

Where Do the Returns to Schooling Come From? Educational Transitions and Labor Market Payoffs

Aleksei Opacic*
Harvard University

July 21, 2026

Abstract

Conventional research on educational effects typically either employs a “years of schooling” measure of education, or dichotomizes attainment as a point-in-time treatment. Yet, such a conceptualization of education is misaligned with the sequential process by which individuals make educational transitions. In this paper, I propose a causal mediation framework for the study of educational effects on outcomes such as earnings. The framework considers the effect of a given educational transition as operating indirectly, via progression through subsequent transitions, as well as directly, net of these transitions. I demonstrate that the average treatment effect (ATE) of education can be additively decomposed into mutually exclusive components that capture these direct and indirect effects. The decomposition has several special properties which distinguish it from conventional mediation decompositions of the ATE, properties which facilitate less restrictive identification assumptions as well as identification of all causal paths in the decomposition. An analysis of the returns to high school completion in the NLSY97 cohort suggests that the payoff to a high school degree stems overwhelmingly from its direct labor market returns. Mediation via college attendance, completion and graduate school attendance is small because of individuals’ low counterfactual progression rates through these subsequent transitions.

Keywords: causal inference, mediation, sequential ignorability, education

*aopacic@g.harvard.edu. Many thanks to Clem Aeppli, Kosuke Imai, Ian Lundberg, Michael Zanger-Tishler, Yi Zhang, and Xiang Zhou, as well as members of the Harvard C.A.R.E.S. lab and a reviewer from the Alexander and Diviya Magaro Peer Pre-Review program, for helpful feedback and conversations. Many thanks also to the three anonymous referees and the associate editor for their invaluable suggestions.

1 Introduction

One of the most resilient social scientific findings across a range of national contexts is the strong association between educational attainment and a variety of life outcomes, including earnings, health, social capital, and family stability (Hout, 2012; Chetty et al., 2023). Conventionally, researchers have taken one of two approaches to evaluating the social and economic returns to education: the first employs a “years of schooling” measure of educational attainment (Angrist and Krueger, 1991, 1992; Kane and Rouse, 1993; Card, 1994; Ashenfelter and Zimmerman, 1997; Card, 1999, 2001; Angrist and Chen, 2011), while the second dichotomizes attainment as a point-in-time treatment. This latter approach has been especially influential in the study of the impact of postsecondary attainment on earnings, where the treatment considered is often an indicator for whether an individual has attended, or graduated from, college (Brand and Xie, 2010; Carneiro et al., 2011; Zimmerman, 2014; Goodman et al., 2017; Smith et al., 2020; Bleemer, 2022; Mountjoy, 2022).

Despite the important insights this literature has made into establishing the causal effect of educational attainment on important social and economic outcomes, extant work has been inattentive to the sequential process by which people make educational transitions (Mare, 1980).¹ At the end of high school, individuals decide whether or not to enroll in college. Among college enrollees, only 60% receive a BA within six years of initial college entry (Snyder et al., 2016), with an even lower proportion for low-income students and students of color (Eller and DiPrete, 2018; Zhou and Pan, 2023). Moreover, amidst higher educational expansion in the US, college graduates must increasingly choose whether to enter the labor market or to enroll in postgraduate education. Increasingly, therefore, educational attainment in the US has become a field of multiple levels with sequential transitions, all of which are independently consequential for individuals’ labor market outcomes, and therefore of independent scientific interest.

¹I use the term “educational transition” to refer both to vertical transitions (e.g. enrollment at a secondary or tertiary institution), as well as to the attainment of a qualification at a given level (e.g. high school graduation or BA completion).

The sequential nature of educational transitions implies that a causal mediation framework can be employed to study the causal paths by which education’s “value-added” occurs. Specifically, we can consider the first transition in a sequence of educational levels of interest as a treatment variable, A , and subsequent transitions as mediators that “transmit” the effects of the treatment and of prior transitions, M_k ($1 \leq k \leq K$). For example, if we are interested in the total effect of high school completion on earnings, we may ask to what extent this total effect operates indirectly, through the effects of college attendance and college completion (putative mediators) on earnings, or directly, through alternative causal pathways. The insight that the total causal effect of education can be decomposed into its direct and indirect effects opens up a range of important research and policy-oriented questions. For example, tracing to what extent an early-stage educational intervention boosts outcomes such as earnings via its promotion of subsequent educational attainment (its indirect effects), or via earnings directly, would enable policy-makers to discern what drives the intervention’s value and to hone subsequent policy (e.g. Hurwitz and Howell, 2014; Sullivan et al., 2019; Castleman et al., 2020; Bird et al., 2021; Dynarski et al., 2021; Black et al., 2023; Turner and Gurantz, 2024). Relatedly, if the early intervention’s effects are heterogeneous across demographic groups, assessing the intervention’s direct and indirect effects could guide researchers to aspects of the educational experience that either promote or inhibit upward mobility. Nevertheless, prior empirical approaches are not well-suited to answering these questions: a “years-of-schooling” approach captures the direct effect of each additional year of schooling, while the dichotomous approach conflates the direct and indirect effects.²

In this article, I introduce a causal mediation framework for analyzing the effects of educational transitions. For the setting of K (≥ 1) monotonic binary or categorical media-

²A further strand of literature, especially prominent in labor economics, explores labor market returns to horizontal aspects of differentiation within a given educational level (e.g. college selectivity, as well as specific colleges) or college types (e.g. Cohodes and Goodman, 2014; Goodman et al., 2017; Mountjoy and Hickman, 2021; Chetty et al., 2023; Eller, 2023). While my proposed framework prioritizes the effects of different levels of education, I discuss in the conclusion how the framework could be extended to accommodate multivariate mediators.

tors, I develop a general formula that decomposes the total effect of any level of education into $K + 1$ monotonic path-specific effects (MPSEs): a direct effect net of K subsequent educational transitions, reflecting the path $A \rightarrow Y$, and K mutually exclusive “continuation” or gross effects, reflecting the paths $A \rightarrow M_1 \rightarrow Y$, $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$, and $A \rightarrow M_1 \cdots \rightarrow M_K \rightarrow Y$. Most importantly, this decomposition exploits a unique characteristic of this empirical setting, in which mediators are characterized by “monotonicity”: that is, where an individual’s potential $k + 1$ mediator value is deterministically zero if that individual’s k th mediator value is 0. The resultant decomposition of the ATE into $K + 1$ monotonic path-specific effects (MPSEs) can be non-parametrically identified under the assumption of sequential ignorability, which allows for the effect of each educational level to be confounded by a distinct set of (observed) intermediate covariates. I introduce several estimation strategies for my proposed decomposition, including a simple linear model-based regression-with-residuals (RWR) procedure, and a non-parametric estimation strategy based on the efficient influence functions (EIFs) of the target parameters (see Chernozhukov et al., 2017; Kennedy, 2022).

This study makes three main contributions. Within the realm of education research, the proposed decomposition draws on important work by Heckman et al. (2018), who present a similar decomposition of the effect of schooling over the early life course, but differs in two important respects. First, I provide nonparametric definitions, identification results, and estimation strategies for decomposing the total effect of schooling through its direct and indirect components. Second, my decomposition accommodates the presence of a distinct set of observed intermediate confounders for each transition. While one limitation of my approach is that I assume away the presence of *unobserved* confounders for each transition, I propose a sensitivity analysis that assesses the robustness of the results to unobserved confounding, under a set of simplifying assumptions.

More broadly, my framework speaks to the burgeoning field of causal mediation analysis in the social, economic, and health sciences, targeted at assessing the causal pathways by

which a treatment affects an outcome. While prior literature overwhelmingly focuses on single-mediator decompositions of the ATE, a growing body of work examines mediation estimands in settings with multiple mediators (Avin et al., 2005; Albert and Nelson, 2011; VanderWeele et al., 2014; Lin and VanderWeele, 2017; Miles et al., 2017; Steen et al., 2017; Vansteelandt and Daniel, 2017; Miles et al., 2020). In particular, in the case of two causally ordered mediators, Daniel et al. (2015) show that the ATE can be decomposed into multiple path-specific effects (PSEs), and outline the assumptions under which some of these effects are identified. Most recently, Zhou (2022b) generalized this framework to the case of K mediators, establishing a set of identifiable PSEs and introducing several regression-based, weighting, and semiparametric efficient estimators. I extend this literature by examining a special empirical setting where the mediators are monotonic. Compared with traditional mediation-based decompositions, monotonicity facilitates PSE identification under weaker identification assumptions, enables identification of all of the causal paths in question, as opposed to just a strict subset of them, and further permits a finer-grained decomposition. The general decomposition also extends previous literature on mediation under monotonicity which has focused exclusively on the case of a single mediator (e.g. Zhou, 2022a).

Finally, I also contribute to a growing parallel literature that proposes a range of non-parametric, and semi-parametric efficient estimators for alternative mediation estimands, based on the efficient influence functions (EIFs) of the causal quantities of interest (e.g. Miles et al., 2020; Farbmacher et al., 2022; Zhou, 2022b), as well as to closely-related work that proposes semi-parametric efficient estimators for dynamic treatment effects (Lewis and Syrgkanis, 2020; Viviano and Bradic, 2021; Bodory et al., 2022).

In the following sections, I first introduce the decomposition for the case of a single intermediate educational transition, before discussing the general case of K intermediate transitions and its identification under the assumption of sequential ignorability (Section 2). In Section 3, I introduce a semiparametric estimation strategy for estimating the proposed decomposition, and in Section 5, I illustrate the proposed framework and methods

using data from the National Longitudinal Survey of Youth (NLSY97) cohort. Section 6 concludes.

2 Monotonic Path-Specific Effects

2.1 A Single Intermediate Transition

I first consider the case of a single intermediate educational transition (monotonic mediator). Suppressing subscripts i , let A denote an indicator for high school graduation (the initial educational transition), M_1 , an indicator for college attendance (a monotonic mediator or transition), and Y , a binary or continuous outcome of interest such as earnings. A single-transition decomposition thus assesses the educational sequence $A \rightarrow M \rightarrow Y$: high school graduation $\rightarrow$ college attendance $\rightarrow$ earnings. In this way, I treat college attendance as a mediator of the total effect of high school graduation on earnings, in relation to which the total effect of high school graduation can be decomposed into an indirect effect (that “flows through” college attendance), and a direct effect (net of college attendance). Following Heckman et al. (2018), I refer to this latter term as the “continuation” value of educational transition A .

Using potential outcomes notation, let $M(a)$ denote an individual’s potential value of the mediator if their treatment status were set to a , and let $Y(a, m)$ denote that individual’s potential outcome if their treatment and mediator statuses were set to a and m , respectively. With our set of potential outcomes ($\{Y(1), Y(0), Y(1, 0), Y(1, 1), Y(0, 1)\}$), we can then define potential outcomes that involve “natural” values of the mediator. For instance, $Y(1, M(0))$ represents the earnings an individual would have if they completed high school ($A = 1$) but their college attendance were fixed to the level it would have been had they not completed high school ($A = 0$).

With this notation, we can define the “natural” decomposition of the average treatment effect as follows (Pearl, 2001; Imai et al., 2010):

$$\begin{aligned}
\text{ATE} &= \mathbb{E}[Y(1) - Y(0)] \\
&= \text{NDE}(a) + \text{NIE}(1 - a),
\end{aligned} \tag{1}$$

for $a = 0, 1$, where $\text{NDE}(a) = \mathbb{E}[Y(1, M(a)) - Y(0, M(a))]$, and $\text{NIE}(a) = \mathbb{E}[Y(a, M(1)) - Y(a, M(0))]$. This decomposition allows a researcher to determine how much of the overall effect of high school completion is due to facilitating access to college, versus directly, net of college attendance.

Potential outcomes such as $Y(1, M(0))$ are central to mediation, but they create challenges in estimation and interpretation due to their “cross-world” nature. In particular, a unit cannot simultaneously complete high school and not complete high school at the same time, so $Y(1, M(0))$ is not observable, even in principle. This creates challenges for identification. In particular, it requires a “cross-world independence” assumption in order to identify the natural effect decomposition of the ATE: $Y(a', m) \perp\!\!\!\perp M(a) \mid A = a, X$.

Many authors have expressed skepticism about this assumption since it requires that there are no (measured *or* unmeasured) post-treatment confounders (confounders that are affected by the treatment and which affect the mediator and outcome). An alternative approach to mediation analysis involves estimating a quantity known as the controlled direct effect (CDE), which captures the causal effect of a treatment when the mediator is fixed to a given value:

$$\text{CDE}(m) = \mathbb{E}[Y(1, m) - Y(0, m)],$$

for $m = 0, 1$. This quantity is attractive because it is identified under a weaker assumption than that required for quantities like $Y(1, M(0))$, one that allows for the existence of observed post-treatment confounders. A drawback of this approach is that it does not quantify the extent of mediation through M and only enables a researcher to rule out the

existence of alternative mediators other than M (see Acharya et al. (2016)).

In the context of educational effects, it is possible to relax the usual cross-world assumption by leveraging a key structural feature of educational transitions. As I formalize in the following section, I assume that educational transitions are characterized by monotonicity. Here, this means that individuals who do not complete high school cannot attend college, or $M(0) = 0$. This sequential nature of educational transitions therefore implies the following *restricted* set of potential outcomes: $\{Y(1), Y(0), Y(1, 0), Y(1, 1)\}$ (i.e., ruling out $Y(0, 1)$). Further, since by the composition assumption $Y(a) = Y(a, M(a))$ (VanderWeele and Vansteelandt, 2009), under monotonicity $Y(0) = Y(0, M(0)) = Y(0, 0)$.

We can then apply these restrictions to Equation 1 as follows (see Zhou, 2022a):

$$\begin{aligned}
\text{ATE} &= \mathbb{E}[Y(1) - Y(0)] \\
&= \underbrace{\mathbb{E}[Y(1, M(0)) - Y(0, M(0))]}_{=\text{NDE}(0)} + \underbrace{\mathbb{E}[Y(1, M(1)) - Y(1, M(0))]}_{=\text{NIE}(1)} \\
&= \mathbb{E}[Y(1, 0) - Y(0, 0)] + \mathbb{E}[Y(1, M(1)) - Y(1, 0)] \\
&= \mathbb{E}[Y(1, 0) - Y(0, 0)] + \mathbb{E}[M(1)[Y(1, 1) - Y(1, 0)]] \\
&= \mathbb{E}[Y(1, 0) - Y(0, 0)] + \mathbb{E}[M(1)]\mathbb{E}[Y(1, 1) - Y(1, 0)] + \text{cov}[M(1), Y(1, 1) - Y(1, 0)]
\end{aligned} \tag{2}$$

$$= \underbrace{\Delta_0}_{A \rightarrow Y} + \underbrace{\pi_1 \Delta_1 + \eta_1}_{A \rightarrow M \rightarrow Y}, \tag{3}$$

where the third equality follows by monotonicity, the fourth, because $Y(1, M(1)) - Y(1, 0) = M(1)[Y(1, 1) - Y(1, 0)] + [1 - M(1)] \cdot [Y(1, 0) - Y(1, 0)] = M(1)[Y(1, 1) - Y(1, 0)]$, and the fifth, by rules of covariance. Here, Δ_0 and Δ_1 denote the direct effects of the first and intermediate transitions on the outcome, $A \rightarrow Y$ and $M \rightarrow Y$, respectively, π_1 denotes the total effect of the first transition on the intermediate transition $A \rightarrow M$, and η_1 denotes the covariance between the effect of the initial transition on completion of the second and

the effect of the second transition on Y . Specifically, η_1 is positive if those who would attend college given high school completion (i.e., $M(1) = 1$) benefit more from college attendance in terms of their later earnings (i.e., have a larger $Y(1, 1) - Y(1, 0)$) than those who do not (i.e., $M(1) = 0$), and negative if the opposite is true. Meanwhile, the composite term $(\pi_1 \Delta_1 + \eta_1)$ captures the average indirect effect of the treatment via the intermediate transition ($A \rightarrow M \rightarrow Y$), comprising the sum of (i) the probability of college enrollment if an individual graduated high school, multiplied by the direct effect of college enrollment, and (ii) the covariance between college enrollment and its direct effect on earnings.

While motivated by the natural effect decomposition of the ATE, the monotonicity constraint on M leads to several differences from Equation 1. First, whereas Equation 1 can be written for $a = 0$ and $a = 1$, Equation 3 is the algebraically unique natural effect decomposition of the ATE under monotonicity. This is because the natural effect decomposition for $a = 0$ hinges on the counterfactual term $Y(1, M(0)) = Y(1, 0)$, whereas the natural effect decomposition for $a = 1$ hinges on the counterfactual term $Y(0, M(1))$. Under monotonicity, only the first of these quantities is well-defined. Second, under mediator monotonicity, $\text{NDE}(0) = \text{CDE}(0)$. Therefore, while in general the CDE does not quantify the extent of mediation through M (since the difference between the ATE and CDE indicates the portions of the total effect due to interaction without mediation, and due to mediation, both with and without interaction (VanderWeele, 2014)) in this special case the CDE completely characterizes the extent of mediation.³ Third, and relatedly, because $Y(1, M(0)) = Y(1, 0)$, Equation 3 does not depend on any cross-world counterfactuals, and therefore can be identified under weaker assumptions than Equation 1.

Finally, the decomposition in Equation 3 can be compared with a “randomized effect”

³An alternative way to see this is that, as VanderWeele (2014) shows, the individual-level natural direct effect can be written as $(Y(1, 0) - Y(0, 0) + M(0)(Y(1, 1) - Y(0, 1) - Y(1, 0) + Y(0, 0)))$, but under monotonicity, $M(0) = 0$ for all individuals, and so the individual-level natural direct effect reduces to the first term. Interestingly, under monotonicity there is no “interaction effect” between the treatment and mediator (since the mediator can only take a value of 1 when the treatment is activated), and so the NIE reduces to pure indirect effect (PIE) discussed in VanderWeele (2014). Hence, under monotonicity, $\text{ATE} = \text{CDE} + \text{PIE}$, and there are no interaction terms in the decomposition.

decomposition of treatment effects. An alternative approach to avoiding cross-world counterfactuals and treatment-induced confounding is to use *randomized intervention analogues* to the natural direct and indirect effects (rNDE and rNIE) (VanderWeele et al., 2014). Rather than setting the mediator to the level it would “naturally” take under an alternative treatment level, these estimands conceptualize interventions that randomly draw the mediator from its population distribution under each treatment regime.

Analogous to the natural effect decomposition in Equation 1, we can define a *randomized effect decomposition* of a treatment effect as follows:

$$\text{rATE} = \text{rNDE}(a) + \text{rNIE}(1 - a), \quad (4)$$

for $a = 0, 1$, where $\text{rNDE}(a) = \mathbb{E}[Y(1, G_{a|X}) - Y(0, G_{a|X})]$ and $\text{rNIE}(a) = \mathbb{E}[Y(a, G_{1|X}) - Y(a, G_{0|X})]$, and rATE is defined to be the sum of these two components. Here, $G_{a|X} \equiv \Pr(M = m|A = a, x)$ denotes a value of the mediator randomly drawn from its conditional distribution under treatment $A = a$ given baseline covariates X . Importantly, Equation 4 decomposes not the ATE but a randomized average treatment effect (rATE), which may differ from the ATE. By doing so, it avoids the cross-world counterfactual $Y(a, M(a'))$ by substituting a randomized intervention on the mediator distribution. This substitution identifies the indirect and direct pathways, even when post-treatment confounders of the mediator–outcome relationship are present.

When $a = 0$, the rNDE captures exactly the NDE under a monotonicity constraint (i.e., $\text{rNDE}(0) = \Delta_0$). This is because $\Pr(M = 1|A = 0, x) = 0$ and $\Pr(M = 0|A = 0, x) = 1$ and so the randomized effect reduces to the CDE. Using results from Yu et al. (2024), the difference between the NIE under monotonicity and the rNIE is equal to

$$\begin{aligned} \pi_1 \Delta_1 + \eta_1 - \text{rNIE} &= \mathbb{E}[\text{cov}[M(1) - M(0), Y(1, 1) - Y(1, 0)|X]] \\ &= \mathbb{E}[\text{cov}[M(1), Y(1, 1) - Y(1, 0)|X]] \text{ (since } M(0) = 0). \end{aligned}$$

This expression highlights that the NIE–rNIE gap measures residual within- X -stratum

covariance between the mediator and its causal effect on Y , i.e., dependence induced by intermediate confounding. Thus, deviations between the ATE and rATE will stem from this within-stratum covariance. The randomized intervention framework parallels the monotonic decomposition in Equation 3, which eliminates logically impossible counterfactuals (e.g., $Y(0, M(1))$) through structural constraints on the mediator rather than redefining the estimand itself.

2.2 Generalization to K Intermediate Transitions

I now generalize the approach introduced in the preceding section to the case of K intermediate transitions. As previously, I denote the treatment (“initial transition”) of high school graduation by A , and use $M_1, \dots, M_K$ to refer to the K subsequent transitions of interest (“intermediate transitions”), where I assume that all of $M_1, \dots, M_K$ are binary and that for any $i < j$, M_i temporally precedes M_j . For instance, we may wish to decompose the total effect of high school completion on earnings via college attendance (M_1), college completion (M_2) and graduate school attendance (M_3). Let an overbar denote a vector of variables, such that $\overline{M}_k = (M_1, M_2, \dots, M_k)$ and $(1, \overline{1}_{k-1}) = (A = 1, M_1 = 1, \dots, M_{k-1} = 1)$. Further, let $[K]$ denote the set $\{0, 1, \dots, K\}$. In addition, I denote by X a vector of pretreatment confounders of the effect of $(A, \overline{M}_k)$ on (M_{k+1}, Y) , and by $\overline{Z}_k = (Z_1, \dots, Z_k)$ a vector of intermediate confounders that may confound the causal effect of M_k on (M_{k+1}, Y) . Using potential outcomes notation, $Y(1, \overline{1}_{k-1}, m_k)$ thus denotes an individual’s potential earnings if they completed, possibly contrary to fact, the treatment and the first $k - 1$ intermediate transitions, and then either completed ($m_k = 1$) or did not complete ($m_k = 0$) the k th *intermediate* transition. Similarly, $M_{k+1}(1, \overline{1}_k)$ denotes an individual’s potential value of the $k + 1$ th intermediate transition were that individual to complete the treatment and the first k intermediate transitions. As is standard in the mediation literature, I make the following composition assumption (VanderWeele and Vansteelandt, 2009):

Assumption 1. Composition: $Y(1, \overline{1}_{k-1}, m_k) = Y(1, \overline{1}_{k-1}, m_k, M_{k+1}(1, \overline{1}_{k-1}, m_k)), \forall k \in [K -$

1], $M_0 \equiv A$.

In words, Assumption 1 states that a person’s potential outcome under $(1, \bar{1}_{k-1}, m_k)$ is equal to their potential outcome under $A = 1, \dots, M_{k-1} = 1, m_k$ and under the value M_{k+1} would naturally take under $A = 1, \dots, M_{k-1} = 1, m_k$. I also invoke the following constraint on units’ potential transition values:

Assumption 2. Monotonicity: $M_{k+1}(M_k = 0) = 0 \forall k \in [K - 1], M_0 \equiv A$.

Informally, Assumption 2 (*monotonicity*) states that an individual’s potential $k + 1$ th transition value is deterministically 0 if that individual fails to complete the prior (k th) transition. It is analogous to a one-sided non-compliance assumption within an instrumental variables (IV) framework, which precludes the presence of both “defiers” as well as “always-takers” principal strata. We can then use this assumption to decompose the ATE of A on Y , which I denote by τ_0 . Specifically, let τ_k denote the gross effect of the k th mediator on Y , i.e.,

$$\tau_k = \mathbb{E}[Y(1, \bar{1}_k) - Y(1, \bar{1}_{k-1}, 0)],$$

let Δ_0 denote the direct effect of A on Y , and let Δ_k denote the direct effect of the k th mediator on Y , i.e.,

$$\Delta_k = \mathbb{E}[Y(1, \bar{1}_{k-1}, 1, 0) - Y(1, \bar{1}_{k-1}, 0)].$$

To explicate my approach, note that the gross effect of the k th mediator, τ_k , includes not only the direct effect $M_k \rightarrow Y$, net of subsequent educational transitions Δ_k , but also the indirect effects of M_k via subsequent transitions ($M \rightsquigarrow Y$, where a squiggly arrow denotes a combination of multiple paths). This insight motivates us to further decompose τ_k into its direct and indirect components. Under the composition assumption, τ_k can be

decomposed as

$$\tau_k = \Delta_k + \pi_{k+1}\tau_{k+1} + \eta_{k+1}, \quad (5)$$

where

$$\begin{aligned} \pi_{k+1} &= \mathbb{E}[M_{k+1}(1, \bar{1}_k)], \\ \eta_{k+1} &= \text{cov}[M_{k+1}(1, \bar{1}_k), Y(1, \bar{1}_{k+1}) - Y(1, \bar{1}_k, 0)]. \end{aligned}$$

For $k = 1, \dots, K-1$, iteratively substituting equation 5 into the corresponding expression for τ_{k-1} yields

$$\tau_0 = \underbrace{\Delta_0}_{A \rightarrow Y} + \sum_{k=1}^K \underbrace{(\Pi_{j=1}^k \pi_j) \Delta_k + (\Pi_{j=1}^{k-1} \pi_j) \eta_k}_{\theta_k \triangleq A \rightarrow M_1 \dots \rightarrow M_k \rightarrow Y}, \quad (6)$$

where $\pi_0 = 1$. Further, $\Delta_K = \tau_K$ is a *gross* or continuation effect, since this latter path is a composite one that contains all residual paths omitted in the decomposition (i.e., through educational transitions subsequent to K , if they exist). Thus, the θ_k terms capture how much of the total effect of high school completion flows through each intermediate transition considered (i.e., via college attendance, via college completion, and via graduate school attendance), while Δ_0 captures that portion of the total effect that operates directly, net of the K intermediate transitions considered.

2.3 A Comparison with Conventional Mediation Analysis with Multiple Causally Ordered Mediators

The decomposition introduced in the previous section has an analog in the context of a mediation-based decomposition of the ATE with multiple ordered mediators, but differs from these conventional mediation analyses in important ways. To illustrate the differences,

consider a binary treatment, A , an outcome of interest, Y , and a vector of pretreatment covariates, X , and let $M_1, M_2, \dots, M_K$ denote K causally ordered mediators, assuming that for any $i < j$, M_i precedes M_j , as above. Moreover, define $M_0 \equiv \emptyset$, and let

$$\overline{M}_k \equiv (M_1, M_2, \dots, M_k), \quad \overline{M}_k(a) \equiv (M_1(a), M_2(a, M_1(a)), \dots, M_k(a, \overline{M}_{k-1}(a))),$$

with $\overline{M}_0(a) \equiv \emptyset$. Using the potential outcomes notation as above, define the following expectation of a nested counterfactual,

$$\psi(a, \overline{M}_k(a^*)) \triangleq \mathbb{E}[Y(a, \overline{M}_k(a^*))].$$

Under Pearl's (2009) nonparametric structural equation model (NPSEM), Zhou (2022b) demonstrates that the ATE of A on Y can be decomposed into $K + 1$ identifiable path-specific effects (PSEs) corresponding to each of the causal paths $A \rightarrow Y$ and $A \rightarrow M_k \rightsquigarrow Y$ ($k \in [K]$) (see also Daniel et al., 2015):

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)] = \underbrace{\psi(1, \overline{M}_K(0)) - \psi(0, \overline{M}_K(0))}_{A \rightarrow Y} + \sum_{k=1}^K \underbrace{(\psi(1, \overline{M}_k(0)) - \psi(1, \overline{M}_{k-1}(0)))}_{A \rightarrow M_k \rightsquigarrow Y}. \quad (7)$$

This decomposition holds algebraically when Assumption 2 does not hold (i.e., when the mediators are not monotonic), and differs from the proposed decomposition (Equation 6) in several ways.⁴

⁴Vansteelandt and Daniel (2017) propose a decomposition of the total effect in settings with multiple, potentially causally ordered mediators using interventional (in)direct effects. A key advantage of this approach is that, like the MPSE decomposition, it permits exposure-induced intermediate confounding by defining effects in terms of stochastic interventions on mediator distributions. When mediators are causally dependent, however, mediation through a given mediator is not, in general, fully summarized by the corresponding interventional indirect effect alone; an additional component capturing exposure-induced changes in the dependence between mediators (their Equation (8)) is required to complete the decomposition. Accordingly, in settings with causally dependent mediators, mediation along a particular sequential path may be zero even when the associated interventional indirect effect is nonzero. By contrast, the framework proposed here exploits monotonicity to restrict the support of downstream mediators to that of upstream transitions, so that all mediation operates along well-defined sequential paths and no separate dependence term is required. Interestingly, under monotonicity, the interventional direct effect in

First and most importantly, the monotonic decomposition departs from conventional mediation decompositions in distinguishing a different set of causal pathways. As shown in the previous section, the monotonic and standard mediation decompositions both begin with a natural effect decomposition in the case of a single mediator. Conventional mediation decompositions then obtain a 2-mediator decomposition by decomposing the direct path $A \rightarrow Y$ net of M_1 into its direct component net of M_2 and indirect path $A \rightarrow M_2 \rightarrow Y$. This is shown in Figure 1, Panel A. Notably, to assess the mediating role of M_1 , only the composite path $A \rightarrow M_1 \rightsquigarrow Y = (A \rightarrow M_1 \rightarrow Y) + (A \rightarrow M_1 \rightarrow M_2 \rightarrow Y)$ is identified. The reason for this is that identification of the pure path-specific effects $A \rightarrow M_1 \rightarrow Y$ or $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$ fails when the *recanting witness criterion* is violated (Avin et al., 2005). Specifically, M_1 acts as a “recanting witness” because it lies on the path $A \rightarrow M_1 \rightarrow Y$ but also has an additional path to Y through M_2 ($M_1 \rightarrow M_2 \rightarrow Y$) that is not contained in the path of interest. The nested counterfactual $\mathbb{E}[Y(a, M_1(a_1), M_2(a_2, M_1(a_{12}))))]$ is identified if and only if $a_1 = a_{12}$. Consequently, the individual PSEs for $A \rightarrow M_1 \rightarrow Y$ and $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$ are not identified, whereas the composite path $A \rightarrow M_1 \rightsquigarrow Y$, which includes all possible paths from M_1 to Y , remains identifiable. More generally, traditional decompositions do not disentangle the mediating effects of M_k that are direct (net of subsequent mediators) and indirect (through different combinations of subsequent mediators).

By contrast, as shown in Figure 1, Panel B, my proposed decomposition under mediator monotonicity begins with the natural effect decomposition with a single mediator as in Equation 1, and then decomposes the path via M_1 into pathways that operate further via M_2 ($A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$) and that operate directly to Y ($A \rightarrow M_1 \rightarrow Y$), leaving the direct effect (Δ_0) untouched. In fact, Δ_0 is not further decomposable under monotonicity because this assumption yields a more restricted set of causal paths. To see why, consider

Vansteelandt and Daniel (2017) coincides exactly with Δ_0 in the MPSE decomposition, echoing the result that the randomized natural direct effect at zero equals Δ_0 , as discussed in Section 2.1.

the pathway $A \rightarrow M_2 \rightarrow Y$ under monotonicity in the two-mediator setting:

$$\mathbb{E}[Y(1, M_1(0), M_2(1, M_1(0))) - Y(1, M_1(0), M_2(0, M_1(0)))] .$$

Under mediator monotonicity, $M_1(0) = 0$ (e.g., without completing high school one cannot attend college), and $M_2(\cdot, 0) = 0$ (e.g., without college attendance one cannot complete a BA). Therefore

$$M_2(0, M_1(0)) = M_2(0, 0) = 0.$$

and the path $A \rightarrow M_2 \rightarrow Y$ equals

$$Y(1, 0, 0) - Y(0, 0, 0) = 0.$$

More generally, consider the path-specific effect attributed to the pathway $A \rightarrow M_k \rightarrow Y$ with $k \geq 2$, defined in the usual way by fixing all earlier mediators at their natural levels under $A = 1$ and varying only M_k with respect to A :

$$\tau_{A \rightarrow M_k \rightarrow Y}(a) = \mathbb{E}[Y(1, \overline{M}_{k-1}(0), M_k(1, \overline{M}_{k-1}(0))) - Y(1, \overline{M}_{k-1}(0), M_k(0, \overline{M}_{k-1}(0)))] ,$$

where $\overline{M}_k(a)$ denotes the vector $(M_1(a), \dots, M_k(a), \overline{M}_{k-1}(a))$. Under monotonicity, fixing earlier mediators at their baseline levels implies that $\overline{M}_k(0) = 0$, and that $M_k(1, \overline{M}_{k-1}(0)) = M_k(0, \overline{M}_{k-1}(0)) = 0$. In words, there is no direct effect of the treatment on later mediators once earlier transitions are held fixed at their natural value under $A = 0$. As a result, the two potential outcomes inside the expectation coincide, and

$$\tau_{A \rightarrow M_k \rightarrow Y}(1) = 0 \quad \text{for all } k \geq 2.$$

Any indirect effect involving M_k must therefore operate through the full causal chain $A \rightarrow M_1 \rightarrow \dots \rightarrow M_k \rightarrow Y$. Figure 2 illustrates the causal pathways defined and identified

under the proposed decomposition in the case of two monotonic mediators.

Second, the PSE decomposition of the ATE in general mediation settings is not algebraically unique, and thus the PSEs defined under alternative decompositions will differ if the effects of the treatment and each mediator vary across levels of the other mediators. In fact, depending on the order in which the paths $A \rightarrow Y$ and $A \rightarrow M_k \rightsquigarrow Y$ are considered, there are $(K + 1)!$ identifiable different ways of decomposing the ATE; the decomposition shown in Equation 7 is just one such decomposition. Consider the case of two causally dependent mediators. In this setting, the causal pathway $A \rightarrow M_2 \rightsquigarrow Y$ can be defined with respect to four different combinations of levels of the treatment and first mediator: under (i) $a = 1$ and $M_1(1)$, (ii) $a = 1$ and $M_1(0)$, (iii) $a = 0$ and $M_1(1)$, or (iv) $a = 0$ and $M_1(0)$. By contrast, as a direct consequence of monotonicity, the MPSE decomposition is the unique PSE decomposition of the ATE.

Finally, the sequential ignorability assumption required to identify the MPSE decomposition is weaker than those required to identify a generic PSE decomposition of the ATE. Specifically, the latter requires Pearl’s (2009) non-parametric structural equation model (NPSEM), which stipulates that $(M_{k+1}(a_{k+1}, \bar{m}_k), \dots, M_K(a_K, \bar{m}_{K-1}), Y(a_{K+1}, \bar{m}_K)) \perp\!\!\!\perp M_k(a_k, \bar{m}_{k-1}^*) \mid X, A, \bar{M}_{k-1}, \forall k \in [K]$. This assumption, sometimes referred to as the “cross-world” independence assumption, is stronger than the sequential ignorability assumption (4) required to identify the MPSE decomposition since it rules out the existence of confounders of the mediators, be they observed or unobserved. By contrast, the MPSE decomposition identification results accommodate *observed* intermediate confounding without altering the substance of the decomposition. I discuss this point in detail in the following section.

2.4 Identification

To identify the causal effects of interest, I rely on a series of sequential ignorability assumptions. While most closely associated with the dynamic treatment effects literature,

which relies on observing a complete set of time-varying confounders in order to identify longitudinal effects (see e.g. Lewis and Syrgkanis, 2020; Viviano and Bradic, 2021; Bodory et al., 2022), these assumptions can be transferred to a mediation context, given the fact that the mediators of interest are all causally ordered. As discussed in the previous section, sequential ignorability identification assumptions are distinct from – and in fact weaker than – the assumptions typically employed in studies of causal mediation.

Let $M_k = \emptyset \forall k < 1$. In order to estimate the decomposition shown in Equation 6, it suffices to identify the expectation of two types of composite counterfactuals ($Y(1, \bar{1}_{k-1}, m_k)$ and $M_{k+1}(1, \bar{1}_k)$), as well as covariance terms of the form $\text{cov}[M_{k+1}(1, \bar{1}_k), Y(1, \bar{1}_{k+1}) - Y(1, \bar{1}_k, 0)] \forall k \in [K - 1]$. I invoke the following three assumptions:

Assumption 3. Consistency: for any unit, if $A = a$, $Y = Y(a)$; if $(A, \bar{M}_{k-1}, M_k) = (1, \bar{1}_{k-1}, m_k)$, then $Y = Y(1, \bar{1}_{k-1}, m_k) \forall k \in [K]$, and if $(A, \bar{M}_k) = (1, \bar{1}_k)$, then $M_{k+1} = M_{k+1}(1, \bar{1}_k) \forall m_{k+1} \in \{0, 1\}, \forall k \in [K - 1]$.

Assumption 4. Sequential ignorability: $(M_1(1), Y(a)) \perp\!\!\!\perp A|X$;
 $Y(1, \bar{1}_{k-1}, m_k) \perp\!\!\!\perp M_k|X, \bar{Z}_k, \bar{M}_{k-1} = \bar{1}_{k-1}$ and $M_{k+1}(1, \bar{1}_k) \perp\!\!\!\perp M_k|X, \bar{Z}_k, \bar{M}_{k-1} = \bar{1}_{k-1}$,
 $\forall m_k \in \{0, 1\}, \forall k \in \{1, \dots, K\}, M_0 \equiv A$.

Assumption 5. Positivity: $p_{A|X}(a|x) > \epsilon > 0$, $p_{M_k|X, A, \bar{Z}_k, \bar{M}_{k-1}}(m_k|x, a, \bar{z}_k, \bar{m}_{k-1} = \bar{1}_{k-1}) > \epsilon > 0 \forall k \in [K]$.

Assumption 3 (*consistency*) states that a unit's observed outcome equals its potential outcome under a given treatment sequence. Note that under Assumption 1 (Composition), if $Y = Y(1, \bar{1}_{k-1}, m_k)$, then $Y = Y(1, \bar{1}_{k-1}, m_k, M_{k+1}(1, \bar{1}_{k-1}, m_k)) = \dots = Y(1, \bar{1}_{k-1}, m_k, M_{k+1}(1, \bar{1}_{k-1}, m_k), \dots, M_K(\cdot))$. In plain words, the $K - k$ mediators after mediator k all take their natural levels. Assumption 4 (*sequential ignorability*) is the no unmeasured confounding assumption for the treatment and all mediators. It is considered plausible when sufficient pre-treatment and intermediate covariates ($X, \bar{Z}_K$) are collected. Finally, Assumption 5 (*positivity*) requires that treatment assignment is not deterministic, and that mediator assignment is not deterministic when the treatment and prior mediators

$\overline{m}_{k-1}$ take on a value of 1.

Under Assumptions 3–5, $\mathbb{E}[Y(1, \overline{1}_{k-1}, m_k)]$ and $\mathbb{E}[M_{k+1}(1, \overline{1}_k)]$ are identified, respectively, as

$$\mathbb{E}[Y(1, \overline{1}_{k-1}, m_k)] = \int_x \int_{\overline{z}_k} \mathbb{E}[Y|x, \overline{z}_k, 1, \overline{1}_{k-1}, m_k] \left[\prod_{j=1}^k dP(z_j|x, \overline{z}_{j-1}, \overline{1}_{j-1}) \right] dP(x) \quad (8)$$

$$\mathbb{E}[M_{k+1}(1, \overline{1}_k)] = \int_x \int_{\overline{z}_k} \mathbb{E}[M_{k+1}|x, \overline{z}_k, 1, \overline{1}_k] \left[\prod_{j=1}^k dP(z_j|x, \overline{z}_{j-1}, \overline{1}_{j-1}) \right] dP(x) \quad (9)$$

For a proof of the above formulas, see Robins (1986). The covariance (η_k) components in the decomposition are then identified as the “residual” terms such as in Equation 5, which follows directly from the fact that all other components in these equations are identified. Thus, for $k \in \{1, \dots, K\}$, we can identify η_k as

$$\eta_k = \tau_{k-1} - \Delta_{k-1} - \pi_k \tau_k. \quad (10)$$

3 Estimation

The identification results outlined above suggest that the proposed decomposition can be estimated via several approaches, including outcome-based modeling, models for the treatment and mediators via inverse probability weighting, as well as doubly robust approaches. This section outlines two complementary estimation strategies (one semiparametric approach and one parametric approach) for implementing the MPSE decomposition. After providing a short summary of the approaches, I detail a semiparametric estimation approach in the main text, and refer readers to Supplementary Material A for further detail on the parametric approach. I also provide a simulation study comparing the performance of the two estimation strategies in Supplementary Material B.

The first estimation approach is a semiparametric, debiased machine learning (DML)

estimator, which uses efficient influence functions and cross-fitting to estimate each component of the monotonic path-specific effect (MPSE) decomposition. The second is a simpler regression-with-residuals (RWR) estimator, which relies on parametric linear models and sequential residualization to translate the θ_k components in Equation 6 directly into regression coefficients that can be read off from these linear models. The RWR approach provides a transparent, low-computational alternative that directly links the decomposition to regression coefficients, making it useful both as a practical estimator in large datasets and as a parametric benchmark against which to compare the DML results. Nevertheless, when X and $\bar{Z}_K$ are high-dimensional, the parametric models underlying RWR may be misspecified, which can in turn introduce bias. By contrast, the DML approach is robust to flexible, high-dimensional models for the treatment, mediators, and outcome.

The DML approach is characterized by two components: first, the use of a Neyman orthogonal estimating equation based on the efficient influence function (EIF) for the target parameters, which makes estimates of the parameter “locally robust” to estimates of the nuisance functions; second, the use of a K -fold cross-fitting algorithm (Chernozhukov et al., 2017).

Let $O = (X, A, \bar{Z}_K, \bar{M}_K, Y)$ denote the observed data, and $\mathcal{P}$ a nonparametric model over O wherein all laws satisfy the positivity assumption described in Section 2. Before proceeding, I define the following auxiliary functions, as introduced in Section 2: $\psi_{km_k} \triangleq \mathbb{E}[Y(\bar{1}_k, m_k)]$ and $\phi_k \triangleq \mathbb{E}[M_{k+1}(\bar{1}_{k+1})]$, for all $k \in [K]$. Using the identification results given in Section 2.4, ψ_{km_k} can be written in terms of expectations of observed data:

$$\psi_{km_k} = \mathbb{E}_X \mathbb{E}_{Z_1|X,1} \dots \mathbb{E}_{Z_k|X,\bar{Z}_{k-1},1,\bar{1}_{k-1}} \mathbb{E}[Y \mid X, \bar{Z}_k, 1, \bar{1}_{k-1}, m_k]. \quad (11)$$

For each $j \in [k]$, we can thus define $\mu_{jm_k}^k(X, \bar{Z}_k)$ iteratively as

$$\begin{aligned}\mu_{km_k}^k(X, \bar{Z}_k) &\triangleq \mathbb{E}[Y \mid X, \bar{Z}_k, 1, \bar{1}_{k-1}, m_k], \\ \mu_{jm_k}^k(X, \bar{Z}_j) &\triangleq \mathbb{E}[\mu_{j+1m_k}^k(X, \bar{Z}_{j+1}) \mid X, \bar{Z}_j, 1, \bar{1}_j] \forall j \in [k-1].\end{aligned}$$

This recursive definition of $\mu_{jm_k}^k(X, \bar{Z}_k)$ is a compact way of expressing the nested expectations in Equation 11. For example, in the case of $k = 2$, the recursion becomes

$$\begin{aligned}\mu_{2m_2}^2(X, Z_1, Z_2) &= \mathbb{E}[Y \mid X, Z_1, Z_2, 1, \bar{1}_1, m_2], \\ \mu_{1m_2}^2(X, Z_1) &= \mathbb{E}_{Z_2 \mid X, Z_1, 1, \bar{1}_1}[\mu_{2m_2}^2(X, Z_1, Z_2)], \\ \mu_{0m_2}^2(X) &= \mathbb{E}_{Z_1 \mid X, 1}[\mu_{1m_2}^2(X, Z_1)].\end{aligned}$$

Thus the counterfactual mean is

$$\psi_{2m_2} = \mathbb{E}_X[\mu_{0m_2}^2(X)].$$

These expressions make explicit that, at each step, we take the expectation of the previous conditional expectation with respect to the distribution of the next confounder, Z_{j+1} given $(X, \bar{Z}_j, 1, \bar{1}_j)$.

Next, let $\pi_{km_k}(X, \bar{Z}_k) \triangleq \Pr[M_k = m_k \mid X, \bar{Z}_k, 1, \bar{1}_{k-1}] \forall k \in [K]$, and $\pi_{01}(X) \triangleq \Pr[A = 1 \mid X]$. The efficient influence function (EIF) of ψ_{km_k} is closely related to the EIF for the g-formula, and can be written as

$$\psi_{km_k}(O) = \sum_{j=0}^{k+1} \varphi_j(O), \tag{12}$$

where

$$\begin{aligned}
\varphi_0(O) &= \mu_{0m_k}^k(X) - \psi_{km_k} \\
\varphi_j(O) &= \frac{A}{\pi_{01}(X)} \left(\prod_{l=1}^{j-1} \frac{M_l}{\pi_{l1}(X, \bar{Z}_l)} \right) (\mu_{jm_k}^k(X, \bar{Z}_j) - \mu_{(j-1)m_k}^k(X, \bar{Z}_{j-1})), \quad j \in \{1, \dots, k\} \\
\varphi_{k+1}(O) &= \frac{A}{\pi_{01}(X)} \left(\frac{\mathbb{I}(M_k = m_k)}{\pi_{km_k}(X, \bar{Z}_k)} \prod_{l=1}^{k-1} \frac{M_l}{\pi_{l1}(X, \bar{Z}_l)} \right) (Y - \mu_{km_k}^k(X, \bar{Z}_k)).
\end{aligned}$$

For a proof, see Rotnitzky et al. (2017). The expression above decomposes the efficient influence function for ψ_{km_k} into $k+2$ components, each corresponding to one layer of the iterated expectation representation in Equation 11. The first term, $\varphi_0(O)$, centers the EIF (around zero) by subtracting the target parameter ψ_{km_k} from its plug-in estimate $\mu_{0m_k}^k(X)$. The next k terms, $\varphi_1(O), \dots, \varphi_k(O)$, are sequential bias-correction terms. Each $\mu_{jm_k}^k(X, \bar{Z}_j)$ is an estimated conditional expectation that appears in the iterated formula for ψ_{km_k} , and estimation error in these nuisance regressions would normally introduce bias. Each $\varphi_j(O)$ therefore takes the form of a weighted residual that subtracts off the discrepancy between two successive levels of the recursion, $\mu_{jm_k}^k(X, \bar{Z}_j) - \mu_{(j-1)m_k}^k(X, \bar{Z}_{j-1})$, and weights it by the inverse probability of the relevant portion of the treatment-mediator history. The last term $\varphi_{k+1}(O)$ plays the same bias-correction role for the outcome regression. It incorporates the residual $Y - \mu_{km_k}^k(X, \bar{Z}_k)$ and weights it by the inverse probability of the full treatment-mediator sequence required to identify ψ_{km_k} . Together, these $k+1$ bias-correction terms ensure that the influence function is orthogonal to first-order errors in all nuisance functions, allowing the resulting DML estimator to achieve semiparametric efficiency under standard regularity conditions.

Since the expression above gives the efficient influence function for ψ_{km_k} under the nonparametric model $\mathcal{P}$, its variance determines the lowest achievable asymptotic variance for any regular, asymptotically linear estimator of ψ_{km_k} . Consequently, the semiparametric efficiency bound for any asymptotically linear estimator of ψ_{km_k} is $\mathbb{E}[(\varphi_{km_k}(O))^2]$.

This EIF motivates an EIF-based estimator for ψ_{km_k} , obtained by solving the empirical

moment condition $\mathbb{P}_n[\varphi_{km_k}(O; \hat{\eta})] = 0$, where $\mathbb{P}_n[\cdot]$ denotes an empirical average, and where $\varphi_{km_k}(O; \hat{\eta})$ denotes the estimated EIF, evaluated using plug-in estimators for the nuisance functions. Specifically,

$$\begin{aligned} \hat{\psi}_{km_k}^{\text{eif}} = & \mathbb{P}_n \left[\frac{A}{\hat{\pi}_{01}(X)} \left(\frac{\mathbb{I}(M_k = m_k)}{\hat{\pi}_{km_k}(X, \bar{Z}_k)} \prod_{l=1}^{k-1} \frac{M_l}{\hat{\pi}_{l1}(X, \bar{Z}_l)} \right) (Y - \hat{\mu}_{km_k}^k(X, \bar{Z}_k)) \right. \\ & + \sum_{j=1}^k \frac{A}{\hat{\pi}_{01}(X)} \left(\prod_{l=1}^{j-1} \frac{M_l}{\hat{\pi}_{l1}(X, \bar{Z}_l)} \right) (\hat{\mu}_{jm_k}^k(X, \bar{Z}_j) - \hat{\mu}_{j-1m_k}^k(X, \bar{Z}_{j-1})) \\ & \left. + \hat{\mu}_{0m_k}^k(X) \right]. \end{aligned} \quad (13)$$

A similar EIF-based estimator can be used for ϕ_k to estimate the π_k terms in Equation 6. This estimator is based on the following nuisance functions for estimation (see Supplementary Material K for details):

$$\begin{aligned} \gamma_k(X, \bar{Z}_k) &\triangleq \mathbb{E}[M_{k+1} \mid X, \bar{Z}_k, \bar{1}_{k+1}], \\ \gamma_j(X, \bar{Z}_j) &\triangleq \mathbb{E}[\gamma_{j+1}(X, \bar{Z}_{j+1}) \mid X, \bar{Z}_j, \bar{1}_{j+1}] \forall j \in [k-1]. \end{aligned}$$

Next, following Kennedy (2022, p. 15), let $\mathbb{IF} : \Psi \rightarrow L_2(\mathbb{P})$ denote the operator mapping the functionals $\{\Delta_k, \pi_k, \eta_k\} : \mathcal{P} \rightarrow \mathbb{R}, \forall k \in [K]$ to their respective influence functions under the nonparametric model $\mathcal{P}$. Because the (Δ_k, τ_k) components of the decomposition are linear in ψ_{km_k} , by linearity of the EIF, $(\mathbb{IF}(\Delta_k), \mathbb{IF}(\tau_k))$ can be expressed as linear combinations of $\varphi_{km_k}(O)$. In particular, $\mathbb{IF}(\tau_k) = \varphi_{(k+1)1}(O) - \varphi_{k,0}(O)$ and $\mathbb{IF}(\Delta_k) = \varphi_{(k+1)0}(O) - \varphi_{k,0}(O)$. The EIFs of η_k and $\theta_k, \forall k \in [K]$ under $\mathcal{P}$ are derived as in Theorem 3.1:

Theorem 3.1. *The EIFs of $\eta_k, \theta_k \forall k \in [1, \dots, K]$ under P are given, respectively, by*

$$\begin{aligned}\mathbb{IF}(\eta_k) &= \mathbb{IF}(\tau_{k-1}) - \mathbb{IF}(\Delta_{k-1}) - \tau_k \mathbb{IF}(\pi_k) - \pi_k \mathbb{IF}(\tau_k), \\ \mathbb{IF}(\theta_k) &= \mathbb{IF}(\Delta_k) \prod_{j=1}^k \pi_j + \Delta_k \sum_{j=1}^k \mathbb{IF}(\pi_j) \prod_{\substack{l=1 \\ l \neq j}}^k \pi_l + \mathbb{IF}(\eta_k) \prod_{j=1}^{k-1} \pi_j + \eta_k \sum_{j=1}^{k-1} \mathbb{IF}(\pi_j) \prod_{\substack{l=1 \\ l \neq j}}^{k-1} \pi_l,\end{aligned}$$

for $k \in \{1, \dots, K\}$, with $\theta_0 = \Delta_0$, and where $\mathbb{RIF}(\phi) = \mathbb{IF}(\phi) + \phi$, denotes the recentered EIF of a parameter (about the truth). Their corresponding EIF-based estimators are (see Supplementary Material K for derivations):

$$\begin{aligned}\hat{\eta}_k^{eif} &= \widehat{\mathbb{RIF}}(\tau_{k-1}) - \widehat{\mathbb{RIF}}(\Delta_{k-1}) - \hat{\tau}_k \widehat{\mathbb{RIF}}(\pi_k) - \hat{\pi}_k \widehat{\mathbb{RIF}}(\tau_k) + \hat{\pi}_k \hat{\tau}_k, \\ \hat{\theta}_k^{eif} &= \widehat{\mathbb{RIF}}(\Delta_k) \prod_{j=1}^k \hat{\pi}_j + \hat{\Delta}_k \sum_{j=1}^k \widehat{\mathbb{RIF}}(\pi_j) \prod_{\substack{l=1 \\ l \neq j}}^k \hat{\pi}_l + \widehat{\mathbb{RIF}}(\eta_k) \prod_{j=1}^{k-1} \hat{\pi}_j + \hat{\eta}_k \widehat{\mathbb{RIF}}(\pi_j) \prod_{\substack{l=1 \\ l \neq j}}^{k-1} \hat{\pi}_l \\ &\quad - k \hat{\Delta}_k \prod_{j=1}^k \hat{\pi}_j - (k-1) \hat{\eta}_k \prod_{j=1}^{k-1} \hat{\pi}_j.\end{aligned}$$

where $\widehat{\mathbb{RIF}}(\phi) = \widehat{\mathbb{IF}}(\phi) + \phi$, and $\widehat{\mathbb{IF}}(\phi)$ denotes the influence function of a parameter evaluated at estimates of its component nuisance functions (see Supplementary Material K for derivations).

When machine learning estimators are used to compute the nuisance functions, in order to ensure the convergence rates outlined in Theorem 3.2 below, one could assume Donsker-type conditions for the nuisance function estimators, which restricts the set of estimators available to use. Alternatively, to expand the class of estimators that can be used for estimating the nuisance functions, sample-splitting can be used. In particular, Chernozhukov et al. (2017) suggest a “cross-fitting” procedure, which comprises the following steps: (1) Randomly split data into J folds: $\{S_1, \dots, S_J\}$; (2) For each fold S_j , use the remaining $(j-1)$ folds (training sample) to fit a flexible machine-learning model for each of the nuisance functions involved in the estimating equations; (3) For each observation in

j (estimation sample), use estimates of the above models to construct a set of estimated RIF functions for $\Delta_k \forall k \in \{0, \dots, K-1\}$, and for $(\pi_k, \tau_k, \eta_k, \theta_k) \forall k \in [K]$; (4) Compute an estimate of the decomposition components by averaging the estimated RIF functions across all subsamples S_1 through S_J . When all nuisance functions are estimated via data-adaptive methods and cross-fitting, the semiparametric efficiency of θ_k^{EIF} is given in the following Theorem:

Theorem 3.2 (Semiparametric efficiency). *Under Assumption 5, and under suitable regularity conditions (e.g. Chernozhukov et al., 2018), then $\hat{\theta}_k^{\text{EIF}}$ is semiparametric efficient if*

$$\sum_{j=k}^{k+1} \left[\sum_{l=0}^j R_n(\hat{\pi}_{l1}) R_n(\hat{\mu}_{l0}^j) \right] + \sum_{j=0}^{k-1} \left[R_n(\hat{\pi}_{j1}) R_n(\hat{\mu}_{j0}^{k-1}) + R_n(\hat{\pi}_{j1}) R_n(\hat{\mu}_{j1}^{k-1}) \right] + \sum_{j=0}^k \left[\sum_{l=0}^j R_n(\hat{\pi}_{l1}) R_n(\hat{\gamma}_l^j) \right]$$

is $o(n^{-1/2})$, where $R_n(\cdot)$ denotes a mapping from a nuisance function to its $L_2(P)$ convergence rate, and where $\hat{\mu}_{l0}^{K+1} \triangleq \hat{\mu}_{l1}^K$.

To gain some intuition for the result in Theorem 3.2, we can focus on $\theta_1 = \pi_1 \Delta_1 + \eta_1$, i.e., the MPSE through M_1 when $K = 1$. Note that estimation of $\theta_1 = \pi_1 \Delta_1 + \eta_1$ requires estimating the following decomposition components: $(\pi_1, \Delta_1, \tau_0, \Delta_0, \tau_1)$. To estimate these components, it suffices to estimate the following quantities: $(\phi_1, \psi_{01}, \psi_{00}, \psi_{10}, \psi_{11})$. In order for $\hat{\theta}_1^{\text{EIF}}$ to be semiparametric efficient, we require that the estimators employed for the set $(\phi_1, \psi_{01}, \psi_{00}, \psi_{10}, \psi_{11})$, i.e., $(\hat{\phi}_1^{\text{EIF}}, \hat{\psi}_{01}^{\text{EIF}}, \hat{\psi}_{00}^{\text{EIF}}, \hat{\psi}_{10}^{\text{EIF}}, \hat{\psi}_{11}^{\text{EIF}})$, are themselves semiparametric efficient. Thus, a sufficient (but not necessary) condition in order for $\hat{\theta}_1^{\text{EIF}}$ to obtain the semiparametric efficiency bound is if, for any two nuisance functions involved in $(\hat{\phi}_1^{\text{EIF}}, \hat{\psi}_{01}^{\text{EIF}}, \hat{\psi}_{00}^{\text{EIF}}, \hat{\psi}_{10}^{\text{EIF}}, \hat{\psi}_{11}^{\text{EIF}})$, the product of their convergence rates is $o(n^{-1/2})$. In this way, $\hat{\theta}_1^{\text{EIF}}$ will obtain the semiparametric efficiency bound if all of its constituent nuisance functions converge at a rate faster than $n^{-1/4}$ (although it will also obtain the efficiency bound under a variety of alternative conditions).

Under the DML estimation procedure, inference for all components of the MPSE decomposition is conducted using the efficient influence functions (EIFs) of the target parameters. When nuisance functions are estimated using cross-fitting and data-adaptive methods, the resulting DML estimators based on the EIFs derived above are asymptotically linear and

converge at a $\sqrt{n}$ -rate. In particular, under standard regularity conditions (Chernozhukov et al., 2017; Kennedy, 2022), each estimator admits the expansion

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}(\mathbb{P}_n - \mathbb{P})[\phi(O)] + o_p(1),$$

where $\phi(O)$ denotes the corresponding EIF. As a result, $\hat{\theta}$ is asymptotically normal with variance $\mathbb{E}[\phi(O)^2]$, which is consistently estimated by the empirical variance of the estimated EIF. Wald-type confidence intervals reported in the main text are constructed using this plug-in variance estimator. For example, inference on τ_1 can be conducted by estimating $\mathbb{P}_n[(\hat{\psi}_{11}^{\text{EIF}} - \hat{\psi}_{10}^{\text{EIF}})^2]/n$.

4 A Simulation Study

In this section, I evaluate the finite-sample performance of my two estimation procedures via a simulation experiment. Supplementary Material B provides further details. Specifically, I compare how the DML estimator proposed in Section 3 (as well as the parametric, RWR estimator described in Appendix A) perform under different degrees of misspecification. Specifically, I consider the setting of two causally ordered monotonic mediators, with post-treatment confounding. I generate simulations of observed data $O = (X_1, X_2, A, Z, M, Y)$ as follows:

$$U_1, U_2, U_3, U_4 \sim \text{MVN}(0_4, I_4)$$

$$X_1 \sim \text{N}((U_1, U_2, U_3, U_4)\beta_{X_1}, 1)$$

$$X_2 \sim \text{N}((U_1, U_2, U_3, U_4)\beta_{X_2}, 1)$$

$$A \sim \text{Bern}(g^{-1}[(1, X)\beta_A])$$

$$Z|A = 1 \sim \text{N}((1, X)\beta_Z, 1)$$

$$M_1|A = 1 \sim \text{Bern}(g^{-1}[(1, X, Z)\beta_{M_1}]) \quad M|A = 0 = 0$$

$$M_2|M_1 = 1 \sim \text{Bern}(g^{-1}[(1, X, Z)\beta_{M_2}]) \quad M_2|M_1 = 0 = 0$$

$$Y \sim \text{N}((1, A, X, AZ, AM_1, AM_1M_2)\beta_Y, 1).$$

The coefficients $(\beta_{X_1}, \beta_{X_2}, \beta_Y)$ are drawn from a $\text{Unif}(-1, 1)$ distribution, while the coefficient β_A is drawn from a $\text{Unif}(-0.5, 0.5)$ distribution. $g(\cdot)$ is a link function as described in the Supplementary Material. Further, $X = (X_1, X_2)$. In order to test how the DML and RWR methods perform when the relevant models are misspecified, I also construct transformations of the observed covariates (X^*) as follows, employing a similar setup to Kang and Schafer (2007):

$$X_1^* = (\exp(X_1/2) - 1)^2$$

$$X_2^* = X_2/(1 + \exp(X_2)) + 10$$

For each simulated dataset, I construct two estimates of the path-specific effects $(\theta_0, \theta_1, \theta_2)$ via the RWR and DML procedures described in Section 3. Standard errors for the coverage rates are computed via the estimated variance of the estimated EIFs for the DML approach, and via the nonparametric bootstrap with 250 replications for the RWR procedure. For the DML estimator, for each component involved in the decomposition, I construct a Neyman-orthogonal “signal” using its EIF. The recentered EIFs for each component are shown in Supplementary Material B.

I run 1000 replications of this DGP and compute the average bias and coverage of nominal 95% confidence intervals for sample sizes of 1000, 1500 and 2000 and using either the “correctly specified” covariates (X_1, X_2, Z) and the “incorrectly specified”, transformed versions (X_1^*, X_2^*, Z) . I calculate the true value of θ_1 by recovering the true values of the parameter set $(\theta_0, \theta_1, \theta_2)$ in each Monte Carlo simulation.

Figure S1 in Supplementary Material B presents the results from this simulation exercise. Under correctly specified models, the DML and RWR estimators perform similarly. In particular, both estimators exhibit negligible finite-sample bias for θ_0 , θ_1 , and θ_2 , with absolute biases on the order of 10^{-3} or smaller, and coverage rates close to the nominal 95% level. Under incorrectly specified models, however, the performance of the two estimators diverges sharply. The DML estimator remains stable: even when supplied with a misspecified feature space, absolute biases remain small for all MPSEs, and coverage rates remain close to nominal, typically between 92% and 95%. By contrast, the RWR estimator performs much more poorly, displaying a large amount of bias that in fact grows with the sample size, a large RMSE, and coverage rates that are not close to nominal. Overall, when the relevant models are correctly specified, both the parametric and semiparametric approaches perform well. Under misspecification, however, the advantages of the semiparametric estimation strategy become clear.

5 Empirical Analysis

To illustrate my approach empirically, I draw on data from the National Longitudinal Survey of Youth 1997 (NLSY97). I parse out the direct effect of high school graduation on adult earnings and its indirect or continuation effects via (i) college attendance, (ii) college graduation, and (iii) graduate school attendance. My analytic sample comprises $N = 7,305$ respondents.

I construct four types of variables: educational transitions, adult earnings, a set of confounders for the effect of high school graduation on subsequent transitions and earn-

ings, and a single set of intermediate confounders for the effect of college completion on subsequent transitions and earnings. My educational transition variables contain a binary treatment denoting whether a respondent had graduated high school by age 22, and three binary mediators denoting whether the respondent had attended a 4-year college by age 22, whether the respondent had received a BA degree by age 29, and whether the respondent had enrolled in a graduate level program by age 29, respectively. I assume that all individuals who make a given educational transition have made all previous educational transitions. Thus, by construction, my coding strategy disallows for cases which violate the monotonicity assumption.⁵

My outcome of interest is logged average annual earnings at ages 32–36. For each respondent and each survey year in this age range, I construct total annual labor-market income by summing wage, salary, and business income, and then compute the (logged) average of these annual totals across this age range. This multi-year average yields a more stable measure of early-adult earnings that smooths over short-term income fluctuations.

⁶ Earnings are adjusted for inflation to 2023 dollars using the personal consumption expenditures (PCE) index. After dropping respondents with missing earnings information, I accommodate those with zero earnings by adding a small constant of \$1,000 to observed earnings (though in Supplementary Material F, I replicate my main analyses under alternative definitions of earnings).

In an effort to satisfy the sequential ignorability assumption (Assumption 4), I include a large array of covariates in my models. This set of covariates is more expansive than those

⁵ Assuming away cases in which an individual makes a particular educational transition without having made *all* previous transitions serves as a reasonable approximation to reality. Among the set of individuals who have non-missing earnings information in the NLSY97 (i.e., those who comprise my analytic sample), 94% of individuals observed to attend graduate school by age 29 also completed a BA by age 29; 93% of respondents who completed a BA by age 29 had attended a 4-year college by age 22 (6% of those who completed a BA by age 29 first attended a 4-year college between ages 23 and 26 inclusive), and 99% of respondents who attended a 4-year college by age 22 had also completed high school.

⁶ Age 36 is the latest age at which earnings are consistently observed in the NLSY97. Because earnings gains associated with graduate education may materialize later in the life course, I also assess robustness to using a later earnings window. In Supplementary Material H, I re-estimate the full MPSE decomposition using the NLSY79 cohort and measure logged earnings over ages 35–44. The qualitative patterns are highly similar, and the contribution of the graduate-education pathway remains modest, for reasons discussed further in the supplement.

used in previous observational studies of returns to education (see in particular Scott-Clayton and Wen, 2019). In particular, in addition to including information on respondent demographics (gender, race, ethnicity, age in 1997), and observed pre-college performance such as overall high school GPA and test score on the Armed Services Vocational Aptitude Battery (ASVAB), I include detailed information on socioeconomic background. Since my proposed decomposition also facilitates the inclusion of a distinct set of observed intermediate confounders for each transition, I include two postsecondary characteristics (Z) to adjust for confounders of the effect of BA completion and graduate school attendance on earnings: field of study and college GPA. To assess the robustness of my main conclusion to forms of unobserved confounding, in Supplementary Material C, I produce a set of “bias-corrected” estimates of the decomposition components under certain assumptions about the nature of the confounding.

A large proportion (just under 50%) of respondents are missing information on covariates X and Z . For my main analyses, I impute missing values on these covariates via multiple imputation to increase efficiency, but in Supplementary Material E, I replicate these analyses restricted to the sample of respondents with complete information. This exercise produces substantively similar results (for covariate means for each of these analytic samples, see Supplementary Material D). After constructing the analytical sample, I apply both the DML estimator described in Section 3 as well as a parametric, regression-with-residuals (RWR) algorithm (described in Supplementary Material A) to implement the proposed decomposition. For the DML approach, I estimate all nuisance functions, using a super learner composed of the Lasso and random forest and, following Chernozhukov et al. (2017), use five-fold cross-fitting. Supplementary Material J gives further details about the particular models required given my assumed data generation process.

I conduct two sets of analyses. First, I assume that positivity holds at every transition of the education sequence, i.e., the estimated propensity of each transition lies strictly within $(0, 1)$. Second, I relax the positivity assumption by allowing units to have transition

propensities of exactly zero or one. In this scenario, the MPSE decomposition is unidentified, and therefore requires a redefinition of the target quantity as the decomposition among the common-support subpopulation across transitions. I present results under the first approach (assuming positivity) in the main text, and compare them with results from the common-support, trimmed sample under the second approach (relaxing positivity) in Supplementary Material G. In practice the two approaches yield substantively similar conclusions: only the college-completion pathway attenuates modestly once units with extreme propensity scores are excluded.

Figure 3 shows my estimates of the average total effect (ATE) on log earnings and its direct and continuation components under both the DML and RWR procedures. Standard errors for RWR estimates are obtained via the non-parametric bootstrap, while standard errors for DML estimates are obtained via the variance of the estimated EIF for each MPSE. Both procedures return similar estimates, though deviate in the estimated magnitude of MPSE θ_1 , and DML estimates come expectedly with a significantly greater amount of precision. The first column shows that the estimated ATE of graduating high school on log earnings under DML (RWR) is 0.68 (0.63), which implies an earnings premium of approximately 96%. The next two columns indicate that the vast majority (69% under DML and 75% under RWR) of the ATE operates directly, i.e. net of college attendance, BA completion and graduate school attendance (MPSE θ_0 , $A \rightarrow Y$). Specifically, high school graduates who do not proceed to college can be expected to earn on average 0.47 log earnings more than high school non-completers under DML, an earnings premium of 59% (RWR returns an identical estimate).

While the majority of the ATE is explained by the direct effect, a non-trivial portion occurs through mediation effects through later transitions. Under DML, the continuation effects of high school graduation via college attendance without BA completion (MPSE θ_1 , $A \rightarrow M_1 \rightarrow Y$) and via BA completion without graduate school participation (MPSE θ_2 , $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$) both mediate roughly 15% of the ATE, and correspond to an

earnings premium of approximately 10%. The RWR estimate of θ_1 is notably lower at 0.03 and is also imprecisely estimated. Under both estimation procedures, the continuation effect via graduate school attendance ($A \rightarrow M_1 \rightarrow M_2 \rightarrow M_3 \rightarrow Y$) is very small and fails to reach conventional levels of significance. In sum, the total effect of high school graduation on earnings is determined overwhelmingly by its direct effect on earnings.

Table 1 shows DML and RWR estimates of the various components (the direct effects (Δ_k), probabilities (π_k) and covariance terms (η_k)) that constitute the continuation effects θ_k . Several points are of note. First, the components in the table offer insights into the economic and educational returns to different educational stages. The direct effects of each educational transition (Δ_k) are highly variable: they are largest for high school graduation and for college completion (both at 0.47 under DML and RWR), and lowest for college attendance and graduate school participation (at 0.2 and 0.14, respectively, under DML). Note that the payoff to graduate school attendance could be depressed by the fact that I observe individuals at a maximum age of only 36, if graduate school earnings premia materialize only much later in the life course. The counterfactual continuation probabilities (π_k) also provide insight into barriers in educational participation. In particular, even if an individual were to complete high school (possibly contrary to fact), that individual would have under a 50% chance of continuing to a 4-year college without further intervention to increase individuals' college application, admissions and enrollment rates. Further, even if individuals were to counterfactually both complete high school and attend a 4-year college, only a very small proportion ($\pi_1 \cdot \pi_2 = 0.24$) would be expected to complete their BA degree without further intervention at the college-level.

Second, the fine-grained nature of the MPSE decomposition enables us to trace the continuation effects to their constituent components. In particular, while the direct effect of high school completion is comparable to the direct effect of BA graduation on earnings, the overall continuation effect via BA completion that it informs (MPSE θ_2 , $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$) is small because θ_2 is approximately (plus the small value of η_2) equal to Δ_2

scaled by the product $\pi_1 \cdot \pi_2 = 0.24$. In words, despite the relatively large direct effect of BA completion on earnings, given individuals' low counterfactual probability of BA completion, this transition is not an important mediating pathway of the total effect of high school completion on earnings. The result is that college attendance without completion mediates high school graduation's earnings effects as much as BA completion, despite the fact that college attendance without completion yields a much smaller earnings return for high school graduates than BA completion without graduate school attendance does for college enrollees.

One instructive point of comparison for these results is instrumental variable (IV) estimates of returns to years of schooling, typically estimated in the range of 6% to 12% (Angrist and Krueger, 1991, 1992; Kane and Rouse, 1993; Card, 1994; Ashenfelter and Zimmerman, 1997; Angrist and Chen, 2011). While my estimate of the overall return to high school graduation (τ_0) could appear large in this light, several factors could reconcile this difference. First, τ_0 captures the direct *and* continuation effects of high school completion (whereas IV estimates of schooling returns capture schooling's direct effects). Further, τ_0 captures the effect of multiple additional years of schooling (as the high school graduates and high school non-completers that form the comparison group differ by multiple years of schooling), as opposed to a single year's additional return. In fact, we can more directly compare my DML estimate of the direct return to high school graduation (Δ_0) of 0.46 (corresponding to an earnings premium of 58%) using the fact that, in the NLSY97, high school non-completers attained on average 3.7 fewer years of schooling than high school completers. An IV estimate of 12%, for example, would therefore imply an earnings return to 3.7 additional years of approximately 52% - broadly in line with my result. Still, to assess the robustness of the above findings to potential violations of Assumption 4 (Sequential Ignorability), I implement a sensitivity analysis in Supplementary Material C. Under the stated assumptions about the pattern of unobserved confounding, my primary finding that the ATE of high school graduation is overwhelmingly mediated via its direct effect remains

highly robust to unobserved confounding.

6 Conclusion

In this article, I have developed a causal mediation framework for analyzing education effects on earnings. First, I have demonstrated that the total effect of any level of education can be decomposed into a direct effect and K mutually exclusive “continuation” effects. All of these effects are identifiable under the assumption of sequential ignorability. Importantly, this property allows for the effect of each educational transition to be confounded by a distinct set of observed covariates—a property which allows for weaker identification conditions compared with conventional mediation-based decompositions of the ATE (Miles et al., 2017; Zhou, 2022b).

Several directions for future research follow naturally from the proposed framework. First, in this paper I have considered a decomposition of the average treatment effect for the case of binary monotonic mediators, but many educational processes are more finely graded. Extensions to settings with categorical or multivalued transitions—such as different types of postsecondary institutions, fields of study, or graduate degrees—would further broaden the applicability of the framework. Supplementary Material I outlines one such extension, but generalizing the framework to categorical and continuous mediators remains an open area for future work.

Second, although the MPSE decomposition relaxes cross-world assumptions and permits observed intermediate confounding, it still relies on sequential ignorability. In practice, this assumption may be difficult to satisfy fully in observational settings, particularly when selection into successive educational transitions is driven by unobserved traits such as motivation or ability. Developing alternative identification results that, for example, exploit instrumental variables—long used in the education literature to address selection into schooling—would be a particularly promising extension of the MPSE framework.

Finally, while this paper emphasizes educational attainment, the monotonic structure

exploited here arises in many other domains characterized by state-dependent transitions, such as family formation, health progression, or criminal justice contact. This characteristic is particularly salient in demographic phenomena. Certain demographic events are rigid in their monotonicity by definition. For example, researchers may be interested in discerning the degree to which positive effects of marriage on outcomes such as earnings and life satisfaction are undermined by the negative effects of divorce and separation (and, in turn, their mitigation via re-marriage) (Kenney, 2004; Sweeney and Phillips, 2004) – monotonic transitions. Similarly, the effect of parenthood on earnings can be seen as operating directly, through the effect of having a first child net of subsequent children, as well as operating indirectly through the effects of having multiple children. A similar perspective may be taken in a criminal justice context: the total effect of early-stage police contact (such as being searched for contraband) on educational and socio-psychological outcomes can be decomposed into path-specific effects via subsequent arrest and incarceration (Weaver and Lerman, 2010; Kirk and Sampson, 2013; Sugie and Turney, 2017). Applying the MPSE framework to these contexts, and comparing the resulting decompositions across domains, may yield new insights into how life transitions shape later outcomes through both direct and sequential mechanisms.

7 Tables and figures

Table 1: Direct Effects (Δ_k), Probabilities (π_k) and Covariance Terms (η_k) Involved in Decomposition via Debiased Machine-Learning (DML) and Regression-With-Residuals (RWR).

	Δ_0	Δ_1	Δ_2	Δ_3	π_1	π_2	π_3	η_1	η_2	η_3
DML	0.466 (0.057)	0.200 (0.031)	0.444 (0.034)	0.138 (0.040)	0.427 (0.008)	0.555 (0.016)	0.314 (0.011)	0.007 (0.007)	0.004 (0.008)	-0.002 (0.008)
RWR	0.469 (0.115)	0.116 (0.093)	0.466 (0.097)	0.163 (0.114)	0.374 (0.015)	0.515 (0.077)	0.159 (0.022)	-0.016 (0.007)	0.083 (0.015)	0.028 (0.051)

Note: The Δ_k parameters capture the average effect of completing the k th mediator *but no subsequent mediator* on earnings, relative to completing the $k - 1$ th mediator. For instance, Δ_0 denotes the effect of completing high school (M_1) but not attending college nor, under Assumption 2, completing any subsequent mediators, relative to attending high school but not completing it ($M_0 \equiv A$). The π_k terms capture the average of individuals’ counterfactual completion status of the k th mediator under completion of all prior mediators $M_0, \dots, M_{k-1}$. For example, π_1 denotes individuals’ average counterfactual college attendance, after—possibly contrary to fact—their completion of high school. Finally, the η_k terms refer to the covariance between individuals’ own counterfactual completion status of the k th mediator, and their own “gross” effect of completing the k th mediator on earnings. To recall, the “gross” effect of the k th mediator captures the effect of completing that mediator, relative to completing only the $k - 1$ th mediator, irrespective of whether that effect operates directly (net of subsequent mediators) or via subsequent transitions. For example, η_1 denotes the covariance between each individual’s counterfactual college attendance status and their gross effect of college attendance on earnings.

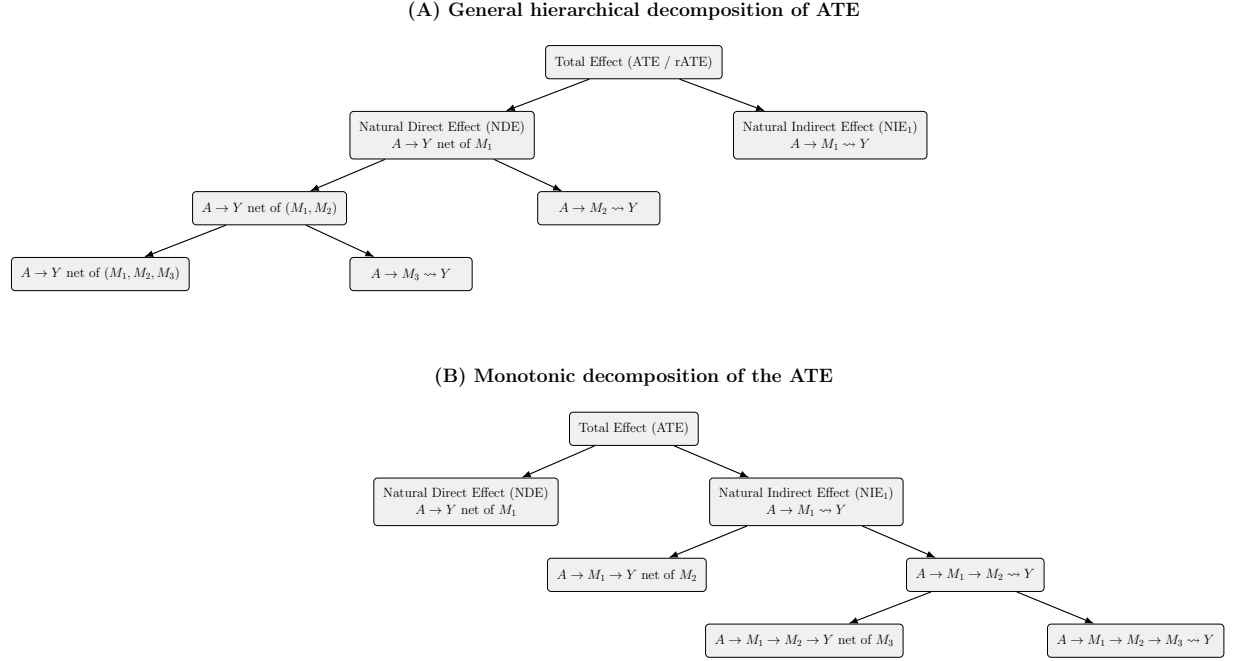


Figure 1: Nested decomposition of the Average Treatment Effect (ATE) into direct and sequential indirect pathways: general hierarchical decomposition (Panel A) and proposed monotonic decomposition (Panel B).

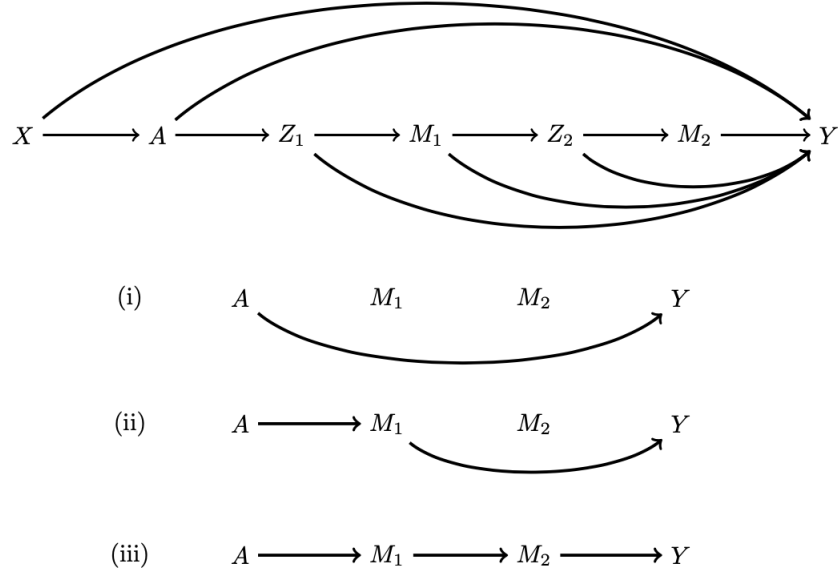


Figure 2: Causal Relationships Between Two Monotonic Mediators Shown in a Directed Acyclic Graph (DAG) and the 3 Monotonic Path Specific Effects (MPSEs). A denotes an initial transition of interest, Y , an outcome, and M_1 and M_2 are two causally ordered, monotonic mediators. The set (X, Z_1, Z_2) captures pre-treatment and intermediate confounders.

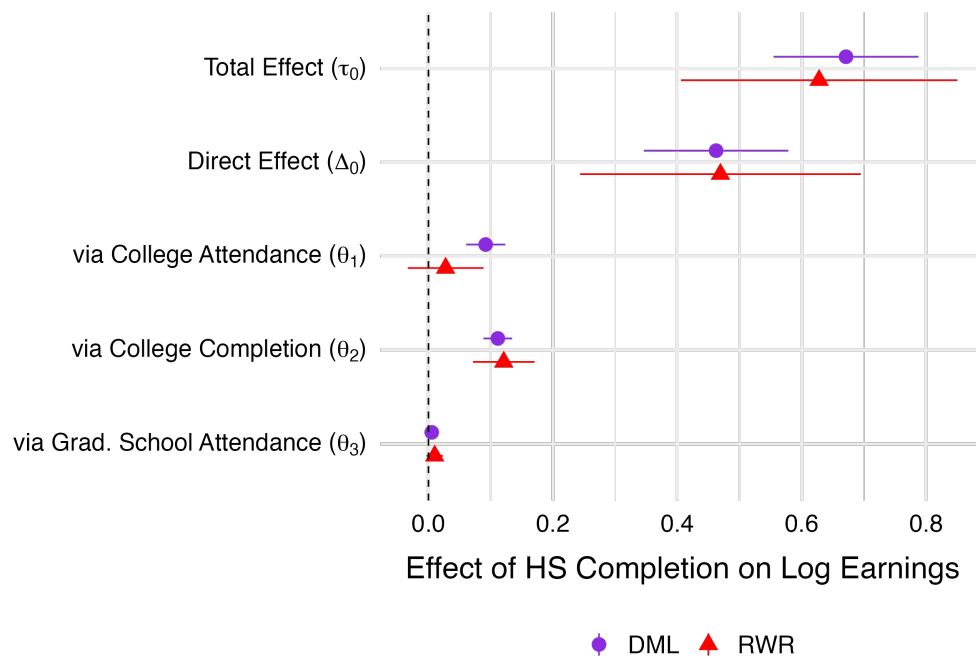


Figure 3: Decomposition of the Average Total Effect (ATE) of High School Graduation on Logged Earnings via Debiased Machine-Learning (DML) and Regression-With-Residuals (RWR).

References

- Avidit Acharya, Matthew Blackwell, and Maya Sen. Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review*, 110(3): 512–529, 2016.
- Jeffrey M Albert and Suchitra Nelson. Generalized causal mediation analysis. *Biometrics*, 67:1028–1038, 2011. ISSN 0006-341X.
- Joshua D Angrist and Stacey H Chen. Schooling and the vietnam-era gi bill: Evidence from the draft lottery. *American Economic Journal: Applied Economics*, 3:96–118, 2011. ISSN 1945-7782.
- Joshua D Angrist and Alan B Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106:979–1014, 1991. ISSN 1531-4650.
- Joshua D Angrist and Alan B Krueger. The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American statistical Association*, 87:328–336, 1992. ISSN 0162-1459.
- Orley Ashenfelter and David J Zimmerman. Estimates of the returns to schooling from sibling data: Fathers, sons, and brothers. *Review of Economics and Statistics*, 79:1–9, 1997. ISSN 0034-6535.
- Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. 2005.
- Kelli A Bird, Benjamin L Castleman, Jeffrey T Denning, Joshua Goodman, Cait Lambertson, and Kelly Ochs Rosinger. Nudging at scale: Experimental evidence from fafsa completion campaigns. *Journal of Economic Behavior & Organization*, 183:105–128, 2021. ISSN 0167-2681.

- Sandra E Black, Jeffrey T Denning, Lisa J Dettling, Sarena Goodman, and Lesley J Turner. Taking it to the limit: Effects of increased student loan availability on attainment, earnings, and financial well-being. *American Economic Review*, 113:3357–3400, 2023. ISSN 0002-8282.
- Zachary Bleemer. Affirmative action, mismatch, and economic mobility after california’s proposition 209. *The Quarterly Journal of Economics*, 137:115–160, 2022. ISSN 0033-5533.
- Hugo Bodory, Martin Huber, and Lukáš Lafférs. Evaluating (weighted) dynamic treatment effects by double machine learning. *The Econometrics Journal*, 25:628–648, 2022. ISSN 1368-4221.
- Jennie E Brand and Yu Xie. Who benefits most from college? evidence for negative selection in heterogeneous economic returns to higher education. *American sociological review*, 75:273–302, 2010. ISSN 0003-1224.
- David Card. Earnings, schooling, and ability revisited. 1994.
- David Card. *The causal effect of education on earnings*, volume 3, pages 1801–1863. Elsevier, 1999. ISBN 1573-4463.
- David Card. Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, 69:1127–1160, 2001. ISSN 0012-9682.
- Pedro Carneiro, James J Heckman, and Edward J Vytlacil. Estimating marginal returns to education. *American Economic Review*, 101:2754–2781, 2011. ISSN 0002-8282.
- Benjamin L Castleman, Denise Deutschlander, and Gabrielle Lohner. Pushing college advising forward: Experimental evidence on intensive advising and college success. *Ed-WorkingPapers. com*, 2020.

- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107:261–265, 2017. ISSN 0002-8282.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.
- Raj Chetty, David J Deming, and John N Friedman. Diversifying society’s leaders? the causal effects of admission to highly selective private colleges, 2023.
- Sarah R Cohodes and Joshua S Goodman. Merit aid, college quality, and college completion: Massachusetts’ adams scholarship as an in-kind subsidy. *American Economic Journal: Applied Economics*, 6:251–285, 2014. ISSN 1945-7782.
- Rhian M Daniel, Bianca L De Stavola, Simon N Cousens, and Stijn Vansteelandt. Causal mediation analysis with multiple mediators. *Biometrics*, 71:1–14, 2015. ISSN 0006-341X.
- Susan Dynarski, C J Libassi, Katherine Micheltore, and Stephanie Owen. Closing the gap: The effect of reducing complexity and uncertainty in college pricing on the choices of low-income students. *American Economic Review*, 111:1721–1756, 2021. ISSN 0002-8282.
- Christina Ciocca Eller. What makes a quality college? re-examining the equalizing potential of higher education in the united states. 2023. ISSN 0002-9602.
- Christina Ciocca Eller and Thomas A DiPrete. The paradox of persistence: Explaining the black-white gap in bachelor’s degree completion. *American Sociological Review*, 83: 1171–1214, 2018. ISSN 0003-1224.
- Helmut Farbmacher, Martin Huber, Lukáš Lafférs, Henrika Langen, and Martin Spindler. Causal mediation analysis with double machine learning. *The Econometrics Journal*, 25: 277–300, 2022. ISSN 1368-4221.

- Joshua Goodman, Michael Hurwitz, and Jonathan Smith. Access to 4-year public colleges and degree completion. *Journal of Labor Economics*, 35:829–867, 2017. ISSN 0734-306X.
- James J Heckman, John Eric Humphries, and Gregory Veramendi. Returns to education: The causal effects of education on earnings, health, and smoking. *Journal of Political Economy*, 126:S197–S246, 2018. ISSN 0022-3808.
- Michael Hout. Social and economic returns to college education in the united states. *Annual review of sociology*, 38:379–400, 2012.
- Michael Hurwitz and Jessica Howell. Estimating causal impacts of school counselors with regression discontinuity designs. *Journal of Counseling & Development*, 92:316–327, 2014. ISSN 0748-9633.
- Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, February 2010. doi: 10.1214/10-STS321.
- Thomas J Kane and Cecilia E Rouse. Labor market returns to two-and four-year colleges: is a credit a credit and do degrees matter? 1993.
- Joseph D.Y. Kang and Joseph L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22, 2007. ISSN 08834237. doi: 10.1214/07-STS227.
- Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.
- Catherine Kenney. Cohabiting couple, filing jointly? resource pooling and us poverty policies. *Family Relations*, 53:237–247, 2004. ISSN 0197-6664.
- David S Kirk and Robert J Sampson. Juvenile arrest and collateral educational damage in the transition to adulthood. *Sociology of education*, 86:36–62, 2013. ISSN 0038-0407.

- Greg Lewis and Vasilis Syrgkanis. Double/debiased machine learning for dynamic treatment effects via g-estimation. *arXiv preprint arXiv:2002.07285*, 2020.
- Sheng-Hsuan Lin and Tyler VanderWeele. Interventional approach for path-specific effects. *Journal of Causal Inference*, 5:20150027, 2017. ISSN 2193-3685.
- Robert D Mare. Social background and school continuation decisions. *Journal of the American Statistical Association*, 75:295–305, 1980. ISSN 0162-1459.
- Caleb H Miles, Ilya Shpitser, Phyllis Kanki, Seema Meloni, and Eric J Tchetgen Tchetgen. Quantifying an adherence path-specific effect of antiretroviral therapy in the nigeria pepfar program. *Journal of the American Statistical Association*, 112:1443–1452, 2017. ISSN 0162-1459.
- Caleb H Miles, Ilya Shpitser, Phyllis Kanki, Seema Meloni, and Eric J Tchetgen Tchetgen. On semiparametric estimation of a path-specific effect in the presence of mediator-outcome confounding. *Biometrika*, 107:159–172, 2020. ISSN 0006-3444.
- Jack Mountjoy. Community colleges and upward mobility. *American Economic Review*, 112:2580–2630, 2022. ISSN 0002-8282.
- Jack Mountjoy and Brent R Hickman. The returns to college (s): Relative value-added and match effects in higher education, 2021.
- Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994. ISSN 1573-4412.
- Judea Pearl. Direct and indirect effects. In J.S. Breese and D. Koller, editors, *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420, San Francisco, CA, 2001. Morgan Kaufmann Publishers.
- Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3, 2009. ISSN 19357516. doi: 10.1214/09-SS057.

- James Robins. 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, 1986. ISSN 0270-0255.
- Andrea Rotnitzky, James Robins, and Lucia Babino. On the multiply robust estimation of the mean of the g-functional. *arXiv preprint arXiv:1705.08582*, 2017.
- Judith Scott-Clayton and Qiao Wen. Estimating returns to college attainment: Comparing survey and state administrative data-based estimates. *Evaluation Review*, 43:266–306, 2019. ISSN 0193-841X.
- Jonathan Smith, Joshua Goodman, and Michael Hurwitz. The economic impact of access to public four-year colleges, 2020.
- Thomas D Snyder, Cristobal de Brey, and Sally A Dillow. Digest of education statistics 2014, 50th edition. nces 2016-006, 2016.
- Johan Steen, Tom Loeys, Beatrijs Moerkerke, and Stijn Vansteelandt. Flexible mediation analysis with multiple mediators. *American journal of epidemiology*, 186:184–193, 2017. ISSN 0002-9262.
- Naomi F Sugie and Kristin Turney. Beyond incarceration: Criminal justice contact and mental health. *American Sociological Review*, 82:719–743, 2017. ISSN 0003-1224.
- Zachary Sullivan, Benjamin L Castleman, and Eric Bettinger. College advising at a national scale: Experimental evidence from the collegepoint initiative. 2019.
- Megan M Sweeney and Julie A Phillips. Understanding racial differences in marital disruption: Recent trends and explanations. *Journal of Marriage and Family*, 66:639–650, 2004. ISSN 0022-2445.
- Lesley J Turner and Oded Gurantz. Experimental estimates of college coaching on post-secondary re-enrollment, 2024.

- Tyler J VanderWeele. A unification of mediation and interaction: a 4-way decomposition. *Epidemiology*, 25(5):749–761, 2014.
- Tyler J VanderWeele and Onyebuchi A Arah. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*, pages 42–52, 2011. ISSN 1044-3983.
- Tyler J VanderWeele and Stijn Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, 2:457–468, 2009. ISSN 1938-7997.
- Tyler J VanderWeele, Stijn Vansteelandt, and James M Robins. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology (Cambridge, Mass.)*, 25:300, 2014.
- Stijn Vansteelandt and Rhian M Daniel. Interventional effects for mediation analysis with multiple mediators. *Epidemiology*, 28:258–265, 2017. ISSN 1044-3983.
- Davide Viviano and Jelena Bradic. Dynamic covariate balancing: estimating treatment effects over time. *arXiv preprint arXiv:2103.01280*, 2021.
- Vesla M Weaver and Amy E Lerman. Political consequences of the carceral state. *American Political Science Review*, 104:817–833, 2010. ISSN 1537-5943.
- Ang Yu, Li Ge, and Felix Elwert. When do natural mediation effects differ from their randomized interventional analogues: Test and theory. *arXiv preprint arXiv:2407.02671*, 2024.
- Xiang Zhou. Attendance, completion, and heterogeneous returns to college: A causal mediation approach. *Sociological Methods & Research*, page 00491241221113876, 2022a. ISSN 0049-1241.

Xiang Zhou. Semiparametric estimation for causal mediation analysis with multiple causally ordered mediators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84:794–821, 2022b. ISSN 1369-7412.

Xiang Zhou and Guanghui Pan. Higher education and the black-white earnings gap. *American Sociological Review*, 88:154–188, 2023. ISSN 0003-1224.

Seth D Zimmerman. The returns to college admission for academically marginal students. *Journal of Labor Economics*, 32:711–754, 2014. ISSN 0734-306X.

Supplemental Materials (to appear online)

A Parametric, regression-with-residuals (RWR) estimation

In this section, I propose a linear regression-with-residuals (RWR) approach for the MPSE decomposition. The approach relies on two steps. The first involves residualizing pre-treatment confounders with respect to their marginal means, and intermediate confounders on all causally prior confounders, i.e., $X^\perp \triangleq X - \mathbb{E}[X]$, and $Z_k^\perp \triangleq M_{k-1}[Z_k - \mathbb{E}[Z_k | X, \bar{Z}_{k-1}, M_{k-1} = 1]]$ for all $k \in [K]$, $M_0 \triangleq A$. For now, we are agnostic about the functional form used for $\mathbb{E}[Z_k | X, \bar{Z}_{k-1}, M_{k-1} = 1]$. The second step involves fitting three sets of models. The first is simply a model for the outcome given pre-treatment covariates and the treatment, namely,

$$\mathbb{E}[Y | X, A] = \lambda_0 + \lambda_1 A + \alpha_1^T X^\perp + \alpha_2^T A X^\perp; \quad (14)$$

The second is a set of models for the outcome given covariates, the treatment and M_k for all $k \in [K]$, i.e.,

$$\begin{aligned} \mathbb{E}[Y | X, \bar{Z}_k, A, \bar{M}_k] = & \beta_{k,0} + c_{k,0}A + \sum_{j=1}^k \beta_{k,j}M_j + \eta_{k,1}^\top X^\perp + c_{k,1}A X^\perp + \sum_{j=1}^{k-1} \eta_{k,j}^T M_j X^\perp \\ & + \sum_{j=1}^k \gamma_{k,j}^T Z_j^\perp + \sum_{j=1}^{k-1} M_j \sum_{l=1}^j \xi_{k,k,l}^\top Z_l^\perp, \end{aligned} \quad (15)$$

while the third is a set of models for each mediator given covariates, the treatment, conditional on the treatment and all prior mediators, i.e., for all $k \in [K - 1]$,

$$\mathbb{E}[M_{k+1} \mid X, \bar{Z}_k, \bar{1}_{k+1}] = \theta_{k,0} + \delta_{k,1}^T X^\perp + \sum_{j=1}^k \delta_{k,j+1}^T Z_j^\perp. \quad (16)$$

These models differ from conventional linear regression in that (i) pre-treatment variables are centered around their marginal means, and (ii) post-treatment confounders $Z_k \forall k \in \{1, \dots, K\}$ are centered around their conditional means given all antecedent variables. Under Assumptions 3-5 in the main text, and assuming that the outcome and mediators are linear in pre- and post-treatment confounders, the treatment, and prior mediators, and that all necessary interaction terms have been accounted for, then the ATE τ_0 can be obtained from the linear model $\mathbb{E}[Y \mid X, A]$ as λ_1 , and coefficients from the models $\mathbb{E}[Y \mid X, A, \bar{Z}_k, \bar{M}_k]$ and $\mathbb{E}[M_{k+1} \mid X, A, \bar{Z}_k, \bar{M}_k]$ yield estimates of the components of the decomposition as follows:

$$\tau_k = \mathbb{E}[Y(\bar{1}_{k+1}) - Y(\bar{1}_k, 0)] = \beta_{k,k}, \forall k \in \{1, \dots, K\},$$

$$\Delta_k = \mathbb{E}[Y(\bar{1}_{k+1}, \underline{0}_{k+2}) - Y(\bar{1}_k, \underline{0}_{k+1})] = \beta_{k+1,k-1}, \forall k \in \{0, \dots, K-1\},$$

$$\pi_{k+1} = \mathbb{E}[M_{k+1}(\bar{1}_{k+1})] = \theta_{k,0}, \forall k \in \{0, \dots, K-1\}.$$

I state the RWR estimation procedure formally in the following algorithm:

Algorithm 1 RWR

1. For each of the baseline confounders, compute $\hat{X}^\perp = X - \mathbb{P}_n[X]$, where $\mathbb{P}_n[\cdot]$ denotes empirical average.
2. Fit $\hat{\mathbb{E}}[Y \mid X, A]$ using the linear specification shown above; an estimate of τ_0 is given by $\hat{\lambda}_1$.
3. For each set of post-treatment confounders Z_k , $k \in \{1, \dots, K\}$, compute $Z_k^\perp =$

$M_{k-1}[Z_k - \mathbb{E}[Z_k \mid X, \bar{Z}_{k-1}, M_{k-1} = 1]]$ where an overbar denotes a vector of variables such that $\bar{Z}_k = (Z_1, \dots, Z_k)$, by fitting a regression of Z_k on X and $\bar{Z}_{k-1}$ among units with $M_{k-1} = 1$ and then calculating the residuals.

4. For each $k \in \{1, \dots, K\}$:

(a) compute least squares estimates of equations 15 and 16, using estimates of $X^\perp$ and $Z_k^\perp$.

(b) compute $\hat{\tau}_k^{\text{RWR}} = \hat{\beta}_{k,k}$, $\hat{\Delta}_{k-1}^{\text{RWR}} = \hat{\beta}_{k,k-1}$, and $\hat{\pi}_k^{\text{RWR}} = \hat{\theta}_{k-1,0}$.

5. Compute the decomposition using $\hat{\tau}_k$, $\hat{\Delta}_k$ and $\hat{\pi}_{k+1}$, and estimating the covariance terms as $\hat{\eta}_k^{\text{RWR}} = \hat{\beta}_{k-1,k} - \hat{\beta}_{k,k-1} - \hat{\beta}_{k,1}\hat{\theta}_{k-1,k}$, and the continuation effects as $\hat{\theta}_k^{\text{RWR}} = (\prod_{j=1}^k \hat{\pi}_j^{\text{RWR}})\hat{\Delta}_k^{\text{RWR}} + (\prod_{j=1}^{k-1} \hat{\pi}_j^{\text{RWR}})\hat{\eta}_k^{\text{RWR}}$.

Standard errors and confidence intervals can then be obtained via the non-parametric bootstrap, or by using their asymptotic analytic variance. Specifically, let $\hat{\theta}_k^* \triangleq (\hat{\beta}_{k,0}, \hat{\beta}_{k,1}, \theta_{k,0})$ denote a set of parameters. Under the above models, we have that $\hat{\theta}^* = \{\hat{\lambda}_1, \theta_1^*, \dots, \theta_K^*\}$ solves $\mathbb{P}_n[g(O; \hat{\theta}^*)] = 0$, where $g(O; \theta^*)$ is the set of stacked moment conditions with solution $\hat{\theta}^*$. Under standard regularity conditions (Newey and McFadden, 1994), under correct specification of the models wherein all residualized quantities are estimated via linear models, the set $\hat{\theta}^*$ is consistent and asymptotically normal, such that $\sqrt{n}(\hat{\theta}^* - \theta^*)$ converges to a mean-zero normal distribution with finite variance $V = G^{-1}\Omega(G^{-1})^\top$, where $\Omega = \mathbb{E}[g(O; \theta^*)g(O; \theta^*)^\top]$, and where $G = \mathbb{E}[\frac{\partial g(O; \theta^*)}{\partial \theta^\top}]$. It follows by a simple application of the Delta Method that the set $\hat{\gamma}_k^* \triangleq (\hat{\tau}_k^{\text{RWR}}, \hat{\Delta}_{k-1}^{\text{RWR}}, \hat{\pi}_k^{\text{RWR}}, \hat{\eta}_k^{\text{RWR}}, \hat{\theta}_k^{\text{RWR}}) \forall k \in [K]$ is also consistent and asymptotically normal.

B Further details on simulation study

In this section, I provide further details on the simulation exercise described in the main text. To recall, the assumed data-generating process is

$$\begin{aligned}
U_1, U_2, U_3, U_4 &\sim \text{MVN}(0_4, I_4) \\
X_1 &\sim \text{N}((U_1, U_2, U_3, U_4)\beta_{X_1}, 1) \\
X_2 &\sim \text{N}((U_1, U_2, U_3, U_4)\beta_{X_2}, 1) \\
A &\sim \text{Bern}(g^{-1}[(1, X)\beta_A]) \\
Z|A = 1 &\sim \text{N}((1, X)\beta_Z, 1) \\
M_1|A = 1 &\sim \text{Bern}(g^{-1}[(1, X, Z)\beta_{M_1}]) \quad M|A = 0 = 0 \\
M_2|M_1 = 1 &\sim \text{Bern}(g^{-1}[(1, X, Z)\beta_{M_2}]) \quad M_2|M_1 = 0 = 0 \\
Y &\sim \text{N}((1, A, X, AZ, AM_1, AM_1M_2)\beta_Y, 1).
\end{aligned}$$

The coefficients $(\beta_{X_1}, \beta_{X_2}, \beta_Y)$ are drawn from a $\text{Unif}(-1, 1)$ distribution, while the coefficient β_A is drawn from a $\text{Unif}(-0.5, 0.5)$ distribution. When evaluating the DML estimation procedure, I set g to be the logistic link: $g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$. While the DML estimator is agnostic to the functional form of g —all nuisance functions, including propensity scores and mediator models, are estimated using flexible machine-learning methods—the RWR estimator relies on parametric linear regressions for both mediator and outcome models. When evaluating the RWR estimator, I therefore set g to be the identity link.⁷

For the DML estimator, for each component involved in the decomposition, I construct

⁷More subtly, even if the outcome model is linear in (M_1, M_2) by construction, the regression of Y on (X, A, M_1, Z) used by RWR is generally misspecified when mediator models are nonlinear. To see this, note that

$$\mathbb{E}[Y | X, A = 1, M_1 = m, Z] = \mathbb{E}[\mathbb{E}[Y | X, A = 1, M_1 = m, M_2, Z] | X, A = 1, M_1 = m, Z],$$

which involves integrating a linear function of M_2 with respect to a nonlinear conditional distribution of $M_2 | X, Z, A = 1, M_1 = 1$. Unless $\mathbb{E}[M_2 | X, Z, A = 1, M_1 = 1]$ is linear in X , this marginal conditional expectation is itself nonlinear in X . Thus, even when the structural outcome model is linear, the reduced-form outcome regression used by RWR will be misspecified unless g is the identity link.

a Neyman-orthogonal “signal” using its EIF. The recentered EIFs for each component are as follows:

$$\begin{aligned}
M_1^*(1) &= \gamma_1(X) + \frac{\mathbb{I}(A=1)}{\pi_0(X,1)} (M_1 - \gamma_1(X)), \\
M_2^*(1,1) &= \gamma_2(X) + \frac{\mathbb{I}(A=1)\mathbb{I}(M_1=1)}{\pi_0(X,1)\pi_1(X,Z,1)} (M_2 - \gamma_2(X)), \\
Y^*(a) &= \mu_0(X,a) + \frac{\mathbb{I}(A=a)}{\pi_0(X,a)} (Y - \mu_0(X,a)), \quad a \in \{0,1\}, \\
Y^*(1,m_1) &= \nu_1(X,m_1) + \frac{\mathbb{I}(A=1)\mathbb{I}(M_1=m_1)}{\pi_0(X,1)\pi_1(X,Z,m_1)} (Y - \mu_1(X,Z,m_1)) \\
&\quad + \frac{\mathbb{I}(A=1)}{\pi_0(X,1)} (\mu_1(X,Z,m_1) - \nu_1(X,m_1)), \quad m_1 \in \{0,1\}, \\
Y^*(1,1,m_2) &= \nu_2(X,m_2) + \frac{\mathbb{I}(A=1)\mathbb{I}(M_1=1)\mathbb{I}(M_2=m_2)}{\pi_0(X,1)\pi_1(X,Z,1)\pi_2(X,Z,m_2)} (Y - \mu_2(X,Z,m_2)) \\
&\quad + \frac{\mathbb{I}(A=1)\mathbb{I}(M_1=1)}{\pi_0(X,1)\pi_1(X,Z,1)} (\mu_2(X,Z,m_2) - \nu_2(X,m_2)), \quad m_2 \in \{0,1\}.
\end{aligned}$$

where

$$\begin{aligned}
\pi_0(X,a) &\triangleq \Pr(A=a \mid X), \\
\pi_1(X,Z,m_1) &\triangleq \Pr(M_1=m_1 \mid X,Z,A=1), \\
\pi_2(X,Z,m_2) &\triangleq \Pr(M_2=m_2 \mid X,Z,A=1,M_1=1), \\
\gamma_1(X) &\triangleq \mathbb{E}[M_1 \mid X, A=1], \\
\gamma_2(X) &\triangleq \mathbb{E}[M_2 \mid X, A=1, M_1=1], \\
\mu_0(X,a) &\triangleq \mathbb{E}[Y \mid X, A=a], \\
\mu_1(X,Z,m_1) &\triangleq \mathbb{E}[Y \mid X, A=1, Z, M_1=m_1], \\
\nu_1(X,m_1) &\triangleq \mathbb{E}[\mu_1(X,Z,m_1) \mid X, A=1], \\
\mu_2(X,Z,m_2) &\triangleq \mathbb{E}[Y \mid X, A=1, Z, M_1=1, M_2=m_2], \\
\nu_2(X,m_2) &\triangleq \mathbb{E}[\mu_2(X,Z,m_2) \mid X, A=1, M_1=1].
\end{aligned}$$

When both the outcome model and the mediator models are linear probability models, these quantities can be derived analytically. For example, under linearity, the continuation effect associated with the second mediator coincides with the coefficient on M_2 in the outcome model, and is used to calculate the MPSE via M_2 . However, under a nonlinear DGP, such closed-form expressions for the θ_k terms no longer exist because the mediator counterfactuals are nonlinear functions of (X, Z) . In this case, I recover the true values of $(\theta_0, \theta_1, \theta_2)$ by Monte Carlo integration under the known data-generating process. Specifically, for each replication, I simulate a large population from the structural equations under the relevant interventions in order to recover the counterfactual mediator probabilities. I then plug these counterfactual probabilities into the corresponding linear expressions implied by the outcome model to obtain the true values of the path-specific effects. These Monte Carlo quantities are treated as the ground truth against which finite-sample bias and coverage are evaluated.

It is worthwhile briefly clarifying how the proposed monotonic path-specific effect (MPSE) decomposition relates to—and departs from—conventional mediation analyses of the ATE with multiple causally ordered mediators. Without mediator monotonicity, the ATE admits an algebraic decomposition into a collection of path-specific effects (PSEs) corresponding to the causal paths $A \rightarrow Y$ and $A \rightarrow M_k \rightsquigarrow Y$ for $k = 1, \dots, K$ (Avin et al., 2005; Daniel et al., 2015; Zhou, 2022b). However, when mediators are causally ordered, only composite effects of the form $A \rightarrow M_k \rightsquigarrow Y$ —which aggregate all downstream pathways from M_k to Y —are generally identifiable.

Under mediator monotonicity, this structure collapses in an important way. Because later mediators are deterministically zero whenever earlier mediators are not realized, path-specific effects of the form $A \rightarrow M_k \rightarrow Y$ for $k \geq 2$ are identically zero (see Section 2.3 in the main text). Consequently, the general PSE decomposition simplifies to a single-mediator natural effect decomposition in which the total indirect effect operates exclusively through the full causal chain $A \rightarrow M_1 \rightarrow \dots \rightarrow M_k \rightarrow Y$. This collapse has two implications for

the simulation results. First, it explains why conventional multi-mediator decompositions do not provide a meaningful benchmark in this setting: once monotonicity is imposed, the distinction between multiple indirect paths disappears, and the estimand reduces to a single composite mediation effect. Second, and more importantly, even this reduced decomposition is not identifiable under standard mediation assumptions when intermediate confounders are present. In the simulation design considered here, these assumptions are deliberately violated by allowing for observed post-treatment confounders Z that affect both later mediators and the outcome.

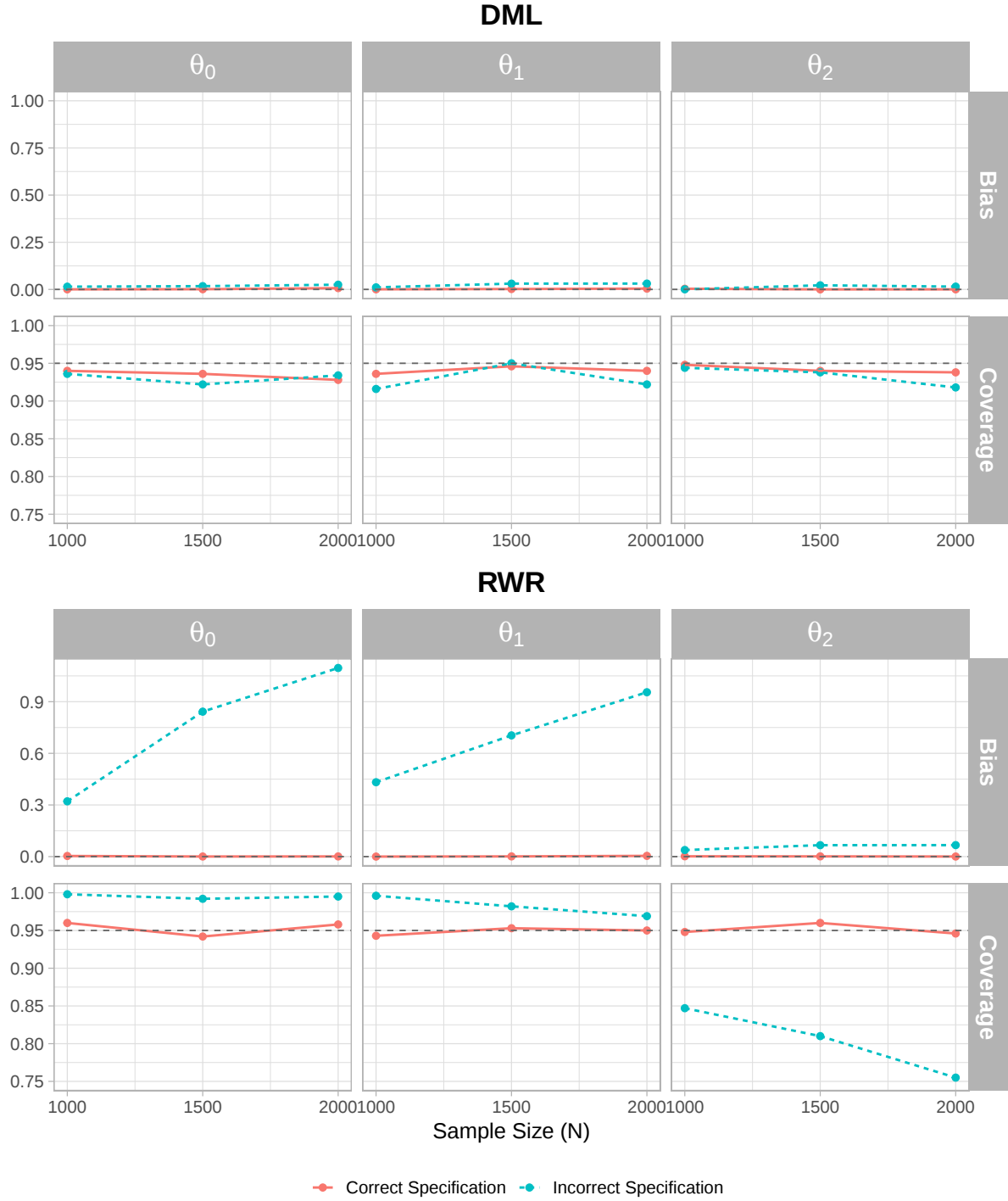


Figure S1: Bias, RMSE, and coverage of DML and RWR estimators for $n = 1000, 1500, 2000$. The red dots show the performance of the DML and RWR estimators when the correct feature matrix is supplied to the estimators; the blue dots show the performance of the two estimators when an incorrect feature matrix is supplied to the estimators.

C Sensitivity analysis

How do my estimates of returns to different educational stages, as presented in the main text, tally with previous findings on the labor market returns to education? While previous work does not estimate quantities analogous to the direct and indirect effects of interest (i.e., the θ_k terms), some prior educational returns estimates are closely related to the net effect (τ_k) terms that inform the total, direct and indirect components of the decomposition. My estimate of the net effect of 4-year college enrollment (τ_1) is large, at 54%, but not implausibly so. While Zimmerman (2014) and Smith et al. (2020) recover college earnings returns at around 20% by age 30 (exploiting admissions discontinuities in the Florida and Georgia state university systems), both of these studies estimate the earnings premium from attending a less selective 4-year college rather than a community college, for the marginally qualified university attendee. By contrast, τ_1 captures the effect of 4-year college enrollment compared with community college *and* no college enrollment, pooling across the less selective colleges examined in Zimmerman (2014) and Smith et al. (2020), as well as over more selective colleges which could have greater earnings effects. Moreover, since τ_1 represents an effect averaged over all individuals, it reflects a return among a broader population than the marginal college-goers examined in previous studies.⁸ As I discuss in the main text, an additional, especially instructive point of comparison are instrumental variable (IV) estimates of returns to years of schooling. These results are in fact quite consistent with those I report in the main text.

Of course, an alternative explanation is that my estimates are upwardly biased by a large degree of unobserved confounding. While the sequential ignorability assumption facilitates

⁸My estimate of Δ_1 (the direct effect of 4-year college attendance on earnings) further tallies with a similar quantity estimated by Scott-Clayton and Wen (2019). On the intensive margin of employment (i.e. dropping respondents with zero observed earnings), the authors estimate a return to college attendance without degree completion of 0.21. While, theoretically, one might expect my estimate—which corresponds to the extensive employment margin (including respondents with zero observed earnings)—to be larger, the fact that it is slightly smaller could reflect several factors, including the richer array of pre-college controls I use in my models, model mis-specification resulting from linearities imposed in prior work and, perhaps most importantly, collider-stratification biases induced by conditioning on BA completion in Scott-Clayton and Wen (2019)’s models (biases that are likely reduced by the inclusion of time-varying controls).

identification of educational effect pathways under a weaker set of conditions than might be typically invoked in mediation settings, it is still strong and fundamentally unverifiable. To assess potential bias of the estimated MPSEs due to unobserved confounders not picked up in my covariate set $(X, \overline{Z}_K)$, I propose a sensitivity analysis for each of the MPSEs.

Assume first that we have a binary unobserved confounder, U , for the treatment-outcome relationship. Assuming that $\alpha_0 = \mathbb{E}[Y|x, a, U = 1] - \mathbb{E}[Y|x, a, U = 0]$ does not depend on x or a , and further that $\beta_0 = \Pr[U = 1|x, A = 1] - \Pr[U = 1|x, A = 0]$ does not depend on x , for $\tau_0 = \mathbb{E}[Y(1) - Y(0)] \triangleq \text{ATE}$, then $\text{bias}(\tau_0) = \alpha\beta$ (VanderWeele and Arah, 2011).

Next, consider an unobserved binary confounder, U_k that affects both M_k and Y for any $k \in \{1, \dots, K\}$. Then, under a weaker instantiation of Assumption 4 (Sequential Ignorability), i.e.,

$$Y(\overline{1}_k, m_k) \perp\!\!\!\perp (A, \overline{M}_k) | X, A, U_k, \overline{Z}_k, \overline{M}_{k-1} \forall k \in [K], \quad (17)$$

which states that potential outcomes under an arbitrary transition sequence are independent of observed treatment and mediator values conditional on observed confounders $(X, \overline{Z}_k)$ and unobserved confounders U_k . Under the following set of assumptions: (Assumption A_k) $\alpha_k = \mathbb{E}[Y|x, \overline{z}_k, \overline{1}_k, m_k, U_k = 1] - \mathbb{E}[Y|x, \overline{z}_k, \overline{1}_k, m_k, U_k = 0]$ does not depend on $(x, \overline{z}_k, \overline{1}_k, m_k)$, and (Assumption B_k), $\beta_k = \Pr[U_k = 1|x, \overline{z}_k, \overline{1}_k, m_k] - \Pr[U_k = 1|x, \overline{z}_k, \overline{1}_k]$ does not depend on $(x, \overline{z}_k)$, we can show that, for any $k \in \{1, \dots, K\}$,

$$\text{bias}(\tau_k) = \alpha_k \beta_k,$$

and, further, that

$$\text{bias}(\Delta_{k-1}) = -\alpha_k \beta_k \pi_k,$$

where $\pi_k = \int_x \int_{\overline{z}_k} \Pr[M_k = 1|x, \overline{z}_k, \overline{1}_k] \prod_{j=1}^k dP(z_j|x, \overline{z}_{j-1}, \overline{1}_{j-1}) dP(x)$, and is estimable

from observed data using the estimation strategies described previously. A contour plot showing bias-adjusted estimates of Δ_{k*} and τ_k then enables assessment of how strong the unobserved confounder would need to be to reduce estimates of the direct and gross effects to zero. I illustrate these techniques in my empirical illustration below.

In order to assess the robustness of my empirical findings in the main text to potential violations of Assumption 4 (Sequential Ignorability), I implement this sensitivity analysis discussed above.

Figure S2 below displays a set of contour plots, which capture the bias-corrected estimates of the (Δ_k, τ_k) terms under varying degrees of confounding (that is, under different values of α_k and β_k). For example, the level set marked “0” corresponds to values of (α_k, β_k) required in order for the unobserved confounder to fully “explain away” estimates (Δ_k, τ_k) (i.e., to reduce their true values to zero). Importantly, each row corresponds to a different set of (α_k, β_k) terms for a given U , such that the top row corresponds to (α_0, β_0) , while the second row corresponds to (α_1, β_1) , and so on.

For simplicity, I consider U to be an unmeasured binary confounder that is (marginally) positively associated with each transition $A, M_1, \dots, M_3$ as well as with adult earnings Y . To benchmark the hypothetical behavior of U , for each plot, I also display the values of (α_k, β_k) that would correspond to a U that behaved similarly to a given confounder that I do observe in the data: an indicator for whether an individual’s test score on the ASVAB is above the median. In each plot, I mark this point and label it “Ability”. For each plot, I also mark the point on the zero contour that corresponds to $\alpha_k = \beta_k$ (i.e, the point at which the unobserved confounder’s associations with the treatment and with the outcome are equal, and reduce the true value of the parameter to zero).

I focus on estimates of τ_0 and Δ_0 to assess how robust my primary conclusion—that the ATE of high school completion is overwhelmingly mediated by high school’s direct effect on earnings—is to unobserved confounding. Bias-adjusted estimates of the ATE τ_0 are presented in the top row of Figure S2. Since U is assumed to be positively associated with

both A and Y , τ_0 is overestimated and suffers from a bias of $\alpha_0\beta_0$. My estimate of τ_0 at 0.67 is nevertheless quite robust: if U had similar effects to ability, the effect would be reduced by 0.06 log points, to 0.61, still implying a high earnings premium to high school completion overall in excess of 84%.

How do my estimates of the direct effect of high school completion Δ_0 (and, in particular, about the proportion of the total effect that is direct) fare under unobserved confounding? The second row of Figure S2 considers bias-adjusted estimates of $\Delta_0 = \theta_0$ under different values of (α_1, β_1) , which correspond to the effects of an unobserved confounder U (marginally) positively associated with both M_1 and with Y . As described above, in this scenario, Δ_0 is affected by a bias of $-\alpha_1\beta_1\pi_1$. Importantly, even if the unobserved confounder U is *marginally* positively associated with high school graduation (A), the conditional association between U and A may be zero or even negative since M_1 is a collider of A and U . In the case that the conditional association between U and A is negative, $-\alpha_1\beta_1\pi_1$ would be positive, implying an overestimation of the direct effect θ_0 . On the plot, I show estimates of U if it behaved similarly to the ability variable. Indeed, despite the fact that ability is marginally positively associated with high school completion (top row of Figure S2), its conditional association—conditional on college attendance—is depressed to zero. Thus, it would take an extreme form of confounding for θ_0 to be largely different from its estimated value of .46. In this way, my primary finding that the ATE of high school graduation is overwhelmingly mediated via its direct effect remains highly robust to patterns of unobserved confounding, under my set of simplifying assumptions.

The bias-factor sensitivity analysis presented above has the drawback of assessing only the influence of a single unmeasured confounder at a time, for a given transition. In practice, however, confounders may be unmeasured not only at a given transition but also at earlier ones. Unfortunately, the bias-factor approach scales poorly here, requiring $2K$ parameters (a sensitivity-parameter pair (α_k, β_k) for each of the K transitions).

Therefore, to assess how sensitive later continuation effects are to multiple-transition confounding, I undertake a calibrated simulation analysis. Specifically, I simulate a data-generating process that mirrors the empirical setting (A = high school completion, M_1 = college attendance, M_2 = BA completion, M_3 = graduate schooling, Y = log earnings), augmented with one unobserved baseline confounder (X_3) and one unobserved intermediate confounder revealed at each subsequent transition (Z_1, Z_2, Z_3). The strengths of these confounders are *calibrated* to one of two observed covariates in the NLSY97 analytic sample, and I then re-estimate Δ_3 while cumulatively omitting them from the adjustment set.

I simulate from an expanded version of the data-generating process presented in the Simulation Study in the main text, additionally drawing a third mediator M_3 and an intermediate confounder Z_3 , and letting Y depend on these variables as well. I report two calibrations, calibrating all four unobserved confounders to the partial associations of (i) AFQT and (ii) high-school GPA with the outcome and with each transition.

For each of $B = 500$ simulated samples, I estimate Δ_3 using the regression-with-residuals (RWR) estimator under five nested adjustment sets: full adjustment (S_0 ; includes X_3, Z_1, Z_2, Z_3), then cumulatively omitting X_3 (S_1), Z_1 (S_2), Z_2 (S_3), and Z_3 (S_4). Under S_0 the estimator recovers the true value of Δ_3 .

Figure S3 presents the results. Under either calibration, even the full history of unobserved confounders—one at every stage, each as strong (at every transition it affects) as AFQT or high-school GPA—inflates the estimate by only 0.012–0.015, i.e., roughly 8–9% of the observed value. The implied true continuation effect remains at approximately .15. The estimate is therefore fairly robust to unobserved confounding of a magnitude compa-

table to that of observed covariates. In any case, because Δ_3 remains a modest component of the total effect, confounding would not alter the central takeaway – that the effect of high-school completion on earnings operates primarily through its direct effect rather than through graduate schooling – and if anything, would reinforce this conclusion.

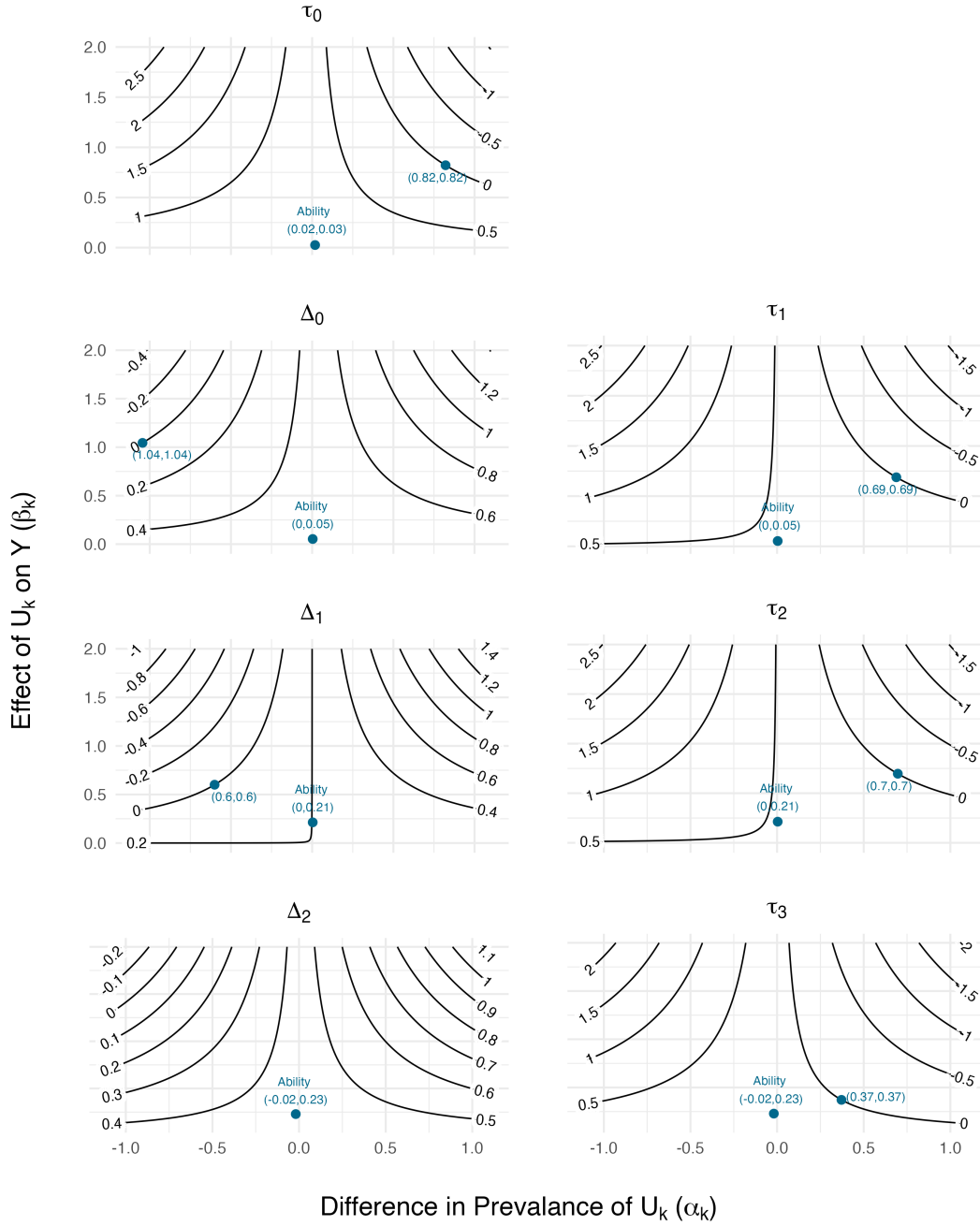


Figure S2: Sensitivity Analysis for the “gross effect” (τ_k) and “direct effect” (Δ_k) terms in decomposition. Each row corresponds to a different set of (α_k, β_k) terms, where $\alpha_k = \mathbb{E}[Y|x, \bar{z}_k, \bar{I}_k, m_k, U_k = 1] - \mathbb{E}[Y|x, \bar{z}_k, \bar{I}_k, m_k, U_k = 0]$ parameterizes the effect of U_k on Y , and $\beta_k = \Pr[U_k = 1|x, \bar{z}_k, \bar{I}_k, m_k] - \Pr[U_k = 1|x, \bar{z}_k, \bar{I}_k]$ parameterizes the effect of M_k on U_k . Each row corresponds to a different set of (α_k, β_k) terms. For example, the top row corresponds to (α_0, β_0) , while the second row corresponds to (α_1, β_1) , and so on.

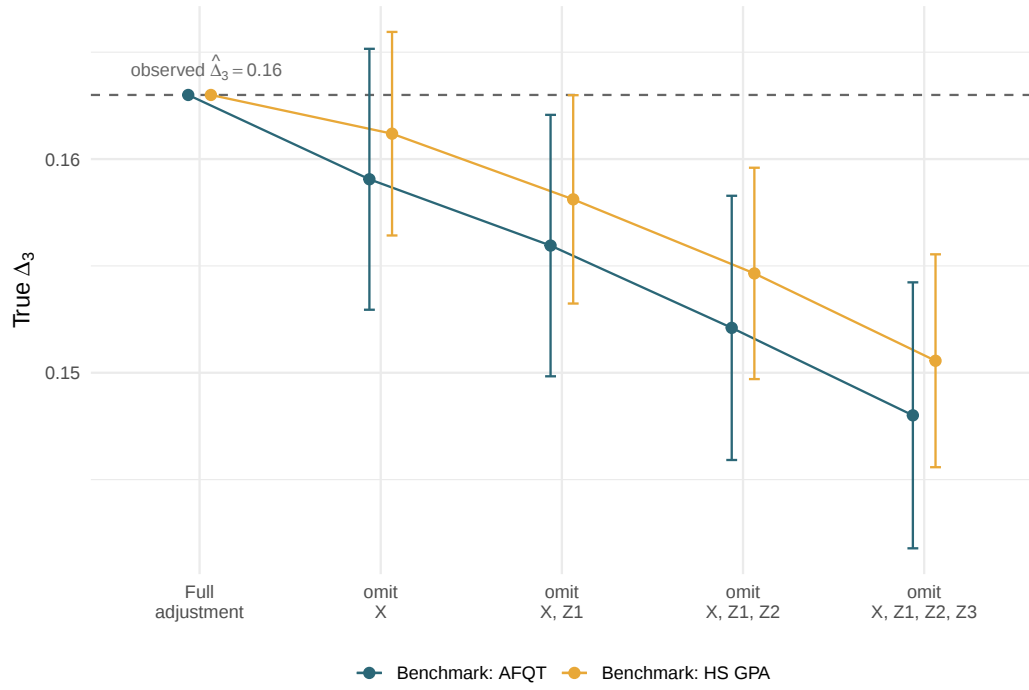


Figure S3: Calibrated sensitivity analysis of Δ_3 to cumulative multiple-transition confounding. The observed estimate $\hat{\Delta}_3$ (dashed line) is held fixed; each point plots the implied true Δ_3 under a scenario that leaves the labeled confounders unmeasured, for confounders calibrated to AFQT (dark teal) and high-school GPA (gold). Bars are 95% Monte Carlo intervals ($B = 500$, $n = 2,000$; RWR estimator).

D Further details on variable construction and education groups

Variable construction

In an effort to satisfy the sequential ignorability assumption (Assumption 4), I include a large array of covariates in my models for the effects of completing educational transitions on labor market outcomes. Figure S4 summarizes my assumed data-generating process for the empirical example. In addition to including information on respondent demographics (gender, race, ethnicity, age at 1997), and observed pre-college performance such as overall high school GPA and test score on the Armed Services Vocational Aptitude Battery (ASVAB), I include detailed information on socioeconomic background (parental education, parental income, parental assets, co-residence with both biological parents, presence of a paternal figure, rural residence, southern residence), an index of substance use, an index of delinquency, and whether the respondent had any children by age 18, and peer and school-level characteristics (measures of peers' college expectations and behaviors). Both parental income and parental asset variables are transformed to 2023 dollars.

Since my proposed decomposition also facilitates the inclusion of a distinct set of observed intermediate confounders for each transition to adjust for selection processes that may confound the causal effects of each transition on earnings (i.e., the $A - Y$ and $M_k - Y$ relationships, for $k \in \{1, \dots, K\}$), I include two postsecondary characteristics (Z) to adjust for confounders of the effect of BA completion and graduate school attendance on earnings, namely, field of study, and college GPA. Specifically, I use college self-reported major field of study, drawing on the NLSY survey instrument asking respondents about their choice of major in each month in which they were enrolled in college, and using a dummy variable to denote whether a respondent majored in a STEM or non-STEM field by age 29. Finally, college GPA is measured using the respondent's cumulative GPA from the Post-Secondary Transcript Study. I treat two of the Z_k sets as empty (namely, Z_1 and Z_3), assuming

that the effects of the first mediator (college attendance, M_1) on subsequent transitions and adult earnings are unconfounded given background characteristics (X), and that the effects of the third mediator (graduate school attendance, M_3) are unconfounded given background characteristics (X) and postsecondary characteristics (Z).⁹

How convincingly do I satisfy the sequential ignorability assumption? Despite the inclusion of a comprehensive set of background covariates in my models, it is possible that observed variables do not perfectly proxy for all important confounders jointly affecting education and earnings. In particular, researchers often argue that important variables, such as students' innate ability, ambition, and detailed forms of socioeconomic advantage, confound observational estimates of educational returns (e.g. Carneiro et al., 2011). While some research suggests that observational estimates of earnings returns may well capture actual returns to education and that the degree of observational bias may be rather small (Card, 1999), it is of course impossible to quantify the true extent of the bias in the estimates I produce. The sensitivity analysis described above provides a step towards this goal.

I note that my assumption of ignorability of M_3 without conditioning on intermediate variables Z_3 is perhaps the strongest assumption I make. For example, many individuals take time off to work before enrolling in graduate school, and labor market experience and earnings gained in the interim period between college completion and graduate school enrollment may confound the latter variable's effects on earnings. Nevertheless, including a measure of labor market characteristics for this period is difficult because some respondents enroll directly in graduate school after BA completion, such that pre-graduate school earnings variables would be undefined for these individuals.

Table S1 shows conditional means of respondent attributes X and Z for the full (im-

⁹To be clear, assuming that Z_1 and Z_3 are empty is not to say that M_1 and M_3 are marginally unconfounded; rather, it means that the set of covariates that confound the effects of M_1 is assumed to be the same as those that confound the effects of A , and that the set of covariates that confound the effects of M_3 are assumed to be the same as those that confound the effects of M_2 . This assumption is in part data-driven, given the few variables observed chronologically post high school graduation and pre college attendance.

puted) and restricted (non-imputed) samples, showing first the mean among the full population of high school goers, and progressively restricting the sample from (i) high school (HS) non-completers, to (ii) HS graduates, to (iii) college attendees and, finally, to (iv) BA completers. Imputed and non-imputed means—shown without and with brackets, respectively—are highly similar across variables. As I progressively restrict the sample to those who attained higher educational levels, variables capturing components of socioeconomic advantage (such as parental income, parental education and household net worth) increase monotonically in value. Background covariates measuring aspects of the school environment (such as peers’ college expectations—which is an indicator for whether over 90% of a respondent’s peers expected to go to college) behave similarly. I also see that students who progress to higher educational levels have higher levels of pre-college ability: HS non-completers have on average an ASVAB Percentile score of 22.3, compared with only in excess of 70 among BA completers. Similarly, college-goers’ average high school GPA is approximately .5 higher than high school graduates overall (regardless of whether or not they proceed to college). Nevertheless, the association between high school GPA and attainment declines at higher educational levels: BA completers have only on average a .11 higher a high school GPA than the pooled group of college goers, irrespective of their BA completion status. At this stage, college GPA appears to matter more: college goers overall have on average a college GPA of 2.77, while BA completers’ average college GPA is 3.07.

Educational groups: raw mean earnings

Table S2 (column 2) presents the proportion of individuals who have attained each level of education constructed above. By age 22, a small, but not insignificant, proportion of individuals who enroll in high school do not complete their studies (13%), and by this same age, just over 40% of individuals have attended a 4-year college. By age 29, 29% of individuals have attained a Bachelor’s degree or higher. These estimates of high school

completion and BA completion align closely both with those reported in previous studies that employ the NLSY97 (e.g. Scott-Clayton and Wen (2019)), as well as with those reported in the Current Population Survey (CPS). Table S2 (columns 3-4) also shows mean log earnings by educational group (column 3), alongside the estimated gap between these means and mean log earnings among high school non-completers (column 4). High school dropouts earn an average of 9.07 log earnings, while groups with higher levels of attainment earn successively more than high school dropouts, though at a decreasing rate. High school graduates earn on average 1.11 log earnings more than high school non-completers, implying an earnings premium in excess of 200% ($\exp(1.11) - 1$), while college goers earn on average 1.5 log earnings more than high school non-completers (or 0.6 log earnings more than high school graduates). At the highest end, graduate school goers earn on average 10.88 log earnings. These educational premia are extremely high, since they reflect both the causal effect of a given educational level as well as the effects of individual, geographic and family factors correlated both with attainment and with adult earnings. To net out these patterns of selection, we need to turn to estimates of the MPSE decomposition, as well as its constituent components.

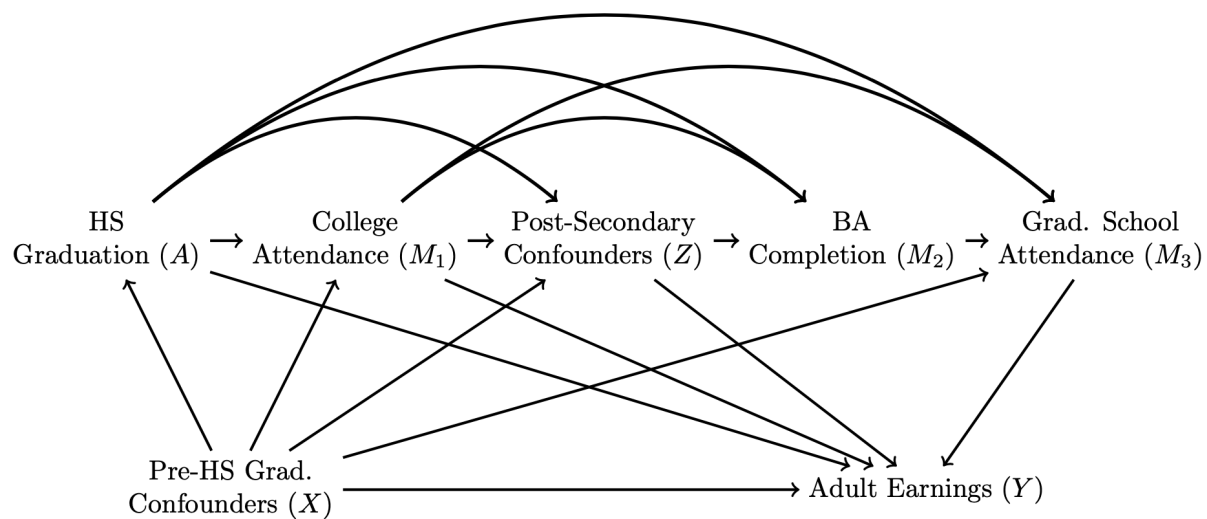


Figure S4: DAG showing the hypothesized causal relationships between high school completion A and adult earnings Y via mediators M_1 , M_2 and M_3 .

Table S1: Conditional means of background and college-level attributes by sample type (imputed dataset and non-imputed dataset).

	Full Population	HS Non-Completers	HS Graduates	College Goers	BA Completers
Female	0.50 (0.50)	0.44 (0.47)	0.51 (0.51)	0.55 (0.56)	0.57 (0.58)
Black	0.17 (0.16)	0.25 (0.24)	0.16 (0.15)	0.13 (0.10)	0.10 (0.09)
Hispanic	0.14 (0.12)	0.20 (0.18)	0.13 (0.12)	0.09 (0.08)	0.08 (0.07)
Parental Income	80,235 (79,243)	43,308 (43,863)	85,898 (83,865)	110,240 (109,636)	118,026 (114,329)
Parental Education	12.85 (12.84)	11.18 (11.24)	13.11 (13.04)	14.06 (14.13)	14.38 (14.34)
Household Net Worth	175,513 (176,097)	61,997 (62,318)	192,920 (190,959)	272,606 (284,029)	306,177 (307,101)
Lived w Biological Parents	0.52 (0.51)	0.30 (0.28)	0.55 (0.55)	0.67 (0.69)	0.72 (0.73)
Father Figure Present	0.75 (0.75)	0.61 (0.61)	0.77 (0.76)	0.84 (0.83)	0.86 (0.86)
Lived in Rural Area	0.28 (0.30)	0.25 (0.27)	0.28 (0.31)	0.28 (0.32)	0.28 (0.31)
Lived in South	0.36 (0.35)	0.44 (0.42)	0.35 (0.34)	0.33 (0.30)	0.31 (0.29)
Children by 18	0.07 (0.07)	0.21 (0.24)	0.05 (0.05)	0.01 (0.01)	0.01 (0.01)
Substance Abuse Score	1.09 (1.10)	1.32 (1.37)	1.06 (1.07)	0.86 (0.82)	0.82 (0.77)
Delinquency Score	1.38 (1.39)	2.07 (2.13)	1.27 (1.29)	0.91 (0.88)	0.81 (0.79)
Peers' College Expectations	0.56 (0.56)	0.40 (0.38)	0.59 (0.58)	0.69 (0.69)	0.72 (0.71)
Property Stolen at School	0.24 (0.23)	0.28 (0.29)	0.23 (0.22)	0.21 (0.20)	0.20 (0.19)
Threatened at School	0.22 (0.23)	0.30 (0.33)	0.20 (0.21)	0.14 (0.13)	0.13 (0.11)
In a Fight at School	0.16 (0.16)	0.32 (0.34)	0.14 (0.13)	0.07 (0.07)	0.06 (0.06)
ASVAB Percentile	48.36 (48.52)	22.27 (21.77)	52.37 (52.02)	67.22 (69.66)	70.66 (71.82)
High School GPA	2.84 (2.81)	2.12 (2.14)	2.95 (2.90)	3.30 (3.33)	3.41 (3.43)
Stem Major				0.17 (0.19)	0.18 (0.19)
College GPA				2.78 (2.86)	3.07 (3.12)
Earnings (\$)	46,505 (45,151)	21,597 (21,566)	46,505 (45,151)	64,984 (64,595)	72,374 (70,212)
Log (Earnings + c)	10.03 (10.06)	9.07 (9.14)	10.03 (10.06)	10.57 (10.65)	10.73 (10.78)

Note: Numbers denote means for the imputed sample (non-imputed sample). Means are adjusted for multiple imputation via Rubin's (1987) method, and all statistics are calculated using NLSY97 sampling weights.

Table S2: Means of observed log earnings by educational participation, and earnings gaps (versus high school non-completers).

Group	Population Proportion	Log Earnings	Gap (vs HS Non-Completers)
HS Non-Completers	0.13	9.07 (0.05)	
HS Graduates	0.87	10.18 (0.02)	1.11 (0.05)
College Goers	0.41	10.57 (0.03)	1.5 (0.05)
BA Completers	0.29	10.73 (0.03)	1.66 (0.06)
Grad. School Goers	0.09	10.88 (0.05)	1.81 (0.07)

Note: The category "High School Non-Completers" captures all individuals who attended high school but did not obtain a high school diploma; "High School Graduates" refers to those individuals who graduated high school, regardless of their subsequent educational experiences (i.e., whether or not they proceeded to college); "College Goers" refers to individuals who attended a 4-year college, irrespective of whether they completed their degree; "BA Completers" denotes individuals who completed a Bachelor's degree, while "Grad. School Goers" captures individuals who participated in a graduate-level degree program. A small constant of \$1,000 is added to observed earnings before taking the log. All statistics are computed with a monotonicity assumption imposed on the observed data (i.e. such that all individuals who complete a given educational level are coded as having completed all prior levels). All statistics are calculated using NLSY97 sampling weights, and standard errors are in parentheses.

E Results without imputation of missing covariates

In the main text, I report estimates of the MPSE decomposition for the ATE of high school completion on logged annual earnings for the full sample of NLSY97 respondents with non-missing educational information and non-missing earnings ($N = 7,305$). A very large number (approximately 50%) of these respondents are missing information on one or more of the covariates (X, Z) used in the models in order to identify the decomposition components. Table S3 summarizes missingness patterns for the background covariates X by respondents' educational attainment—which are all self-reported by respondents. Missingness is generally modest for most pretreatment covariates, though certain variables (e.g., household net worth and ASVAB percentile) exhibit higher rates of non-response (at 25% and 20% of the sample overall, respectively). Non-response is more common among respondents with lower levels of schooling.

Table S4 clarifies the nature of missingness in the intermediate covariates Z (STEM major and college GPA). These intermediate covariates are observed only for respondents who attend college, but even within this group missingness is substantial. One source of missingness is transcript non-receipt: these variables are derived from the NLSY97 Postsecondary Transcript Study, and are therefore observed only for respondents for whom a transcript was successfully obtained. Transcript non-response is the primary driver of missingness: institutions sometimes did not supply transcripts, supplied incomplete records, or provided degree-program information without course-level grades.¹⁰ Another potential source of missingness may be that students drop out of college before declaring a major. Consistent with this, among college-goers who do not complete a BA, approximately 10% are missing information on degree major, compared with 1.5% of BA completers. This pattern raises a natural concern for my strategy to impute missing values for STEM. If missingness reflects a student's failure to declare a major, then this missingness should be reflected as *true*

¹⁰See *NLSY97 Appendix 12: Postsecondary Transcript Study Documentation* (<https://www.nlsinfo.org/content/cohorts/nlsy97/other-documentation/codebook-supplement/appendix-12-post-secondary-transcript-study>).

missingness in an additional level of the STEM variable, rather than being approached as a missing data problem. At the same time, it is difficult to disentangle true non-declaration of a major from missingness caused by institutional non-compliance with transcript requests; students who leave college early may also disproportionately attend institutions that are less responsive to the NLSY’s Postsecondary Transcript Study. As a result, the observed missingness conflates measurement limitations with underlying educational progression in ways that cannot be fully separated.

To assess the sensitivity of my primary conclusions to these issues, I undertake two exercises. First, I replicate my DML and RWR estimates on a non-imputed analytic sample, dropping observations with missing values ($N = 3,735$). Figure S5 below shows the results of this exercise. For both estimation procedures, results under multiple imputation and non-imputation are highly similar. As is to be expected, imputation reduces standard errors significantly, especially for the parametric RWR procedure. Further, the greatest variability between imputed and non-imputed results comes from effects pertaining to high school completion, perhaps because patterns of missingness are correlated with educational attainment. Despite this, because the total effect τ_0 and direct effect θ_0 are similarly attenuated in the imputed sample, the overall conclusion about the importance of the direct effect in explaining the ATE remains unaffected.

Second, to assess whether missingness in field of study reflects meaningful differences in educational progression versus measurement limitations, I re-estimated the entire MPSE decomposition excluding the STEM indicator from the intermediate confounder set Z . The resulting estimates (presented in Table S5) are nearly identical to those reported in the main text, indicating that the main empirical findings are not sensitive to whether field of study is included in the intermediate confounder set. Nevertheless, the question of whether major declaration itself constitutes an important intermediate transition (prior to BA completion) is an important direction for future research.

Table S3: Proportion of missing background covariates by educational level

		Parental income	Parental education	Household net worth	ASVAB percentile	High school GPA	Peers' expectations (75th)	Peers' expectations (90th)	Property stolen at school
HS Non-Completers	1144	0.058	0.069	0.233	0.302	0.097	0.021	0.021	0.034
HS Graduates	3443	0.033	0.036	0.238	0.203	0.008	0.015	0.015	0.011
College Goers	885	0.040	0.030	0.270	0.159	0.016	0.006	0.006	0.005
BA Completers	1246	0.019	0.023	0.255	0.127	0.009	0.007	0.007	0.006
Grad. School Goers	587	0.029	0.020	0.307	0.148	0.019	0.007	0.007	0.005

Table S4: Proportion of missing intermediate covariates (transcript-based) by educational level

	N	STEM major	College GPA
HS Non-Completers	1144	1.000	1.000
HS Graduates	3443	1.000	1.000
College Goers	885	0.103	0.357
BA Completers	1246	0.015	0.272
Grad. School Goers	587	0.019	0.274

Table S5: Direct Effects (Δ_k), Probabilities (π_k) and Covariance Terms (η_k) Involved in Decomposition via Debiased Machine-Learning (DML), without STEM degree.

	Δ_0	Δ_1	Δ_2	Δ_3	π_1	π_2	π_3	η_1	η_2	η_3
DML	0.462	0.197	0.442	0.145	0.427	0.554	0.313	0.006	0.007	-0.001
	(0.059)	(0.034)	(0.043)	(0.039)	(0.009)	(0.014)	(0.017)	(0.007)	(0.007)	(0.009)

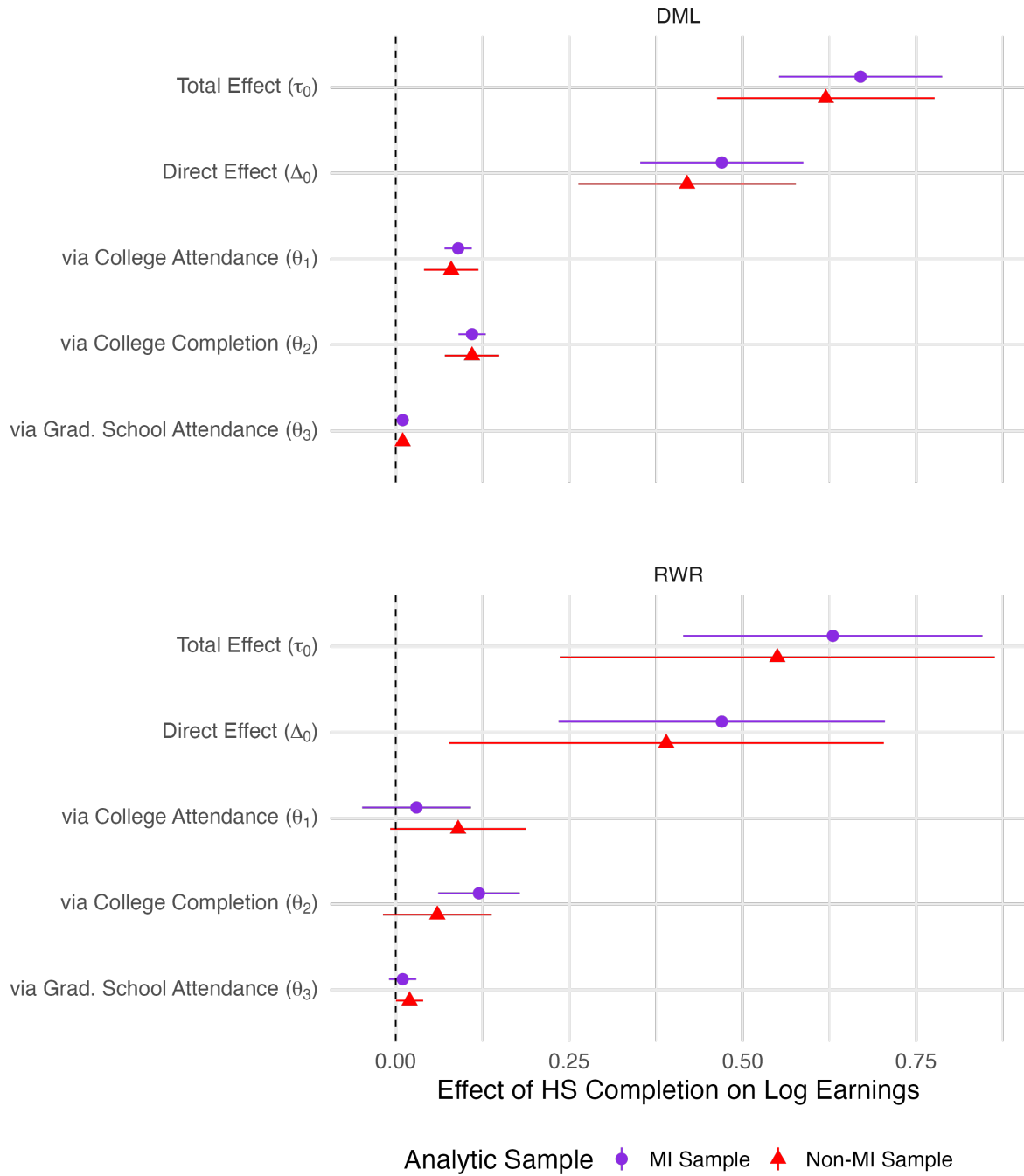


Figure S5: Decomposition of the Average Total Effect (ATE) of High School Graduation on Logged Earnings Under Multiple Imputation (MI) and Under Dropping Observations with Missing (X, Z) Values. Results with multiple imputation (purple lines) are reproduced from the main text ($N = 7,305$); results without multiple imputation (red lines) employ a sample restricted to respondents with observed values for all covariates used ($N = 3,735$).

F Results under alternative definitions of earnings

In the main text, I report estimates of the MPSE decomposition components for the ATE of high school completion on logged annual earnings. Logged annual earnings in the main text are defined as the log of observed annual earnings plus a small constant of \$1,000, in order to accommodate respondents with zero observed annual earnings. In order to assess the sensitivity of the reported results to the choice of this constant, I replicate the main analyses under alternative definitions of earnings. Figure S6 reports estimates of the direct and indirect effects under a series of different constants c added to pre-logged annual earnings, for $c \in \{10, 100, 1000\}$, while Figure S7 shows estimates of these direct and indirect effects for observed annual earnings in dollar values. Beginning with Figure S6, we see that, while for the indirect effects θ_1 , θ_2 , and θ_3 , both DML and RWR estimates are quite consistent under these different constants, estimates of the total effect τ_0 as well as the direct effect Δ_0 are quite sensitive to the choice of constant. Specifically, lower constant values correspond with large increases in the DML estimate of τ_0 from 0.67 ($c = 1000$) to 1.20 ($c = 10$), and of Δ_0 from 0.47 ($c = 1000$) to 0.89 ($c = 10$). This is because individuals with less than a high school degree are more likely than their higher-educated counterparts to have zero or low earnings, making their logged earnings rather sensitive to the choice of constant. Nevertheless, because the total effect τ_0 and direct effect θ_0 are similarly affected by the change in constant value, the importance of the direct effect in explaining the ATE of high school completion is reinforced. In particular, the proportion of the total effect that is direct is estimated to be roughly 70% under each of $c = 10, 100, 1000$. Turning next to Figure S7, under DML, we estimate that high school completion increases earnings in expectation by roughly \$16,500, corresponding to an earnings return of approximately 53% relative to a baseline of \$31,300 without high school graduation ($\mathbb{E}[Y(0)]$). Almost half ($\frac{7824}{16454} \cdot 100 = 47.5\%$) of the total effect is estimated to operate directly, with 29% and 23% mediated via college attendance and college completion, respectively.

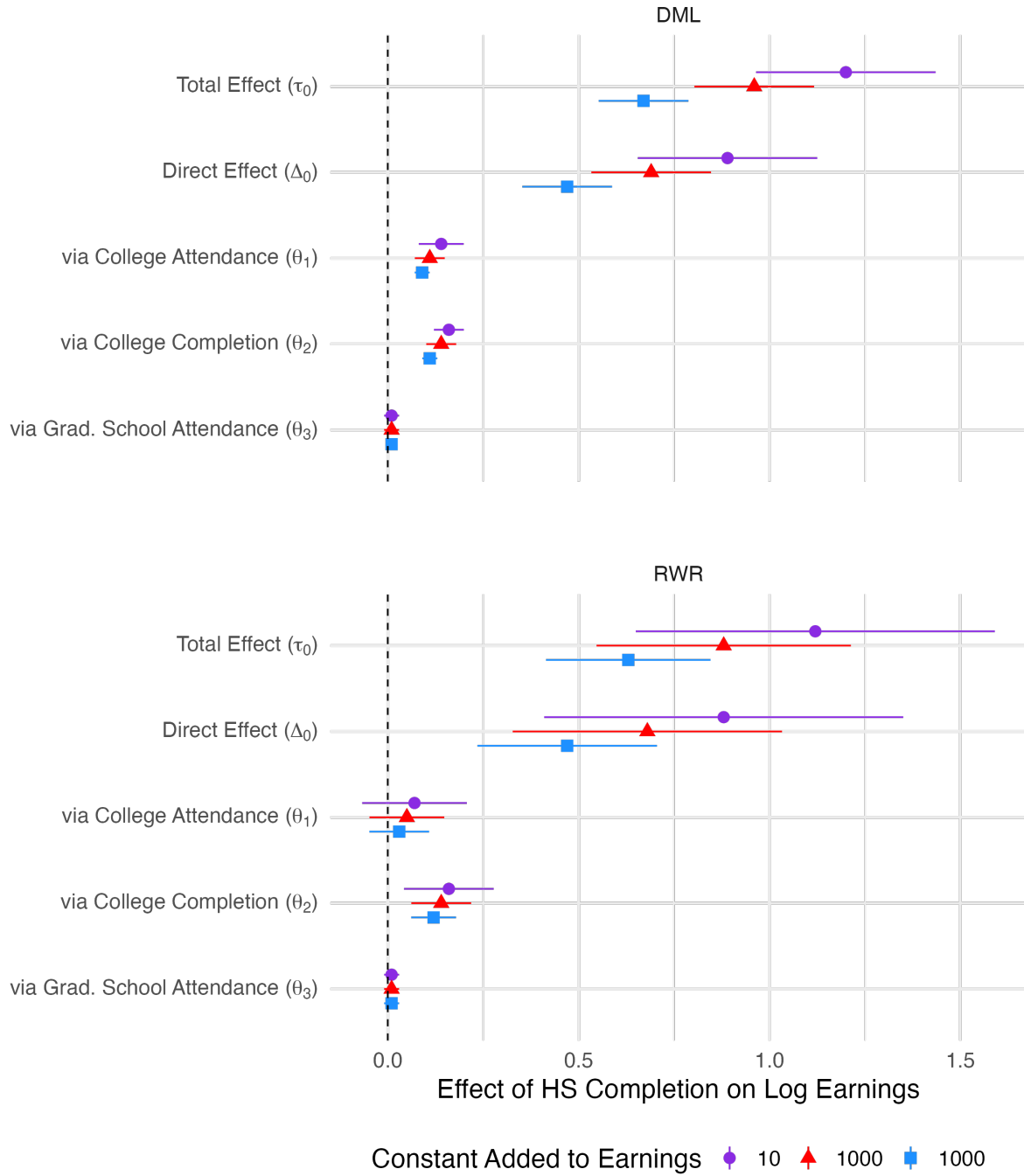


Figure S6: Decomposition of the Average Total Effect (ATE) of High School Graduation on Logged Earnings Under Alternative Definitions of Earnings. The figure shows estimates of the total effect (τ_0) as well as the continuation effects $\Delta_0, \tau_1, \dots, \tau_K$ when constants of 10, 100 and 1000, respectively, are added to raw annual earnings (in dollar amounts) before taking the log.

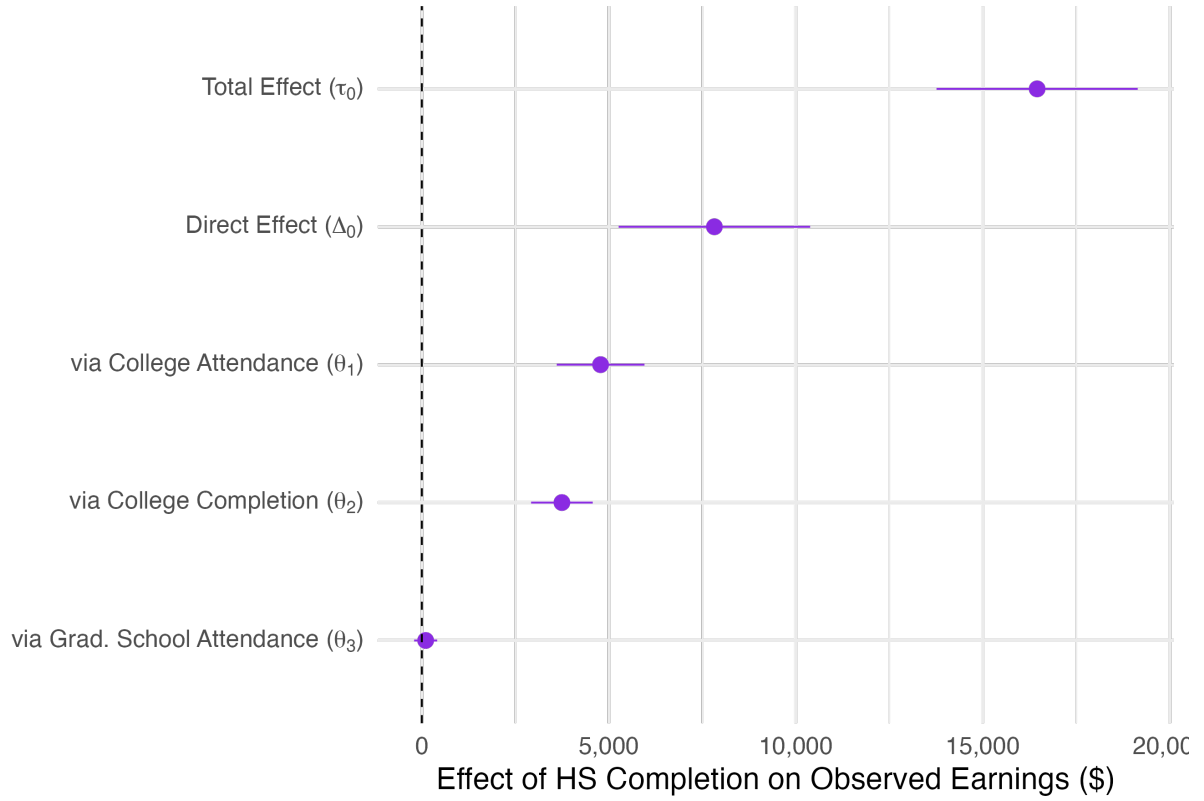


Figure S7: Decomposition of the Average Total Effect (ATE) of High School Graduation on Logged Earnings with Different Definitions of Earnings. The figure shows estimates of the total effect (τ_0) as well as the continuation effects $\Delta_0, \tau_1, \dots, \tau_K$ when constants of 10, 100 and 1000 are added to raw annual earnings (in dollar amounts) before taking the log.

G Results under different positivity assumptions

Under the positivity assumption maintained in the main text, the decomposition components (the Δ , π , and η terms, and hence τ_0 and $\theta_1, \dots, \theta_3$) are nonparametrically identified, and I estimate them by both parametric RWR and DML.

In a second analysis, I relax the positivity assumption by allowing some units to have transition propensities of exactly zero or one. These units need to be excluded from the analysis, but because the true propensities are unknown, I operationalize this stage by stage, trimming units whose estimated propensity $\hat{\pi}(X_i)$ falls below $\min_{i: T_i=1} \hat{\pi}(X_i)$ or above $\max_{i: T_i=0} \hat{\pi}(X_i)$ for that transition’s indicator T , retaining a unit only if it lies within the common-support region of every transition it reaches. Concretely, I trim below/above $\min_{i: A_i=1} \hat{\pi}(X_i) / \max_{i: A_i=0} \hat{\pi}(X_i)$ at high-school completion; $\min_{i: M_{1i}=1, A_i=1} \hat{\pi}(X_i) / \max_{i: M_{1i}=0, A_i=1} \hat{\pi}(X_i)$ at college attendance; $\min_{i: M_{2i}=1, M_{1i}=1, A_i=1} \hat{\pi}(X_i) / \max_{i: M_{2i}=0, M_{1i}=1, A_i=1} \hat{\pi}(X_i)$ at BA completion; and $\min_{i: M_{3i}=1, M_{2i}=1, M_{1i}=1, A_i=1} \hat{\pi}(X_i) / \max_{i: M_{3i}=0, M_{2i}=1, M_{1i}=1, A_i=1} \hat{\pi}(X_i)$ at graduate-school attendance. The propensity scores $\hat{\pi}(X_i)$ are estimated using a super learner composed of the Lasso and random forests.

Averaging over the imputed datasets, these bounds equal (0.162, 0.995) at high-school completion, (0.022, 0.979) at college attendance, (0.092, 0.965) at BA completion, and (0.166, 0.583) at graduate-school attendance, leading me to exclude, respectively, 590 completers and 4 non-completers; 94 college attendees and 16 non-attendees; 28 BA completers and 8 non-completers; and 3 graduate-school attendees and 5 non-attendees—626 of 7,305 units in total, leaving a common-support sample of 6,679. Note that this trimmed sample no longer yields an estimate of the ATE, but rather a conditional treatment effect (and corollary decomposition) for units that have a non-zero probability of attaining or not attaining each transition.

Comparing results from the full sample (assuming positivity) and the common-support sample (relaxing positivity) in Table S6 and Figure S8 shows that trimming has little effect on the substantive conclusions. Under both parametric RWR and DML the total effect

of high-school completion on log earnings is essentially unchanged (RWR: 0.628 vs. 0.653; DML: 0.675 vs. 0.643), as is the direct effect θ_0 , which dominates the decomposition and accounts for roughly 70–75% of the total (RWR: 0.469 vs. 0.507; DML: 0.466 vs. 0.464). The continuation effects are largely stable: in particular, the graduate-school pathway θ_3 is negligible throughout. The largest change among the underlying components is in Δ_3 , the graduate-school gross-effect, which falls from 0.138 to 0.080 under DML (it is essentially unchanged under RWR); this estimate is imprecise, however, and its standard error roughly doubles (from 0.040 to 0.075). More generally, standard errors are larger in the trimmed sample.

Overall, excluding units with extreme estimated propensity scores does not materially alter the main conclusion that the effect of high-school completion on earnings is predominantly direct, with secondary contributions operating through college completion and graduate school.

Table S6: Direct effects (Δ_k), transition probabilities (π_k), and covariance terms (η_k) in the decomposition, estimated by DML and parametric RWR, assuming positivity (full sample) and relaxing positivity (common-support / trimmed sample). Standard errors, adjusted for multiple imputation via Rubin’s (1987) method, in parentheses.

	Δ_0	Δ_1	Δ_2	Δ_3	π_1	π_2	π_3	η_1	η_2	η_3
DML, Assuming	0.466	0.200	0.444	0.138	0.427	0.555	0.314	0.007	0.004	-0.002
	(0.057)	(0.031)	(0.034)	(0.040)	(0.007)	(0.008)	(0.008)	(0.008)	(0.016)	(0.011)
DML, Relaxing	0.464	0.239	0.464	0.080	0.360	0.519	0.307	0.003	-0.006	0.000
	(0.062)	(0.040)	(0.045)	(0.075)	(0.023)	(0.018)	(0.011)	(0.008)	(0.015)	(0.021)
RWR, Assuming	0.469	0.116	0.466	0.163	0.374	0.515	0.159	-0.016	0.083	0.028
	(0.115)	(0.093)	(0.097)	(0.114)	(0.007)	(0.015)	(0.051)	(0.015)	(0.077)	(0.022)
RWR, Relaxing	0.507	0.176	0.462	0.176	0.320	0.479	0.211	-0.002	0.040	0.014
	(0.100)	(0.106)	(0.124)	(0.174)	(0.019)	(0.021)	(0.066)	(0.013)	(0.085)	(0.029)

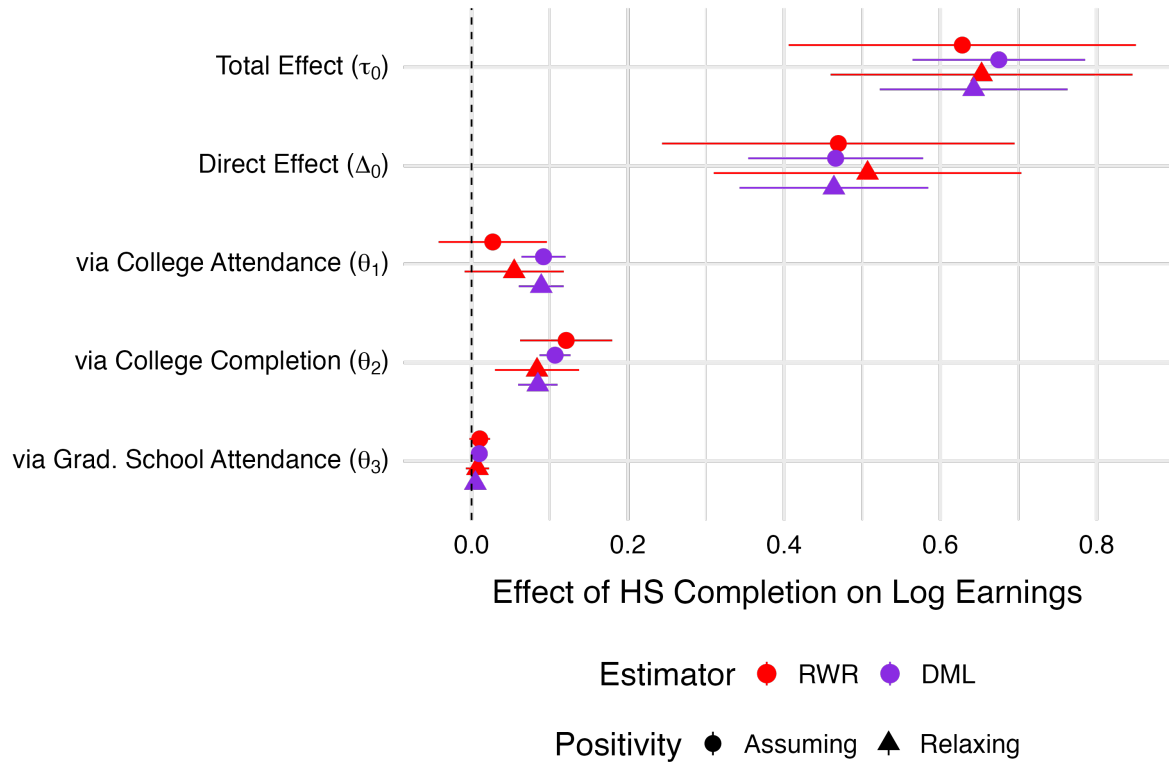


Figure S8: Total effect (τ_0), direct effect (θ_0), and pathway-specific continuation effects via college attendance (θ_1), college completion (θ_2), and graduate-school attendance (θ_3) of high-school completion on log earnings, estimated by parametric RWR and DML, assuming positivity and relaxing positivity (common-support sample). Bars are 95% confidence intervals.

H Results for the NLSY79 cohort

Because NLSY97 respondents are observed only through age 36, the main analysis uses logged average earnings over ages 32–36. To evaluate whether later-life earnings alter the decomposition—especially for graduate schooling, whose returns may rise after early adulthood—I also report results using the NLSY79 cohort, for whom we can observe earnings up until age 44.

All variables are defined analogously, although several differences between the two surveys are of note. First, the NLSY79 contains a narrower set of baseline covariates. Whereas the NLSY97 includes high school GPA and a rich set of school characteristics (disciplinary environment, parental assets, school safety, and peer-context measures), the NLSY79 does not. Consequently, the covariate vector in this replication consists of a more limited set of demographic and family background characteristics: indicators of gender and race/ethnicity; parental socioeconomic status (income, educational attainment, and occupation); family structure and sibship size; a cognitive ability measure based on the Armed Forces Qualification Test (AFQT); and the respondent’s expectations for their own educational attainment. The NLSY79 also provides several proxies for cultural resources in the home at age 14—specifically, whether the household regularly received magazines or newspapers and whether any household member held a library card. These measures capture elements of the early educational environment but are considerably narrower than the school-context and academic performance variables available in the NLSY97.

Second, the NLSY79 lacks transcript-based measures of college major or GPA, so there are no intermediate covariates Z between the BA and graduate-school transitions. Identification of the MPSE components therefore relies on stronger assumptions than in the NLSY97 analysis, as academic performance and field-of-study sorting cannot be adjusted for when isolating the contribution of BA and post-BA educational pathways.

Figures S9 and S10 report estimates of the direct and indirect components of the MPSE decomposition for the NLSY79 cohort, for earnings at age 32–36 (top panel) and at age

35-44 (bottom panel). Figure S9 presents results when a constant $c \in \{10, 100, 1000\}$ is added to annual earnings prior to logging, while Figure S10 shows the decomposition for observed earnings in dollar values. Table S7 also summarizes the estimated direct effects (Δ_k), stage-specific transition probabilities (π_k), and covariance components (η_k) for the NLSY79 cohort for earnings at age 32-36 and at age 35-44.¹¹

Within the NLSY79 cohort, across the earlier and later earnings windows, results are remarkably stable for the log-transformed specifications. As shown in Figure S9, estimates of τ_0 and θ_k are nearly indistinguishable across $\log(Y+10)$, $\log(Y+100)$, and $\log(Y+1000)$, with only mild divergence when the largest constant is applied. By contrast, decompositions based on raw earnings (Figure S10) exhibit larger differences across age windows. The larger total effect τ_0 at ages 35-44 than at ages 32-36, for example, results from the fact that the same proportional earnings premium associated with high school completion translates into a larger dollar difference as overall earnings rise with age. Even so, the ordering and relative magnitudes of the direct and indirect pathways remain consistent across specifications.

Importantly, the graduate-school pathway contributes almost identically to the total effect at ages 32-36 and 35-44, even though one might reasonably expect its contribution to increase at later ages as the graduate-school earnings premium becomes larger.

Table S7 clarifies the source of this stability. The continuation effect (Δ_3) nearly doubles between the two windows, but the covariance component η_3 declines, indicating a weaker alignment between the propensity to attend graduate school and the incremental graduate-school earnings premium at later ages. Early in adulthood, individuals most likely to pursue graduate education tend to experience the largest immediate gains from doing so, generating a larger covariance term. By ages 35-44, graduate-school earnings advantages

¹¹There are several differences in the estimated effects in the NLSY79 versus the NLSY97 cohorts. Notably, across specifications, the total effect of high school completion is smaller in the NLSY79 cohort ($\tau_0 = 0.4$ for $\log(Y+1000)$), compared with $\tau_0 = 0.675$ in the NLSY97), consistent with lower marginal returns to schooling for the earlier cohort. The contribution of college attendance to the total effect is also estimated to be negligible ($\theta_1 = 0.03$) compared with in the NLSY97 cohort ($\theta_1 = 0.092$), which may reflect the fact that mediation via “college attendance” in this decomposition captures any postsecondary enrollment, as opposed to mediation via 4-year attendance in the NLSY97, as the NLSY79 public use dataset does not distinguish clearly between two-year and four-year colleges.

are more broadly distributed across degree holders rather than being concentrated among those with the highest propensity to attend. Because the overall probability of traversing the entire trajectory from high school completion to graduate school remains small, these offsetting forces produce a nearly unchanged θ_3 across age windows.

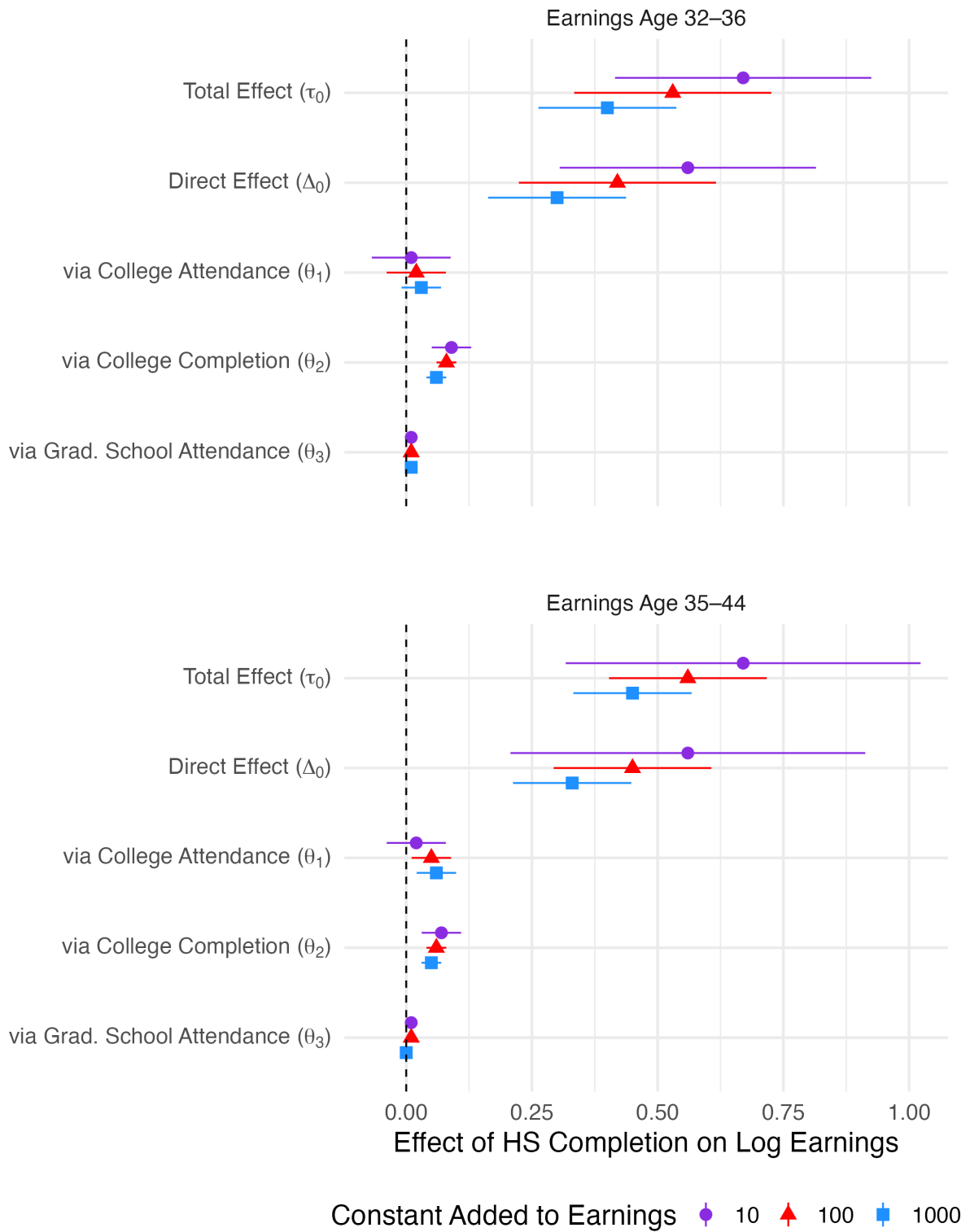


Figure S9: Decomposition of the Average Total Effect (ATE) of high school graduation on logged earnings under alternative definitions of earnings. The figure shows estimates of the total effect (τ_0) as well as the continuation effects $\Delta_0, \tau_1, \dots, \tau_K$ when constants of 10, 100, and 1000 are added to annual earnings (in dollars) prior to taking the log.

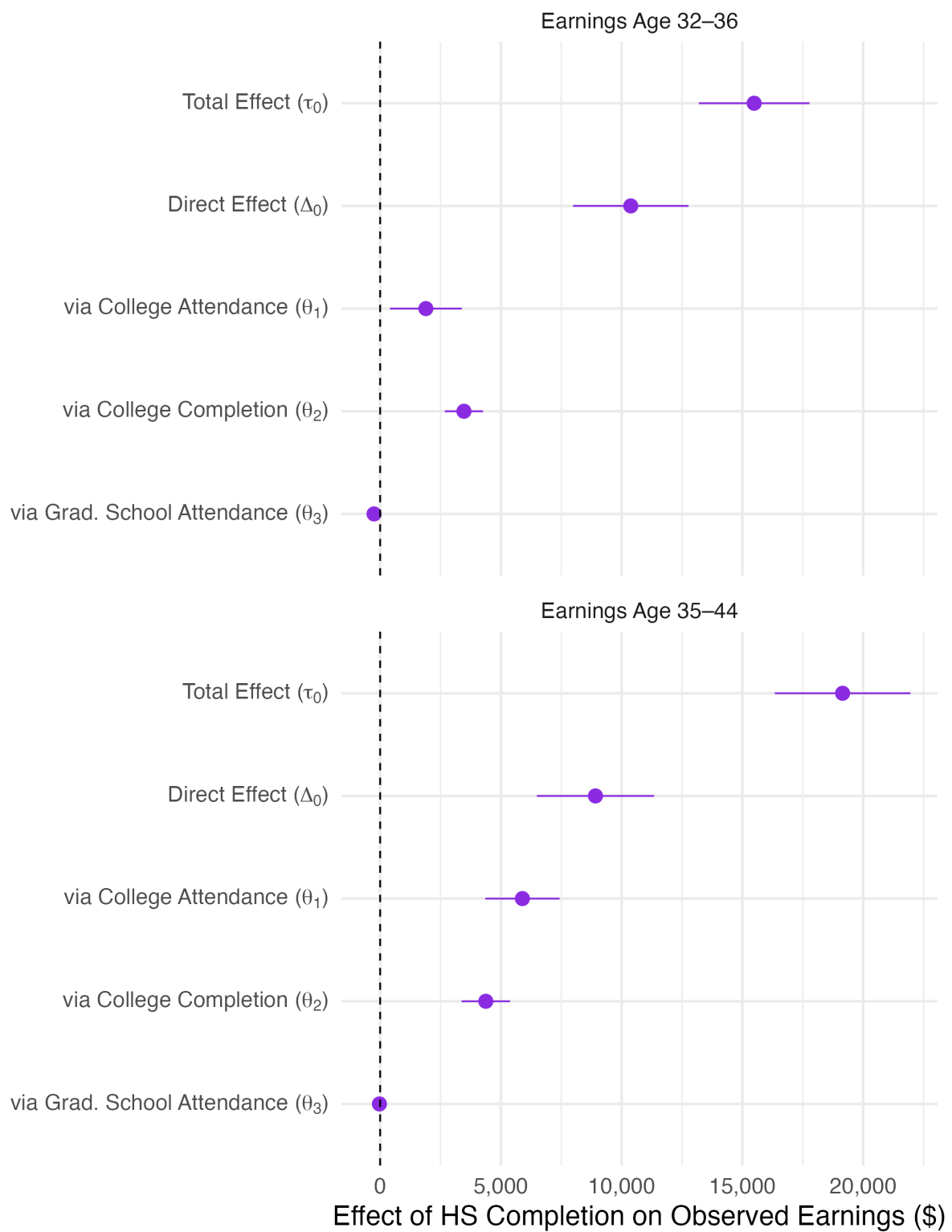


Figure S10: Decomposition of the Average Total Effect (ATE) of high school graduation on observed annual earnings in dollars (no transformation). The figure shows estimates of the total effect (τ_0) as well as the continuation effects $\Delta_0, \tau_1, \dots, \tau_K$ using raw annual earnings.

Table S7: Direct Effects (Δ_k), Probabilities (π_k) and Covariance Terms (η_k) Involved in Decomposition via Debiased Machine-Learning (DML) for NLSY79 cohort.

	Δ_0	Δ_1	Δ_2	Δ_3	π_1	π_2	π_3	η_1	η_2	η_3
Earnings Age 32–36	0.300	0.070	0.520	0.150	0.480	0.330	0.120	-0.010	-0.040	0.020
Earnings Age 35–44	0.330	0.140	0.440	0.290	0.480	0.330	0.120	-0.010	-0.030	-0.010
	(0.070)	(0.040)	(0.040)	(0.050)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
	(0.060)	(0.040)	(0.040)	(0.040)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)

I Extension to categorical mediators

The main text considers a decomposition of the ATE in the case of binary monotonic mediators (i.e., educational transitions), but the framework naturally extends to settings with categorical transitions. Such an extension is especially appealing in the context of U.S. higher education, where individuals follow increasingly variegated pathways. In the early 2010s, fewer than 40% of high school graduates enrolled directly in a four-year college, whereas roughly 30% entered a two-year college, and close to one third of these later transferred to a four-year program. Nearly half of all BA recipients nationally had attended a two-year college at some point in their educational careers. A similar situation arises at the transition to postgraduate education: “graduate school” encompasses multiple substantively distinct routes (e.g. Master’s degrees, professional programs, PhDs), each featuring different patterns of selection and labor-market returns. Dichotomizing such transitions discards precisely the heterogeneity that is often of substantive interest.

To illustrate the extension, consider a single intermediate transition M_k that now takes values in a finite set of mutually exclusive and exhaustive categories,

$$\mathcal{H}_k = \{h_1, \dots, h_H\},$$

rather than a binary indicator. For each category $h \in \mathcal{H}_k$, let $M_k(h)$ denote the potential mediator value under category h , and let $Y(1_k, h)$ denote the potential outcome under completion of all prior transitions $1_k = (A = 1, M_1 = 1, \dots, M_{k-1} = 1)$ with the mediator M_k set to h .

The net effect of transition k can then be written as

$$\tau_k = \Delta_k + \sum_{h \in \mathcal{H}_{k+1}} (\pi_{k+1,h} \Delta_{k+1,h} + \eta_{k+1,h}), \quad (18)$$

where

$$\Delta_{k,h} \equiv \mathbb{E}[Y(1_k, h) - Y(1_k, h_0)], \quad h_0 \in \mathcal{H}_k \text{ (baseline),}$$

$$\pi_{k,h} \equiv \mathbb{E}[\mathbf{1}\{M_k(1_k) = h\}],$$

$$\eta_{k,h} \equiv \text{cov}[\mathbf{1}\{M_k(1_k) = h\}, Y(1_k, h) - Y(1_k, h_0)].$$

These terms generalize the binary case: $\Delta_{k,h}$ is the direct effect of category h relative to h_0 ; $\pi_{k,h}$ is the counterfactual probability of attaining h given completion of all prior transitions; and $\eta_{k,h}$ captures the alignment between the payoff $Y(1_k, h) - Y(1_k, h_0)$ and the propensity to realize category h . The direct effects $\Delta_{k,h}$ and path probabilities $\pi_{k,h}$ remain identified under Assumptions 3–5. However, when $|\mathcal{H}_k| > 2$, the covariance terms are no longer separately nonparametrically identified. In the binary case,

$$\tau_k = \Delta_k + \pi_k \Delta_k + \eta_k$$

contains a single η_k and one identifying restriction. By contrast, (18) contains multiple unknowns $\{\eta_{k,h}\}$ but only one restriction, implying that the vector of covariance components cannot be uniquely recovered without additional structure—for example, assuming homogeneity across categories, proportionality, or a parametric model for effect heterogeneity.

Graduate school example (NLSY97) In my empirical application, the $k = 3$ transition corresponds to the education decision following BA completion. I disaggregate this transition into three categories:

$$\mathcal{H}_3 = \{\text{BA-only, MA, PhD}\},$$

with BA-only as the baseline h_0 . Among BA completers in the NLSY97, roughly 73% do not pursue any graduate degree, 22% complete a Master’s degree, and only 5% complete a

professional or PhD degree.¹² Stage-specific causal contrasts are:

$$\Delta_{3,h} = \mathbb{E}[Y(1_3, h) - Y(1_3, \text{BA-only})], \quad h \in \{\text{MA}, \text{PhD}\}.$$

The continuation value that enters the full MPSE decomposition is

$$\theta_3 = \left(\prod_{j=1}^2 \pi_j \right) \sum_{h \in \{\text{MA}, \text{PhD}\}} (\pi_{3,h} \Delta_{3,h} + \eta_{3,h}). \quad (19)$$

Figure S11 displays the resulting decomposition for MA and professional/PhD pathways, and Table S8 reports the estimated components. The continuation effects for the graduate-school transition are small for both pathways, with $\hat{\theta}_{3,\text{MA}} = 0.0119$ and $\hat{\theta}_{3,\text{PhD}} = 0.0077$, corresponding to earnings effects of high school graduation via these transitions of 1.2% and 0.8%, respectively. Although the continuation effect is slightly lower for the PhD/professional pathway, this masks a much larger underlying causal contrast: the net effect of completing a PhD or professional postgraduate program relative to not pursuing postgraduate study is $\hat{\Delta}_{3,\text{PhD}} = 0.521$ (a 68% earnings premium), compared to a more modest $\hat{\Delta}_{3,\text{MA}} = 0.213$ (a 24% premium) for Master’s degrees. The similarity in the overall continuation effects instead reflects stark differences in the counterfactual probabilities of each pathway: only about 2% of BA completers pursue a PhD-level degree ($\hat{\pi}_{3,\text{PhD}} \approx 0.021$), whereas Master’s attainment is roughly six times more common ($\hat{\pi}_{3,\text{MA}} \approx 0.134$).

Table S8: Estimated components for MA and PhD pathways in the graduate-school transition.

	$\Delta_{3,MA}$	$\Delta_{3,PhD}$	$\pi_{3,MA}$	$\pi_{3,PhD}$	$\eta_{3,MA}$	$\eta_{3,PhD}$
DML	0.213	0.521	0.134	0.021	0.022	0.022
	(0.032)	(0.032)	(0.007)	(0.002)	(0.008)	(0.008)

¹²Professional (DDS, MD and JD) and PhD degree holders are pooled due to small cell sizes.

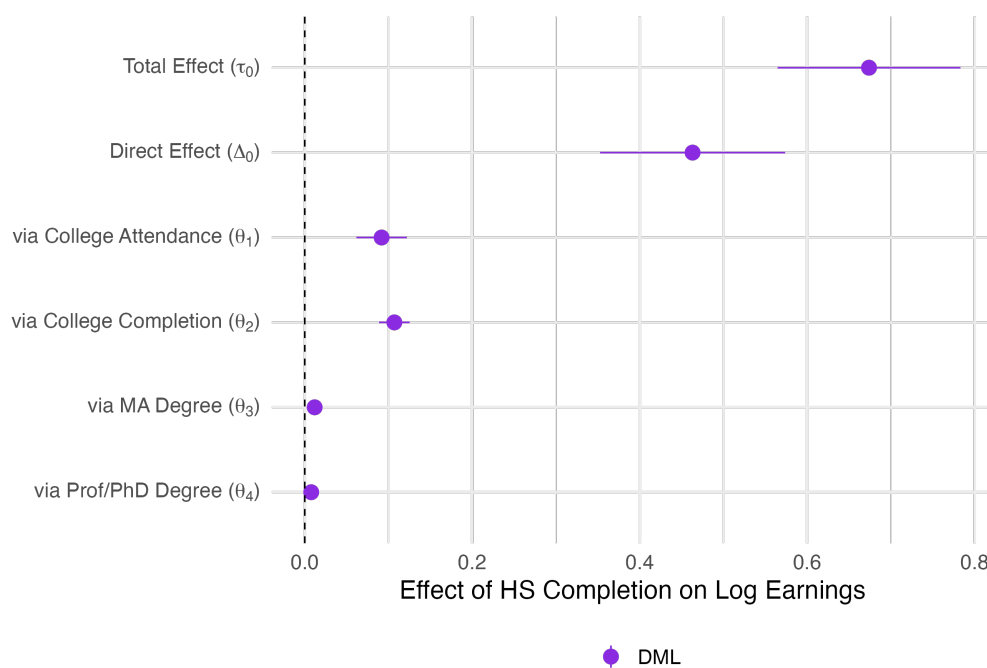


Figure S11: Continuation effects for Master's and Professional/PhD pathways in the NLSY97.

J Description of EIFs used in empirical illustration

For each component involved in the MPSE, I construct a Neyman-orthogonal “signal” using its EIF, whose exact form depends on whether each set of intermediate confounders is empty or not. Figure S4 in the main text shows a potential data-generating process for the direct and indirect (continuation) effects of high school graduation on adult earnings, via three transitions: college attendance (M_1), BA completion (M_2), and graduate school attendance (M_3). I assume that a set of pre-college characteristics serve as confounders for the $A - (M_1, M_2, M_3, Y)$ relationships, and that a set of post-secondary confounders Z confound the $M_2 - (M_3, Y)$ relationships.

Under these assumptions for the various sets of confounders, my MPSE decomposition implies that, in the case of the four transitions (one treatment and three mediators), it suffices to estimate the following three sets of parameters: (i) four direct effects $\Delta_k, k \in \{0, \dots, 3\}$, where $\Delta_3 = \tau_3$, (ii) four gross effects $\tau_k, k \in \{0, \dots, 3\}$, where $\tau_0 = \text{ATE}$, (iii) three mediator terms, $\pi_k, k \in \{1, \dots, 3\}$. All components in the three-mediator decomposition can then be estimated as functions of these parameters. For each of these target parameters, I construct a Neyman-orthogonal signal using its efficient influence function. Because of my assumed data-generating process, which maintains that there is only a single set of intermediate confounders (as opposed to a separate set of confounders for each mediator), the EIF for each estimand involved in the decomposition simplifies somewhat. Specifically, the recentered EIFs for each component in the decomposition are shown below:

$$\begin{aligned}
M_1^*(1) &= \gamma_1(X) + \frac{\mathbb{I}(A=1)}{\pi_0(X,1)}(M_1 - \gamma_1(X)), \\
M_2^*(1,1) &= \gamma_2(X) + \frac{\mathbb{I}(A=1)\mathbb{I}(M_1=1)}{\pi_0(X,1)\pi_1(X,1)}(M_2 - \gamma_2(X)), \\
M_3^*(1,1,1) &= \mathbb{E}[\gamma_3(X, Z)|X, A=1, M_1=1] \\
&\quad + \frac{\mathbb{I}(A=1)\mathbb{I}(M_1=1)}{\pi_0(X,1)\pi_1(X,1)}(\gamma_3(X, Z) - \mathbb{E}[\gamma_3(X, Z)|X, A=1, M_1=1]) \\
&\quad + \frac{\mathbb{I}(A=1)\mathbb{I}(M_1=1)\mathbb{I}(M_2=1)}{\pi_0(X,1)\pi_1(X,1)\pi_2(X, Z, 1)}(M_3 - \gamma_3(X, Z)), \\
Y^*(a) &= \mu_0(X, a) + \frac{\mathbb{I}(A=a)}{\pi_0(X, a)}(Y - \mu_0(X, a)), \text{ for } a \in \{0, 1\} \\
Y^*(1, m_1) &= \mu_1(X, m_1) + \frac{\mathbb{I}(A=1)\mathbb{I}(M_1=m_1)}{\pi_0(X, 1)\pi_1(X, m_1)}(Y - \mu_1(X, m_1)), \text{ for } m_1 \in \{0, 1\} \\
Y^*(1, 1, m_2) &= \mathbb{E}[\mu_2(X, Z, m_2)|X, A=1, M_1=1] \\
&\quad + \frac{\mathbb{I}(A=1)\mathbb{I}(M_1=1)}{\pi_0(X, 1)\pi_1(X, 1)}(\mu_2(X, Z, m_2) - \mathbb{E}[\mu_2(X, Z, m_2)|X, A=1, M_1=1]) \\
&\quad + \frac{\mathbb{I}(A=1)\mathbb{I}(M_1=1)\mathbb{I}(M_2=m_2)}{\pi_0(X, 1)\pi_1(X, 1)\pi_2(X, Z, m_2)}(Y - \mu_2(X, Z, m_2)), \text{ for } m_2 \in \{0, 1\} \\
Y^*(1, 1, 1, m_3) &= \mathbb{E}[\mu_3(X, Z, m_3)|X, A=1, M_1=1] \\
&\quad + \frac{\mathbb{I}(A=1)\mathbb{I}(M_1=1)}{\pi_0(X, 1)\pi_1(X, 1)}(\mu_3(X, Z, m_3) - \mathbb{E}[\mu_3(X, Z, m_3)|X, A=1, M_1=1]) \\
&\quad + \frac{\mathbb{I}(A=1)\mathbb{I}(M_1=1)\mathbb{I}(M_2=1)\mathbb{I}(M_3=m_3)}{\pi_0(X, 1)\pi_1(X, 1)\pi_2(X, Z, 1)\pi_3(X, Z, m_3)}(Y - \mu_3(X, Z, m_3)) \text{ for } m_3 \in \{0, 1\},
\end{aligned}$$

where

$$\pi_0(X, a) \triangleq \Pr[A = a \mid X]$$

$$\pi_1(X, m_1) \triangleq \Pr[M_1 = m_1 \mid X, A = 1]$$

$$\pi_2(X, Z, m_2) \triangleq \Pr[M_2 = m_2 \mid X, A = 1, M_1 = 1, Z]$$

$$\pi_3(X, Z, m_3) \triangleq \Pr[M_3 = m_3 \mid X, A = 1, M_1 = 1, Z, M_2 = 1]$$

$$\gamma_1(X) \triangleq \mathbb{E}[M_1 \mid X, A = 1]$$

$$\gamma_2(X) \triangleq \mathbb{E}[M_2 \mid X, A = 1, M_1 = 1]$$

$$\gamma_3(X, Z) \triangleq \mathbb{E}[M_3 \mid X, A = 1, M_1 = 1, Z, M_2 = 1]$$

$$\mu_0(X, a) \triangleq \mathbb{E}[Y \mid X, A = a]$$

$$\mu_1(X, m_1) \triangleq \mathbb{E}[Y \mid X, A = 1, M_1 = m_1]$$

$$\mu_2(X, Z, m_2) \triangleq \mathbb{E}[Y \mid X, A = 1, M_1 = 1, Z, M_2 = m_2]$$

$$\mu_3(X, Z, m_3) \triangleq \mathbb{E}[Y \mid X, A = 1, M_1 = 1, Z, M_2 = 1, M_3 = m_3].$$

K Proofs and technical details

K.1 EIFs for η_k and θ_k terms (Theorem 3.1)

Under Assumptions 3-5, the covariance component η_k is identified as $\eta_k = \tau_{k-1} - \Delta_{k-1} - \pi_k \tau_k$. Following Kennedy (2022, p. 15), I let $\mathbb{IF} : \Psi \rightarrow L_2(\mathbb{P})$ denote the operator mapping the functionals $\{\Delta_k, \pi_k, \eta_k\} : \mathcal{P} \rightarrow \mathbb{R}, \forall k \in [K]$ to their respective influence functions under the nonparametric model $\mathcal{P}$. First, by linearity of the EIF, $\mathbb{IF}(\eta_k)$ is given by

$$\mathbb{IF}(\eta_k) = \mathbb{IF}(\tau_{k-1}) - \mathbb{IF}(\Delta_{k-1}) - \mathbb{IF}(\pi_k \tau_k).$$

Since $\mathbb{IF}(\pi_k \tau_k)$ can be written as follows $\mathbb{IF}(\pi_k \tau_k) = \tau_k \mathbb{IF}(\pi_k) + \pi_k \mathbb{IF}(\tau_k)$, $\mathbb{IF}(\eta_k)$ can be written as

$$\begin{aligned} \mathbb{IF}(\eta_k) &= \mathbb{IF}(\tau_{k-1}) - \mathbb{IF}(\Delta_{k-1}) - (\tau_k \mathbb{IF}(\pi_k) + \pi_k \mathbb{IF}(\tau_k)) \\ &= \mathbb{IF}(\tau_{k-1}) - \mathbb{IF}(\Delta_{k-1}) - \tau_k \mathbb{IF}(\pi_k) - \pi_k \mathbb{IF}(\tau_k). \end{aligned}$$

Noticing that we can rewrite this expression as

$$\begin{aligned} \mathbb{IF}(\eta_k) &= \mathbb{RIF}(\tau_{k-1}) - \tau_{k-1} - \mathbb{RIF}(\Delta_{k-1}) + \Delta_{k-1} - \tau_k \mathbb{RIF}(\pi_k) + \tau_k \pi_k - \pi_k \mathbb{RIF}(\tau_k) + \pi_k \tau_k \\ &= \mathbb{RIF}(\tau_{k-1}) - \mathbb{RIF}(\Delta_{k-1}) - \tau_k \mathbb{RIF}(\pi_k) - \pi_k \mathbb{RIF}(\tau_k) + \pi_k \tau_k - \eta_k, \end{aligned}$$

where $\mathbb{RIF}(\phi) = \mathbb{IF}(\phi) + \phi$, we can obtain the corresponding EIF-based estimator for η_k by solving the empirical moment condition implied by setting the average of the above equation equal to 0, and plugging in the set of estimated nuisance functions:

$$\hat{\eta}_k^{\text{eif}} = \widehat{\mathbb{RIF}}(\tau_{k-1}) - \widehat{\mathbb{RIF}}(\Delta_{k-1}) - \tau_k \widehat{\mathbb{RIF}}(\pi_k) - \pi_k \widehat{\mathbb{RIF}}(\tau_k) + \pi_k \tau_k,$$

where $\mathbb{R}\hat{\mathbb{I}}\mathbb{F}(\phi) = \hat{\mathbb{I}}\mathbb{F}(\phi) + \phi$, and $\hat{\mathbb{I}}\mathbb{F}(\phi)$ denotes the influence function of a parameter evaluated at estimates of its component nuisance functions.

Turning next to the influence functions for the continuation effects θ_k , $k \in \{1, \dots, K\}$, following the same logic as the above, we can write the EIF of θ_k , $\mathbb{I}\mathbb{F}(\theta_k)$, as

$$\mathbb{I}\mathbb{F}(\theta_k) = \mathbb{I}\mathbb{F}(\Delta_k) \prod_{j=1}^k \pi_j + \Delta_k \sum_{j=1}^k \mathbb{I}\mathbb{F}(\pi_j) \prod_{l:l \neq j}^k \pi_l + \mathbb{I}\mathbb{F}(\eta_k) \prod_{j=1}^{k-1} \pi_j + \eta_k \sum_{j=1}^{k-1} \mathbb{I}\mathbb{F}(\pi_j) \prod_{l:l \neq j}^{k-1} \pi_l.$$

Rewriting this expression as

$$\begin{aligned} \mathbb{I}\mathbb{F}(\theta_k) &= \mathbb{R}\mathbb{I}\mathbb{F}(\Delta_k) \prod_{j=1}^k \pi_j + \Delta_k \sum_{j=1}^k \mathbb{R}\mathbb{I}\mathbb{F}(\pi_j) \prod_{l:l \neq j}^k \pi_l + \mathbb{R}\mathbb{I}\mathbb{F}(\eta_k) \prod_{j=1}^{k-1} \pi_j + \eta_k \sum_{j=1}^{k-1} \mathbb{R}\mathbb{I}\mathbb{F}(\pi_j) \prod_{l:l \neq j}^{k-1} \pi_l \\ &\quad - k\Delta_k \prod_{j=1}^k \pi_j - (k-1)\eta_k \prod_{j=1}^{k-1} \pi_j - \theta_k, \end{aligned}$$

we obtain the corresponding EIF-based estimator for θ_k as:

$$\begin{aligned} \hat{\theta}_k^{\text{eif}} &= \widehat{\mathbb{R}\mathbb{I}\mathbb{F}}(\Delta_k) \prod_{j=1}^k \hat{\pi}_j + \hat{\Delta}_k \sum_{j=1}^k \widehat{\mathbb{R}\mathbb{I}\mathbb{F}}(\pi_j) \prod_{l:l \neq j}^k \hat{\pi}_l + \widehat{\mathbb{R}\mathbb{I}\mathbb{F}}(\eta_k) \prod_{j=1}^{k-1} \hat{\pi}_j + \hat{\eta}_k \widehat{\mathbb{R}\mathbb{I}\mathbb{F}}(\pi_j) \prod_{l:l \neq j}^{k-1} \hat{\pi}_l \\ &\quad - k\hat{\Delta}_k \prod_{j=1}^k \hat{\pi}_j - (k-1)\hat{\eta}_k \prod_{j=1}^{k-1} \hat{\pi}_j. \end{aligned}$$

K.2 Semiparametric efficiency (Theorem 3.2)

In this section, I establish the conditions required for the semiparametric efficiency of all terms featured in the decomposition. Before proceeding, I establish some notational preliminaries. Let $\|g\| = (\int g^\top g dP)^{1/2}$ denote the $L_2(P)$ norm, and let $R_n(\cdot)$ denote a mapping from a nuisance function to its $L_2(P)$ convergence rate. Let $\hat{\varphi}_{km_k}^{\text{EIF}} = \mathbb{P}_n[m(O; \hat{\eta})]$, where $m(O; \hat{\eta})$ is the quantity inside $\mathbb{P}_n[\cdot]$ in equation 13, and $\hat{\eta} = (\hat{\pi}_0, \dots, \hat{\pi}_K, \hat{\mu}_0, \dots, \hat{\mu}_K)$. For the purposes of the technical proofs in this appendix, I adopt a slightly more compact notation than that used in the main text. Specifically, I define $M_0 \equiv A$, and let $\bar{1}_k \equiv (A = 1, M_1 = 1, \dots, M_{k-1} = 1)$. We have that

$$\begin{aligned} \hat{\varphi}_{km_k}^{\text{EIF}} - \varphi_{km_k} &= \mathbb{P}_n[m(O; \hat{\eta})] - P[m(O; \eta)] \\ &= \mathbb{P}_n[m(O; \eta)] + \underbrace{P[m(O; \hat{\eta}) - m(O; \eta)]}_{\triangleq R_2(\hat{\eta})} + (\mathbb{P}_n - P)[m(O; \hat{\eta}) - m(O; \eta)], \end{aligned} \tag{20}$$

where $Pg = \int g dP$ denotes the expectation of function g at the truth. The first term in equation 20 is a sample average, and can be analyzed with the central limit theorem. It has an asymptotic variance of $\mathbb{E}[(\varphi_{km_k}(O; \eta))^2]$. The last term is an empirical process term that will be $o_p(n^{-1/2})$ if either the nuisance functions fall in a Donsker class or if cross-fitting is used to induce independence between $\hat{\eta}$ and O . Thus, $\hat{\theta}^{\text{EIF}}$ will be asymptotically normal and semiparametric efficient if $R_2(\hat{\eta})$ is $o(n^{-1/2})$. To analyze this term, I first note that

$$\begin{aligned}
P[m(O; \eta)] &= P \left[\frac{A}{\pi_{01}(X)} \left(\frac{\mathbb{I}(M_k = m_k)}{\pi_{km_k}(X, \bar{Z}_k)} \prod_{j=1}^{k-1} \frac{M_j}{\pi_{j1}(X, \bar{Z}_j)} \right) (Y - \mu_{km_k}^k(X, \bar{Z}_k)) \right. \\
&\quad \left. + \sum_{j=1}^k \frac{A}{\pi_{01}(X)} \left(\prod_{l=1}^{j-1} \frac{M_l}{\pi_{l1}(X, \bar{Z}_l)} \right) (\mu_{jm_k}^k(X, \bar{Z}_k) - \mu_{j-1m_k}^k(X, \bar{Z}_{j-1})) + \mu_0(X) \right] \\
&= P \left[\frac{A}{\pi_{01}(X)} \left(\frac{\mathbb{I}(M_k = m_k)}{\pi_{km_k}(X, \bar{Z}_k)} \prod_{j=1}^{k-1} \frac{M_j}{\pi_{j1}(X, \bar{Z}_j)} \right) \cdot \right. \\
&\quad \left(\underbrace{\mathbb{E}[Y - \mu_{km_k}^k(X, \bar{Z}_k) | X, \bar{Z}_k, \bar{1}_k, m_k]}_{=0} \right) \\
&\quad + \sum_{j=1}^k \frac{A}{\pi_{01}(X)} \left(\prod_{l=1}^{j-1} \frac{M_l}{\pi_{l1}(X, \bar{Z}_l)} \right) \cdot \\
&\quad \left(\underbrace{\mathbb{E}[\mu_{jm_k}^k(X, \bar{Z}_j) - \mu_{(j-1)m_j}^k(X, \bar{Z}_{j-1}) | X, \bar{Z}_{j-1}, \bar{1}_j]}_{=0} \right) + \mu_0(X) \Big] \\
&= P \left[\mu_0(X) \right].
\end{aligned}$$

Plugging this result into $R_2(\hat{\eta})$, we have that

$$\begin{aligned}
R_2(\hat{\eta}) &= P[m(O; \hat{\eta}) - m(O; \eta)] \\
&= P \left[\sum_{j=0}^{k-1} \frac{A}{\hat{\pi}_{01}(X)} \left(\prod_{l=1}^{j-1} \frac{M_j}{\hat{\pi}_{l1}(X, \bar{Z}_l)} \right) \cdot \right. \\
&\quad (\hat{\pi}_{j1}(X, \bar{Z}_j) - \pi_{j1}(X, \bar{Z}_j)) (\hat{\mu}_{jm_k}^k(X, \bar{Z}_j) - \mu_{jm_k}^k(X, \bar{Z}_j)) \\
&\quad + \frac{A}{\hat{\pi}_{01}(X)} \left(\prod_{l=1}^{k-1} \frac{M_j}{\hat{\pi}_{l1}(X, \bar{Z}_l)} \right) \cdot \\
&\quad \left. (\hat{\pi}_{km_k}(X, \bar{Z}_j) - \pi_{km_k}(X, \bar{Z}_j)) (\hat{\mu}_{km_k}^k(X, \bar{Z}_k) - \mu_{km_k}^k(X, \bar{Z}_k)) \right] \\
&= \sum_{j=0}^k O_p(\|\hat{\pi}_{j1}(X, \bar{Z}_j) - \pi_{j1}(X, \bar{Z}_j)\| \cdot \|\hat{\mu}_{jm_k}^k(X, \bar{Z}_j) - \mu_{jm_k}^k(X, \bar{Z}_j)\|),
\end{aligned}$$

where the last equality results from the positivity assumption that $\hat{\pi}_{k1}(X, \bar{Z}_k)$ is bounded

away from zero, for all $k \in [K]$, and from the Cauchy-Schwartz inequality. Then, assuming that the empirical process term is of order $o_p(n^{-1/2})$, we can write $\hat{\varphi}_{km_k}^{\text{EIF}} - \varphi_{km_k}$ as

$$\begin{aligned} \hat{\psi}_{km_k}^{\text{EIF}} - \psi_{km_k} &= \mathbb{P}_n[m(O; \eta) - \psi_{km_k}] \\ &+ \sum_{j=0}^k O_p(\|\hat{\pi}_{j1}(X, \bar{Z}_j) - \pi_{j1}(X, \bar{Z}_j)\|) \cdot O_p(\|\hat{\mu}_{jm_k}^k(X, \bar{Z}_j) - \mu_{jm_k}^k(X, \bar{Z}_j)\|) \\ &+ o_p(n^{-1/2}). \end{aligned}$$

Thus, letting $R_n(k, m_k) \triangleq \sum_{j=0}^k R_n(\hat{\pi}_j) R_n(\hat{\mu}_{jm_k}^k)$, $\hat{\varphi}_{km_k}^{\text{EIF}}$ is consistent if $R_n(k, m_k) = o(1)$ and it is semiparametric efficient if $R_n(k, m_k) = o(n^{-1/2})$. Clearly, then, $\hat{\Delta}_k^{\text{rEIF}}$ is consistent if $\sum_{j=k}^{k+1} R_n(j, 0) = o(1)$ and it is semiparametric efficient if $\sum_{j=k}^{k+1} R_n(j, 0) = o(n^{-1/2})$. Similarly, $\hat{\tau}_k^{\text{rEIF}}$ is consistent if $\sum_{j=0}^1 R_n(k, j) = o(1)$ and it is semiparametric efficient if $\sum_{j=0}^1 R_n(k, j) = o(n^{-1/2})$.

Turning next to $\phi_k \triangleq \mathbb{E}[M_{k+1}(\bar{1}_{k+1})]$, for all $k \in [K-1]$, we can similarly define $\gamma_k(X, \bar{Z}_k)$ iteratively as

$$\begin{aligned} \gamma_k(X, \bar{Z}_k) &\triangleq \mathbb{E}[M_{k+1} \mid X, \bar{Z}_k, \bar{1}_{k+1}] \\ \gamma_j(X, \bar{Z}_j) &\triangleq \mathbb{E}[\gamma_{j+1}(X, \bar{Z}_{j+1}) \mid X, \bar{Z}_j, \bar{1}_{j+1}] \quad \forall j \in [k-1]. \end{aligned}$$

Similarly to the previous case, the EIF of ϕ_k is equal to

$$\begin{aligned} \varphi_k(O) &= \frac{A}{\pi_{01}(X)} \left(\prod_{l=1}^k \frac{M_l}{\pi_{l1}(X, \bar{Z}_l)} \right) (Y - \gamma_k(X, \bar{Z}_k)) \\ &+ \sum_{j=0}^k \frac{A}{\pi_{01}(X)} \left(\prod_{l=1}^{j-1} \frac{M_l}{\pi_{l1}(X, \bar{Z}_l)} \right) (\gamma_j(X, \bar{Z}_j) - \gamma_{j-1}(X, \bar{Z}_{j-1})) \\ &+ \gamma_0(X) - \phi_k. \end{aligned}$$

Following similar arguments to the above, we have that

$$\begin{aligned}\hat{\phi}_k^{\text{EIF}} - \phi_k &= \mathbb{P}_n[m_2(O; \eta) - \phi_k] \\ &+ \sum_{j=0}^k O_p(\|\hat{\pi}_{j1}(X, \bar{Z}_j) - \pi_{j1}(X, \bar{Z}_j)\|) \cdot O_p(\|\hat{\gamma}_j(X, \bar{Z}_j) - \gamma_j(X, \bar{Z}_j)\|) \\ &+ o_p(n^{-1/2}),\end{aligned}$$

where $m_2(O; \hat{\eta}) = \varphi_k + \phi_k$. Thus, letting $R_n(k, \gamma) \triangleq \sum_{j=0}^k R_n(\hat{\pi}_j)R_n(\hat{\gamma}_j)$, $\hat{\phi}_k^{\text{EIF}}$ is consistent if $R_n(k, \gamma) = o(1)$ and it is semiparametric efficient if $R_n(k, \gamma) = o(n^{-1/2})$. This result implies that if all nuisance functions are consistently estimated and converge at faster than $n^{1/4}$ rates, then $\hat{\phi}_k^{\text{EIF}}$ is semiparametric efficient. I first establish the following lemma:

Lemma K.1. *Let X_n and Y_n denote two convergent sequences, where $X_n = O_p(n^{-1/2})$ and $Y_n = o_p(n^{-1/2})$. Then, (a) $X_n Y_n = o_p(n^{-1/2})$, and (b) $X_n X_n = o_p(n^{-1/2})$.*

Proof. (a) $X_n = O_p(n^{-1/2}) = n^{-1/2}O_p(1) = o_p(1)$. Thus, $X_n Y_n = o_p(1)o_p(n^{-1/2}) = o_p(n^{-1/2})$. (b) $X_n X_n = O_p(n^{-1/2})O_p(n^{-1/2}) = O_p(n^{-1}) = n^{-1/2}O_p(n^{-1/2}) = n^{-1/2}o_p(1)$ (by (a)). Thus, $X_n X_n = o_p(n^{-1/2})$. $\square$

Using this lemma, I establish rate conditions for the semiparametric efficiency of $\eta_k = \tau_{k-1} - \Delta_{k-1} - \pi_k \tau_k$. We can analyze the asymptotic behavior of $\hat{\eta}_k = \hat{\tau}_{k-1} - \hat{\Delta}_{k-1} - \hat{\pi}_k \hat{\tau}_k$ via a distributional expansion of each plug-in estimator:

$$\begin{aligned}
\hat{\eta}_k &= \hat{\tau}_{k-1} - \hat{\Delta}_{k-1} - \hat{\pi}_k \hat{\tau}_k \\
&= (\tau_{k-1} + \mathbb{P}_n[\tau_{k-1}^{\text{EIF}}] + \tau_{k-1}^{\text{EP}} + \tau_{k-1}^{\text{R2}}) - (\Delta_{k-1} + \mathbb{P}_n[\Delta_{k-1}^{\text{EIF}}] + \Delta_{k-1}^{\text{EP}} + \Delta_{k-1}^{\text{R2}}) \\
&\quad - [\pi_k + \mathbb{P}_n[\pi_k^{\text{EIF}}] + \pi_k^{\text{EP}} + \pi_k^{\text{R2}}][\tau_k + \mathbb{P}_n[\tau_k^{\text{EIF}}] + \tau_k^{\text{EP}} + \tau_k^{\text{R2}}] \\
&= (\tau_{k-1} + \mathbb{P}_n[\tau_{k-1}^{\text{EIF}}] + o_p(n^{-1/2}) + \tau_{k-1}^{\text{R2}}) - (\Delta_{k-1} + \mathbb{P}_n[\Delta_{k-1}^{\text{EIF}}] + o_p(n^{-1/2}) + \Delta_{k-1}^{\text{R2}}) \\
&\quad - [\pi_k + \mathbb{P}_n[\pi_k^{\text{EIF}}] + o_p(n^{-1/2}) + \pi_k^{\text{R2}}][\tau_k + \tau_k^{\text{EIF}} + o_p(n^{-1/2}) + \tau_k^{\text{R2}}] \\
&= [\tau_{k-1} - \Delta_{k-1} - \pi_k \tau_k] + \mathbb{P}_n[(\tau_{k-1} - \Delta_{k-1} - \pi_k \tau_k)^{\text{EIF}}] \\
&\quad + \tau_{k-1}^{\text{R2}} + \Delta_{k-1}^{\text{R2}} + \pi_k^{\text{R2}} + \tau_k^{\text{R2}} + O_p(n^{-1/2})O_p(n^{-1/2}) + O_p(n^{-1/2})o_p(n^{-1/2}) + o_p(n^{-1}) + o_p(n^{-1/2}) \\
&= [\tau_{k-1} - \Delta_{k-1} - \pi_k \tau_k] + \mathbb{P}_n[(\tau_{k-1} - \Delta_{k-1} - \pi_k \tau_k)^{\text{EIF}}] \\
&\quad + \tau_{k-1}^{\text{R2}} + \Delta_{k-1}^{\text{R2}} + \pi_k^{\text{R2}} + \tau_k^{\text{R2}} + o_p(n^{-1/2}),
\end{aligned}$$

where the penultimate equality follows from Theorem 3.1, and the final equality follows from Lemma K.1.

Thus, for any $k \in \{1, \dots, K\}$, $\hat{\eta}_k = \hat{\tau}_{k-1} - \hat{\Delta}_{k-1} - \hat{\pi}_k \hat{\tau}_k$ is semiparametric efficient if $\tau_{k-1}^{\text{R2}} + \Delta_{k-1}^{\text{R2}} + \pi_k^{\text{R2}} + \tau_k^{\text{R2}} = o_p(n^{-1/2})$.

For the continuation terms $(\theta_k = (\Pi_{j=1}^k \pi_j) \Delta_k + (\Pi_{j=1}^{k-1} \pi_j) \eta_k)$, I proceed by induction. Let $\hat{\Delta}_* = \Delta_* + \mathbb{P}_n[\Delta_*^{\text{EIF}}] + \Delta_*^{\text{EP}} + \Delta_*^{\text{R2}}$ and $\hat{\eta}_* = \eta_* + \mathbb{P}_n[\eta_*^{\text{EIF}}] + \eta_*^{\text{EP}} + \eta_*^{\text{R2}}$ be asymptotically linear, where $*$ $\in \{1, \dots, K\}$. For $k = 1$, we can asymptotically expand $\hat{\pi}_1 \hat{\Delta}_*$ as

$$\begin{aligned}
\hat{\pi}_1 \hat{\Delta}_* &= (\pi_1 + \mathbb{P}_n[\pi_1^{\text{EIF}}] + \pi_1^{\text{EP}} + \pi_1^{\text{R2}})(\Delta_* + \mathbb{P}_n[\Delta_*^{\text{EIF}}] + \Delta_*^{\text{EP}} + \Delta_*^{\text{R2}}) \\
&= (\pi_1 + \mathbb{P}_n[\pi_1^{\text{EIF}}] + o_p(n^{-1/2}) + \pi_1^{\text{R2}})(\Delta_* + \mathbb{P}_n[\Delta_*^{\text{EIF}}] + o_p(n^{-1/2}) + \Delta_*^{\text{R2}}) \\
&= \pi_1 \Delta_* + \mathbb{P}_n[\pi_1^{\text{EIF}} \Delta_* + \Delta_*^{\text{EIF}} \pi_1] + \pi_1^{\text{R2}} + \Delta_*^{\text{R2}} + o_p(n^{-1/2}) \quad (\text{by Lemma K.1}) \\
&= \sum_{j=1}^k \pi_j \Delta_* + \mathbb{P}_n[(\sum_{j=1}^k \pi_j \Delta_*)^{\text{EIF}}] + \Delta_*^{\text{R2}} + \sum_{j=1}^k \pi_j^{\text{R2}} + o_p(n^{-1/2}) \quad (\text{by Lemma K.1}),
\end{aligned}$$

and expand $(\Pi_{j=1}^{k-1} \pi_j) \eta_*$, similarly, as

$$\begin{aligned}
\sum_{j=1}^{k-1} \hat{\pi}_j \hat{\eta}_* &= \eta_* + \mathbb{P}_n[\eta_*^{\text{EIF}}] + \tau_{*-1}^{\text{R2}} + \Delta_{*-1}^{\text{R2}} + \pi_*^{\text{R2}} + \tau_*^{\text{R2}} + o_p(n^{-1/2}) \\
&= \sum_{j=1}^{k-1} \pi_j \eta_* + \mathbb{P}_n[(\sum_{j=1}^{k-1} \pi_j \eta_*)^{\text{EIF}}] \\
&\quad + \tau_{*-1}^{\text{R2}} + \Delta_{*-1}^{\text{R2}} + \pi_*^{\text{R2}} + \tau_*^{\text{R2}} + \sum_{j \in \{1, \dots, k^*-1\}: j \neq *} \pi_j^{\text{R2}} + o_p(n^{-1/2}),
\end{aligned}$$

following a similar logic to above. Now, assume that, for $k^* \in \{1, \dots, K\}$, $(\Pi_{j=1}^{k^*} \hat{\pi}_j) \hat{\Delta}_* = \Delta_k \prod_{j=1}^{k^*} \pi_j + \mathbb{P}_n[(\Delta_* \prod_{j=1}^{k^*} \pi_j)^{\text{EIF}}] + \Delta_*^{\text{R2}} + \sum_{j=1}^{k^*} \pi_j^{\text{R2}} + o_p(n^{-1/2})$ and, further, that $(\Pi_{j=1}^{k^*-1} \hat{\pi}_j) \hat{\eta}_* = \Pi_{j=1}^{k^*-1} \pi_j \eta_* + \mathbb{P}_n[(\Pi_{j=1}^{k^*-1} \pi_j \eta_*)^{\text{EIF}}] + \tau_{*-1}^{\text{R2}} + \Delta_{*-1}^{\text{R2}} + \pi_*^{\text{R2}} + \tau_*^{\text{R2}} + \sum_{j \in \{1, \dots, k^*-1\}: j \neq *} \pi_j^{\text{R2}} + o_p(n^{-1/2})$.

Then, by induction, we have that

$$\begin{aligned}
(\Pi_{j=1}^{k^*+1} \hat{\pi}) \hat{\Delta}_* &= \left[(\pi_{k^*+1} + \mathbb{P}_n[\pi_{k^*+1}^{\text{EIF}}] + o_p(n^{-1/2}) + \pi_{k^*+1}^{\text{R2}}) \right] \\
&\quad \left[\Delta_* \prod_{j=1}^{k^*} \pi_j + \mathbb{P}_n[(\Delta_* \sum_{j=1}^{k^*} \pi_j)^{\text{EIF}}] + \Delta_*^{\text{R2}} + \sum_{j=1}^{k^*} \pi_j^{\text{R2}} + o_p(n^{-1/2}) \right] \\
&= \Delta_* \prod_{j=1}^{k^*+1} \pi_j + \mathbb{P}_n[(\Delta_* \prod_{j=1}^{k^*+1} \pi_j)^{\text{EIF}}] + \Delta_*^{\text{R2}} \\
&\quad + \sum_{j=1}^{k^*+1} \pi_j^{\text{R2}} + O_p(n^{-1/2}) O_p(n^{-1/2}) \\
&\quad + O_p(n^{-1/2}) o_p(n^{-1/2}) + o_p(n^{-1}) + o_p(n^{-1/2}) \\
&= \Delta_* \sum_{j=1}^{k^*+1} \pi_j + \mathbb{P}_n[(\Delta_* \prod_{j=1}^{k^*+1} \pi_j)^{\text{EIF}}] + \Delta_*^{\text{R2}} + \sum_{j=1}^{k^*+1} \pi_j^{\text{R2}} + o_p(n^{-1/2}),
\end{aligned}$$

and that

$$\begin{aligned}
(\Pi_{j=1}^{k^*} \hat{\pi}_j) \hat{\Delta}_* &= \left[(\pi_{k^*+1} + \mathbb{P}_n[\pi_{k^*+1}^{\text{EIF}}] + o_p(n^{-1/2}) + \pi_{k^*+1}^{\text{R2}}) \right] \\
&\quad \left[\Pi_{j=1}^{k^*-1} \pi_j \eta_* + \mathbb{P}_n[(\Pi_{j=1}^{k^*-1} \pi_j \eta_*)^{\text{EIF}}] + \tau_{*-1}^{\text{R2}} + \Delta_{*-1}^{\text{R2}} + \pi_*^{\text{R2}} + \tau_*^{\text{R2}} + \sum_{j \in \{1, \dots, k^*-1\}: j \neq *} \pi_j + o_p(n^{-1/2}) \right] \\
&= \Pi_{j=1}^{k^*} \pi_j \eta_* + \mathbb{P}_n[(\Pi_{j=1}^{k^*} \pi_j \eta_*)^{\text{EIF}}] + \tau_{*-1}^{\text{R2}} + \Delta_{*-1}^{\text{R2}} + \pi_*^{\text{R2}} + \tau_*^{\text{R2}} + \sum_{j \in \{1, \dots, k^*\}: j \neq *} \pi_j^{\text{R2}} + o_p(n^{-1/2})
\end{aligned}$$

It follows that, for any $k \in \{1, \dots, K\}$,

$$(\Pi_{j=1}^k \hat{\pi}_j) \hat{\Delta}_k = \Delta_k \sum_{j=1}^k \pi_j + \Pi_{j=1}^{k-1} \pi_j \eta_k + \mathbb{P}_n[(\Delta_k \sum_{j=1}^k \pi_j + \Pi_{j=1}^{k-1} \pi_j \eta_k)^{\text{EIF}}] + \sum_{j=k-1}^k \Delta_j^{\text{R2}} + \sum_{j=k-1}^k \tau_j^{\text{R2}} + \sum_{j=1}^k \pi_j^{\text{R2}}.$$

Thus, $\hat{\theta}_k = (\Pi_{j=1}^k \hat{\pi}_j) \hat{\Delta}_k + (\Pi_{j=1}^{k-1} \hat{\pi}_j) \hat{\eta}_k$ is semiparametric efficient if $\sum_{j=k-1}^k \Delta_j^{\text{R2}} + \sum_{j=k-1}^k \tau_j^{\text{R2}} + \sum_{j=1}^k \pi_j^{\text{R2}} = o(n^{-1/2})$. Theorem 3.2 then follows immediately by recognizing the rate conditions required for each of the constituent functionals of $(\pi_k, \Delta_k, \tau_k)$ to be semiparametric efficient.

L Derivation of RWR procedures

For simplicity, throughout the following I let $Z_0 = X$ and $M_0 = A$. I assume the following linear specification of the outcome model:

$$\begin{aligned} \mathbb{E}[Y \mid \bar{Z}_k, A, \bar{M}_k] &= \beta_{k,0} + c_{k,0}A + \sum_{j=1}^k \beta_{k,j}M_j + \eta_{k,1}^\top X^\perp + c_{k,1}AX^\perp + \sum_{j=1}^{k-1} \eta_{k,j}^\top M_j X^\perp + \sum_{j=1}^k \gamma_{k,j}^\top Z_j^\perp \\ &\quad + \sum_{j=1}^{k-1} M_j \sum_{l=1}^j \xi_{k,k,l}^\top Z_l^\perp, \end{aligned} \tag{21}$$

where $Z_k^\perp = Z_k - \mathbb{E}[Z_k \mid \bar{Z}_{k-1}, M_{k-1} = 1_{k-1}]$, $\forall k \in [0, \dots, K]$. In the following derivations, I use the fact that, $\forall k \in \{1, \dots, K\}$,

$$\begin{aligned} &\int z_k^\perp dP(z_k \mid \bar{z}_{k-1}, m_{k-1} = 1) \\ &= \mathbb{E}[Z_k - \mathbb{E}[Z_k \mid \bar{z}_{k-1}, m_{k-1} = 1] \mid \bar{z}_{k-1}, m_{k-1} = 1] \\ &= 0. \end{aligned}$$

Letting $X = Z_0$, the above also implies that $\int z_0^\perp dP(z_0) = \mathbb{E}[Z_0 - \mathbb{E}[Z_0]] = 0$. Under sequential ignorability and assuming linearity of the outcome with respect to all antecedent variables, we have that

$$\begin{aligned}
\Delta_{k-1} &= \int \mathbb{E}[Y | \overline{M}_{k-1} = \overline{1}_{k-1}, \overline{z}_k, M_k = 0] \prod_{j=0}^k dP(z_j | \overline{z}_{j-1}, m_{j-1} = 1) \\
&\quad - \int \mathbb{E}[Y | \overline{M}_{k-2} = \overline{1}_{k-2}, \overline{z}_k, M_{k-1} = 0] \prod_{j=1}^k dP(z_j | \overline{z}_{j-1}, m_{j-1} = 1) \\
&= \int \left[\beta_{k,0} + c_{k,0} + \sum_{j=1}^{k-1} \beta_{k,j} + \eta_{k,1}^\top X^\perp + c_{k,1} X^\perp + \sum_{j=1}^{k-2} \eta_{k,j}^T X^\perp + \sum_{j=1}^k \gamma_{k,j}^T Z_j^\perp + \sum_{j=1}^{k-2} \sum_{l=1}^j \xi_{k,k,l}^\top Z_l^\perp \right. \\
&\quad \left. - (\beta_{k,0} + c_{k,0} + \sum_{j=1}^{k-2} \beta_{k,j} M_j + \eta_{k,1}^\top X^\perp + c_{k,1} A X^\perp + \sum_{j=1}^{k-2} \eta_{k,j}^T X^\perp + \sum_{j=1}^k \gamma_{k,j}^T Z_j^\perp + \sum_{j=1}^{k-3} \sum_{l=1}^j \xi_{k,k,l}^\top Z_l^\perp) \right] \\
&\quad \prod_{j=0}^k dP(z_j | \overline{z}_{j-1}, m_{j-1} = 1) \\
&= \beta_{k,k-1}.
\end{aligned}$$

Further, for $\tau_k \forall k \in \{1, \dots, K\}$ we have that

$$\begin{aligned}
\tau_k &= \int \mathbb{E}[Y | A = 1, \overline{M}_k = \overline{1}_k, \overline{z}_k] \prod_{j=0}^k dP(z_j | \overline{z}_{j-1}, m_{j-1} = 1) \\
&\quad - \int \mathbb{E}[Y | A = 1, \overline{M}_{k-1} = \overline{1}_{k-1}, \overline{z}_k, M_k = 0] \prod_{j=0}^k dP(z_j | \overline{z}_{j-1}, m_{j-1} = 1) \\
&= \int \left[(\beta_{k,0} + c_{k,0} + \sum_{j=1}^k \beta_{k,j} + \eta_{k,1}^\top X^\perp + c_{k,1} X^\perp + \sum_{j=1}^{k-1} \eta_{k,j}^T X^\perp + \sum_{j=1}^k \gamma_{k,j}^T Z_j^\perp + \sum_{j=1}^{k-1} \sum_{l=1}^j \xi_{k,j,l}^\top Z_l^\perp) \right. \\
&\quad \left. - (\beta_{k,0} + c_{k,0} + \sum_{j=1}^{k-1} \beta_{k,j} + \eta_{k,1}^\top X^\perp + c_{k,1} X^\perp + \sum_{j=1}^{k-2} \eta_{k,j}^T X^\perp + \sum_{j=1}^k \gamma_{k,j}^T Z_j^\perp + \sum_{j=1}^{k-2} \sum_{l=1}^j \xi_{k,j,l}^\top Z_l^\perp) \right] \\
&\quad \prod_{j=0}^k dP(z_j | \overline{z}_{j-1}, m_{j-1} = 1) \\
&= \beta_{k,k}.
\end{aligned}$$

Finally, I assume that

$$\mathbb{E}[M_{k+1} \mid A = 1, \bar{Z}_k, \bar{M}_k = \bar{1}_k] = \theta_0 + \sum_{k=0}^k \delta_{k+1}^T Z_k^\perp \quad .$$

Then:

$$\begin{aligned} \mathbb{E}[M(\bar{1}_{k+1})] &= \theta_0 + \int \left[\sum_{k=0}^k \delta_{k+1}^T Z_k^\perp \prod_{j=0}^k dP(z_j | \bar{z}_{j-1}, m_{j-1} = 1) \right] \\ &= \theta_0. \end{aligned}$$

M Derivation of bias formulae for sensitivity analysis

In this section, I derive the bias formulae for the set (τ_k, Δ_k) for all $k \in [K]$, where K denotes the number of mediators considered in the decomposition, under a sequence of simplifying assumptions. Assume first that we have a binary unobserved confounder, U , for the treatment-outcome relationship. Assuming that $\alpha_0 = \mathbb{E}[Y|x, a, U = 1] - \mathbb{E}[Y|x, a, U = 0]$ does not depend on x or a , and further that $\beta_0 = \Pr[U = 1|x, A = 1] - \Pr[U = 1|x, A = 0]$ does not depend on x , for $\tau_0 = \mathbb{E}[Y(1) - Y(0)] \triangleq \text{ATE}$, I then have that $\text{bias}(\tau_0) = \alpha\beta$ (VanderWeele and Arah, 2011).

Next, consider an unobserved confounder, U_k that affects both M_k and Y for any $k \in \{1, \dots, K\}$. Then, under a weaker iteration of Assumption 4 (Sequential Ignorability), i.e.,

$Y(\bar{1}_k, m_k) \perp\!\!\!\perp (A, \bar{M}_k)|X, A, U_k, \bar{Z}_k, \bar{M}_{k-1} \forall k \in [K], \mathbb{E}[Y(\bar{1}_k, m_k)]$ is identified as

$$\mathbb{E}[Y(\bar{1}_k, m_k)] = \int_x \int_{\bar{z}_k} \mathbb{E}[Y|x, \bar{z}_k, \bar{1}_k, m_k, u_k] [dP(u_k|x, \bar{z}_k, \bar{1}_{k-1}) \prod_{j=1}^k dP(z_j|x, \bar{z}_{j-1}, \bar{1}_{j-1})] dP(x).$$

By contrast, under Assumption 4, my estimator of $\mathbb{E}[Y(\bar{1}_k, m_k)]$, $\tilde{\mathbb{E}}[Y(\bar{1}_k, m_k)]$, converges to

$$\tilde{\mathbb{E}}[Y(\bar{1}_k, m_k)] = \int_x \int_{\bar{z}_k} \mathbb{E}[Y|x, \bar{z}_k, \bar{1}_k, m_k, u_k] [dP(u_k|x, \bar{z}_k, \bar{1}_k) \prod_{j=1}^k dP(z_j|x, \bar{z}_{j-1}, \bar{1}_{j-1})] dP(x).$$

I invoke the following three assumptions: (Assumption A_k) $\alpha_k = \mathbb{E}[Y|x, \bar{z}_k, \bar{1}_k, m_k, U_k = 1] - \mathbb{E}[Y|x, \bar{z}_k, \bar{1}_k, m_k, U_k = 0]$ does not depend on $(x, \bar{z}_k, \bar{1}_k, m_k)$; (Assumption B_k) $\beta_k = \Pr[U_k = 1|x, \bar{z}_k, \bar{1}_k, m_k] - \Pr[U_k = 1|x, \bar{z}_k, \bar{1}_k]$ does not depend on $(x, \bar{z}_k)$; Assumption (C) U_k is binary. Taking the difference between the quantities in the above two equations thus gives that, for any $m_k \in \{0, 1\}$, for any $k \in [K]$, we have that

$$\begin{aligned}
\text{bias}(\tilde{\mathbb{E}}[Y(\bar{1}_k, m_k)]) &= \int (\mathbb{E}[Y|x, \bar{z}_k, \bar{1}_k, m_k, U_k = 1] - \mathbb{E}[Y|x, \bar{z}_k, \bar{1}_k, m_k, U_k = 0]) \cdot \\
&\quad (\Pr[U_k = 1|x, \bar{z}_k, \bar{1}_k, m_k] - \Pr[U_k = 1|x, \bar{z}_k, \bar{1}_k]) \prod_{j=1}^k dP(z_j|x, \bar{z}_{j-1}, \bar{1}_{j-1}) dP(x).
\end{aligned} \tag{22}$$

Consider first $\text{bias}(\Delta_{k-1}) = \text{bias}(\tilde{\mathbb{E}}[Y(\bar{1}_k, 0) - Y(\bar{1}_{k-1}, 0)])$. Under mediator monotonicity (Assumption 1), I immediately have that $\Pr[U_k = 1|x, \bar{z}_k, \bar{1}_k, m_k] - \Pr[U_k = 1|x, \bar{z}_k, \bar{1}_k]$, and thus that $\text{bias}(\Delta_{k-1}) = \text{bias}(\tilde{\mathbb{E}}[Y(\bar{1}_k, 0)])$, which can be written as

$$\begin{aligned}
\text{bias}(\tilde{\mathbb{E}}[Y(\bar{1}_k, 0)]) &= \int (\mathbb{E}[Y|x, \bar{z}_k, \bar{1}_k, 0, U_k = 1] - \mathbb{E}[Y|x, \bar{z}_k, \bar{1}_k, 0, U_k = 0]) \\
&\quad (\Pr[U_k = 1|x, \bar{z}_k, \bar{1}_k, 0] - \Pr[U_k = 1|x, \bar{z}_k, \bar{1}_k]) \prod_{j=1}^k dP(z_j|x, \bar{z}_{j-1}, \bar{1}_{j-1}) dP(x) \\
&= \int (\mathbb{E}[Y|x, \bar{z}_k, \bar{1}_k, 0, U_k = 1] - \mathbb{E}[Y|x, \bar{z}_k, \bar{1}_k, 0, U_k = 0]) \cdot \\
&\quad (\Pr[U_k = 1|x, \bar{z}_k, \bar{1}_k, 0] - (\Pr[U_k = 1|x, \bar{z}_k, \bar{1}_{k+1}] \Pr[M_k = 1|x, \bar{z}_k, \bar{1}_k] \\
&\quad + \Pr[U_k = 1|x, \bar{z}_k, \bar{1}_k, 0] - \Pr[U_k = 1|x, \bar{z}_k, \bar{1}_k, 0] \Pr[M_k = 1|x, \bar{z}_k, \bar{1}_k])) \\
&\quad \prod_{j=1}^k dP(z_j|x, \bar{z}_{j-1}, \bar{1}_{j-1}) dP(x) \\
&= - \int (\mathbb{E}[Y|x, \bar{z}_k, \bar{1}_k, 0, U_k = 1] - \mathbb{E}[Y|x, \bar{z}_k, \bar{1}_k, 0, U_k = 0]) \cdot \\
&\quad ((\Pr[U_k = 1|x, \bar{z}_k, \bar{1}_{k+1}] - \Pr[U_k = 1|x, \bar{z}_k, \bar{1}_k, 0]) \Pr[M_k = 1|x, \bar{z}_k, \bar{1}_k]) \\
&\quad \prod_{j=1}^k dP(z_j|x, \bar{z}_{j-1}, \bar{1}_{j-1}) dP(x).
\end{aligned}$$

Next, applying assumptions A_k and B_k , we can write

$$\text{bias}(\tilde{\mathbb{E}}[Y(\bar{1}_{k+1}, 0)]) = -\alpha_k \beta_k \int_x \int_{\bar{z}_k} \Pr[M_k = 1|x, \bar{z}_k, \bar{1}_k] \prod_{j=1}^k dP(z_j|x, \bar{z}_{j-1}, \bar{1}_{j-1}) dP(x).$$

Second, to compute $\text{bias}(\tau_k) = \text{bias}(\mathbb{E}[Y(\bar{1}_{k+1}) - Y(\bar{1}_k, 0)])$ for any $k \in \{1, \dots, K\}$, beginning with Equation 22 and applying assumptions A_k and B_k once again, we have that

$$\text{bias}(\tau_k) = \alpha_k \beta_k.$$