Attendance, Completion, and Heterogeneous Returns to College: A Causal Mediation Approach*

Xiang Zhou

Harvard University

June 6, 2022

Abstract

A growing body of social science research investigates whether the economic payoff to a college education is heterogeneous — in particular, whether disadvantaged youth can benefit more from attending and completing college relative to their more advantaged peers. Scholars, however, have employed different analytical strategies and reported mixed findings. To shed light on this literature, I propose a causal mediation approach to conceptualizing, evaluating, and unpacking the causal effects of college on earnings. By decomposing the total effect of attending a four-year college into several direct and indirect components, this approach not only clarifies the mechanisms through which college attendance boosts earnings, but illuminates the ways in which the postsecondary system may be *both an equalizer and a stratifier*. The total effect of college attendance, its direct and indirect components, and their heterogeneity across different subpopulations are all identified under the assumption of sequential ignorability. I introduce a debiased machine learning (DML) method for estimating all quantities of interest, along with a set of bias formulas for sensitivity analysis. I illustrate the proposed framework and methodology using data from the National Longitudinal Survey of Youth, 1997 cohort.

*Direct all correspondence to Xiang Zhou, Department of Sociology, Harvard University, 33 Kirkland Street, Cambridge MA 02138; email: xiang_zhou@fas.harvard.edu. The author thanks Paul Bauer, Derick Baum, Richard Breen, Aleksei Opacic, Ang Yu, and two anonymous reviewers for helpful comments on previous versions of this paper.

Introduction

Education is long perceived as a ticket to the American dream, a pathway to economic success regardless of a person's circumstances of birth. Back in 1848, Horace Mann portrayed education as "a great equalizer of the conditions of men" (Mann 1848). In his 2020 presidential campaign, Joe Biden envisioned a plan for higher education so that it serves as a gateway to economic opportunity for everyone, "regardless of their parents' income or the color of their skin."[1] Biden's emphasis on the role of higher education in social mobility is echoed by public opinion — the vast majority of Americans believe that nowadays a college education is "necessary to get ahead" (Hanson and Zogby 2010).

Echoing the public and political discourse on the role of higher education in equalizing opportunities, a growing body of social science research has investigated whether the economic payoff to a college education is heterogeneous — in particular, whether disadvantaged youth can benefit more from attending and completing college relative to their more advantaged peers. If so, it would be apt for us to characterize higher education as an "equalizer," in which case inducing more youth into college would potentially reduce inequality and improve intergenerational mobility.

This body of research, however, has yielded mixed findings (Hout 2012). On the one hand, several studies suggest that the economic payoff to a college education may be greater for students from disadvantaged backgrounds than for their more advantaged peers (e.g., Card 1993; Attewell et al. 2007; Maurin and McNally 2008; Brand and Xie 2010; Zimmerman 2014; Giani et al. 2020). These studies have variously measured (dis)advantage using race/ethnicity, parental income, or the propensity score, i.e., the probability of attending or completing a four-year college given an array of observed pre-college characteristics. In particular, Brand and Xie (2010) find that young people with the lowest propensity scores — typically students from minority and low-income backgrounds — appear to benefit the most from a bachelor's degree (henceforth BA degree), a pattern they call "negative selection." On the other hand, economic studies that pay close attention to unobserved sorting into college suggest a theory of "positive selection," i.e., individuals self-select into college on the basis of

[1] https://joebiden.com/beyondhs/

their anticipated payoffs to attending college, and those most likely to attend college reap the highest economic returns from it (Willis and Rosen 1979; Carneiro et al. 2011; but see Zhou and Xie 2020 for a reanalysis and reinterpretation of Carneiro et al.'s data). More recently, by modeling the earnings return to college as a flexible function of the propensity score, scholars have reported a more nuanced, U-shaped pattern of college effects, especially among men (Cheng et al. 2021; see also Zhou and Xie 2016). In addition, a related strand of research on intergenerational income mobility suggests that once selection processes are adjusted for, the association between parental income and child income is about as strong among college graduates as among non-graduates, a finding that casts doubt on the equalizing potential of a college degree (Zhou 2019; Fiel 2020; but see Karlson 2019).

While it is beyond the scope of this paper to fully reconcile the seemingly incongruent findings on heterogeneous college effects, I highlight an important distinction that has so far received insufficient attention in this body of research, namely, the distinction between attending college and completing a BA degree. In fact, almost all previous research on the economic payoff to higher education has treated college as a dichotomous variable, that is, whether a young adult with a high-school diploma or equivalent has attended (e.g., Carneiro et al. 2011; Zimmerman 2014), or graduated from (e.g., Brand and Xie 2010; Cheng et al. 2021), a four-year college by a certain age. Such a dichotomous approach has several limitations. First, it fails to distinguish the "direct effect" of college attendance (short of a BA degree) from its "continuation value," i.e., its effect on earnings via the possibility it creates for attaining higher levels of education, particularly a BA degree (Heckman et al. 2018). This distinction is consequential because patterns of effect heterogeneity may differ sharply between the direct effect of college attendance and its continuation value. In particular, whereas the direct effect of college attendance may be equalizing, i.e., being larger among more disadvantaged students (Giani et al. 2020), its continuation value may be disequalizing, i.e., favoring students from more advantaged backgrounds. The latter is plausible because minority and low-income college-goers are much less likely to complete college relative to their white and more affluent peers (Bowen et al. 2009; Ciocca Eller and DiPrete 2018). In this case, a dichotomous approach based on either attendance or completion would obscure the opposing patterns of effect heterogeneity associated with different

stages of the educational pipeline.

Second, studies that focus on the effect of a BA degree on earnings often conflate high-school graduates and college dropouts under the umbrella of "non-graduates," and compare college graduates with non-graduates that are similar on a set of pre-college characteristics. This practice may lead to bias because it adjusts only for selection into college, but not *selection out of college*. A college dropout and a college graduate who share the same pre-college characteristics may differ substantially in their postsecondary characteristics, such as college quality, college GPA, and field of study. To the extent that these postsecondary characteristics affect both the chance of college completion and earnings, they are confounders of their causal relationship, which, if not adjusted for, will lead to biased estimates. Moreover, treating high-school graduates and college dropouts as a whole may engender spurious patterns of effect heterogeneity. For example, if we aim to examine heterogeneous effects of a college degree across students with different income backgrounds, high-income non-graduates may be more likely than low-income non-graduates to have attended college in the first place. Thus, if college experience per se (short of a BA degree) boosts earnings — for example, through its effects on human capital, social capital, and career-related information (see Giani et al. 2020 for a detailed discussion) — the estimated effect of a BA degree among high-income youth might be smaller than that among low-income youth simply because the comparison group for high-income college graduates is, on average, more likely to have enjoyed the benefits of a college experience.

To overcome the limitations of the dichotomous approach, I introduce a causal mediation framework for studying the effects of higher education on earnings and their heterogeneity across individuals with different backgrounds. Specifically, by treating BA completion as a mediator that transmits the effect of college attendance on earnings (see Figure 1), the proposed framework enables us to decompose the average total effect of attending a four-year college into four distinct components: (i) the direct effect of college attendance (short of a BA degree) on earnings, (ii) the probability of BA completion given college attendance, (iii) the net effect of BA completion on earnings, and (iv) a residual component reflecting the covariance between BA completion and its net effect on earnings. Each of these components may follow a distinct pattern of effect heterogeneity. For example, the direct effect

of college attendance (i) and the net effect of BA completion (iii) may both follow a pattern of negative selection (Brand and Xie 2010), but the opposite is likely true for the probability of BA completion given college attendance (ii). Thus, the proposed decomposition not only clarifies the mechanisms through which college attendance boosts earnings, but, more importantly, illuminates the ways in which the postsecondary system may be *both an equalizer and a stratifier*.

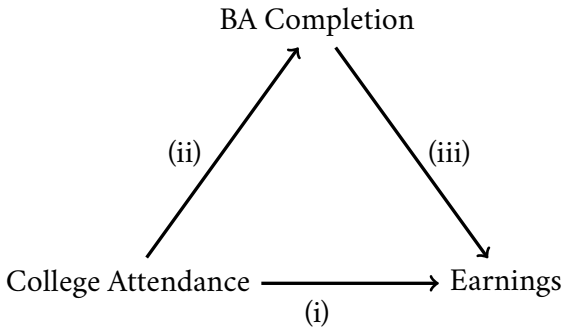

Figure 1: Direct and Indirect Effects of College on Earnings.
Note: Factors that may confound the relationships between college attendance, BA completion, and earnings are omitted.

When we observe a rich set of individual-, family-, and contextual-level characteristics that may affect a person's selection into and out of college, it is reasonable to entertain the assumption of sequential ignorability (Robins 1997), which, in our context, means that (a) given observed pre-college characteristics, no unobserved confounding exists for the effect of college attendance on BA completion and earnings, and (b) among college goers, given observed pre-college and postsecondary characteristics, no unobserved confounding exists for the effect of BA completion on earnings. I show that under sequential ignorability, the total effect of college attendance, its direct and indirect components, and their heterogeneity across different subpopulations are all identified.

Despite the identification result, given the large number of pre-college and postsecondary characteristics we will likely need to adjust for, estimation methods based purely on parametric models may suffer from model uncertainty and large biases due to model misspecification (e.g., Young 2009). To minimize model dependency while preserving statistical efficiency, I introduce a debiased machine learning (DML; Chernozhukov et al. 2018; Semenova and Chernozhukov 2021; Zhou 2020) method for estimating all quantities of interest. Through the use of flexible machine learning methods, care-

fully constructed estimating equations, and sample splitting, the DML estimators are not only robust to model misspecification but also immune to the regularization and overfitting biases that often afflict machine learning estimators of statistical parameters. I illustrate the proposed framework and DML method using data from the National Longitudinal Survey of Youth, 1997 cohort (NLSY97).

Unpacking Heterogeneous College Effects

A Causal Decomposition

We consider completion of a BA degree as an intermediate variable, i.e., a mediator, that transmits the effect of college attendance on earnings. Thus, the total effect of attending a four-year college on earnings can be decomposed into a direct effect of college attendance (short of a BA degree) and an indirect effect that operates through BA completion. The latter component is sometimes referred to as the "continuation value" of college attendance (e.g., Heckman et al. 2018), and it is governed by a person's likelihood of BA completion given college attendance as well as the net effect of BA completion on earnings. Specifically, for individual i, let A_i denote a binary indicator of attending a four-year college, M_i a binary indicator of BA completion, and Y_i labor market earnings. In addition, using the potential-outcomes notation (Rubin 1974), let $M_i(a)$ denote individual i's potential status of BA completion if her college attendance status was set to a, and let $Y_i(a, m)$ denote individual i's potential earnings if her college attendance status was set to a and BA completion status set to m. The total effect (TE) of college attendance on earnings can then be expressed as

$$
\begin{aligned}
\text{TE}_i &= Y_i\big(1, M_i(1)\big) - Y_i(0, M_i(0)) \\
&= Y_i\big(1, M_i(1)\big) - Y_i(0, 0) \quad \text{(because } M_i(0) = 0\text{)} \\
&= \underbrace{Y_i(1,0) - Y_i(0,0)}_{\substack{\text{direct effect of} \\ \text{college attendance}}} + \underbrace{M_i(1)\,\underbrace{\big(Y_i(1,1) - Y_i(1,0)\big)}_{\substack{\text{net effect of} \\ \text{BA completion}}}}_{\substack{\text{indirect effect via} \\ \text{BA completion}}}. \qquad (1)
\end{aligned}
$$

Thus, for individual i, the total effect of college attendance is governed by three components: the direct effect of college attendance ($Y_i(1,0) - Y_i(0,0)$), whether the person would complete a BA degree given college attendance ($M_i(1)$), and the net effect of BA completion ($Y_i(1,1) - Y_i(1,0)$). The product of the latter two components constitutes the indirect effect of college via BA completion.

Since for each individual i, only one of the three potential outcomes $Y_i(0,0)$, $Y_i(1,0)$, and $Y_i(1,1)$ is observed, neither the direct effect of college attendance nor the net effect of BA completion can be computed at the individual level. We thus focus on the population- and group-level means of these effects. First, taking the expectation of equation (1) yields a population-level decomposition:

$$\begin{aligned}\underbrace{\mathbb{E}[\text{TE}_i]}_{:=\Delta_{\text{tot}}} =& \underbrace{\mathbb{E}[Y_i(1,0) - Y_i(0,0)]}_{:=\Delta_{\text{att}}} + \underbrace{\mathbb{E}[M_i(1)]}_{:=\pi_{\text{comp}}} \cdot \underbrace{\mathbb{E}[Y_i(1,1) - Y_i(1,0)]}_{:=\Delta_{\text{comp}}} \\ &+ \underbrace{\text{Cov}[M_i(1), Y_i(1,1) - Y_i(1,0)]}_{:=\Delta_{\text{cov}}} \\ =& \Delta_{\text{att}} + \underbrace{\pi_{\text{comp}}\Delta_{\text{comp}} + \Delta_{\text{cov}}}_{\Delta_{\text{ind}}}. \end{aligned} \tag{2}$$

Here, Δ_{tot} represents the average total effect of college on earnings, Δ_{att} represents the average direct effect of college attendance on earnings, and Δ_{ind} represents the average indirect effect via BA completion. The indirect effect Δ_{ind} equals $\pi_{\text{comp}}\Delta_{\text{comp}} + \Delta_{\text{cov}}$, where π_{comp} represents the probability of BA completion if a person attended college, Δ_{comp} represents the average net effect of BA completion on earnings, and Δ_{cov} is a component reflecting the covariance between BA completion and its net effect on earnings. Intuitively, Δ_{cov} is positive if those who would complete a BA degree given college attendance (i.e., $M_i(1) = 1$) can benefit more from a BA degree (i.e., larger $Y_i(1,1) - Y_i(1,0)$) than those who would not complete a BA degree given college attendance (i.e., $M_i(1) = 0$), and negative if the opposite is true. According to the positive selection thesis (Willis and Rosen 1979; Carneiro et al. 2011), a positive Δ_{cov} may arise if college goers possess knowledge about their individual-specific payoffs to a BA degree and decide whether to pursue a BA degree on the basis of their anticipated payoffs. A positive Δ_{cov} may also arise for structural (rather than individual) reasons, for example, if the financial and cognitive resources of middle- and upper-class students allow them to both com-

plete college at a higher rate and reap higher economic returns from a BA degree relative to their less advantaged peers.

To see how each of the above components varies across individuals with different backgrounds, we can evaluate the conditional expectation of equation (1) given S, some indicator of pre-college advantage. Analogous to the population-level decomposition (2), we have

$$\begin{aligned}\underbrace{\mathbb{E}[\text{TE}_i|S_i=s]}_{:=\Delta_{\text{tot}}(s)} =& \underbrace{\mathbb{E}[Y_i(1,0)-Y_i(0,0)|S_i=s]}_{:=\Delta_{\text{att}}(s)} + \underbrace{\mathbb{E}[M_i(1)|S_i=s]}_{:=\pi_{\text{comp}}(s)} \cdot \underbrace{\mathbb{E}[Y_i(1,1)-Y_i(1,0)|S_i=s]}_{:=\Delta_{\text{comp}}(s)} \\ &+ \underbrace{\text{Cov}[M_i(1), Y_i(1,1)-Y_i(1,0)|S_i=s]}_{:=\Delta_{\text{cov}}(s)} \\ =& \Delta_{\text{att}}(s) + \pi_{\text{comp}}(s)\Delta_{\text{comp}}(s) + \Delta_{\text{cov}}(s), \end{aligned} \tag{3}$$

where $\Delta_{\text{tot}}(s)$, $\Delta_{\text{att}}(s)$, $\pi_{\text{comp}}(s)$, $\Delta_{\text{comp}}(s)$, and $\Delta_{\text{cov}}(s)$ represent the same components in equation (2) among individuals with $S_i = s$.

The group-level decomposition (3) enables us to quantify the equalizing and stratifying roles of higher education. Specifically, the negative selection thesis (Brand and Xie 2010) suggests that the direct effect of college attendance $\Delta_{\text{att}}(s)$ and the net effect of BA completion $\Delta_{\text{comp}}(s)$ may be particularly large among individuals from disadvantaged backgrounds, contributing to the equalizing role of higher education. On the other hand, ample empirical evidence indicates that college graduation rates are much higher among students from more advantaged backgrounds relative to their less privileged peers (e.g., Bowen et al. 2009). Thus the component $\pi_{\text{comp}}(s)$ is likely an increasing function of s, contributing to the stratifying role of higher education. Furthermore, as noted earlier, the positive selection thesis suggests that college students may possess knowledge about their idiosyncratic payoffs to a BA degree and act on it. If such a pattern of self-selection is present and if it is stronger among more advantaged youth than among less advantaged youth (e.g., due to unequal access to information about their idiosyncratic returns to a BA degree or unequal capacities to act on such information), then the within-group covariance component $\Delta_{\text{cov}}(s)$ may also be an increasing function of s, contributing to the stratifying role of higher education. Given these competing forces,

an expansion in college enrollment would have the potential to reduce inequality if the equalizing roles of college (e.g., those associated with $\Delta_{\text{att}}(s)$ and $\Delta_{\text{comp}}(s)$) outweigh its stratifying roles (e.g., those associated with $\pi_{\text{comp}}(s)$ and $\Delta_{\text{cov}}(s)$).

Identification

Since the average total effect and its direct and indirect components all depend on potential outcomes, they cannot be directly estimated from data. We first need to identify these quantities — i.e., write them as functions of observed data only — under appropriate assumptions. In particular, the quantities of interest outlined in the previous section are all identified under the assumption of sequential ignorability (Robins 1997), which, simply speaking, means that given observed covariates, no unobserved confounding exists for the causal relationships among college attendance, BA completion, and earnings. Specifically, if we use X to denote a set of observed pre-college characteristics that may confound the causal effects of college attendance and BA completion on earnings, and Z to denote a set of observed postsecondary characteristics (e.g., college GPA) that may additionally confound the causal effect of BA completion on earnings, the sequential ignorability assumption states that (a) conditional on pre-college characteristics X, college attendance is independent of both potential earnings and potential college completion status (i.e., $\big(M(1), Y(0,0), Y(1,0), Y(1,1)\big) \perp\!\!\!\perp A|X$), and (b) conditional on pre-college characteristics X and postsecondary characteristics Z, BA completion is independent of potential earnings among college goers (i.e., $\big(Y(1,0), Y(1,1)\big) \perp\!\!\!\perp M|X, A=1, Z$). Figure 2 contains a directed acyclic graph (DAG) relating college attendance, BA completion, and earnings to potential pre-college and postsecondary confounders of their relationships. Here, the X and Z vectors are assumed to capture a broad range of potential confounders of the A-M, A-Y, and M-Y relationships, such as socioeconomic background, cognitive and noncognitive skills, motivation, personality traits, and social capital. Under this DAG, sequential ignorability will hold if all of these potential confounders are observed and accurately measured. In practice, however, some of these confounders (e.g., motivation) are likely unobserved or imperfectly measured. Thus, in most (if not all) empirical studies, we should view sequential ignorability as a working assumption and con-

duct a sensitivity analysis (see below) to assess the direction and magnitude of potential bias due to unobserved confounding.

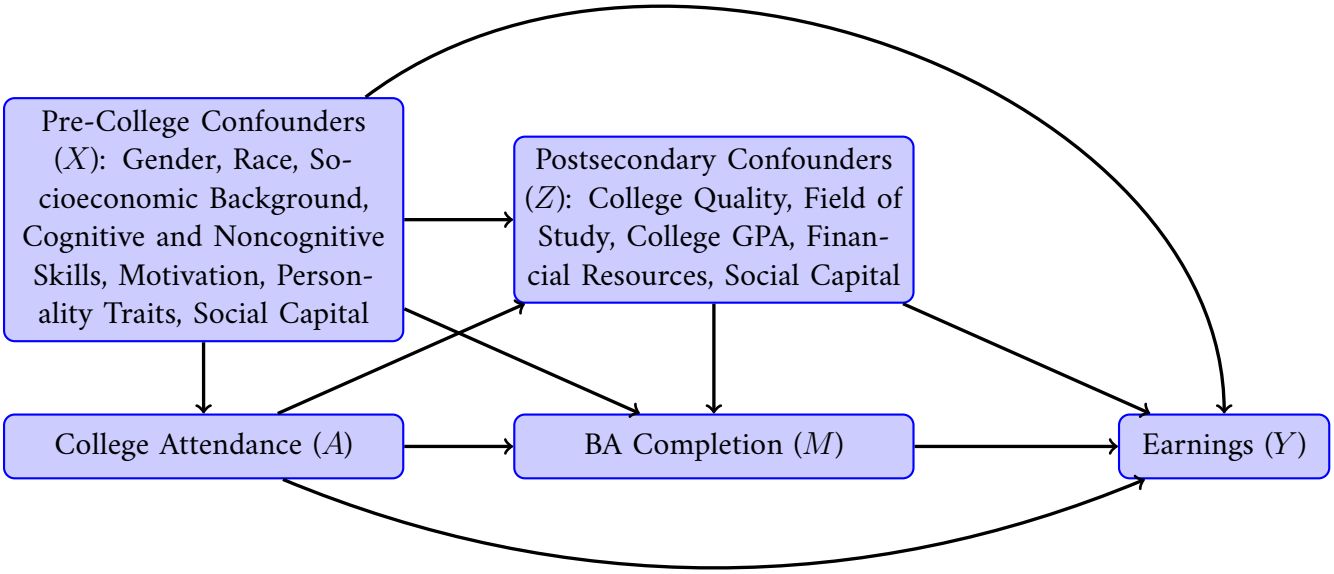

Figure 2: Hypothesized Causal Relationships in a Direct Acyclic Graph.

Equation (2) implies that to identify the total effect of college attendance (Δ_{tot}) and its various components (Δ_{att}, π_{comp}, Δ_{comp}, Δ_{cov}), it suffices to identify the following expected potential outcomes: $\mathbb{E}[M(1)]$, $\mathbb{E}[Y(0,0)]$, $\mathbb{E}[Y(1,M(1))]$, $\mathbb{E}[Y(1,0)]$, and $\mathbb{E}[Y(1,1)]$. Here I omit the subscript i for conciseness. Under sequential ignorability, these quantities are identified via Robins's (1986; 1997) g-formula:

$$\mathbb{E}[M(1)] = \int \mathbb{E}[M|x, A=1]dP(x), \tag{4}$$

$$\mathbb{E}[Y(0,0)] = \int \mathbb{E}[Y|x, A=0]dP(x), \tag{5}$$

$$\mathbb{E}[Y(1,M(1))] = \int \mathbb{E}[Y|x, A=1]dP(x), \tag{6}$$

$$\mathbb{E}[Y(1,0)] = \iint \mathbb{E}[Y|x, A=1, z, M=0]dP(z|x, A=1)dP(x), \tag{7}$$

$$\mathbb{E}[Y(1,1)] = \iint \mathbb{E}[Y|x, A=1, z, M=1]dP(z|x, A=1)dP(x), \tag{8}$$

where $P(u)$ denotes the cumulative distribution function of a random variable U. It is easy to see that π_{comp} is identified by equation (4), Δ_{tot} identified by equation (6) minus equation (5), Δ_{att} identified by equation (7) minus equation (5), Δ_{comp} identified by equation (8) minus equation (7), and Δ_{cov} identified by $\Delta_{\text{tot}} - \Delta_{\text{att}} - \pi_{\text{comp}}\Delta_{\text{comp}}$. Components of the group-level decomposition are identified analogously, except that all quantities involved in equations (4)-(8) should now be conditioned on $S_i = s$.

The sequential ignorability assumption is weaker than the ignorability assumption previously invoked for studying the effect of a college degree on earnings. For example, Brand and Xie (2010) used a dichotomous approach that directly compares college graduates with non-graduates that are similar on a set of pre-college characteristics. This approach implicitly assumes that conditional on pre-college characteristics X, BA completion status is independent of potential earnings under completion and non-completion (i.e., $\big(Y(A,0), Y(1,1)\big) \perp\!\!\!\perp M|X$). This assumption is stronger than sequential ignorability because it rules out (a) a direct effect of college attendance on earnings (the arrow $A \rightarrow Y$ in Figure 2) *and* (b) postsecondary characteristics that may confound the effect of BA completion on earnings (the noncausal path $M \leftarrow Z \rightarrow Y$ in Figure 2). By contrast, sequential ignorability allows for both (a) and (b).

The sequential ignorability assumption is also weaker than the ignorability assumption required for identifying the natural direct and indirect effects (NDE and NIE) in a generic causal mediation analysis (VanderWeele and Vansteelandt 2009; Imai et al. 2010). The latter requires a conditional independence relationship between the so-called cross-world counterfactuals given pretreatment confounders X, namely, $M(a) \perp\!\!\!\perp Y(a^*, m)|X$, which rules out post-treatment confounding of the mediator-outcome relationship. By contrast, all components of our effect decomposition are identified under sequential ignorability, which allows for observed post-treatment confounders Z. To understand this result, note that Δ_{att} and Δ_{comp} correspond to a controlled direct effect (CDE; Pearl 2001; Robins 2003) and a controlled mediator effect (CME; Zheng and Zhou 2015), both of which are identified under sequential ignorability. Interestingly, in our context, because $M_i(0) = 0$, $\Delta_{\text{att}} = \mathbb{E}[Y_i(1,0) - Y_i(0,0)] = \mathbb{E}[Y_i(1, M_i(0)) - Y_i(0, M_i(0))]$ is also the NDE, and

$\Delta_{\text{ind}} = \mathbb{E}[Y_i(1) - Y_i(1,0)] = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(1, M_i(0))]$ is the NIE. Hence, equation (2) is a more fine-grained decomposition of the ATE than the two-component decomposition routinely considered in causal mediation analysis; yet, due to the constraint $M_i(0) = 0$, all components are identified under sequential ignorability.

Nonetheless, sequential ignorability is still a strong and unverifiable assumption, which can be violated whenever unobserved confounders exist for any of the causal relationships involved. For example, in my empirical illustration below, some of the potential confounders depicted in Figure 2 such as motivation are not directly measured. This is a common scenario in observational studies of college effects, or for that matter, in observational studies in general. Thus, in practice, it is prudent to view sequential ignorability as a working assumption and report the sensitivity of estimated causal effects to potential violations of sequential ignorability (e.g., Breen et al. 2015). Later in this section, I outline a bias factor approach for performing sensitivity analysis in our context.

Estimation

Equations (4)-(8) and their group-level counterparts can be estimated via a variety of methods, such as g-computation (Robins 1986, 1997), sequential g-estimation (Vansteelandt 2009; Joffe and Greene 2009), regression-with-residuals (Zhou and Wodtke 2019; Wodtke and Zhou 2020), inverse probability weighting (IPW; VanderWeele 2009), and residual balancing (Zhou and Wodtke 2020) (see Zhou 2020 for an overview of various estimation methods). Yet, all of these methods rely on the correct specification of at least two parametric models about A, M, Z, or Y (implicitly or explicitly). Given the large number of pre-college covariates and postsecondary characteristics we are likely to encounter in practice, estimators based purely on parametric models may suffer large biases due to model misspecification. To minimize model dependency, I now introduce a debiased machine learning (DML; Chernozhukov et al. 2018; Semenova and Chernozhukov 2021; Zhou 2020) method for estimating quantities (4)-(8) and their group-level counterparts.

In our context, the DML approach is characterized by three key elements: a sample-splitting technique called cross-fitting, the construction of a "Neyman-orthogonal signal" for each of the target

parameters in equations (4)-(8), and, when estimating the group-level decomposition (3), a linear model of the Neyman-orthogonal signal on our measure of pre-college advantage S. Specifically, it involves the following steps:

1. Randomly partition the analytical sample $\mathcal{I}$ into J equal-sized subsamples: $\mathcal{I}_1, \mathcal{I}_2 \ldots \mathcal{I}_J$, where J is recommended to be a small number such as 5 (Chernozhukov et al. 2018);

2. For each subsample $\mathcal{I}_j$,

   (a) Use the observations in $\mathcal{I} \backslash \mathcal{I}_j$ (i.e., all observations but those in $\mathcal{I}_j$) to fit a flexible machine learning model for each of the following "nuisance functions":[2] $\Pr[A = 1|x]$, $\Pr[M = 1|x, A = 1]$, $\Pr[M = 1|x, A = 1, z]$, $\mathbb{E}[Y|x, a]$, $\mathbb{E}[Y|x, A = 1, z, m]$, and $\mathbb{E}_{Z|x,A=1}\mathbb{E}[Y|X, A = 1, Z, m]$;

   (b) For each observation in $\mathcal{I}_j$, use estimates of the above models to construct a set of "Neyman-orthogonal signals," one for each potential outcome: $M^*(1)$, $Y^*(0, 0)$, $Y^*\big(1, M(1)\big)$, $Y^*(1, 0)$, and $Y^*(1, 1)$.

3. In the full sample, use the above signals for potential outcomes to construct the corresponding signals for Δ_{tot}, Δ_{att}, π_{comp}, and Δ_{comp}. For example, the signal for Δ_{tot} is given by $Y^*\big(1, M(1)\big) - Y^*(0, 0)$, the signal for Δ_{att} is given by $Y^*(1, 0) - Y^*(0, 0)$, and so on. The sample averages of these signals constitute the DML estimates of the corresponding quantities, and the covariance component is estimated by $\hat{\Delta}_{\text{cov}} = \hat{\Delta}_{\text{tot}} - \hat{\Delta}_{\text{att}} - \hat{\pi}_{\text{comp}}\hat{\Delta}_{\text{comp}}$.

4. To assess effect heterogeneity by pre-college advantage (e.g., $\Delta_{\text{comp}}(s)$), fit a linear model of the corresponding signal (constructed in step 3) on S. The heterogeneity of the covariance component by pre-college advantage is estimated by $\hat{\Delta}_{\text{cov}}(s) = \hat{\Delta}_{\text{tot}}(s) - \hat{\Delta}_{\text{att}}(s) - \hat{\pi}_{\text{comp}}(s)\hat{\Delta}_{\text{comp}}(s)$.

In step 2(b), the Neyman-orthogonal signals are plug-in estimates of the recentered efficient influence functions for the expectations of the corresponding potential outcomes (Semenova and Chernozhukov 2021). Their analytical expressions are given in Supplementary Material A. These signals

[2] A nuisance function is a function that is not of our primary interest but necessary for constructing estimators of our target quantities (i.e., the components in equations (2) and (3)).

satisfy several interesting properties, which, when combined with cross-fitting, yield estimators that are not only robust to model misspecification but also consistent and asymptotically normal under mild conditions.

Below, I use the estimand $\mathbb{E}[Y(0,0)]$, i.e., average potential earnings under non-college-attendance, to illustrate the logic of the DML method. As noted above, under sequential ignorability, $\mathbb{E}[Y(0,0)]$ equals $\int \mathbb{E}[Y|x, A = 0]dP(x)$ (equation 5). To simplify exposition, let us denote this quantity by θ. Its Neyman-orthogonal signal can be written as

$$\varphi(O; \hat{\mu}, \hat{\pi}) = \hat{\mu}(X, 0) + \frac{1 - A}{1 - \hat{\pi}(X)}\big(Y - \hat{\mu}(X, 0)\big), \tag{9}$$

where $O = (X, A, Z, M, Y)$ denotes observed data, $\mu(X, A) := \mathbb{E}[Y|X, A]$ is the conditional mean of earnings given X and A, $\pi(X) := \Pr[A = 1|X]$ is the propensity score of attending college given X, and $\hat{\mu}(X, A)$ and $\hat{\pi}(X)$ denote the empirical estimates of $\mu(X, A)$ and $\pi(X)$, respectively. In equation (9), we use the notation "$\varphi(O; \hat{\mu}, \hat{\pi})$" to highlight that the signal depends on both the observed data O and the estimated models $\hat{\mu}$ and $\hat{\pi}$. This quantity is useful because $\mathbb{E}[\varphi(O; \mu, \pi)] = \theta$, suggesting that we can estimate θ by first estimating the outcome and propensity score models μ and π and then taking a sample mean of $\varphi(O; \hat{\mu}, \hat{\pi})$:

$$\hat{\theta}_{\text{DML}} = \mathbb{P}_n[\varphi(O; \hat{\mu}, \hat{\pi})], \tag{10}$$

where $\mathbb{P}_n[\cdot] := n^{-1}\sum_i[\cdot]$ denotes the operation of computing a sample mean. In this sense, $\varphi(O_i; \hat{\mu}, \hat{\pi})$ can be interpreted as the "contribution" of observation i to the estimator $\hat{\theta}_{\text{DML}}$. Second, the conditional mean of $\varphi(O; \mu, \pi)$ given $S = s$ is $\int \mathbb{E}[Y|x, A = 0]dP(x|s)$, which, under sequential ignorability, equals $\mathbb{E}[Y(0,0)|S = s]$. Thus, we can estimate the latter by averaging $\varphi(O; \hat{\mu}, \hat{\pi})$ among members of group s. When S is continuous, however, we cannot estimate such conditional means nonparametrically. Thus, in step 4 of the above procedure, we fit a linear model of the signal $\varphi(O; \hat{\mu}, \hat{\pi})$ on S, which can be seen as a first-order approximation of $\mathbb{E}[Y(0,0)|S = s]$ when S is continuous but is equivalent to taking a group-specific average of $\varphi(O; \hat{\mu}, \hat{\pi})$ when S is discrete.

To understand the property of the estimator $\hat{\theta}_{\text{DML}}$, it is best to consider the following decomposition of $\sqrt{n}(\hat{\theta}_{\text{DML}} - \theta)$ (Kennedy 2016),

$$\begin{aligned}\sqrt{n}(\hat{\theta}_{\text{DML}} - \theta) =& \sqrt{n}\big(\mathbb{P}_n[\varphi(O; \hat{\mu}, \hat{\pi})] - \mathbb{P}[\varphi(O; \mu, \pi)]\big) \\ =& \underbrace{\sqrt{n}(\mathbb{P}_n - \mathbb{P})[\varphi(O; \mu, \pi)]}_{A} + \underbrace{\sqrt{n}\mathbb{P}[\varphi(O; \hat{\mu}, \hat{\pi}) - \varphi(O; \mu, \pi)]}_{B} \\ &+ \underbrace{\sqrt{n}(\mathbb{P}_n - \mathbb{P})[\varphi(O; \hat{\mu}, \hat{\pi}) - \varphi(O; \mu, \pi)]}_{C}, \end{aligned} \tag{11}$$

where $\mathbb{P}g = \int g d\mathbb{P}$ denotes the expectation of a function g of observed data under the true distribution $\mathbb{P}$, where $g(\cdot)$ is treated as fixed. In equation (11), term A has a mean of zero and variance of $\text{Var}[\varphi(O; \mu, \pi)]$. By the central limit theorem, it converges to $N\big(0, \text{Var}[\varphi(O; \mu, \pi)]\big)$. Thus, by Slutsky's theorem, $\sqrt{n}(\hat{\theta}_{\text{DML}} - \theta)$ will also converge to $N\big(0, \text{Var}[\varphi(O; \mu, \pi)]\big)$ if terms B and C are asymptotically negligible, i.e., if they converge to zero in probability. The latter condition can also be written as $B = o_p(1)$ and $C = o_p(1)$.

First, it can be shown that term B is in the order of $\sqrt{n}O_p(||\hat{\mu} - \mu|| \cdot ||\hat{\pi} - \pi||)$, where $|| \cdot ||$ denotes the $L_2(\mathbb{P})$-norm of a function with respect to probability measure $\mathbb{P}$. The multiplicative structure of $||\hat{\mu} - \mu|| \cdot ||\hat{\pi} - \pi||$ facilitates the use of machine learning methods to estimate the μ and π functions. To see this connection, note that due to the data-driven nature of machine learning algorithms, they generally do not provide $\sqrt{n}$-consistent estimates of the underlying functions, such as μ and π. However, $\sqrt{n}$-consistency is not required of either $\hat{\mu}$ or $\hat{\pi}$ for $\sqrt{n}||\hat{\mu} - \mu|| \cdot ||\hat{\pi} - \pi||$ to converge to zero. In fact, provided that both $\hat{\mu}$ and $\hat{\pi}$ are consistent at a faster-than-$n^{1/4}$ rate, $\sqrt{n}||\hat{\mu} - \mu|| \cdot ||\hat{\pi} - \pi||$ will be $\sqrt{n}o_p(n^{-1/4})o_p(n^{-1/4}) = o_p(1)$, rendering term B asymptotically negligible. This condition, unlike the $\sqrt{n}$-consistency required for the nuisance functions in conventional estimators such as IPW, is achievable for many machine learning methods such as Lasso (Chernozhukov et al. 2018). Second, it can be shown that when cross-fitting is used, term C is in the order of $O_p(||\varphi(O; \hat{\mu}, \hat{\pi}) - \varphi(O; \mu, \pi)||)$, which will also be asymptotically negligible if both $\hat{\mu}$ and $\hat{\pi}$ are consistent.

In sum, when cross-fitting is used in combination with estimating equation (10), the resulting estimator will be consistent and asymptotically normal provided that $\sqrt{n}||\hat{\mu}-\mu|| \cdot ||\hat{\pi}-\pi|| = o_p(1)$, a condition achievable even when the μ and π functions are estimated with flexible machine learning methods. This property of the DML approach makes it highly attractive in our context, in which the rich sets of background characteristics (X) and postsecondary characteristics (Z) (see the next section) make it unrealistic for us to correctly specify parametric models for college attendance and college completion, which would be required to justify conventional methods such as IPW. Since the asymptotic variance of $\hat{\theta}_{\text{DML}}$ is $\text{Var}[\varphi(O;\mu,\pi)]$, we can construct a plug-in estimate of the standard error as $\sqrt{\widehat{\text{Var}}[\varphi(O;\hat{\mu},\hat{\pi})]/n}$.

Although the above reasoning is for the estimand $\mathbb{E}[Y(0,0)]$, the same logic applies to our DML estimators of the other expected potential outcomes (i.e., equations 4, 6-8) and the causal effects Δ_{tot}, Δ_{att}, π_{comp}, and Δ_{comp} (Zhou 2020). Their standard errors can all be estimated through the empirical variances of the corresponding Neyman-orthogonal signals. Estimates of the group-level causal effects $\Delta_{\text{tot}}(s)$, $\Delta_{\text{att}}(s)$, $\pi_{\text{comp}}(s)$, and $\Delta_{\text{comp}}(s)$ are given by the predicted values of the linear models in Step 4, and their standard errors can be estimated through the robust ("sandwich") estimator of the corresponding regression coefficients (Semenova and Chernozhukov 2021). The covariance components Δ_{cov} and $\Delta_{\text{cov}}(s)$, as noted above, are estimated using the plug-in estimators $\hat{\Delta}_{\text{cov}} = \hat{\Delta}_{\text{tot}} - \hat{\Delta}_{\text{att}} - \hat{\pi}_{\text{comp}}\hat{\Delta}_{\text{comp}}$ and $\hat{\Delta}_{\text{cov}}(s) = \hat{\Delta}_{\text{tot}}(s) - \hat{\Delta}_{\text{att}}(s) - \hat{\pi}_{\text{comp}}(s)\hat{\Delta}_{\text{comp}}(s)$. Their standard errors can be estimated through the empirical variances of their influence functions, which are detailed in Supplementary Material B.

A Bias Factor Approach to Sensitivity Analysis

For a generic causal mediation analysis, VanderWeele and Arah (2011) and VanderWeele (2010) introduced a bias factor approach for assessing the sensitivity of estimated total, direct, and indirect effects to unobserved confounding. In our context, this approach can be adapted to derive a set of bias formulas for the total effect of college (Δ_{tot}), the direct effect of attendance (Δ_{att}), and the net effect of completion (Δ_{comp}).

First, let us consider the total effect of college attendance (Δ_{tot}), which may be confounded by unobserved individual characteristics that affect both college attendance (A) and earnings (Y). For analytical tractability, we consider a binary unobserved confounder U, say a personality trait that predisposes a person to prefer cognitive tasks over noncognitive tasks, that affects both college attendance and earnings. Under some simplifying assumptions regarding the homogeneity of the U-A and U-Y relationships, the bias for the estimated Δ_{tot} is given by (see Supplementary Material C)

$$\text{bias}[\Delta_{\text{tot}}] = \alpha_{\text{tot}}\beta_{\text{tot}}, \tag{12}$$

where α_{tot} denotes the difference in the prevalence of U between high school graduates ($A = 0$) and college goers ($A = 1$) given pre-college covariates X, and β_{tot} denotes the average difference in earnings between those with and without U given college attendance status A and pre-college covariates X.

Second, unobserved confounders may exist for the causal effect of BA completion (M) and earnings (Y). In this case, while the total effect of college attendance may still be unbiased, the direct effect of college attendance (Δ_{att}) and the net effect of BA completion (Δ_{comp}) can be over- or underestimated. To explore the direction and magnitude of potential bias, let us again consider a binary unobserved confounder U, say availability of a supportive social network, that affects both BA completion and earnings but may itself be affected by college attendance (A). Under some simplifying assumptions regarding the homogeneity of the U-M and U-Y relationships, the biases for the estimated Δ_{att} and Δ_{comp} are given by (see Supplementary Material C)

$$\text{bias}[\Delta_{\text{att}}] = -\pi_{\text{comp}}\alpha_{\text{comp}}\beta_{\text{net}}, \tag{13}$$

$$\text{bias}[\Delta_{\text{comp}}] = \alpha_{\text{comp}}\beta_{\text{net}}, \tag{14}$$

where π_{comp} is the probability of BA completion given college attendance (see equation 2), α_{comp} denotes the difference in the prevalence of U between college dropouts ($A = 1, M = 0$) and college graduates ($A = M = 1$) given both pre-college and postsecondary characteristics (X and Z), and β_{net}

denotes the net difference in earnings between those with and without the unobserved characteristic U given X, A, M, and Z.

The above formulas can also be used to assess the sensitivity of group-level causal effects $\Delta_{\text{tot}}(s)$, $\Delta_{\text{att}}(s)$, and $\Delta_{\text{comp}}(s)$. In this case, the sensitivity parameters α_{tot}, β_{tot}, π_{comp}, α_{comp}, β_{net} are group-specific, i.e., depending on $S = s$. It is clear that if these sensitivity parameters are identical between individuals with different values of S, estimated patterns of effect heterogeneity will be unaffected. In other words, our estimates of effect heterogeneity will be biased only if there are group differences in these sensitivity parameters. For example, if we found that low-propensity college goers benefit more from completing college than high-propensity college goers, potential bias in this finding would be $\alpha_{\text{comp}}^{\text{low propensity}}\beta_{\text{net}}^{\text{low propensity}} - \alpha_{\text{comp}}^{\text{high propensity}}\beta_{\text{net}}^{\text{high propensity}}$. In the next section, we illustrate this approach by applying it to our estimates from the NLSY97 data.

Empirical Illustration

Data, Measures, and Implementation

Below I illustrate the proposed methods using data from the National Longitudinal Survey of Youth, 1997 cohort (NLSY97).[3] The NLSY97 began with a nationally representative sample of 8,984 men and women at ages 12-17 in 1997. These individuals were interviewed annually through 2011 and biennially thereafter. I limit my analytical sample to respondents who had completed at least a high-school diploma or GED by age 22 and had valid earnings information at ages 30-33, the oldest ages for which data for the youngest respondents in NLSY97 are available ($n = 6,576$).

I construct five sets of variables, each corresponding to a node in Figure 2: college attendance (A),

[3]Previous studies on the economic returns to college have often used data from the NLSY79 (e.g., Brand and Xie 2010; Carneiro et al. 2011). I use the NLSY97 to illustrate the proposed methodology for two reasons. First, compared with the NLSY79, the NLSY97 traces the educational and labor market outcomes of a much younger cohort, making the results from my analyses more relevant to the experience of current and future cohorts of American youth. Second, compared with the NLSY79, the NLSY97 provides a richer set of postsecondary characteristics (Z) that we can adjust for (e.g., college GPA) when estimating the causal effect of a BA degree, making the sequential ignorability assumption more plausible.

BA completion (M), earnings (Y), pre-college characteristics (X), and postsecondary characteristics (Z). Specifically, college attendance (A) denotes whether the respondent had attended a four-year college by age 22, and BA completion (M) denotes whether the respondent had received a BA degree by age 29. A respondent is coded as a *college goer* (i.e., $A = 1$) if she had either attended a four-year college by age 22 or received a BA degree by age 29, and as a *high school graduate* otherwise (i.e., $A = 0$). Among college goers, a respondent is coded as a *college graduate* (i.e., $M = 1$) if she had received a BA degree by age 29, and as a *college dropout/stopout* (i.e., $M = 0$) otherwise. Earnings denote the natural logarithm of the respondent's average annual earnings at ages 30-33 (inflation-adjusted to 2019 dollars). To accommodate respondents with zero earnings (due to unemployment, labor force nonparticipation, and incarceration), I add a small constant (1,000 dollars) to the respondent's average annual earnings before taking the log transformation. To assess the robustness of my findings to this measurement choice, I have conducted parallel analyses using the percentile rank of earnings as the outcome. The results are similar to those reported below (see Supplementary Material D).

Guided by the DAG in Figure 2, I include in the vector of pre-college characteristics (X) several groups of variables: (a) basic demographic variables (gender, race, ethnicity, age in 1997); (b) socioeconomic background (parental education, parental income, parental assets, co-residence with both biological parents, presence of a paternal figure, rural residence, southern residence); (c) cognitive and noncognitive skills (percentile score on the Armed Services Vocational Aptitude Battery test, high school GPA, an index of substance use [ranging from 0 to 3], an index of delinquency [ranging from 0 to 10], whether the respondent had any children by age 18); and (d) peer and school-level characteristics (college expectation among peers, and three dummy variables denoting whether the respondent ever had property stolen at school, was ever threatened at school, and was ever in a fight at school). In particular, parental education is measured using mother's years of schooling; when mother's years of schooling is unavailable, it is measured using father's years of schooling. Parental income is measured as the average annual parental income from 1997 to 2001. Both parental income and parental assets are inflation-adjusted to 2019 dollars.

Similarly, following the DAG in Figure 2, I include in the vector of postsecondary characteristics

(Z) several variables pertaining to college quality as well as the respondent's field of study, college GPA, and the amounts of loans that the student has taken to finance college. In each survey wave of the NLSY97, respondents were asked to report, if any, the names of the colleges in which they were currently or most recently enrolled. Since some respondents attended more than one college, I focus on the college in which the respondent had been enrolled for the longest time by age 29. The college characteristics include: (a) college type, which is a trichotomous variable denoting whether the college is a public institution, a private not-for-profit institution, or a for-profit institution; (b) college selectivity, operationalized as three dummy variables denoting whether the college is one of the "most competitive," "highly competitive," and "very competitive" colleges in Barron's Profile of American Colleges 2000; (c) graduation rate, operationalized as the percentage of students graduating within six years of enrollment measured in 2002; and (d) "upward mobility rate," measured as the percentage of students who reach the top quintile of the income distribution among those with parents in the bottom quintile of the income distribution. Data on graduation rates and upward mobility rates come from the Department of Education's Integrated Postsecondary Education Data System (IPEDS) and the Opportunity Insights project (Chetty et al. 2020), respectively. In each survey wave, respondents who were currently or recently enrolled in college were also asked to report their major field of study. I use a dummy variable to denote whether the field of study in which the respondent had majored for the longest time by age 29 is a STEM field. College GPA is measured as the respondent's cumulative GPA from the Post-Secondary Transcript Study (PSTRAN). Finally, I include two variables representing the total amounts of loans that the respondent had taken from family and friends and from other sources (including the federal government) to pay for college by age 29. Previous studies suggest that educational debt affects both the likelihood of college completion (e.g., Dwyer et al. 2012) and labor market outcomes (e.g., Minicozzi 2005). In my analytical sample, some components of the pre-college characteristics (X) and postsecondary characteristics (Z) contain a small fraction of missing values. They are handled by multivariate imputation via chained equations, with ten imputed data sets. The standard errors of our parameter estimates are adjusted using Rubin's (1987) method.

After constructing the analytical sample, I apply the DML algorithm to implement the decompo-

sitions (2) and (3). To examine effect heterogeneity by pre-college advantage, I compare individuals with different estimated propensity scores of attending college.[4] Previous research has advocated the use of the propensity score as a summary index of pre-college advantage in socioeconomic and academic resources (Brand and Xie 2010; Xie et al. 2012). Thus, heterogeneous returns to college between individuals with lower and higher propensity scores signify the equalizing versus stratifying roles of college. Given that recent research has reported U-shaped patterns of effect heterogeneity by the propensity score (Zhou and Xie 2016; Cheng et al. 2021), I discretize the estimated propensity score into its quintiles and report quintile-specific estimates of all quantities of interest. Following Chernozhukov et al. (2018), I use five-fold cross-fitting, meaning that $J = 5$. All nuisance functions, including the propensity score of college attendance, are estimated using a super learner (Van der Laan et al. 2007) composed of Lasso and random forest.[5] The NLSY sampling weights are used in the estimation of all nuisance functions and target parameters.

Results

Table 1 reports estimates of the average total effect (ATE) and its direct and indirect components (i.e. equation 2). The first column shows that the estimated ATE of attending a four-year college on log earnings is 0.39, implying a 47.7% earnings premium ($e^{0.39} - 1 = 0.477$). The next two columns indicate that the bulk of the ATE is indirect, i.e., through the possibility of completing a BA degree. Without completing a BA degree, the average direct effect of college attendance (Δ_{att}) is estimated at 0.14, or a 15% earnings premium ($e^{0.14} - 1 = 0.15$) relative to high school graduates. The last three columns show estimates of the three components that compose the indirect effect via BA completion: the probability of BA completion given attendance (π_{comp}), the net effect of BA completion (Δ_{comp}),

[4] In my analyses, the estimated propensity scores are treated as given. Thus, standard errors reported for the propensity-score-specific estimates of total, direct, and indirect effects should be viewed as approximate standard errors because they do not account for estimation uncertainty for the propensity score.

[5] A super learner is a weighted average of different machine learning methods designed to minimize prediction error. The algorithm is implemented in the R package SuperLearner (Polley and van der Laan 2017).

Table 1: Decomposition of the Average Total Effect (ATE) of College Attendance on Log Earnings.

Total Effect (Δ_{tot})	Direct Effect (Δ_{att})	Indirect Effect (Δ_{ind})	Completion Prob. (π_{comp})	Completion Effect (Δ_{comp})	Covariance Term (Δ_{cov})
0.39 (0.05)	0.14 (0.06)	0.25 (0.04)	0.57 (0.01)	0.47 (0.07)	-0.02 (0.02)

Note: Numbers in parentheses are estimates of standard errors, which are constructed using the empirical variances of the corresponding influence functions and adjusted for multiple imputation via Rubin's (1987) method.

and the covariance between BA completion and its net effect on earnings (Δ_{cov}). Among them, the covariance component is very small; thus the indirect effect ($\Delta_{\text{ind}} = \pi_{\text{comp}}\Delta_{\text{comp}} + \Delta_{\text{cov}}$) is largely determined by the product of π_{comp} and Δ_{comp} ($0.57 * 0.47 = 0.27$). In particular, the estimated net effect of BA completion implies an earning premium of 60% ($e^{0.47} - 1 = 0.60$) for BA holders compared with college dropouts/stopouts. The sum of the estimated Δ_{att} and Δ_{comp} is 0.61, which can be interpreted as the *joint effect* of attending and completing a four-year college on earnings. In other words, the earnings premium associated with attending and completing a four-year college as opposed to not attending college is about 84% ($e^{0.61} - 1 = 0.84$).

Figure 3 shows estimates of the total effect and its various components in each of the propensity score quintiles. We find suggestions of nonlinearity in several components, such as the total and direct effects of attendance, although estimation uncertainty prevents us from reaching a definitive conclusion. However, several patterns are discernible for the lowest-propensity individuals, i.e., those in the first quintile. On the one hand, their estimated direct effect of attendance is particularly large (0.45), much larger than those for the other quintiles, whose direct effect estimates are all relatively small and statistically indistinguishable from zero. On the other hand, their estimated indirect effect via BA completion is exceptionally small; in fact, it is negative. This finding is counterintuitive if we construe the indirect effect as reflecting the path $A \to M \to Y$ in Figure 2. Since both the effect of A on M (i.e., the probability of BA completion given attendance) and the effect of M on Y (the net effect of BA completion) are positive, how can the indirect effect of A on Y via M be negative? This is due to the (estimated) covariance component for the lowest-propensity group ($\hat{\Delta}_{\text{cov}}(s)$), which is not only negative but larger in absolute value than $\hat{\pi}_{\text{comp}}(s)\hat{\Delta}_{\text{comp}}(s)$, rendering the indirect effect

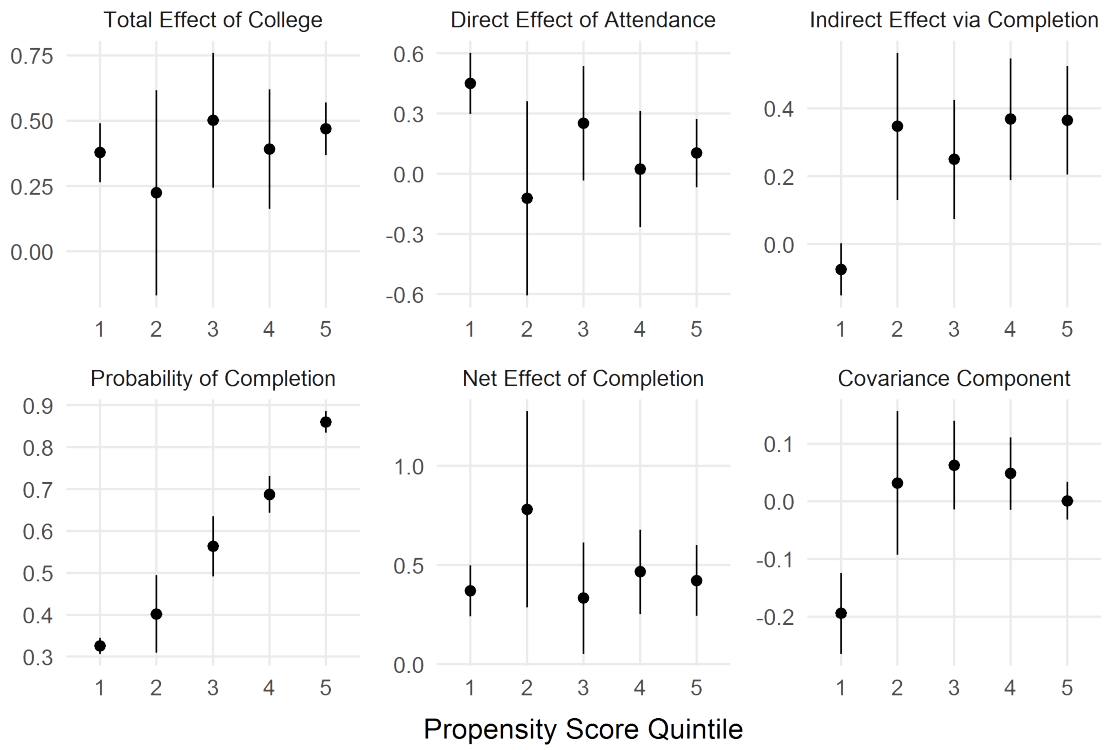

Figure 3: Estimates of the Total Effect and Its Components by Propensity Score Quintile. Note: Line ranges represent 95% confidence intervals.

estimate ($\hat{\Delta}_{\text{ind}}(s) = \hat{\pi}_{\text{comp}}(s)\hat{\Delta}_{\text{comp}}(s) + \hat{\Delta}_{\text{cov}}(s)$) negative. Substantively, the negative covariance means that among the lowest-propensity individuals, those who would benefit more from completing college are less likely to complete college given attendance, a pattern we might call "negative selection among the least advantaged." As a result of their particularly large direct effect and exceptionally small indirect effect, the total effect of college among the lowest-propensity individuals appears comparable to that for their more advantaged peers (e.g., those in the fourth and fifth quintiles). Clearly, without the effect decomposition, the sharp and countervailing patterns of effect heterogeneity between the lowest-propensity individuals and their more advantaged peers would be obscured.

Sensitivity Analyses

Due to data limitations, some of the theoretical constructs depicted in Figure 2, such as motivation, personality traits, and social capital, are not directly captured in the X and Z vectors. Below, I illus-

trate how the bias factor approach to sensitivity analysis described earlier can be employed to assess the direction and magnitude of potential biases due to such unobserved confounders. In particular, let us consider the total effect of college (Δ_{tot}), the direct effect of attendance (Δ_{att}), and the net effect of completion (Δ_{comp}) for individuals in the lowest and highest propensity score quintiles. First, to the extent that an unobserved confounder U (e.g., a personality trait that predisposes a person to prefer cognitive tasks over noncognitive tasks) affects both college attendance and earnings, the bias for our total effect estimate is given by $\alpha_{\text{tot}}\beta_{\text{tot}}$ (equation 12). Given the symmetry of the bias formula, let us consider only cases where U is positively associated with log earnings, i.e., $\beta_{\text{tot}} > 0$, while leaving the sign of α_{tot} unconstrained. Columns 3-4 of Table 2 report the bias-adjusted estimates of Δ_{tot} for the lowest- and highest-propensity individuals across a range of potential values of α_{tot} and β_{tot}. Given that an unobserved characteristic that boosts earnings is likely also positively associated with college attendance, we may focus on the lower part of Table 2, where α_{tot} and β_{tot} are both positive. In this case, although our estimates of Δ_{tot} will be upwardly biased, they are quite robust to unobserved confounding for both groups. For example, even if the unobserved characteristic increases log earnings by 0.3 (given X and A) and its prevalence differs by as much as 30 percentage points between high school graduates and college goers (given X), the bias-adjusted estimates of the total effect are still sizable — 0.29 and 0.38 for the least and the most advantaged groups, respectively.

Second, if an unobserved confounder exists for the effect of BA completion on earnings (e.g., social capital accumulated during college), the biases for our estimates of the direct effect of attendance and the net effect of completion are given by $-\pi_{\text{comp}}\alpha_{\text{comp}}\beta_{\text{net}}$ and $\alpha_{\text{comp}}\beta_{\text{net}}$ (equations 13 and 14), respectively. Columns 5-8 of Table 2 report the bias-adjusted estimates of Δ_{att} and Δ_{comp} across a range of potential values of α_{comp} and β_{net}. When assessing bias$[\Delta_{\text{att}}]$ for lowest- and highest-propensity individuals, I replace π_{comp} with its DML estimate for the corresponding group. Given the symmetry of these formulas, let us consider only cases where $\beta_{\text{net}} > 0$. Since an unobserved characteristic that boosts earnings is likely positively associated with BA completion, it is reasonable to assume that α_{comp} is also positive. Thus, we may focus on the lower panels of Columns 5-8, which suggest that the direct effect of attendance (Δ_{att}) is likely underestimated and the net effect of BA completion (Δ_{comp})

Table 2: Sensitivity Results for the Total Effect of College (Δ_{tot}), the Direct Effect of Attendance (Δ_{att}), and the Net Effect of BA Completion (Δ_{comp}) for Individuals in the Lowest and Highest Propensity Score (PS) Quintiles.

Sensitivity Parameters		Total Effect (Δ_{tot})		Direct Effect of Attendance (Δ_{att})		Net Effect of Completion (Δ_{comp})	
α	β	1st PS Quintile	5th PS Quintile	1st PS Quintile	5th PS Quintile	1st PS Quintile	5th PS Quintile
0	0	0.38	0.47	0.45	0.10	0.37	0.42
-0.3	0.1	0.41	0.50	0.44	0.08	0.40	0.45
-0.3	0.2	0.44	0.53	0.43	0.05	0.43	0.48
-0.3	0.3	0.47	0.56	0.42	0.03	0.46	0.51
-0.2	0.1	0.40	0.49	0.44	0.09	0.39	0.44
-0.2	0.2	0.42	0.51	0.44	0.07	0.41	0.46
-0.2	0.3	0.44	0.53	0.43	0.05	0.43	0.48
-0.1	0.1	0.39	0.48	0.45	0.10	0.38	0.43
-0.1	0.2	0.40	0.49	0.44	0.09	0.39	0.44
-0.1	0.3	0.41	0.50	0.44	0.08	0.40	0.45
0.1	0.1	0.37	0.46	0.45	0.11	0.36	0.41
0.1	0.2	0.36	0.45	0.46	0.12	0.35	0.40
0.1	0.3	0.35	0.44	0.46	0.13	0.34	0.39
0.2	0.1	0.36	0.45	0.46	0.12	0.35	0.40
0.2	0.2	0.34	0.43	0.46	0.14	0.33	0.38
0.2	0.3	0.32	0.41	0.47	0.16	0.31	0.36
0.3	0.1	0.35	0.44	0.46	0.13	0.34	0.39
0.3	0.2	0.32	0.41	0.47	0.16	0.31	0.36
0.3	0.3	0.29	0.38	0.48	0.18	0.28	0.33

Note: The sensitivity parameters α and β refer to α_{tot} and β_{tot} for the total effect of college and to α_{comp} and β_{net} for the direct effect of attendance and the net effect of BA completion.

is likely overestimated. In general, our estimates of the BA completion effect are fairly robust for both the lowest- and highest-propensity groups. The estimated direct effect of attendance, on the other hand, is much more robust for the least advantaged youth than for the most advantaged youth.

As noted earlier, if the sensitivity parameters are constant across the population, our findings of effect heterogeneity will be unchanged. However, the sensitivity parameters may differ between less and more advantaged individuals. Several processes may be at work. On the one hand, it is possible that low-propensity students who attend and complete college disproportionately possess some

unobserved trait, such as motivation, that boosts both educational attainment and earnings, and that high-propensity youth who do not attend or complete college disproportionately face some unobserved barrier to educational attainment that also affects earnings. If so, biases due to the "imbalance" of unobserved confounders between treated and untreated individuals (i.e., the α parameters in equations 12-14) will be larger for both the lowest- and highest-propensity youth than for their medium-propensity peers. On the other hand, unobserved traits such as motivation might have a greater effect on earnings among low-propensity youth than among high-propensity youth, whose advantaged socioeconomic backgrounds might dilute the influence of other factors. If so, biases due to the "impact" of unobserved confounders between treated and untreated individuals (i.e., the β parameters in equations 12-14) will be larger for the low-propensity individuals than for high-propensity individuals. Thus, compared with the first process, the second process is more likely to induce differential selection bias between the lowest- and highest-propensity groups. To be concrete, let us consider the direct effect of attendance, for which our estimated effect heterogeneity will be subject to a differential selection bias of $\pi_{\text{comp}}^{\text{5th quintile}}\alpha_{\text{comp}}^{\text{5th quintile}}\beta_{\text{net}}^{\text{5th quintile}} - \pi_{\text{comp}}^{\text{1st quintile}}\alpha_{\text{comp}}^{\text{1st quintile}}\beta_{\text{net}}^{\text{1st quintile}}$. In this particular case, however, the differential selection bias would have to reach 0.35 to explain away the difference between the lowest- and highest-propensity individuals in their estimated Δ_{att} (0.45 versus 0.10). Considering the range of plausible values for our sensitivity parameters and the associated biases, it is highly unlikely that unobserved confounding plays a significant role in driving the observed effect heterogeneity in Δ_{att}. By contrast, our estimated differences in Δ_{tot} and Δ_{comp} between the first and fifth propensity quintiles are much smaller and consequently more sensitive to unobserved confounding.

Concluding Remarks

Higher education can be a double-edged sword in shaping inequality. It may serve as an equalizer if disadvantaged youth can benefit more from the experience of attending college and from obtaining a college degree than do their more advantaged peers. On the other hand, it reflects and reinforces

preexisting inequalities. In the United States, minority and low-income students are much less likely than their white and more affluent peers to attend a four-year college, and, even when they do, they are less likely to graduate with a BA degree by their late twenties. In this paper, I have developed a potential-outcomes approach to conceptualizing, evaluating, and unpacking the causal effects of college on earnings. By decomposing the total effect of attending a four-year college into several direct and indirect components, this approach not only helps unveil the mechanisms through which college attendance boosts earnings, but illuminates and quantifies the equalizing and stratifying roles of college. Moreover, under the assumption of sequential ignorability, I have introduced a robust and efficient method for estimating all quantities of interest, along with a set of bias formulas for assessing the sensitivity of estimates to unobserved confounding.

Applying the proposed framework and methodology to data from the NLSY97, I find evidence of both equalizing and stratifying roles of higher education. In particular, the estimated direct effect of college attendance is markedly larger among individuals from the lowest propensity score quintile than among their more advantaged peers. Yet, this equalizing effect is offset by the stratifying effects associated with unequal likelihoods of completing college (given attendance) and unequal covariances between BA completion and its net effect on earnings. The latter component is especially intriguing, as it reflects not an inequality in BA attainment or earnings returns per se, but *an inequality in sorting*: whereas more advantaged college-goers may be well informed about their idiosyncratic payoffs to a BA degree and poised to act on such information, their less advantaged peers may lack such information or the capacity to act on it, leading to a pattern of "negative selection among the least advantaged." As a result of these stratifying forces, the estimated total effect of attending a four-year college for the least advantaged youth is no larger than that for their more advantaged peers.

Methodologically, the causal decomposition and the associated methods for estimation and sensitivity analysis constitute a new framework for analyzing the effects of higher education on earnings. Unlike the conventional practice of dichotomizing postsecondary attainment as either "college goers" versus "high school graduates" or "college graduates" versus "non-graduates," the new framework treats BA completion as a mediator that transmits the effect of college attendance on earnings. This

approach not only maps more closely onto the sequential process by which people make educational transitions (Mare 1980), but enables us, for the first time, to isolate the equalizing and stratifying roles of higher education. Moreover, it opens up new possibilities for future research on the nexus between education and earnings inequality. For example, while the present paper has focused on the effects of college attendance and BA completion, the methodological framework can be generalized to incorporate more educational transitions, such as high school attendance→high school graduation→college attendance→ college graduation→postgraduate attendance→postgraduate degree, where the effect of each transition may show a distinct pattern of heterogeneity (Torche 2011). Moreover, even within the journey from college enrollment to BA completion, the same approach can be applied to assess the roles of important milestones, such as persistence through the first year, and the extent to which they differ between more and less advantaged students. Future research can also adapt the proposed effect decomposition to unpack the economic payoff to attending a two-year college, which comprises not only a direct effect of attendance and an indirect effect via potential attainment of an AA degree, but also an indirect effect via potential transfer to a four-year institution and the associated prospect of attaining a BA degree. Given that two-year colleges currently enroll more than a third of all undergraduate students and that nearly half of all students completing a BA degree had some experience within a two-year institution (Ma and Baum 2016), the relationships between two-year college attendance, eventual educational attainment, and earnings inequality constitute an important avenue for future research.

Author's Note

Replication materials are available in Open Science Framework: https://osf.io/psr3j/.

References

Attewell, Paul, David Lavin, Thurston Domina, and Tania Levey. 2007. *Passing the Torch: Does Higher Education for the Disadvantaged Pay Off across the Generations?* Russell Sage Foundation.

Bowen, William G, Matthew M Chingos, and Michael S McPherson. 2009. *Crossing the Finish Line: Completing College at America's Public Universities*. Princeton, NJ: Princeton University Press.

Brand, Jennie E and Yu Xie. 2010. "Who Benefits Most from College? Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education." *American Sociological Review* 75:273–302.

Breen, Richard, Seong-soo Choi, and Anders Holm. 2015. "Heterogeneous Causal Effects and Sample Selection Bias." *Sociological Science* 2:351–369.

Card, David. 1993. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling."

Carneiro, Pedro, James J Heckman, and Edward J Vytlacil. 2011. "Estimating Marginal Returns to Education." *American Economic Review* 101:2754–81.

Cheng, Siwei, Jennie Brand, Xiang Zhou, Yu Xie, and Michael Hout. 2021. "Heterogeneous Returns to College over the Life Course." *Science Advances* .

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21:C1–C68.

Chetty, Raj, John N Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. 2020. "Income Segregation and Intergenerational Mobility across Colleges in the United States." *The Quarterly Journal of Economics* 135:1567–1633.

Ciocca Eller, Christina and Thomas A. DiPrete. 2018. "The Paradox of Persistence: Explaining the Black-White Gap in Bachelor's Degree Completion." *American Sociological Review* 83:1171–1214.

Dwyer, Rachel E, Laura McCloud, and Randy Hodson. 2012. "Debt and Graduation from American Universities." *Social Forces* 90:1133–1155.

Fiel, Jeremy E. 2020. "Great Equalizer or Great Selector? Reconsidering Education as a Moderator of Intergenerational Transmissions." *Sociology of Education* 93:353–371.

Giani, Matt S, Paul Attewell, and David Walling. 2020. "The Value of an Incomplete Degree: Heterogeneity in the Labor Market Benefits of College Non-completion." *The Journal of Higher Education* 91:514–539.

Hanson, Sandra L and John Zogby. 2010. "The Polls-Trends: Attitudes about the American Dream." *Public Opinion Quarterly* 74:570–584.

Heckman, James J, John Eric Humphries, and Gregory Veramendi. 2018. "Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking." *Journal of Political Economy* 126:S197–S246.

Hout, Michael. 2012. "Social and Economic Returns to College Education in the United States." *Annual Review of Sociology* 38:379–400.

Imai, Kosuke, Luke Keele, Teppei Yamamoto, et al. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25:51–71.

Joffe, Marshall M and Tom Greene. 2009. "Related Causal Frameworks for Surrogate Outcomes." *Biometrics* 65:530–538.

Karlson, Kristian Bernt. 2019. "College as Equalizer? Testing the Selectivity Hypothesis." *Social Science Research* 80:216–229.

Kennedy, Edward H. 2016. "Semiparametric Theory and Empirical Processes in Causal Inference." In *Statistical Causal Inferences and Their Applications in Public Health Research*, pp. 141–167. Springer.

Ma, Jennifer and Sandy Baum. 2016. "Trends in Community Colleges: Enrollment, Prices, Student Debt, and Completion." *College Board Research Brief* 4:1–23.

Mann, Horace. 1848. "Twelfth Annual Report to the Massachusetts Board of Education." *The Republic and the School: Horace Mann and the Education of Free Men* .

Mare, Robert D. 1980. "Social Background and School Continuation Decisions." *Journal of the American Statistical Association* 75:295–305.

Maurin, Eric and Sandra McNally. 2008. "Vive La Revolution! Long-Term Educational Returns of 1968 to the Angry Students." *Journal of Labor Economics* 26:1–33.

Minicozzi, Alexandra. 2005. "The Short Term Effect of Educational Debt on Job Decisions." *Economics of Education Review* 24:417–430.

Pearl, Judea. 2001. "Direct and Indirect Effects." In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420. Morgan Kaufmann Publishers Inc.

Polley, EC and M SuperLearner van der Laan. 2017. "Super Learner Prediction: R package, version 2.0-21."

Robins, James M. 1986. "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period-Application to Control of the Healthy Worker Survivor Effect." *Mathematical Modelling* 7:1393–1512.

Robins, James M. 1997. "Causal Inference from Complex Longitudinal Data." In *Latent Variable Modeling and Applications to Causality*, pp. 69–117. Springer.

Robins, James M. 2003. "Semantics of Causal DAG Models and the Identification of Direct and Indirect effects." *Highly Structured Stochastic Systems* pp. 70–81.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:688–701.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: John Wiley & Sons.

Semenova, Vira and Victor Chernozhukov. 2021. "Debiased machine learning of conditional average treatment effects and other causal functions." *The Econometrics Journal* 24:264–289.

Torche, Florencia. 2011. "Is a College Degree Still the Great Equalizer? Intergenerational Mobility across Levels of Schooling in the United States." *American Journal of Sociology* 117:763–807.

Van der Laan, Mark J, Eric C Polley, and Alan E Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6.

VanderWeele, Tyler J. 2009. "Marginal Structural Models for the Estimation of Direct and Indirect effects." *Epidemiology* 20:18–26.

VanderWeele, Tyler J. 2010. "Bias Formulas for Sensitivity Analysis for Direct and Indirect Effects." *Epidemiology* 21:540.

VanderWeele, Tyler J and Onyebuchi A Arah. 2011. "Bias Formulas for Sensitivity Analysis of Unmeasured Confounding for General Outcomes, Treatments, and Confounders." *Epidemiology* pp. 42–52.

VanderWeele, Tyler J and Stijn Vansteelandt. 2009. "Conceptual Issues Concerning Mediation, Interventions and Composition." *Statistics and its Interface* 2:457–468.

Vansteelandt, Stijn. 2009. "Estimating Direct Effects in Cohort and Case–control Studies." *Epidemiology* 20:851–860.

Willis, Robert J. and Sherwin Rosen. 1979. "Education and Self-selection." *Journal of Political Economy* 87:S7–S36.

Wodtke, Geoffrey T and Xiang Zhou. 2020. "Effect decomposition in the presence of treatment-induced confounding: A regression-with-residuals approach." *Epidemiology* 31:369–375.

Xie, Yu, Jennie Brand, and Ben Jann. 2012. "Estimating Heterogeneous Treatment Effects with Observational Data." *Sociological Methodology* 42:314–347.

Young, Cristobal. 2009. "Model Uncertainty in Sociological Research: An Application to Religion and Economic growth." *American Sociological Review* 74:380–397.

Zheng, Cheng and Xiao-Hua Zhou. 2015. "Causal Mediation Analysis in the Multilevel Intervention and Multicomponent Mediator Case." *Journal of the Royal Statistical Society: Series B: Statistical Methodology* pp. 581–615.

Zhou, Xiang. 2019. "Equalization or Selection? Reassessing the "Meritocratic Power" of a College Degree in Intergenerational Income Mobility." *American Sociological Review* 84:459–485.

Zhou, Xiang. 2020. "Some Doubly and Multiply Robust Estimators of Controlled Direct Effects." *arXiv preprint arXiv:2011.09569* .

Zhou, Xiang and Geoffrey T Wodtke. 2019. "A Regression-with-residuals Method for Estimating Controlled Direct Effects." *Political Analysis* 27:360–369.

Zhou, Xiang and Geoffrey T Wodtke. 2020. "Residual Balancing: a Method of Constructing Weights for Marginal Structural Models." *Political Analysis* 28:487–506.

Zhou, Xiang and Yu Xie. 2016. "Propensity Score–based Methods Versus MTE-based Methods in Causal Inference: Identification, Estimation, and Application." *Sociological Methods & Research* 45:3–40.

Zhou, Xiang and Yu Xie. 2020. "Heterogeneous Treatment Effects in the Presence of Self-selection: a Propensity Score Perspective." *Sociological Methodology* 50:350–385.

Zimmerman, Seth D. 2014. "The Returns to College Admission for Academically Marginal Students." *Journal of Labor Economics* 32:711–754.

Supplementary Materials for *Attendance, Completion, and Heterogeneous Returns to College: A Causal Mediation Approach*

Xiang Zhou

A Neyman Orthogonal Signals for DML Estimation

For each of our target parameters in equations (4)-(8), I construct a Neyman-orthogonal signal using its efficient influence function in the nonparametric model over observed data $O = (X, A, Z, M, Y)$. Specifically, these signals are

$$M^*(1) = \hat{\mathbb{E}}[M|X, A = 1] + \frac{A}{\hat{\pi}(X)}\big(M - \hat{\mathbb{E}}[M|X, A = 1]\big), \tag{15}$$

$$Y^*(0, 0) = \hat{\mathbb{E}}[Y|X, A = 0] + \frac{1 - A}{1 - \hat{\pi}(X)}\big(Y - \hat{\mathbb{E}}[Y|X, A = 0]\big), \tag{16}$$

$$Y^*(1, M(1)) = \hat{\mathbb{E}}[Y|X, A = 1] + \frac{A}{\hat{\pi}(X)}\big(Y - \hat{\mathbb{E}}[Y|X, A = 1]\big), \tag{17}$$

$$Y^*(1, 0) = \hat{\nu}_{10}(X) + \frac{A}{\hat{\pi}(X)}\big(\hat{\mu}_{10}(X, Z) - \hat{\nu}_{10}(X)\big) + \frac{A(1 - M)}{\hat{\pi}(X)\big(1 - \hat{\gamma}(X, Z)\big)}\big(Y - \hat{\mu}_{10}(X, Z)\big), \tag{18}$$

$$Y^*(1, 1) = \hat{\nu}_{11}(X) + \frac{A}{\hat{\pi}(X)}\big(\hat{\mu}_{11}(X, Z) - \hat{\nu}_{11}(X)\big) + \frac{AM}{\hat{\pi}(X)\hat{\gamma}(X, Z)}\big(Y - \hat{\mu}_{11}(X, Z)\big), \tag{19}$$

where

$$\pi(X) := \Pr[A = 1|X],$$

$$\gamma(X, Z) := \Pr[M = 1|X, A = 1, Z],$$

$$\mu_{am}(X, Z) := \mathbb{E}[Y|X, A = a, Z, M = m],$$

$$\nu_{am}(X) := \mathbb{E}[\mu_{am}(X, Z)|X, A = a].$$

The Neyman orthogonality of the signals (15)-(17) is given in Chernozhukov et al. (2018). For a proof of the Neyman orthogonality of the signals (18)-(19), see Zhou (2020). In the above equations, $\hat{\mathbb{E}}[M|X, A = 1]$, $\hat{\mathbb{E}}[Y|X, A = 0]$, $\hat{\mathbb{E}}[Y|X, A = 1]$, $\hat{\pi}(X)$, $\hat{\gamma}(X, Z)$, $\hat{\mu}_{am}(X, Z)$, $\hat{\nu}_{am}(X)$ are all nuisance functions estimated from the "training sample" $\mathcal{I}\backslash\mathcal{I}_j$ in each cross-fitting iteration. The signals (15)-(19) are then used to construct the corresponding signals for Δ_{tot}, Δ_{att}, π_{comp}, Δ_{comp}. For example, the signal for Δ_{tot} is given by $Y^*\big(1, M(1)\big) - Y^*(0, 0)$, the signal for Δ_{att} is given by $Y^*(1, 0) - Y^*(0, 0)$, and so on.

B Inference for the Covariance Components $\triangle_{\text{cov}}$ and $\triangle_{\text{cov}}(s)$

As noted in the main text, the covariance component Δ_{cov} is estimated using the plug-in estimator $\hat{\Delta}_{\text{cov}} = \hat{\Delta}_{\text{tot}} - \hat{\Delta}_{\text{att}} - \hat{\pi}_{\text{comp}}\hat{\Delta}_{\text{comp}}$, where $\hat{\Delta}_{\text{tot}}$, $\hat{\Delta}_{\text{att}}$, $\hat{\pi}_{\text{comp}}$, and $\hat{\Delta}_{\text{comp}}$ are estimated by the sample means of their corresponding Neyman-orthogonal signals. For each of these components, we can decompose its asymptotic error in a way akin to equation (11). For example, for $\hat{\Delta}_{\text{tot}}$, we have

$$\begin{aligned}\sqrt{n}(\hat{\Delta}_{\text{tot}} - \Delta_{\text{tot}}) = \underbrace{\sqrt{n}(\mathbb{P}_n - \mathbb{P})[\varphi_{\text{tot}}(O;\mu,\pi)]}_{A} + \underbrace{\sqrt{n}\mathbb{P}[\varphi_{\text{tot}}(O;\hat{\mu},\hat{\pi}) - \varphi_{\text{tot}}(O;\mu,\pi)]}_{B} \\ + \underbrace{\sqrt{n}(\mathbb{P}_n - \mathbb{P})[\varphi_{\text{tot}}(O;\hat{\mu},\hat{\pi}) - \varphi_{\text{tot}}(O;\mu,\pi)]}_{C},\end{aligned} \tag{20}$$

where $\varphi_{\text{tot}}(O;\mu,\pi) = Y^*\big(1, M(1)\big) - Y^*(0,0)$ is the Neyman-orthogonal signal for Δ_{tot}. Since terms B and C are assumed to be asymptotically negligible, the above expression implies that $\hat{\Delta}_{\text{tot}}$ is asymptotically linear with influence function $\phi_{\text{tot}}(O) = \varphi_{\text{tot}}(O;\mu,\pi) - \mathbb{P}\varphi_{\text{tot}}(O;\mu,\pi)$ (i.e., the demeaned Neyman-orthogonal signal). That is,

$$\hat{\Delta}_{\text{tot}} = \Delta_{\text{tot}} + \mathbb{P}_n\phi_{\text{tot}}(O) + o_p(n^{-1/2}),$$

where $\mathbb{P}_n[\cdot] = n^{-1}\sum_{i=1}^{n}[\cdot]$, and "$\mu$" and "$\pi$" are omitted from $\phi_{\text{tot}}(O)$ to simplify notation. The influence function $\phi_{\text{tot}}(O)$ is essential because its variance captures the asymptotic variance of $\hat{\Delta}_{\text{tot}}$:

$$\begin{aligned}\sqrt{n}(\hat{\Delta}_{\text{tot}} - \Delta_{\text{tot}}) &= \sqrt{n}\big(\mathbb{P}_n\phi_{\text{cov}}(O) + o_p(n^{-1/2})\big) \\ &= \sqrt{n}\mathbb{P}_n\phi_{\text{cov}}(O) + o_p(1) \\ &\xrightarrow{d} N(0, \text{Var}(\phi_{\text{cov}}(O))),\end{aligned}$$

where the last line is due to the central limit theorem and Slutsky's theorem. Similarly, we can show that $\hat{\Delta}_{\text{att}}$, $\hat{\pi}_{\text{comp}}$, and $\hat{\Delta}_{\text{comp}}$ are all asymptotically linear with the corresponding influence functions given by their demeaned Neyman-orthogonal signals. Denoting these influence functions by $\phi_{\text{att}}(O)$,

$\phi_\pi(O)$, and $\phi_{\text{comp}}(O)$, we have

$$
\begin{aligned}
\hat{\Delta}_{\text{cov}} =& \hat{\Delta}_{\text{tot}} - \hat{\Delta}_{\text{att}} - \hat{\pi}_{\text{comp}}\hat{\Delta}_{\text{comp}} \\
=& \big(\Delta_{\text{tot}} + \mathbb{P}_n\phi_{\text{tot}}(O) + o_p(n^{-1/2})\big) - \big(\Delta_{\text{att}} + \mathbb{P}_n\phi_{\text{att}}(O) + o_p(n^{-1/2})\big) \\
& - \big(\pi_{\text{comp}} + \mathbb{P}_n\phi_\pi(O) + o_p(n^{-1/2})\big)\big(\Delta_{\text{comp}} + \mathbb{P}_n\phi_{\text{comp}}(O) + o_p(n^{-1/2})\big) \\
=& \underbrace{\Delta_{\text{tot}} - \Delta_{\text{att}} - \pi_{\text{comp}}\Delta_{\text{comp}}}_{=\Delta_{\text{cov}}} + \mathbb{P}_n[\underbrace{\phi_{\text{tot}}(O) - \phi_{\text{att}}(O) - \phi_\pi(O)\Delta_{\text{comp}} - \phi_{\text{comp}}(O)\pi_{\text{comp}}}_{:=\phi_{\text{cov}}(O)}] + o_p(n^{-1/2}).
\end{aligned}
\tag{21}
$$

Hence, $\hat{\Delta}_{\text{cov}}$ is also asymptotically linear with influence function $\phi_{\text{cov}}(O)$. The asymptotic variance of $\hat{\Delta}_{\text{cov}}$ is thus $\text{Var}[\phi_{\text{cov}}(O)]$, which can be estimated by its empirical analog $\widehat{\text{Var}}[\hat{\phi}_{\text{cov}}(O)]$.

The group-level covariance component $\Delta_{\text{cov}}(s)$ is estimated using the plug-in estimator $\hat{\Delta}_{\text{cov}}(s) = \hat{\Delta}_{\text{tot}}(s) - \hat{\Delta}_{\text{att}}(s) - \hat{\pi}_{\text{comp}}(s)\hat{\Delta}_{\text{comp}}(s)$. Here, $\hat{\Delta}_{\text{tot}}(s)$, $\hat{\Delta}_{\text{att}}(s)$, $\hat{\pi}_{\text{comp}}(s)$, and $\hat{\Delta}_{\text{comp}}(s)$ are the predicted values of their corresponding regression models. As shown in Semenova and Chernozhukov (2021), these predicted values are also asymptotically linear with influence functions in the form of $s^T\mathbb{E}[SS^T]S\epsilon$, where S is a column vector of regressors and ϵ is the error term of the regression model. Hence, we have a group-level counterpart of equation (21). The standard error of $\hat{\Delta}_{\text{cov}}(s)$ can then be estimated through the empirical variance of its influence function.

C Bias Formulas for Sensitivity Analysis

First, let us consider a binary unobserved confounder U that affects both college attendance (A) and earnings (Y) and make the following simplifying assumptions: (A1) $\mathbb{E}[Y|x, U = 1, a] - \mathbb{E}[Y|x, U = 0, a]$ does not depend on x and a; (A2) $\Pr[U = 1|x, A = 1] - \Pr[U = 1|x, A = 0]$ does not depend on x (VanderWeele and Arah 2011). For any $a \in \{0, 1\}$, we have

$$\begin{aligned}\mathbb{E}[Y(a)] &= \int \mathbb{E}[Y|x, u, a]dP(x, u) \\ &= \int \big(\mathbb{E}[Y|x, U = 1, a] \Pr[U = 1|x] + \mathbb{E}[Y|x, U = 0, a] \Pr[U = 0|x]\big)dP(x),\end{aligned}$$

where $Y(a) := Y(a, M(a))$. Without adjusting for U, our estimator for $\mathbb{E}[Y(a)]$ will converge to

$$\begin{aligned}\mathbb{E}^*[Y(a)] &= \int \mathbb{E}[Y|x, a]dP(x) \\ &= \int \big(\mathbb{E}[Y|x, U = 1, a] \Pr[U = 1|x, a] + \mathbb{E}[Y|x, U = 0, a] \Pr[U = 0|x, a]\big)dP(x).\end{aligned}$$

Taking the difference between $\mathbb{E}^*[Y(a)]$ and $\mathbb{E}[Y(a)]$ yields

$$\text{bias}\big[\mathbb{E}[Y(a)]\big] = \int \big(\mathbb{E}[Y|x, U = 1, a] - \mathbb{E}[Y|x, U = 0, a]\big)\big(\Pr[U = 1|x, a] - \Pr[U = 1|x]\big)dP(x). \tag{22}$$

Substituting $a = 0, 1$ into equation (22), taking the difference between $\text{bias}\big[\mathbb{E}[Y(1)]\big]$ and $\text{bias}\big[\mathbb{E}[Y(0)]\big]$, and applying assumptions A1 and A2, we obtain

$$\begin{aligned}\text{bias}[\Delta_{\text{tot}}] &= \underbrace{\big(\Pr[U = 1|x, A = 1] - \Pr[U = 1|x, A = 0]\big)}_{:=\alpha_{\text{tot}}} \underbrace{\big(\mathbb{E}[Y|x, U = 1, a] - \mathbb{E}[Y|x, U = 0, a]\big)}_{:=\beta_{\text{tot}}} \\ &= \alpha_{\text{tot}}\beta_{\text{tot}}.\end{aligned}$$

Next, consider a binary unobserved confounder U that affects both BA completion (M) and earnings (Y) and make the following simplifying assumptions (B1) $\mathbb{E}[Y|x, a, z, U = 1, m] -$

$\mathbb{E}[Y|x, a, z, U = 0, m]$ does not depend on x, a, z, m; (B2) $\Pr[U = 1|x, A = 1, z, M = 1] - \Pr[U = 1|x, A = 1, z, M = 0]$ does not depend on x and z. For any $(a, m) \in \{(0, 0), (1, 0), (1, 1)\}$, we have

$$\begin{aligned}\mathbb{E}[Y(a, m)] &= \int \mathbb{E}[Y|x, a, z, u, m]dP(z, u|x, a)dP(x) \\ &= \int \big(\mathbb{E}[Y|x, a, z, U = 1, m] \Pr[U = 1|x, a, z] + \\ &\quad \mathbb{E}[Y|x, a, z, U = 0, m] \Pr[U = 0|x, a, z]\big)dP(z|a, x)dP(x).\end{aligned}$$

Without adjusting for U, our estimator for $\mathbb{E}[Y(a, m)]$ will converge to

$$\begin{aligned}\mathbb{E}^*[Y(a, m)] &= \int \mathbb{E}[Y|x, a, z, m]dP(z|x, a)dP(x) \\ &= \int \big(\mathbb{E}[Y|x, a, z, U = 1, m] \Pr[U = 1|x, a, z, m] \\ &\quad + \mathbb{E}[Y|x, a, z, U = 0, m] \Pr[U = 0|x, a, z, m]\big)dP(z|a, x)dP(x).\end{aligned}$$

Taking the difference between $\mathbb{E}^*[Y(a, m)]$ and $\mathbb{E}[Y(a, m)]$ yields

$$\begin{aligned}\text{bias}\big[\mathbb{E}[Y(a, m)]\big] = \int &\big(\mathbb{E}[Y|x, a, z, U = 1, m] - \mathbb{E}[Y|x, a, z, U = 0, m]\big) \\ &\cdot \big(\Pr[U = 1|x, a, z, m] - \Pr[U = 1|x, a, z]\big)dP(z|a, x)dP(x). \qquad (23)\end{aligned}$$

Since $M = 0$ when $A = 0$, $\Pr[U = 1|x, A = 0, z, M = 0] = \Pr[U = 1|x, A = 0, z]$. Therefore, $\text{bias}\big[\mathbb{E}[Y(0, 0)]\big] = 0$. Substituting $a = 1$ and $m = 0$ into equation (23) and applying assumptions B1 and B2, we obtain

$$\begin{aligned}\text{bias}[\Delta_{\text{att}}] = &- \underbrace{\big(\Pr[U = 1|x, A = 1, z, M = 1] - \Pr[U = 1|x, A = 1, z, M = 0]\big)}_{:=\alpha_{\text{comp}}} \\ &\cdot \underbrace{\big(\mathbb{E}[Y|x, a, z, U = 1, m] - \mathbb{E}[Y|x, a, z, U = 0, m]\big)}_{:=\beta_{\text{net}}} \int \Pr[M = 1|x, A = 1, z]dP(z|A = 1, x)dP(x) \\ = &- \alpha_{\text{comp}}\beta_{\text{net}} \int \Pr[M = 1|x, A = 1]dP(x)\end{aligned}$$

$$= - \pi_{\text{comp}} \alpha_{\text{comp}} \beta_{\text{net}}.$$

Substituting $a = 1$ and $m = 0, 1$ into equation (23), taking the difference between bias$\big[\mathbb{E}[Y(1,1)]\big]$ and bias$\big[\mathbb{E}[Y(1,0)]\big]$, and applying assumptions B1 and B2, we obtain

$$\begin{aligned}
\text{bias}[\Delta_{\text{comp}}] = & \underbrace{\big(\Pr[U=1|x,A=1,z,M=1] - \Pr[U=1|x,A=1,z,M=0]\big)}_{:=\alpha_{\text{comp}}} \\
& \cdot \underbrace{\big(\mathbb{E}[Y|x,a,z,U=1,m] - \mathbb{E}[Y|x,a,z,U=0,m]\big)}_{:=\beta_{\text{net}}} \\
= & \alpha_{\text{comp}} \beta_{\text{net}}.
\end{aligned}$$

D Results on Percentile Ranks of Earnings

Table D1 and Figure D1 report results when the outcome is measured by the percentile rank of earnings, paralleling Table 1 and Figure 3 in the main text. We can see that the two sets of results are highly consistent.

Table D1: Decomposition of the Average Total Effect (ATE) of College Attendance on Earnings Rank.

Total Effect (Δ_{tot})	Direct Effect (Δ_{att})	Indirect Effect (Δ_{ind})	Completion Prob. (π_{comp})	Completion Effect (Δ_{comp})	Covariance Term (Δ_{cov})
9.79 (0.95)	4.01 (1.04)	5.78 (0.62)	0.57 (0.01)	10.51 (1.03)	-0.20 (0.37)

Note: Numbers in parentheses are estimates of standard errors, which are constructed using the empirical variances of the corresponding influence functions and adjusted for multiple imputation via Rubin's (1987) method.

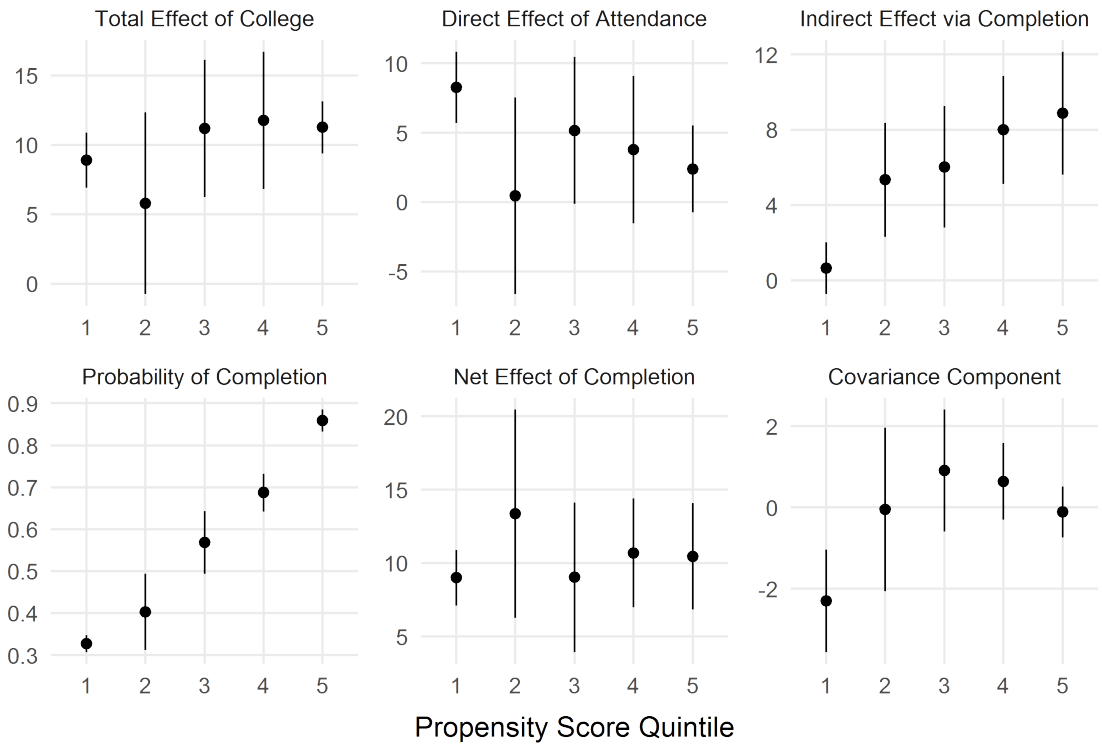

Figure D1: Estimates of the Total Effect and Its Components by Propensity Score Quintile. Note: Line ranges represent 95% confidence intervals.