Thus, "Graduate HS" refers to the decision node of whether or not the agent will graduate high school, and refers to the base state of not graduating high school. The error bars and significance levels for the estimated ATE Equation (13) 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 socio-emotional endowments above (below) the median in the overall population. These categories are not mutually exclusive, as some people may be high-ability in one dimension but low-ability in another. 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.

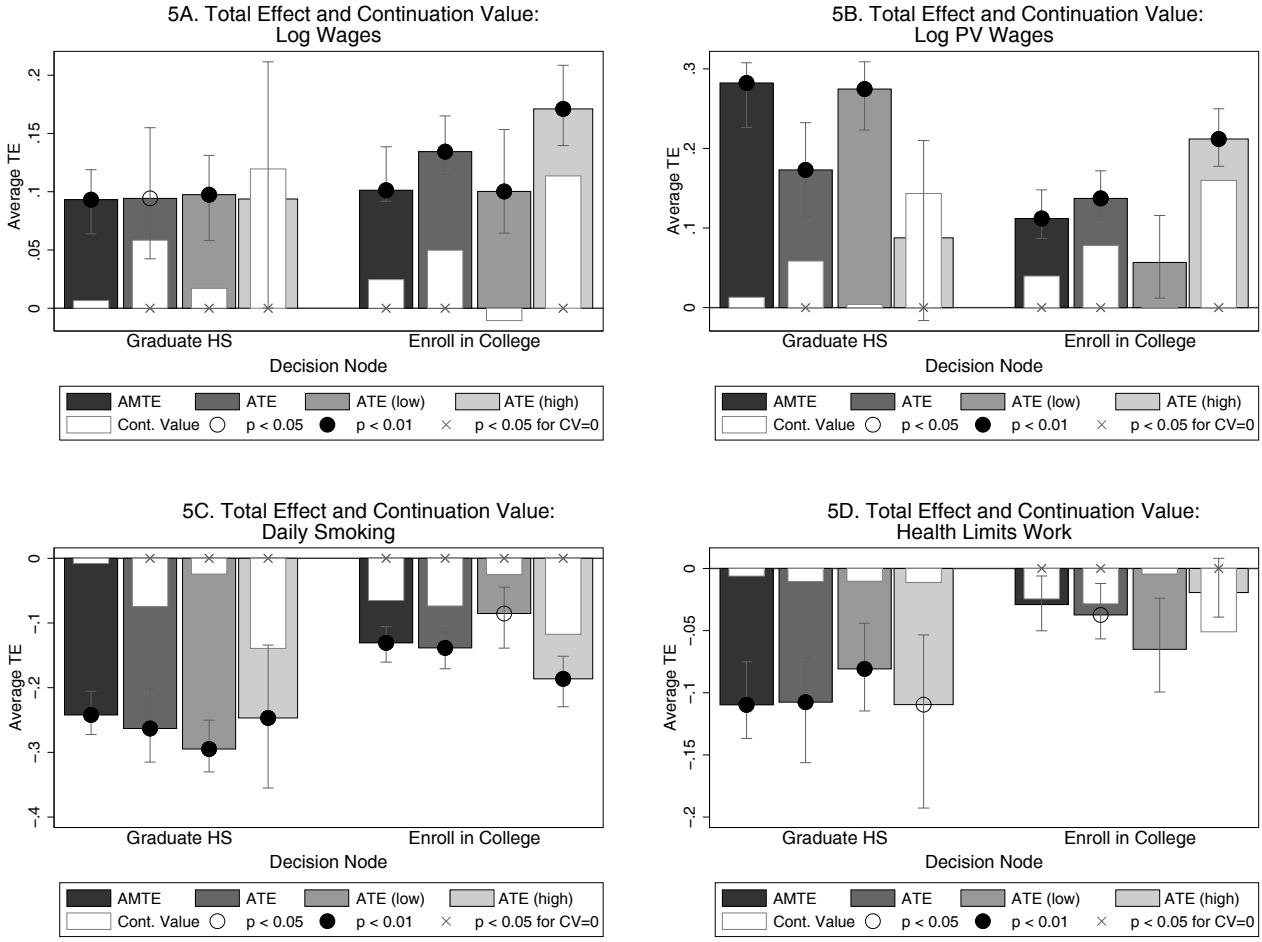
6.2.1 Continuation Values

We next decompose the node-specific average treatment effects, just discussed, into continuation value components. Figure 5 presents (in the white bars) the estimated continuation value components, while the bars behind the white boxes are the full treatment effects from Figure 4. We only display continuation values for nodes where these are possible.

For all outcomes except "Health Limits Work," there are large continuation values for high-ability individuals. While returns to high school are roughly the same across ability levels, the mechanisms producing these effects are different. The benefits for low-ability persons come through direct values. The benefits for high-ability persons come through continuation values. For most nodes and treatment effects, the continuation values are statistically significant as indicated by the "x" in Figure 5.[80]

[80]Note that if direct effects are negative, continuation values may be larger than treatment effects.

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 socio-emotional endowments. Low-ability individuals are those in the bottom 50% of the distributions of both cognitive and socio-emotional 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 defined.

6.3 The Effects on Cognitive and Non-Cognitive Endowments on Treatment Effects

Disaggregating the treatment effects for "high-" and "low-" endowment $\boldsymbol{\theta}$ individuals in Figure 4 is a coarse approach. A byproduct of our analysis is that we can determine the contribution of cognitive and non-cognitive endowments ($\boldsymbol{\theta}$) to the explanation of estimated

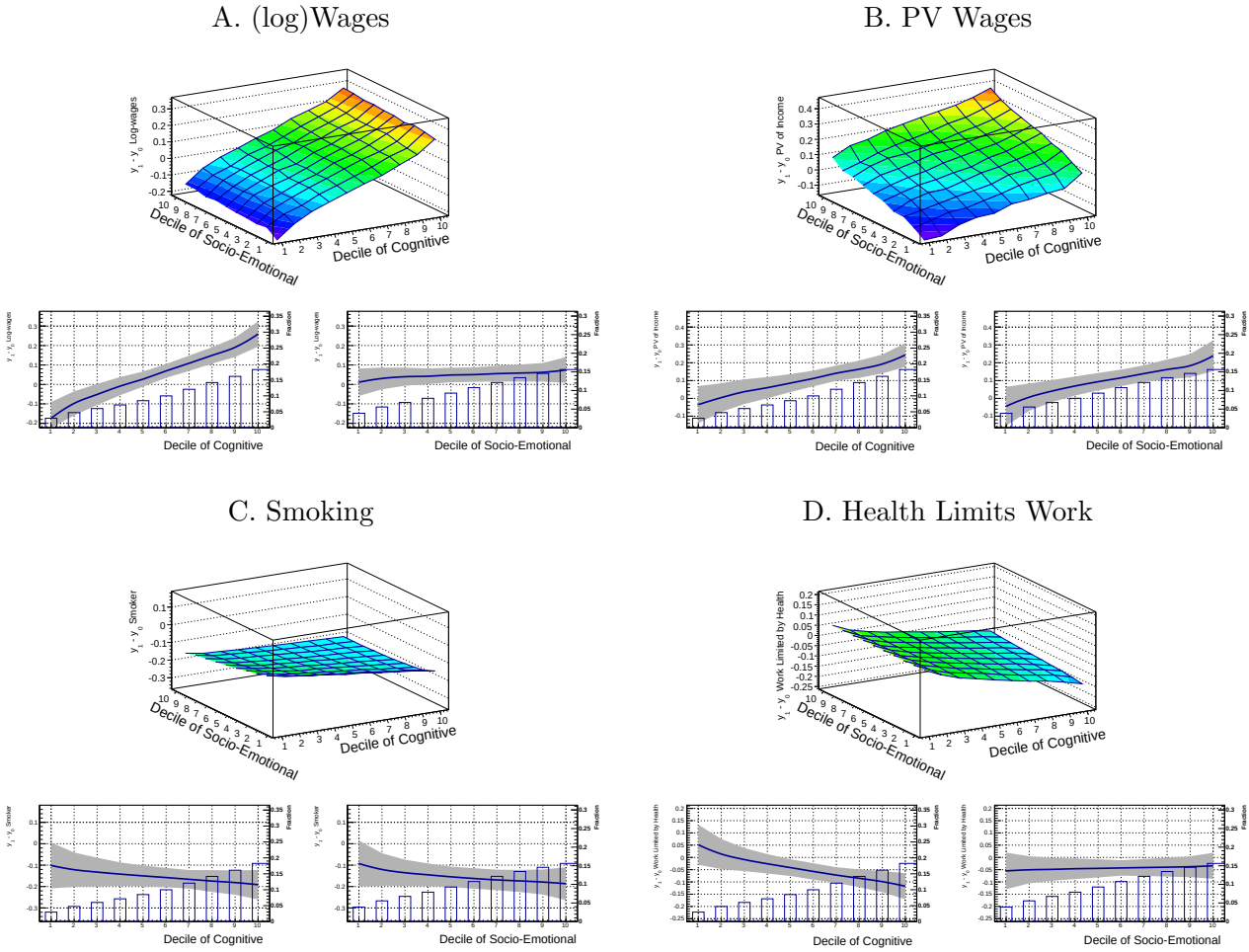
treatment effects. We can decompose the overall effects of $\boldsymbol{\theta}$ into their contribution to the causal effects at each node and the contribution of endowments to attaining that node. We find substantial contributions of $\boldsymbol{\theta}$ to each component at each node.

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 non-cognitive endowments.[81,82] Treatment effects, in general, depend on both measures of ability. Moreover, different outcomes depend on the two dimensions of ability in different ways. For example, the treatment effect of graduating college is increasing in both dimensions for the present values of wages, but the reductions in health limitations with education depend mostly on cognitive endowments.

[81] Web Appendix A.7 reports a full set of results across all nodes.

[82] These figures 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



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 $Y_{\text{some college}}$ and $Y_{\text{four-year degree}}$ denote the outcomes associated with attaining some college and graduating with a four-year degree, respectively. For each outcome, the first figure (top) presents $E(Y_{\text{four-year degree}} - Y_{\text{some college}}|d^C, d^{SE})$ where d^C and d^{SE} denote the cognitive and socio-emotional deciles computed from the marginal distributions of cognitive and socio-emotional endowments. The second figure (bottom left) presents $E(Y_{\text{four-year degree}} - Y_{\text{some college}}|d^C)$ so that the socio-emotional 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 $E(Y_{\text{four-year degree}} - Y_{\text{some college}}|d^{SE})$ 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 socio-emotional endowment.

6.4 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—a feature 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.[83]

The graphs provide a nice summary of our main findings for all dynamic treatment effects for college graduation. There are strong causal effects for all outcomes. There is also substantial heterogeneity among persons. Sorting on gains is pronounced for wage outcomes but less so for health and smoking. This is consistent with the analysis in Appendix A.15.3, where we report estimates of sorting on gains.

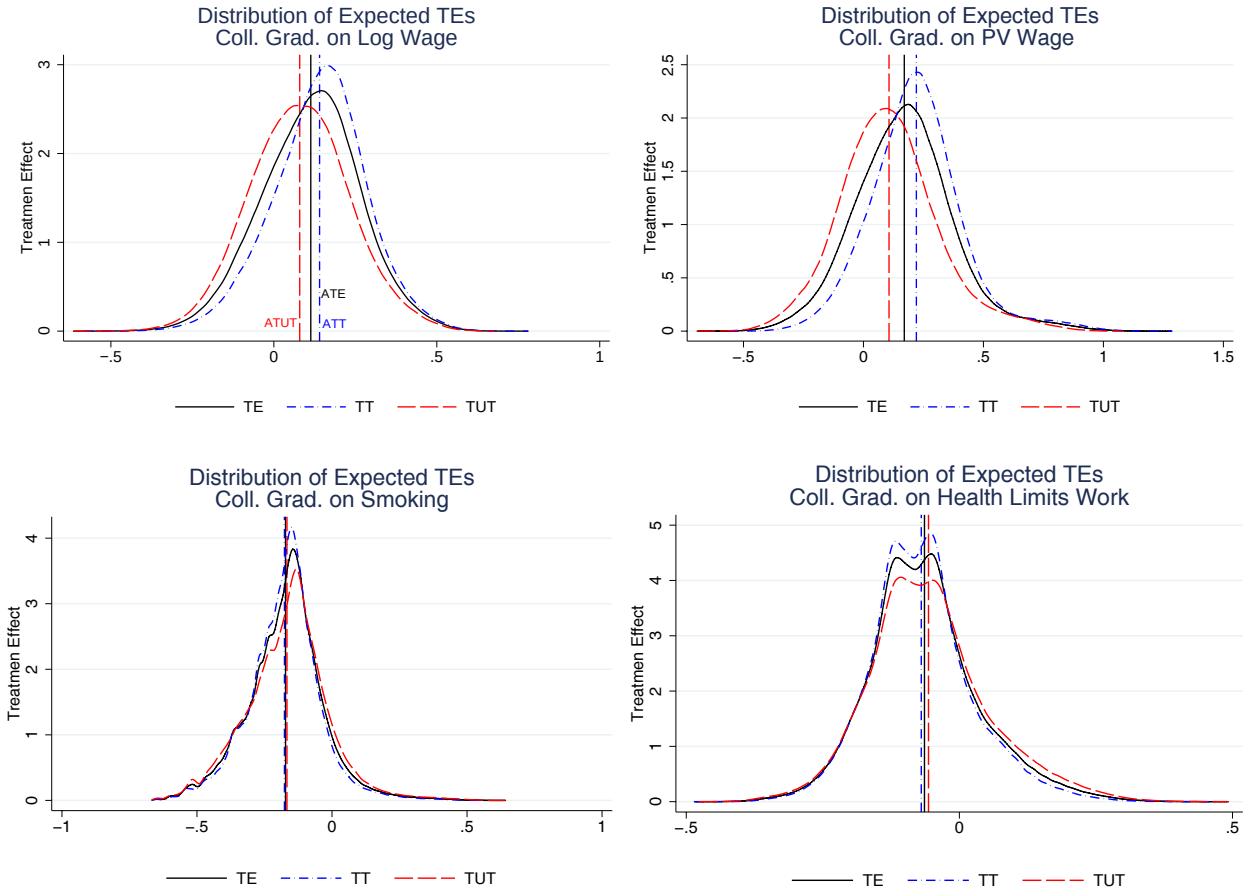
A byproduct of our analysis is that we can test the rank-invariance of counterfactual outcomes across states. The assumption of rank invariance is the basis for the numerous analyses based on quantile treatment effects.[85] It implies that the Spearman correlations are 1 across any pair of counterfactual states. In our simulations, we find that the Spearman correlations are large but are also not 1. They are between 0.70 and 0.85 for log wages, 0.60 and 0.90 for present value of wages, and notably smaller for smoking and health limitations.[86] Rank invariance is an especially poor assumption for those outcomes.

[83]Web Appendix A.9 reports a full set of distributions of treatment effects for all outcomes. Expectations are computed over the idiosyncratic error terms (ω_s^k).[84] Variation in the expected treatment effect comes from the variation in observed variables ($\boldsymbol{X}$) and the unobserved endowments ($\boldsymbol{\theta}$).

[85]See, e.g., Bitler et al. (2006).

[86]See Web Appendix Tables A22–A25.

Figure 7: Distributions of Expected Treatment Effects: College Graduation



Notes: Distributions of treatment effects including continuation values for those who reach the educational choice. The vertical lines represent the average treatment effects (ATE, ATT, and ATUT) for each of the distributions.

6.5 Taking Stock of the Becker-Chiswick-Mincer Model

We have tested many features of the widely-used model of Equation (1) to determine its robustness. Some features of Mincer model (1) are broadly consistent with our estimated structural model. While OLS-regression-adjusted versions are not linear in years of schooling (see the evidence in Web Appendix A.8), our estimated ATEs are roughly consistent with linearity for most outcomes.

The correlation between ρ_i and S_i is a centerpiece of the modern IV literature.[87] It varies across transitions (see Web Appendix A.13). However, ρ_i turns out to be node-specific ($\rho_{s,s',i}$)

[87] This is the correlated random coefficient model. See Heckman and Vytlacil (1998).

and not the same across transitions.

Sorting on gains, measured either for specification as in (19) ($COV(\rho_{s,s'}, D_s) \neq 0$), or for specification as in (21) ($COV(\rho_{j-1,j}, Q_j) \neq 0$), reveals that there is positive sorting on wage gains only at the higher levels of education. Our estimated correlation patterns are consistent with our evidence on sorting gains presented in Web Appendix A.15.

6.6 Summarizing Our Analysis of Causal Effects of Education

In this section and in our Web Appendix, we have analyzed a variety of economically interpretable treatment effects. We reach the following broad conclusions.

(1) There are substantial causal benefits for all outcomes analyzed from education, except for GED certification.

(2) Continuation values are an important component of causal effects for most outcomes except health limits work.

(3) There are substantial benefits from graduating high school that are especially strong for the less able, many of whom currently do not graduate. This suggests strong gains from programs promoting high school graduation.

(4) For the wage outcomes we study, there is evidence on sorting on gains from graduating college for high-ability persons.[88] There are no causal effects of college graduation for low-ability persons. College graduation is not for all.

(5) There are strong benefits of education for those at the margin of indifference at all nodes. These are largely direct effects with little contribution from continuation values.

(6) We estimate strong causal effects for the non-monetary outcomes studied. They are particularly strong for high school graduation. There is little evidence of sorting on gains in either non-monetary outcome examined. Continuation values are largely absent for our measure of health. For smoking, continuation values are most pronounced among higher-

[88]Part of the relationship between ability and returns to college could operate through college quality. For example, Dillon and Smith (2015) show that ability is an important determinant of college quality and that college quality improves wages even after controlling for ability.

ability persons. Selection bias is less empirically important for smoking, but is substantial for health limits work.

7 Policy Simulations from Our Model

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. We consider two policy experiments: (i) a tuition subsidy; and (ii) an increase in the cognitive and non-cognitive endowments of those at the bottom of the endowment distribution. The first policy experiment is similar to what is estimated by LATE only in the special case where the instrument in LATE corresponds to the exact policy experiment. The second policy experiment is of interest because early childhood programs boost these endowments (Heckman et al., 2013a). Neither set of counterfactuals generated can be estimated from instrumental variable estimands. We ignore general equilibrium effects in 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 of the power of our methodology, we simulate the response to a policy intervention that provides a one standard deviation subsidy to early college tuition (approximately $850 dollars per year of college). Column 1 of Table 2 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 we do not find evidence that college tuition affects high school graduation rates, the subsidy only induces high school graduates to change their college enrollment decisions and does not affect high school graduation decisions. Those

induced to enroll may then go on to graduate with a four-year degree.[89] Columns 2 and 3 of Table 2 decompose the PRTE into the average gains for those induced to enroll and then go on to earn four-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 four-year degrees.

Figure 8 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,[90] 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 decomposes the effect of those induced to switch into effects for those who go on to graduate with four-year degrees and effects for those who do not. Those induced to switch in the top deciles are more likely to go on to graduate with a four-year college degree.

Table 2: PRTE: Standard Deviation Decrease in Tuition

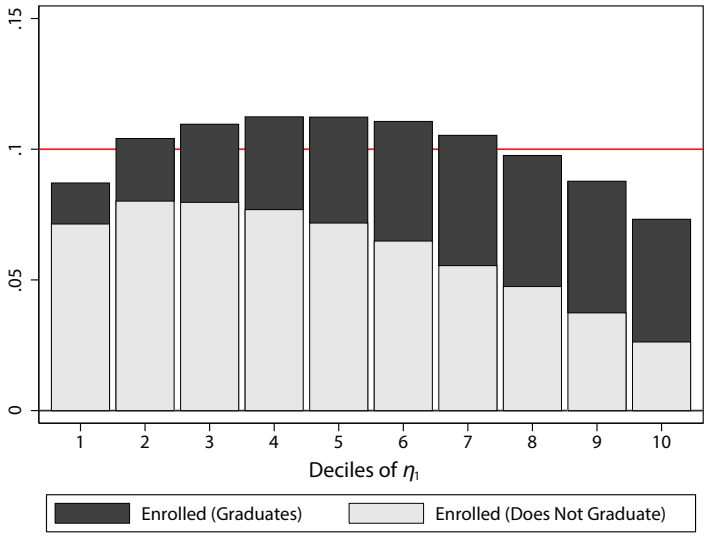
	PRTE		Four-Year Degree		No Four-Year Degree	
Log Wages	0.125	(0.023)	0.143	(0.027)	0.114	(0.027)
PV Log Wages	0.129	(0.03)	0.138	(0.033)	0.123	(0.028)
Health Limits Work	-0.036	(0.022)	-0.025	(0.021)	-0.043	(0.023)
Smoking	-0.131	(0.029)	-0.166	(0.030)	-0.108	(0.030)

Notes: The table shows the policy-relevant treatment effect (PRTE) of reducing tuition for the first two years of college by a standard deviation (approx. $850 per annum). The PRTE is the average treatment effect of those induced to change educational choices as a result of the policy:
$PRTE^k_{p,p'} := \int\int\int E(Y^k(p') - Y^k(p)|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{\theta} = \overline{\boldsymbol{\theta}}) dF_{\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}, \overline{\boldsymbol{\theta}}|S(p) \neq S(p'))$. 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 four-year college degrees. Column 3 shows the PRTE for individuals induced to enroll but who do not complete four-year degrees.

[89] Models were estimated that include tuition as a determinant of the high school graduation decision. However, the 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. We do not impose the requirement that agents know and act on publicly available information.

[90] The unobservable is the bundle $\eta_1 = -(\boldsymbol{\theta}'\boldsymbol{\lambda}_1 - \nu_1)$.

Figure 8: 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 η_2, where $\eta_1 = -(\boldsymbol{\theta}'\boldsymbol{\lambda}_1 - \nu_1)$. η_1 is the unobserved component of the educational choice model. The deciles are conditional on $Q_1 = 1$, so η_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 four-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 four-year degree. For outcomes such as smoking, the benefits are larger for those who graduate with a four-year degree. The large gains for marginal individuals induced to enroll is consistent with the empirical literature, that finds large psychic costs are necessary to justify college schooling choices, and the failure of agents to respond to strong monetary incentives.

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.[91] Given a PRTE of 0.13 for log present value of wage income, the average gains for those induced to enroll is \$36,401 in year 2000 dollars. 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).

[91] However, a limitation of our model is that we can only estimate the monetary costs and do not estimate psychic costs.

7.2 Boosting Cognitive and Non-Cognitive Endowments

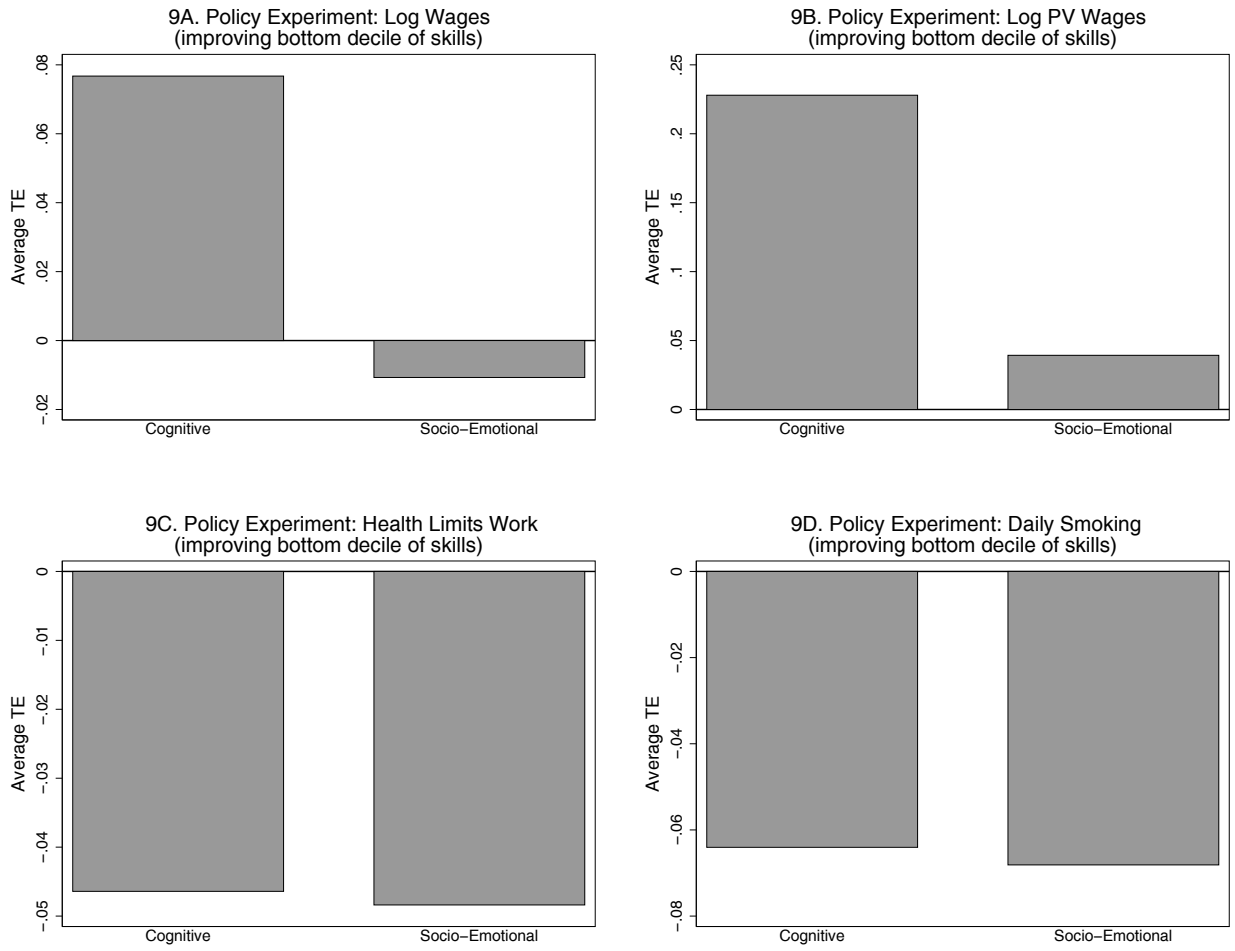
Using simulation methods, it is possible to construct counterfactual policy simulations unrelated to any particular set of instruments. For example, some early childhood programs have been shown to have lasting impacts on the cognitive and non-cognitive endowments of low-ability children (see Heckman et al., 2013a). We simulate two policy experiments: (i) increasing the cognitive endowment of those in the lowest decile; and (ii) increasing the non-cognitive endowment of those in the lowest decile.[92]

The panels of Figure 9 show the average gains for increasing the cognitive or non-cognitive endowments 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 has substantial effects on health.[93]

[92]The details of how these simulations were conducted are presented in Web Appendix A.11. Our model does not address general equilibrium effects of such a change in the endowment distribution.

[93]We present additional policy simulations in Web Appendix A.11.

Figure 9: Policy Experiments



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

8 Testing the Two-Factor Assumption

Throughout this paper, we have assumed that selection of outcomes occurs on the basis of a two-component vector $\boldsymbol{\theta}$, where the components can be proxied by our measures of cognitive and non-cognitive endowments. An obvious objection to this approach is that there may be unproxied endowments that affect both choices and outcomes that we do not measure. For example, one could imagine that a component of the idiosyncratic error terms

in the educational choices (ν_j) represent taste for schooling. This could generate correlations between the unobservables in the different educational choices and bias our results.

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 $D_j, j = 1, \ldots, \overline{s}, k = 1, \ldots, 4$. If an important third factor common to both outcome and choice equations has been omitted, the sample fit should be poor. In fact, we find a good fit.

Cunha and Heckman (2016) 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 evidence is consistent with their findings.

9 Comparisons with Simple Treatment Effect Estimators

In this paper, we have exploited the assumption of conditional independence of outcomes and choices given $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}$. This raises the question of how similar our results would be if we had used simple matching and regression methods.

Table 3 presents two sets of estimates for the models discussed in detail at the base of the table. The first four columns of numbers are node-specific linear-regression estimators of Equation (11), using as regressors the background variables reported in Table 1 but not the "exclusion restriction" variables. The first column of estimates come from a model without any control for $\boldsymbol{\theta}$. The estimates for the other three models control for $\boldsymbol{\theta}$ in various ways as noted at the base of Table 3.

We use two versions of nearest-neighbor matching estimators based on the full set of control variables listed in Table 1. Details of the matching procedures are given at the base of Table 3 and in Web Appendix A.18.

Table 3: Average Treatment Effects - Comparison of Estimates from Our Model to Those from Simpler Methods

	Linear Regression				Matching		Model
HS Grad.	OLS	OLS-P	OLS-F	OLS-FI	NNM(3)-F	PSM-F	ATE_j^k *
Wages	0.205	0.073	0.155	0.159	0.098	0.132	0.094
SE	(0.025)	(0.026)	(0.025)	(0.035)	(0.037)	(0.051)	(0.056)
PV-Wage	0.380	0.213	0.318	0.277	0.196	0.226	0.173
SE	(0.030)	(0.031)	(0.030)	(0.041)	(0.053)	(0.058)	(0.059)
Smoking	-0.327	-0.246	-0.281	-0.301	-0.260	-0.271	-0.263
SE	(0.028)	(0.029)	0.028	0.041	(0.058)	(0.060)	(0.056)
Health-Limits-Work	-0.178	-0.115	-0.151	-0.150	-0.048	-0.095	-0.108
SE	(0.022)	(0.024)	(0.023)	(0.033)	(0.029)	(0.036)	(0.042)
Coll. Enroll	OLS	OLS-P	OLS-F	OLS-FI	NNM(3)-F	PSM-F	ATE_j^k
Wages	0.223	0.121	0.186	0.190	0.177	0.207	0.134
SE	(0.023)	(0.024)	(0.024)	(0.023)	(0.029)	(0.031)	(0.025)
PV-Wage	0.221	0.109	0.176	0.171	0.188	0.226	0.137
SE	(0.027)	(0.029)	(0.028)	(0.027)	(0.030)	(0.032)	(0.029)
Smoking	-0.177	-0.138	-0.165	-0.170	-0.129	-0.144	-0.139
SE	(0.026)	(0.028)	(0.027)	(0.028)	(0.029)	(0.058)	(0.028)
Health-Limits-Work	-0.085	-0.037	-0.066	-0.057	-0.029	-0.042	-0.037
SE	(0.020)	(0.022)	(0.021)	(0.021)	(0.022)	(0.029)	(0.022)
Coll. Grad	OLS	OLS-P	OLS-F	OLS-FI	NNM(3)-F	PSM-F	ATE_j^k
Wages	0.210	0.146	0.184	0.185	0.173	0.143	0.114
SE	(0.032)	(0.034)	(0.033)	(0.035)	(0.041)	(0.051)	(0.037)
PV-Wage	0.243	0.163	0.208	0.228	0.191	0.269	0.171
SE	(0.037)	(0.040)	(0.038)	(0.037)	(0.039)	(0.042)	(0.040)
Smoking	-0.209	-0.171	-0.195	-0.192	-0.132	-0.161	-0.172
SE	(0.032)	(0.035)	(0.033)	(0.035)	(0.039)	(0.039)	(0.043)
Health-Limits-Work	-0.085	-0.069	-0.078	-0.077	-0.048	-0.051	-0.064
SE	(0.024)	(0.026)	(0.025)	(0.026)	(0.026)	(0.027)	(0.031)

Notes: We estimate the ATE inclusive of continuation values for each outcome and and educational choice using a variety of methods. All models are estimated for populations that reach the node being analyzed ($Q_j = 1$), inclusive of those who go on to further schooling in order to make them comparable to the ATE from our model that includes continuation values (Equation (21)). All OLS models use the full set of controls listed in Table 1. "OLS" estimates a linear model using a schooling dummy (Q_{j+1}), and controls ($Y = Q_{j+1}b_j + \boldsymbol{X}'\boldsymbol{\beta} + \epsilon$). "OLS-P" estimates a linear model using a schooling dummy, a vector of controls, and three measures of abilities arrayed in a vector $\boldsymbol{A}$: summed ASVAB scores, GPA, and an indicator of risky behavior ($Y = Q_{j+1}b_j + \boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{A}'\boldsymbol{\alpha} + \epsilon$). All models ending in "-F" are estimated using Bartlett factor scores (Bartlett, 1937, 1938) estimated using our measurement system, but using the built-in routine for estimating factor models in STATA via maximum likelihood, not accounting for schooling at the time of the test. "OLS-F" estimates the model $Y = Q_{j+1}b_j + \boldsymbol{X}'\boldsymbol{\beta} + \hat{\boldsymbol{\theta}}'\boldsymbol{\alpha} + \epsilon$ where $\hat{\boldsymbol{\theta}}$ are the Bartlett factor scores described above. "OLS-FI" is similar to "OLS-F" except that Q_{j+1} is interacted with the $\boldsymbol{X}$ and $\hat{\boldsymbol{\theta}}$ allowing the coefficients on the controls and abilities to vary by education level. "NNM(3)-F" is the estimated treatment effect from nearest-neighbor matching with 3 neighbors. Neighbors are matched on their Bartlett cognitive factor, Bartlett non-cognitive factor, and an index constructed from their observed characteristics ($\boldsymbol{Z}$) generating choices as described in Web Appendix A.18. "PSM-F" presents the estimated average treatment effect from propensity score matching where propensity scores are estimated using Bartlett cognitive factors, Bartlett non-cognitive factors, the full set of control variables, and the full set of node-specific instruments. "ATE_j^k" presents the estimated average treatment effect from the model presented in this paper (inclusive of continuation value), corresponding to Equation (13).

The OLS estimates differ greatly from model estimates when there is no adjustment for ability. Controlling for ability has substantial effects on the estimated average treatment effects. Across schooling nodes, all of the estimates that control for $\boldsymbol{\theta}$ are "within the ball park" of the estimates produced from our model, although some discrepancies are substantial. This is good news for applied economists mainly interested in using simple methods to estimate node-specific average treatment effects. However, these simple methods do not estimate decision rules, do not enable analysts to estimate AMTE and PRTE, or address many of the other questions addressed in this paper.[94]

10 Summary and Conclusion

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

The early literature on human capital ignored the dynamics of schooling choices. This paper develops and estimates a robust dynamic model of schooling and its causal consequences for earnings, health, and smoking. Our 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 rationality and information processing assumptions maintained in the dynamic discrete choice literature on education.

We use our dynamic choice model to estimate causal effects arising from multiple levels of schooling rather than just the binary comparisons typically featured in the literature on

[94] Table A70 of the Web Appendix compares OLS estimates of dynamic treatment effects and continuation values with our model estimates. The OLS estimates are "within ballpark" for smoking and health limits work, but they are wide off the mark for wages and PV wages.

treatment effects and in many structural papers.[95] By estimating a sequential model of schooling in a unified framework, we are able to analyze the *ex post* returns to education for people at different margins of choice and analyze a variety of economically interesting policy counterfactuals. We are able to characterize who benefits from education for a variety of market and non-market outcomes.

We decompose the benefits of schooling at different levels into direct components and indirect components arising from continuation values. We estimate substantial continuation value components of graduating high school and completing college 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 of dynamic treatment effects underestimate the full benefits of education. For low-ability individuals, there are substantial direct effects of graduating high school, but little continuation value.

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. Both cognitive and non-cognitive endowments affect schooling choices and outcomes at each level of schooling.

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.[96] They roughly approximate our model estimates of average treatment effects, provided we condition on endowments of cognitive and non-cognitive skills. However, these simple methods *do not* identify the treatment effects for persons at the margins of different choices (the average marginal treatment effects).[97] We test the empirical foundations of the Mincer model and find it wanting. A richer specification of the schooling earnings decision is warranted to generate empirically supported estimates of causal effects.

We use our estimated model to conduct two policy experiments. We determine the groups

[95] See, e.g., Willis and Rosen (1979).

[96] IV estimates are very different from our model estimates. See Heckman et al. (2016).

[97] We can roughly approximate continuation values using simple methods. See Table A70 in the Web Appendix.

that benefit from a tuition reduction policy and what those benefits are. We also examine how the impact of boosts in cognitive and non-cognitive skills affects educational choices and outcomes.

Our analysis enriches the pioneering analysis of Becker (1964). The early research on human capital was casual about agent heterogeneity. It ignored selection bias and sorting gains from schooling. Later work by Griliches (1977) focused on selection bias ("ability bias"), but ignored sorting gains. In this paper, we quantify both components of outcome equations. We find evidence of selection bias at all levels of schooling for all outcomes and sorting gains at higher levels of schooling for wage outcomes.

Our findings thus support the basic insights of Becker (1964). Schooling has strong causal effects on market and non-market outcomes. Both cognitive and non-cognitive endowments affect schooling choices and outcomes. People sort into schooling based on realized incremental gains.

References

Abbring, Jaap H., and James J. Heckman, 2007. "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation." In *Handbook of Econometrics*, vol. 6B, edited by James J. Heckman, and Edward E. Leamer, chap. 72. Amsterdam: Elsevier Science B. V., pp. 5145–5303.

Adda, Jérôme, and Russell W. Cooper, 2003. *Dynamic Economics: Quantitative Methods and Applications*. Cambridge, MA: The MIT Press.

Almlund, Mathilde, Angela Duckworth, James J. Heckman, and Tim Kautz, 2011. "Personality Psychology and Economics." In *Handbook of the Economics of Education*, vol. 4, edited by Eric A. Hanushek, Stephen Machin, and Ludger Wößmann, chap. 1. Amsterdam: Elsevier, pp. 1–181.

Altonji, Joseph G., 1993. "The Demand for and Return to Education When Education Outcomes Are Uncertain." *Journal of Labor Economics* 11 (1):48–83.

Angrist, Joshua D., and Guido W. Imbens, 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association* 90 (430):431–442.

Angrist, Joshua D., and Jörn-Steffan Pischke, 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Arcidiacono, Peter, and Robert A. Miller, 2011. "Conditional Choice Probability Estimation of Dynamic Discrete Choice Models with Unobserved Heterogeneity." *Econometrica* 7 (6):1823–1868.

Bamberger, Gustavo, 1987. "Occupational Choice: the Role of Undergraduate Education." Ph.D. thesis, University of Chicago, Graduate School of Business.

Bartlett, Maurice S., 1937. "The Statistical Conception of Mental Factors." *British Journal of Psychology* 28 (1):97–104.

———, 1938. "Methods of Estimating Mental Factors." *Nature* 141:609–610.

Becker, Gary S., 1962. "Investment in Human Capital: A Theoretical Analysis." *Journal of Political Economy* 70 (5, Part 2):9–49.

———, 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Chicago: University of Chicago Press for the National Bureau of Economic Research, 1st edn.

Becker, Gary S., and Barry R. Chiswick, 1966. "Education and the Distribution of Earnings." *American Economic Review* 56 (1/2):358–369.

Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes, 2006. "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments." *American Economic Review* 96 (4):988–1012.

Blevins, Jason R., 2014. "Nonparametric Identification of Dynamic Decision Processes with Discrete and Continuous Choices." *Quantitative Economics* 5 (3):531–554.

Borghans, Lex, Angela L. Duckworth, James J. Heckman, and Bas ter Weel, 2008. "The Economics and Psychology of Personality Traits." *Journal of Human Resources* 43 (4):972–1059.

Cameron, Stephen V., and James J. Heckman, 1993. "The Nonequivalence of High School Equivalents." *Journal of Labor Economics* 11 (1, Part 1):1–47.

———, 2001. "The Dynamics of Educational Attainment for Black, Hispanic, and White Males." *Journal of Political Economy* 109 (3):455–499.

Cameron, Stephen V., and Christopher Taber, 2004. "Estimation of Educational Borrowing Constraints Using Returns to Schooling." *Journal of Political Economy* 112 (1):132–182.

Card, David, 1999. "The Causal Effect of Education on Earnings." In *Handbook of Labor Economics*, vol. 3A, edited by Orley C. Ashenfelter, and David Card, chap. 30. Amsterdam: Elsevier Science B.V., pp. 1801–1863.

———, 2001. "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems." *Econometrica* 69 (5):1127–1160.

Carneiro, Pedro, Karsten Hansen, and James J. Heckman, 2003. "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice." *International Economic Review* 44 (2):361–422.

Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil, 2010. "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin." *Econometrica* 78 (1):377–394.

———, 2011. "Estimating Marginal Returns to Education." *American Economic Review* 101 (6):2754–2781.

Carneiro, Pedro, Costas Meghir, and Matthias Parey, 2013. "Maternal Education, Home Environments, and the Development of Children and Adolescents." *Journal of the European Economic Association* 11 (S1):123–160.

Cawley, John, Karen Conneely, James J. Heckman, and Edward J. Vytlacil, 1997. "Cognitive Ability, Wages, and Meritocracy." In *Intelligence, Genes, and Success : Scientists Respond to The Bell Curve*, edited by Bernie Devlin, Stephen E. Fienberg, Daniel P. Resnick, and Kathryn Roeder, chap. 8. New York: Springer Verlag, pp. 179–192.

Comay, Yochanan, Arie Melnik, and Moshe A. Pollatschek, 1973. "The Option Value of Education and the Optimal Path for Investment in Human Capital." *International Economic Review* 14 (2):421–435.

Cunha, Flávio, and James J. Heckman, 2016. "Decomposing Trends in Inequality in Earnings into Forcastable and Uncertain Components." Forthcoming, *Journal of Labor Economics*.

Cunha, Flávio, James J. Heckman, and Salvador Navarro, 2007. "The Identification and Economic Content of Ordered Choice Models with Stochastic Cutoffs." *International Economic Review* 48 (4):1273–1309.

Cutler, David M., and Adriana Lleras-Muney, 2010. "Understanding Differences in Health Behaviors by Education." *Journal of Health Economics* 29 (1):1–28.

Dillon, Eleanor Wiske, and Jeffrey Andrew Smith, 2015. "The Consequences of Academic Match between Students and Colleges." IZA Discussion Paper 9080, Institute for the Study of Labor, Bonn.

Dothan, Uri, and Joseph Williams, 1981. "Education as an Option." *The Journal of Business* 54 (1):117–139.

Eckstein, Zvi, and Kenneth I. Wolpin, 1989. "The Specification and Estimation of Dynamic Stochastic Discrete Choice Models: A Survey." *Journal of Human Resources* 24 (4):562–598.

Eisenhauer, Philipp, James J. Heckman, and Stefano Mosso, 2015a. "Estimation of Dynamic Discrete Choice Models by Maximum Likelihood and the Simulated Method of Moments." *International Economic Review* 56 (2):331–357.

Eisenhauer, Philipp, James J. Heckman, and Edward J. Vytlacil, 2015b. "Generalized Roy Model and Cost-Benefit Analysis of Social Programs." *Journal of Political Economy* 123 (2):413–433.

Geweke, John, and Michael Keane, 2001. "Computationally Intensive Methods for Integration in Econometrics." In *Handbook of Econometrics*, vol. 5, edited by James J. Heckman, and Edward E. Leamer, chap. 56. Amsterdam: Elsevier Science B. V., pp. 3463–3568.

Griliches, Zvi, 1977. "Estimating the Returns to Schooling: Some Econometric Problems." *Econometrica* 45 (1):1–22.

Grossman, Michael, 2000. "The Human Capital Model." In *Handbook of Health Economics*, vol. 1, edited by Anthony J. Culyer, and Joseph P. Newhouse, chap. 7. Amsterdam: Elsevier Science B. V., pp. 347–408.

Haavelmo, Trygve, 1943. "The Statistical Implications of a System of Simultaneous Equations." *Econometrica* 11 (1):1–12.

Heckman, James J., 1981. "The Empirical Content of Alternative Models of Labor Earnings." Unpublished manuscript, University of Chicago, Department of Economics.

———, 2001. "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture." *Journal of Political Economy* 109 (4):673–748.

———, 2008. "Econometric Causality." *International Statistical Review* 76 (1):1–27.

———, 2010. "Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy." *Journal of Economic Literature* 48 (2):356–398.

Heckman, James J., Pedro Carneiro, and Edward Vytlacil, 2011a. "Estimating Marginal Returns to Education." *American Economic Review* 101 (6):2754–2871.

Heckman, James J., John Eric Humphries, and Tim Kautz, editors, 2014. *The Myth of Achievement Tests: The GED and the Role of Character in American Life*. Chicago: University of Chicago Press.

Heckman, James J., John Eric Humphries, and Gregory Veramendi, 2016. "Dynamic Treatment Effects." *Journal of Econometrics* 191 (2):276–292.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd, 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65 (2):261–294.

Heckman, James J., Lance J. Lochner, and Petra E. Todd, 2006a. "Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond." In *Handbook of the*

Economics of Education, vol. 1, edited by Eric A. Hanushek, and Frank Welch, chap. 7. Amsterdam: Elsevier, pp. 307–458.

———, 2008. "Earnings Functions and Rates of Return." *Journal of Human Capital* 2 (1):1–31.

Heckman, James J., and Salvador Navarro, 2007. "Dynamic Discrete Choice and Dynamic Treatment Effects." *Journal of Econometrics* 136 (2):341–396.

Heckman, James J., and Rodrigo Pinto, 2015. "Causal Analysis after Haavelmo." *Econometric Theory* 31 (1):115–151.

Heckman, James J., Rodrigo Pinto, and Peter A. Savelyev, 2013a. "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103 (6):2052–2086.

Heckman, James J., Susanne M. Schennach, and Benjamin Williams, 2011b. "Matching with Error-Laden Covariates." Unpublished manuscript, University of Chicago, Department of Economics.

———, 2013b. "Matching on Proxy Variables." Unpublished Manuscript, University of Chicago, Department of Economics.

Heckman, James J., Jeffrey A. Smith, and Nancy Clements, 1997. "Making the Most Out Of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies* 64 (4):487–535.

Heckman, James J., Jora Stixrud, and Sergio Urzúa, 2006b. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24 (3):411–482.

Heckman, James J., and Sergio Urzúa, 2010. "Comparing IV With Structural Models: What Simple IV Can and Cannot Identify." *Journal of Econometrics* 156 (1):27–37.

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

Heckman, James J., and Edward J. Vytlacil, 1998. "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return Is Correlated with Schooling." *Journal of Human Resources* 33 (4):974–987.

———, 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proceedings of the National Academy of Sciences* 96 (8):4730–4734.

———, 2005. "Structural Equations, Treatment Effects and Econometric Policy Evaluation." *Econometrica* 73 (3):669–738.

———, 2007a. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In *Handbook of Econometrics*, vol. 6B, edited by James J. Heckman, and Edward E. Leamer, chap. 70. Amsterdam: Elsevier Science B. V., pp. 4779–4874.

———, 2007b. "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Economic Estimators to Evaluate Social Programs and to Forecast Their Effects in New Environments." In *Handbook of Econometrics*, vol. 6B, edited by James J. Heckman, and Edward E. Leamer, chap. 71. Amsterdam: Elsevier Science B. V., pp. 4875–5143.

Imbens, Guido W., and Joshua D. Angrist, 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2):467–475.

Kane, Thomas J., and Cecilia E. Rouse, 1993. "Labor Market Returns to Two- and Four-Year

Colleges: Is a Credit a Credit and Do Degrees Matter?" Working Paper 4268, National Bureau of Economic Research.

Keane, Michael P., and Kenneth I. Wolpin, 1997. "The Career Decisions of Young Men." *Journal of Political Economy* 105 (3):473–522.

Kling, Jeffrey R., 2001. "Interpreting Instrumental Variables Estimates of the Returns to Schooling." *Journal of Business and Economic Statistics* 19 (3):358–364.

Lochner, Lance, 2011. "Non-Production Benefits of Education: Crime, Health, and Good Citizenship." Working Paper 16722, National Bureau of Economic Research.

Manski, Charles F., 2004. "Measuring Expectations." *Econometrica* 72 (5):1329–1376.

McMahon, Walter W., 2000. "Externalities, Non-Market Effects, and Trends in Returns to Educational Investments." In *The Appraisal of Investments in Educational Facilities*. Paris: European Investment Bank/OECD, pp. 51–83.

———, 2009. *Higher Learning, Greater Good*. Baltimore, MD: Johns Hopkins University Press.

Mincer, Jacob, 1974. *Schooling, Experience, and Earnings*. New York: Columbia University Press for National Bureau of Economic Research.

Oreopoulos, Philip, and Uros Petronijevic, 2013. "Making College Worth It: A Review of Research on the Returns to Higher Education." Working Paper 19053, National Bureau of Economic Research.

Oreopoulos, Philip, and Kjell G. Salvanes, 2011. "Priceless: The Nonpecuniary Benefits of Schooling." *Journal of Economic Perspectives* 25 (1):159–184.

Quandt, Richard E., 1958. "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes." *Journal of the American Statistical Association* 53 (284):873–880.

———, 1972. "A New Approach to Estimating Switching Regressions." *Journal of the American Statistical Association* 67 (338):306–310.

Rust, John, 1994. "Structural Estimation of Markov Decision Processes." In *Handbook of Econometrics*, vol. 4, edited by Robert F. Engle, and Daniel L. McFadden, chap. 51. New York, NY: North-Holland, pp. 3081–3143.

Schennach, Susanne M., Halbert White, and Karim Chalak, 2012. "Local Indirect Least Squares and Average Marginal Effects in Nonseparable Structural Systems." *Journal of Econometrics* 166 (2):282–302.

Stange, Kevin M., 2012. "An Empirical Investigation of the Option Value of College Enrollment." *American Economic Journal: Applied Economics* 4 (1):49–84.

Vytlacil, Edward J., 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica* 70 (1):331–341.

Weisbrod, Burton A., 1962. "Education and Investment in Human Capital." *Journal of Political Economy* 70 (5, Part 2: Investment in Human Beings):106–123.

White, Halbert, and Karim Chalak, 2009. "Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning." *Journal of Machine Learning Research* 10:1759–1799.

Willis, Robert J., and Sherwin Rosen, 1979. "Education and Self-Selection." *Journal of Political Economy* 87 (5, Part 2):S7–S36.