Attendance, Completion, and Heterogeneous Returns to College*  

Xiang Zhou  
Harvard University  
August 27, 2021  

Abstract  
A growing body of social science research has investigated whether the economic payoff to a college education is heterogeneous — in particular, whether socioeconomically 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 sequential 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 disequalizer. The total effect of college attendance, its direct and indirect components, and their heterogeneity by socioeconomic background 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.
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.” 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 socioeconomically 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

1https://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 (Chen et al. 2018; 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 influence of parental income on 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. 2018), 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 than 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
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 sequential approach, which draws upon the language and logic of causal mediation analysis, 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 dis-
tinct pattern of effect heterogeneity by socioeconomic background. 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 disequalizer.

![Diagram of College Attendance, BA Completion, and Earnings](image)

Figure 1: The Effects of College on Earnings in a Direct Acyclic Graph.

When we observe a large number of individual-, family-, and school-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 by socioeconomic background 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 meth-
ods, carefully 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 methodology using data from the National Longitudinal Survey of Youth, 1997 cohort (NLSY97). I first evaluate the total effect of attending a four-year college on log earnings and its direct and indirect components via BA completion. I then examine how the total effect and its direct and indirect components vary across individuals with different socioeconomic backgrounds, where socioeconomic background is measured by (a) parental income and (b) the propensity score of attending college. I find evidence of equalizing effects for both college attendance (i) and BA completion (iii). The estimated direct effect of college attendance, for example, is markedly larger among individuals from the lowest propensity score quintile than among the rest of the population. However, in terms of both parental income and the propensity score, the modest equalizing effects of attendance and completion are outweighed by the disequalizing effects associated with unequal likelihoods of completing a BA degree (ii) and unequal magnitudes of the covariance component (iv). Consequently, the total effect of attending a four-year college appears disequalizing, i.e., greater for more advantaged youth than for their less advantaged peers (in terms of both parental income and the propensity score). I discuss the policy implications of these findings toward the end of the paper.

**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

$$\text{TE}_i = Y_i(1, M_i(1)) - Y_i(0, M_i(0))$$

$$= Y_i(1, M_i(1)) - Y_i(0, 0) \quad \text{(because } M_i(0) = 0)$$

$$= Y_i(1, 0) - Y_i(0, 0) + M_i(1) \left( Y_i(1, 1) - Y_i(1, 0) \right).$$

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 averages of these effects. First, taking the expectation of equation (1) yields a population-level decomposition:

$$\mathbb{E}[\text{TE}_i] = \mathbb{E}[Y_i(1, 0) - Y_i(0, 0)] + \mathbb{E}[M_i(1)] \cdot \mathbb{E}[Y_i(1, 1) - Y_i(1, 0)]$$

$$+ \text{Cov}[M_i(1), Y_i(1, 1) - Y_i(1, 0)]$$

$$= \Delta_{\text{tot}} = \Delta_{\text{att}} + \pi_{\text{comp}} \Delta_{\text{comp}} + \Delta_{\text{cov}}.$$
Here, $\Delta_{\text{tot}}$ represents the average total effect of college on earnings, $\Delta_{\text{att}}$ represents the average direct effect of college attendance on earnings, and $\Delta_{\text{ind}}$ represents the average indirect effect via BA completion. The indirect effect $\Delta_{\text{ind}}$ equals $\pi_{\text{comp}}\Delta_{\text{comp}} + \Delta_{\text{cov}}$, where $\pi_{\text{comp}}$ represents the probability of BA completion if a person attended college, $\Delta_{\text{comp}}$ represents the average net effect of BA completion on earnings, and $\Delta_{\text{cov}}$ is a component reflecting the covariance between BA completion and its net effect on earnings. Intuitively, $\Delta_{\text{cov}}$ 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 $\Delta_{\text{cov}}$ 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 $\Delta_{\text{cov}}$ 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 complete 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 socioeconomic backgrounds, we can evaluate the conditional expectation of equation (1) given $S_i$, some indicator of socioeconomic background. Analogous to the population-level decomposition (2), we have

\[
\begin{align*}
\mathbb{E}[TE_i | S_i = s] &= \mathbb{E}[Y_i(1, 0) - Y_i(0, 0) | S_i = s] + \mathbb{E}[M_i(1) | S_i = s] \cdot \mathbb{E}[Y_i(1, 1) - Y_i(1, 0) | S_i = s] \\
&\quad + \text{Cov}[M_i(1), Y_i(1, 1) - Y_i(1, 0) | S_i = s] \\
&= \Delta_{\text{att}}(s) + \pi_{\text{comp}}(s)\Delta_{\text{comp}}(s) + \Delta_{\text{cov}}(s), \quad (3)
\end{align*}
\]

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 disequalizing 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 par-
ticularly large among individuals from disadvantaged backgrounds, contributing to the equalizing
role of higher education. On the other hand, ample empirical evidence indicates that college grad-
uation rates are much higher among students from more advantaged backgrounds relative to their
less privileged peers. Thus the component $\pi_{\text{comp}}(s)$ is likely an increasing function of socioeconomic
background, contributing to the disequalizing role of higher education. Furthermore, as we 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 un-
equal access to information), then the within-group covariance component $\Delta_{\text{cov}}(s)$ may also be an
increasing function of socioeconomic background, contributing to the disequalizing role of higher
education. Given these competing forces, an expansion in college enrollment would have the poten-
tial to reduce inequality by socioeconomic background if the equalizing effects of college (e.g., those
associated with $\Delta_{\text{att}}(s)$ and $\Delta_{\text{comp}}(s)$) outweigh its disequalizing effects (e.g., those associated with
$\pi_{\text{comp}}(s)$ and $\Delta_{\text{cov}}(s)$). In the next section, we evaluate this hypothesis using data from the NLSY97.

In addition to assessing how the causal effects of college attendance and BA completion differ
by socioeconomic background, we can also evaluate a set of “potential earnings disparities” between
individuals with more and less advantaged backgrounds. For example, if $S$ denotes parental income,
then $\mathbb{E}[Y_i(1, 1) \mid S_i = s]$ would characterize how potential earnings for a college graduate vary as a
function of parental income. Similarly, $\mathbb{E}[Y_i(1, M_i(1)) \mid S_i = s]$ would characterize how potential
earnings for a college goer (regardless of completion status) vary as a function of parental income.
Such potential earnings disparities may be referred to as “controlled disparities,” akin to the concept of
“controlled mobility” introduced in Zhou (2019) (see Jackson and VanderWeele (2018) and Lundberg
(2020) for more expansive discussions of this concept). As noted in Lundberg (2020), these quantities
should best be interpreted locally, i.e., as earnings inequality that would arise among a small sample of
high- and low-income youth if their college attendance/completion status was fixed at a given level.²

**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 pre-college characteristics that may confound the causal effects of college attendance and BA completion on earnings, and $Z$ to denote a set of postsecondary characteristics (e.g., college GPA) that may 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., $(M(1), Y(0, 0), Y(1, 0), Y(1, 1)) \perp \perp A|X)$, and (b) conditional on pre-college characteristics $X$ and postsecondary characteristics $Z$, BA completion is independent of potential earnings among college goers (i.e., $(Y(1, 0), Y(1, 1)) \perp \perp M|X, A = 1, Z$). The sequential ignorability assumption is satisfied in Figure 2, which contains a directed acyclic graph (DAG) visualizing the hypothesized causal relationships among the variables defined previously.

Equation (2) implies that to identify the total effect of college attendance ($\Delta_{\text{tot}}$) and its various components ($\Delta_{\text{att}}, \pi_{\text{comp}}, \Delta_{\text{comp}}, \Delta_{\text{cov}}$), it suffices to identify the following expected potential outcomes: $E[M(1)], E[Y(0, 0)], E[Y(1, M(1))], E[Y(1, 0)],$ and $E[Y(1, 1)]$. Here we omit the subscript $i$ for conciseness. Under the sequential ignorability assumption, these quantities are identified via

²Alternatively, these quantities could be interpreted *globally*, i.e., as inequality that would arise if the educational status of all youth was fixed at a given level. Compared with the global interpretation, the local interpretation is both more realistic and more immune to potential violations of the stable unit treatment value assumption (see Lundberg 2020).
Robins’s (1986; 1997) g-formula:

\[ E[M(1)] = \int E[M|x, A = 1]dP(x), \]  
\[ E[Y(0, 0)] = \int E[Y|x, A = 0]dP(x), \]  
\[ E[Y(1, M(1))] = \int E[Y|x, A = 1]dP(x), \]  
\[ E[Y(1, 0)] = \int \int E[Y|x, A = 1, z, M = 0]dP(z|x, A = 1)dP(x), \]  
\[ E[Y(1, 1)] = \int \int E[Y|x, A = 1, z, M = 1]dP(z|x, A = 1)dP(x), \]

where \( P(u) \) denotes the cumulative distribution function of a random variable \( U \). It is easy to see that \( \pi_{\text{comp}} \) is identified by equation (4), \( \Delta_{\text{tot}} \) identified by equation (6) minus equation (5), \( \Delta_{\text{att}} \) identified by equation (7) minus equation (5), \( \Delta_{\text{comp}} \) identified by equation (8) minus equation (7), and \( \Delta_{\text{cov}} \) 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 $X$, both college attendance and BA completion are independent of potential earnings (i.e., $(Y(0, 0), Y(1, 0), Y(1, 1)) \perp \perp (A, M) | X$). 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).

Nonetheless, sequential ignorability is still a strong and unverifiable assumption, which can be violated whenever unobserved confounders (e.g., motivation, preference for cognitive tasks over physical tasks, etc.) exist for any of the causal relationships involved. Thus, in practice, it is prudent to report the sensitivity of estimated causal effects to potential violations of sequential ignorability (e.g., Breen et al. 2015). In Section 1, I outline a bias factor approach (VanderWeele 2010, VanderWeele and Arah 2011) 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 (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 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 equations (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 socioeconomic background $S$. Specifically, it involves the following steps:

1. Randomly partition the analytical sample $\mathcal{I}$ into $J$ equal-sized subsamples: $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_J$;

2. For each subsample $\mathcal{I}_j$,
   
   (a) Use the observations in $\mathcal{I} \setminus \mathcal{I}_j$ (i.e., all observations but those in $\mathcal{I}_j$) to fit a flexible machine learning model for each of the following "nuisance functions":
       $\Pr[A = 1|x]$, $\Pr[M = 1|x, A = 1]$, $\Pr[M = 1|x, A = 1, z]$, $\mathbb{E}[Y|x, a]$, $\mathbb{E}[Y|x, A = 1, z, m]$, 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 "signals," one for each potential outcome: $M^*(1)$, $Y^*(0, 0)$, $Y^*(1, M(1))$, $Y^*(1, 0)$, and $Y^*(1, 1)$ (see Supplementary Material A for detailed expressions of these signals).

3. In the full sample, use the above signals for potential outcomes to construct the corresponding signals for $\Delta_{\text{tot}}$, $\Delta_{\text{att}}$, $\Delta_{\text{comp}}$, $\Delta_{\text{comps}}$, and $\Delta_{\text{cov}}$. The sample averages of these signals constitute the DML estimates of the corresponding quantities.

4. To assess effect heterogeneity by socioeconomic background (e.g., $\Delta_{\text{comp}}(s)$), fit a linear model of the corresponding signal (constructed in step 3) on $S$. Similarly, to assess the potential earn-

---

3 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).

4 Technically, these signals are plug-in estimates of the recentered efficient influence functions for the expectations of the corresponding potential outcomes (see Semenova and Chernozhukov 2021). For each observation, its signal for a potential outcome can be interpreted as its “contribution” to the mean of the potential outcome.
ings disparities associated with a given level of education (e.g., college graduate), fit a linear model of the signal for the corresponding potential outcome (e.g., $Y^\ast(1, 1)$) on $S$.

In step 2(b), the signals are constructed using the efficient influence functions of the corresponding estimands, making our estimates not only robust to model misspecification, but also consistent and asymptotically normal under mild conditions (Zhou 2020). Moreover, the nuisance functions and the target parameters are estimated in different subsamples ($\mathcal{I} \setminus \mathcal{I}_j$ and $\mathcal{I}_j$, respectively), potential bias due to overfitting is removed. For all quantities of interest, asymptotically valid standard errors can be constructed using sample variances of the estimated influence functions.

**A Bias Factor Approach to Sensitivity Analysis**

In the context of 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 used to derive a set of bias formulas for the total effect of college ($\Delta_{\text{tot}}$), the direct effect of attendance ($\Delta_{\text{att}}$), and the net effect of completion ($\Delta_{\text{comp}}$).

First, let us consider the total effect of college attendance ($\Delta_{\text{tot}}$), 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 whether a person has a strong interpersonal skill, that affects both college attendance and earnings. Under some simplifying assumptions regarding the homogeneity of the relationships between $U$, $A$, and $Y$, the bias for the estimated $\Delta_{\text{tot}}$ is given by (see Supplementary Material B)

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

(9)

where $\alpha_{\text{tot}}$ 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 $\beta_{\text{tot}}$ 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 ($\Delta_{\text{att}}$) and the net effect of BA completion ($\Delta_{\text{comp}}$) 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. Under some simplifying assumptions regarding the homogeneity of the relationships between $U$, $A$, $M$, $Z$ and $Y$, the biases for the estimated $\Delta_{\text{att}}$ and $\Delta_{\text{comp}}$ are given by (see Supplementary Material B)

\[
\text{bias}[\Delta_{\text{att}}] = \alpha_{\text{att}} \beta_{\text{net}},
\]

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

where $\alpha_{\text{att}}$ denotes the difference in the prevalence of $U$ between high school graduates ($A = M = 0$) and college dropouts ($A = 1, M = 0$) given both pre-college and postsecondary characteristics ($X$ and $Z$), $\alpha_{\text{comp}}$ 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 $\beta_{\text{net}}$ 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 $\alpha_{\text{tot}}$, $\beta_{\text{tot}}$, $\alpha_{\text{att}}$, $\alpha_{\text{comp}}$, $\beta_{\text{net}}$ are group-specific, i.e., depending on $S = s$. It is clear that if these sensitivity parameters are identical between individuals from different socioeconomic backgrounds, 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). 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 ($A$), 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, 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 C).

Pre-college characteristics ($X$) include basic demographic variables (gender, race, ethnicity, age at 1997), socioeconomic background (parental education, parental income, parental assets, co-residence
with both biological parents, presence of a paternal figure, rural residence, southern residence), ability and behavior (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 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 asset are inflation-adjusted to 2019 dollars.

Postsecondary characteristics ($Z$) include several variables pertaining to the college attended by the respondent as well as the respondent’s field of study and college GPA. 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 many 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. Finally, college GPA is measured as the respondent’s cumulative GPA from the Post-Secondary Transcript Study (PSTRAN). 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 described in Section to implement the decompositions (2) and (3). To examine effect heterogeneity, I use two indicators of socioeconomic background ($S$): the percentile rank of parental income and the estimated propensity score of attending college given all pre-college covariates.\(^5\) Previous research has advocated the use of the propensity score as a uni-dimensional summary of socioeconomic background (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 disequalizing roles of college. To obtain an overall picture of effect heterogeneity between more and less advantaged youth, I first fit a linear model for each of the signals (e.g., $Y^*(1, 0) - Y^*(0, 0)$) on parental income rank/the propensity score. Furthermore, given that previous research has reported U-shaped patterns of effect heterogeneity by the propensity score (Zhou and Xie 2016; Cheng et al. 2018), I also 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.\(^6\)

\(^5\) In my analyses, the estimated propensity scores are taken 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.

\(^6\) 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.

<table>
<thead>
<tr>
<th>Total Effect (Δ_tot)</th>
<th>Direct Effect (Δ_att)</th>
<th>Indirect Effect (Δ_ind)</th>
<th>Completion Prob. (π_comp)</th>
<th>Completion Effect (Δ_comp)</th>
<th>Covariance Term (Δ_cov)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.39 (0.05)</td>
<td>0.12 (0.07)</td>
<td>0.27 (0.03)</td>
<td>0.57 (0.01)</td>
<td>0.49 (0.07)</td>
<td>-0.01 (0.02)</td>
</tr>
</tbody>
</table>

Note: Numbers in parentheses are estimates of standard errors constructed using sample variances of the estimated efficient influence functions and adjusted for multiple imputation via Rubin’s (1987) method.

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 0.12, or a 12.7% earnings premium ($e^{0.12} - 1 = 0.127$) 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 ($Δ_{comp}$), 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 ($Δ_{ind} = π_{comp}Δ_{comp} + Δ_{cov}$) is largely determined by the product of $π_{comp}$ and $Δ_{comp}$ ($0.57 * 0.49 = 0.28$). In particular, the estimated net effect of BA completion implies an earning premium of 63.2% ($e^{0.49} - 1 = 0.632$) 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% ($e^{0.61} - 1 = 0.840$).

Figure 3 reports how estimates of the total effect and its various components vary by the percentile rank of parental income. Specifically, the upper panels show heterogeneity in the total effect of college ($Δ_{tot}(s)$), the direct effect of attendance ($Δ_{att}(s)$), and the indirect effect via BA completion ($Δ_{ind}(s)$).
Figure 3: Estimated Causal Effects as Linear Functions of Parental Income Rank. 
Note: Ribbons represent 95% confidence intervals.

The total effect of college exhibits a positive, albeit not statistically significant, slope, suggesting that high-income youth on average may benefit more from attending a four-year college relative to low-income youth. In other words, the overall effect of college appears to be *disequalizing*. From the next two panels, we can see that the direct effect of attendance does not vary much (linearly) by parental income, but the indirect effect via BA completion seems stronger for high-income youth than for low-income youth. The overall disequalizing effect of college, therefore, stems chiefly from the disequalizing effect associated with BA completion.

The lower panels reveal why the indirect effect via BA completion is disequalizing by showing effect heterogeneity in each of its three components. We can see that heterogeneity in the indirect effect via BA completion is driven by several competing forces. First, the net effect of BA completion ($\Delta_{\text{comp}}(s)$) is equalizing, as it tends to be stronger among low-income youth than among high-income youth. However, the equalizing effect of BA completion on earnings is offset by the acute inequality associated with the probability of BA completion ($\pi_{\text{comp}}(s)$). As shown in the lower left panel, the
likelihood of completing a BA degree given college attendance is more than twice as large for the highest-income students as for the lowest-income students. In addition, the covariance component ($\Delta_{\text{cov}}(s)$) also manifests itself as an increasing function of parental income, contributing to the disequalizing effect of college. As noted earlier, this pattern may reflect a stronger degree of self-selection into BA attainment (based on anticipated gains from it) among high-income youth than among low-income youth.

As noted earlier, we can also assess potential earnings disparities associated with a given level of education by regressing the signal for the corresponding potential outcome on $S$. The results from this exercise are shown in Figure 4, where the left panel shows how the potential earnings for a high school graduate ($Y(0, 0)$) vary by parental income rank versus those for a college goer ($Y(1, M(1))$), and the right panel reports the potential earnings for a high school graduate ($Y(0, 0)$), for a college dropout/stopout ($Y(1, 0)$), and for a college graduate ($Y(1, 1)$). We can see that at all levels of education, potential earnings depend substantially on parental income, reflecting a "long
shadow of a person's socioeconomic background beyond her educational attainment. Moreover, the left panel suggests that potential earnings depend more strongly on parental income rank for a college goer (dashed line) than for a high school graduate (solid line), echoing the disequalizing effect of college shown in Figure 3. On the other hand, as shown in the right panel, the dependence of potential earnings on parental income rank is somewhat weaker for a college graduate (dotted line) than for a college dropout/stopout (dotdash line) or for a high school graduate (solid line), echoing the equalizing effect of BA completion shown in Figure 3. According to the local interpretation of these potential disparities (see Section ), our findings suggest that among a small sample of low- and high-income youth, earnings inequality by parental income would be higher if everyone attended college than if none attended college, but it would be lower if everyone attended and completed college than if none attended college.

Figure 5 shows how estimates of the total effect and its various components vary by the estimated propensity score. We can see that patterns of effect heterogeneity are similar, except that the direct
effect of attendance now also appears equalizing, i.e., larger among low-propensity individuals than among high-propensity individuals. Yet, because of the strong disequalizing effect associated with the probability of BA completion (lower left panel) and the covariance component (lower right panel), the overall effect of college attendance is still disequalizing (upper left panel). Figure 6 shows how these equalizing and disqualizing effects translate into potential earnings disparities between low- and high-propensity individuals. We can see that the dependence of potential earnings on the propensity score is stronger for a college goer than for a high school graduate (dashed line versus solid line in the left panel), but weaker for a college graduate than for a high school graduate (dotted line versus solid line in the right panel).

To see potential nonlinear patterns of effect heterogeneity, we now turn to Figure 7, which 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 direct effect of attendance, the net effect of completion, and the covariance component, although estimation uncertainty, especially
that associated with the second propensity score quintile, 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 (0.46), much larger than those for the other quintiles, whose direct effect estimates are all 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 $A \rightarrow M \rightarrow Y$ 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 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 col-
lege 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 is 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**

We now illustrate the bias factor approach to sensitivity analysis introduced in Section . In particular, let us consider the total effect of college ($\Delta_{tot}$), the direct effect of attendance ($\Delta_{att}$), and the net effect of completion ($\Delta_{comp}$) for individuals in the lowest and highest propensity score quintiles. First, in the presence of an unobserved confounder $U$ that affects both college attendance and earnings, the bias for our total effect estimate is given by $\alpha_{tot}\beta_{tot}$ (equation 9). Given the symmetry of the bias formula, let us consider only cases where $U$ is positively associated with log earnings, i.e., $\beta_{tot} > 0$, while leaving the sign of $\alpha_{tot}$ unconstrained. Columns 4-5 of Table 2 report the bias-adjusted estimates of $\Delta_{tot}$ for the lowest- and highest-propensity individuals across a range of potential values of $\alpha_{tot}$ and $\beta_{tot}$. 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 $\alpha_{tot}$ and $\beta_{tot}$ are both positive. We find that in this case, although our estimates of $\Delta_{tot}$ will be upwardly biased, they are highly 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 0.37 for the least and the most advantaged groups, respectively.

Second, in the presence of an unobserved confounder for the effect of BA completion on earnings, the biases for our estimates of the direct effect of attendance and the net effect of completion are given by $\alpha_{att}\beta_{net}$ and $\alpha_{comp}\beta_{net}$ (equations 10 and 11), respectively. Given the symmetry of these formulas,
Table 2: Sensitivity Results for the Total Effect of College ($\Delta_{\text{tot}}$), the Direct Effect of Attendance ($\Delta_{\text{att}}$), and the Net Effect of BA Completion ($\Delta_{\text{comp}}$) for Individuals in the Lowest and Highest Propensity Score (PS) Quintiles.

<table>
<thead>
<tr>
<th>Sensitivity Parameters</th>
<th>Bias</th>
<th>Total Effect</th>
<th>Direct Effect of Attendance</th>
<th>Net Effect of Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1st PS Quintile</td>
<td>5th PS Quintile</td>
<td>1st PS Quintile</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$\beta$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0.38</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>-0.3</td>
<td>0.1</td>
<td>0.41</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td>-0.3</td>
<td>0.2</td>
<td>0.44</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>-0.3</td>
<td>0.3</td>
<td>0.47</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>-0.2</td>
<td>0.1</td>
<td>0.40</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>-0.2</td>
<td>0.2</td>
<td>0.42</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>-0.2</td>
<td>0.3</td>
<td>0.44</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>-0.1</td>
<td>0.1</td>
<td>0.39</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>-0.1</td>
<td>0.2</td>
<td>0.40</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>-0.1</td>
<td>0.3</td>
<td>0.41</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>0.37</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
<td>0.36</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>0.1</td>
<td>0.3</td>
<td>0.35</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>0.36</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>0.34</td>
<td>0.42</td>
<td>0.42</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3</td>
<td>0.32</td>
<td>0.40</td>
<td>0.40</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>0.35</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td>0.3</td>
<td>0.2</td>
<td>0.32</td>
<td>0.40</td>
<td>0.40</td>
</tr>
<tr>
<td>0.3</td>
<td>0.3</td>
<td>0.29</td>
<td>0.37</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Note: The sensitivity parameters $\alpha$ and $\beta$ refer to $\alpha_{\text{tot}}$ and $\beta_{\text{tot}}$ for the total effect of college, to $\alpha_{\text{att}}$ and $\beta_{\text{net}}$ for the direct effect of attendance, and to $\alpha_{\text{comp}}$ and $\beta_{\text{net}}$ for the net effect of BA completion.

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 $\alpha_{\text{comp}}$ is also positive. The sign of $\alpha_{\text{att}}$, however, may be either positive or negative. For example, if $U$ represents a positive peer influence, it should be more prevalent among individuals with a higher level of education. Yet, conditioning on college completion status ($M=0$), which is a common consequence of $A$ and $U$ (i.e., a collider variable), may reduce or even reverse their association, due to particularly low values of $U$ among college dropouts/stopouts. Thus, we assess bias[$\Delta_{\text{att}}$] for both positive and negative values.
of $\alpha_{\text{att}}$. The second and third panels of Table 2 report the bias-adjusted estimates of $\Delta_{\text{att}}$ and $\Delta_{\text{comp}}$ across a range of potential values of $\alpha_{\text{att}}$, $\alpha_{\text{comp}}$, and $\beta_{\text{net}}$. We can see that the direct effect of attendance will be underestimated if $\alpha_{\text{att}} < 0$ and overestimated if $\alpha_{\text{att}} > 0$. The net effect of BA completion is likely overestimated, assuming $\alpha_{\text{comp}} \geq 0$. 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, however, appears robust only among the least advantaged youth.

As noted earlier, if the sensitivity parameters are constant across the population, our findings about effect heterogeneity will be unchanged. However, the sensitivity parameters may differ between less and more advantaged individuals. For example, peer influence might be more crucial in shaping college attendance, college completion, and earnings among low-propensity youth than among high-propensity youth. In this case, estimates of $\Delta_{\text{tot}}$, $\Delta_{\text{att}}$, and $\Delta_{\text{comp}}$ may suffer larger amounts of bias among individuals in the lowest propensity score quintile than among those in the highest propensity score quintile. For example, for the direct effect of attendance, our findings about effect heterogeneity will be subject to a differential selection bias of $\alpha_{\text{att}}^{1\text{st quintile}} \beta_{\text{net}}^{1\text{st quintile}} - \alpha_{\text{att}}^{5\text{th quintile}} \beta_{\text{net}}^{5\text{th quintile}}$. In this particular case, however, the differential selection bias would have to reach 0.4 to explain away the difference between the lowest- and highest-propensity individuals in their estimated $\Delta_{\text{att}}$ (0.46 versus 0.06). 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 $\Delta_{\text{att}}$. By contrast, our estimated differences in $\Delta_{\text{tot}}$ and $\Delta_{\text{comp}}$ 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 socioeconomically 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 disequalizing effects 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 equalizing effects associated with both college attendance and BA completion. In particular, the direct effect of college attendance on earnings is largest among the lowest-propensity youth, i.e., those who are least likely to attend college given their background characteristics. Yet, these equalizing effects are outweighed by the disequalizing 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 socioeconomically 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 disequalizing forces, the total effect of college is also disequalizing, i.e., larger for high-income/high-propensity youth than for their low-income/low-propensity peers.

The above findings have several implications for higher education policies aimed at reducing inequality and improving intergenerational mobility. First, given the overall disequalizing effect of college, a blanket expansion of the postsecondary system that simply induces more youth into the college campus will be insufficient to reduce inequality by socioeconomic background. By contrast, given the equalizing effect of a BA degree on earnings, efforts aimed at boosting graduation rates
and shortening time to BA completion should be more effective at disrupting the intergenerational persistence of (dis)advantages. Closing gaps in graduation rates and in time to completion between less and more advantaged students will also lead to a reduction of earnings inequality. In short, educational interventions at the postsecondary level will be most effective if they are designed to both increase overall graduation rates and narrow gaps in BA completion between students from less and more advantaged backgrounds.

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 an intermediate step 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 disequalizing 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 be confounded by a different set of factors. 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.
References


Supplementary Materials

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

\begin{align}
M^*(1) &= \hat{E}[M|X, A = 1] + \frac{A}{\hat{\pi}(X)}(M - \hat{E}[M|X, A = 1]), \quad (12) \\
Y^*(0, 0) &= \hat{E}[Y|X, A = 0] + \frac{1 - A}{1 - \hat{\pi}(X)}(Y - \hat{E}[Y|X, A = 0]), \quad (13) \\
Y^*(1, M(1)) &= \hat{E}[Y|X, A = 1] + \frac{A}{\hat{\pi}(X)}(Y - \hat{E}[Y|X, A = 1]), \quad (14) \\
Y^*(1, 0) &= \hat{\nu}_{10}(X) + \frac{A}{\hat{\pi}(X)}(\hat{\mu}_{10}(X, Z) - \hat{\nu}_{10}(X)) + \frac{A(1 - M)}{\hat{\pi}(X)(1 - \hat{\gamma}(X, Z))}(Y - \hat{\mu}_{10}(X, Z)), \quad (15) \tag{15} \\
Y^*(1, 1) &= \hat{\nu}_{11}(X) + \frac{A}{\hat{\pi}(X)}(\hat{\mu}_{11}(X, Z) - \hat{\nu}_{11}(X)) + \frac{AM}{\hat{\pi}(X)\hat{\gamma}(X, Z)}(Y - \hat{\mu}_{11}(X, Z)), \quad (16) \tag{16}
\end{align}

where

\begin{align}
\hat{\pi}(X) &=: \text{Pr}[A = 1|X], \\
\hat{\gamma}(X, Z) &=: \text{Pr}[M = 1|X, A = 1, Z], \\
\hat{\mu}_{am}(X, Z) &=: \mathbb{E}[Y|X, A = a, Z, M = m], \\
\hat{\nu}_{am}(X) &=: \mathbb{E}[\hat{\mu}_{am}(X, Z)|X, A = a]. 
\end{align}

The Neyman orthogonality of the signals (12)-(14) is given in Chernozhukov et al. (2018). For a proof of the Neyman orthogonality of the signals (15)-(16), see Zhou (2020). In the above equations, $\hat{E}[M|X, A = 1], \hat{E}[Y|X, A = 0], \hat{E}[Y|X, A = 1], \hat{\pi}(X), \hat{\gamma}(X, Z), \hat{\mu}_{am}(X, Z), \hat{\nu}_{am}(X)$ are all nuisance functions estimated from the “training sample” $\mathcal{I} \setminus \mathcal{I}_j$ in each cross-fitting iteration. The
signals (12)-(16) are then used to construct the corresponding signals for $\Delta_{\text{tot}}$, $\Delta_{\text{att}}$, $\pi_{\text{comp}}$, $\Delta_{\text{comp}}$. For example, the signal for $\Delta_{\text{att}}$ is given by $Y^*(1, 0) - Y^*(0, 0)$. The signal for $\Delta_{\text{cov}}$ is constructed using its efficient influence function, which is obtained by applying the delta method to $\Delta_{\text{cov}} = \Delta_{\text{tot}} - \Delta_{\text{att}} - \pi_{\text{comp}} \Delta_{\text{comp}}$.

### B 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

$$
\mathbb{E}[Y(a)] = \int \mathbb{E}[Y|x, u, a]dP(x, u)
= \int \left( \mathbb{E}[Y|x, U = 1, a] \Pr[U = 1|x] + \mathbb{E}[Y|x, U = 0, a] \Pr[U = 0|x] \right) dP(x),
$$

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

$$
\mathbb{E}^*[Y(a)] = \int \mathbb{E}[Y|x, a]dP(x)
= \int \left( \mathbb{E}[Y|x, U = 1, a] \Pr[U = 1|x, a] + \mathbb{E}[Y|x, U = 0, a] \Pr[U = 0|x, a] \right) dP(x).
$$

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

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

(17)
Substituting $a = 0, 1$ into equation (17), taking the difference between bias[$\mathbb{E}[Y(1)]$] and bias[$\mathbb{E}[Y(0)]$], and applying assumptions A1 and A2, we obtain

$$\text{bias}[\Delta_{\text{tot}}] = \left( \Pr[U = 1|x, A = 1] - \Pr[U = 1|x, A = 0] \right) \left( \mathbb{E}[Y|x, U = 1, a] - \mathbb{E}[Y|x, U = 0, a] \right) = \alpha_{\text{tot}} \beta_{\text{tot}}.$$  

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) $\Pr[U = 1|x, A = 1, z, M = 0] - \Pr[U = 1|x, A = 0, z, M = 0]$ does not depend on $x$ and $z$; (B3) $\Pr[U = 1|x, A = 1, z, M = 0]$ does not depend on $x$ and $z$ (VanderWeele 2010). For any $(a, m) \in \{(0, 0), (1, 0), (1, 1)\}$, we have

$$\mathbb{E}[Y(a, m)] = \int \mathbb{E}[Y|x, a, z, m] dP(z|x, a)dP(x) = \int (\mathbb{E}[Y|x, a, z, U = 1, m] \Pr[U = 1|x, a, z] + \mathbb{E}[Y|x, a, z, U = 0, m] \Pr[U = 0|x, a, z]) dP(z|a, x)dP(x).$$

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

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

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

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

Substituting \( a = 0, 1 \) and \( m = 0 \) into equation (18), taking the difference between \( \text{bias} \left[ \mathbb{E}[Y(1,0)] \right] \) and \( \text{bias} \left[ \mathbb{E}[Y(0,0)] \right] \), and applying assumptions B1 and B2, we obtain

\[
\text{bias}[\Delta_{\text{att}}] = \left( \Pr[U = 1|x, A = 1, z, M = 0] - \Pr[U = 1|x, A = 0, z, M = 0] \right)
\]
\[
\cdot \left( \mathbb{E}[Y|x, a, z, U = 1, m] - \mathbb{E}[Y|x, a, z, U = 0, m] \right) \\
= \alpha_{\text{att}} \beta_{\text{net}}. 
\]

Substituting \( a = 1 \) and \( m = 0, 1 \) into equation (18), taking the difference between \( \text{bias} \left[ \mathbb{E}[Y(1,1)] \right] \) and \( \text{bias} \left[ \mathbb{E}[Y(1,0)] \right] \), and applying assumptions B1 and B3, we obtain

\[
\text{bias}[\Delta_{\text{comp}}] = \left( \Pr[U = 1|x, A = 1, z, M = 1] - \Pr[U = 1|x, A = 1, z, M = 0] \right)
\]
\[
\cdot \left( \mathbb{E}[Y|x, a, z, U = 1, m] - \mathbb{E}[Y|x, a, z, U = 0, m] \right) \\
= \alpha_{\text{comp}} \beta_{\text{net}}. 
\]

C  Results on Percentile Ranks of Earnings

Table C1 and Figures C1-C5 report results when the outcome is measured by the percentile rank of earnings, paralleling Table 1 and Figures 3-7 in the main text. We can see that the two sets of results are broadly consistent.

It should be noted that because a person’s earnings rank is a function of both her own earnings and the earnings of everyone else, it depends not only on her own education but also everyone
Table C1: Decomposition of the Average Total Effect (ATE) of College Attendance on Earnings Rank.

<table>
<thead>
<tr>
<th>Total Effect ($\Delta_{tot}$)</th>
<th>Direct Effect ($\Delta_{att}$)</th>
<th>Indirect Effect ($\Delta_{ind}$)</th>
<th>Completion Prob. ($\pi_{comp}$)</th>
<th>Completion Effect ($\Delta_{comp}$)</th>
<th>Covariance Term ($\Delta_{cov}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.72 (0.95)</td>
<td>3.80 (1.14)</td>
<td>5.92 (0.70)</td>
<td>0.57 (0.01)</td>
<td>10.70 (1.18)</td>
<td>-0.18 (0.39)</td>
</tr>
</tbody>
</table>

Note: Numbers in parentheses are estimates of standard errors constructed using sample variances of the estimated efficient influence functions and adjusted for multiple imputation via Rubin’s (1987) method.

else’s. Thus the stable unit treatment value assumption (SUTVA; Rubin 1986) is mechanically violated. Nonetheless, we can still interpret our estimates of the causal effects of college (Table C1 and Figures C1, C3, C5) and the potential earnings disparities (Figures C2 and C4) from the perspective of local interventions, i.e., interventions applied to a small random sample of potential college goers (Lundberg 2020).
Figure C2: Estimates of Potential Earnings Disparities as Linear Functions of Parental Income Rank. Note: Ribbons represent 95% confidence intervals.
Figure C3: Estimated Causal Effects as Linear Functions of the Propensity Score.  
Note: Ribbons represent 95% confidence intervals.
Figure C4: Estimates of Potential Earnings Disparities as Linear Functions of the Propensity Score. Note: Ribbons represent 95% confidence intervals.
Figure C5: Estimates of the Total Effect and Its Components by Propensity Score Quintile. Note: Ribbons represent 95% confidence intervals.