Nonparametric Causal Decomposition of Group Disparities

Ang Yu,[*] Felix Elwert[†]

July 26, 2024

Abstract

We introduce a new nonparametric causal decomposition approach that identifies the mechanisms by which a treatment variable contributes to a group-based outcome disparity. Our approach distinguishes three mechanisms: group differences in 1) treatment prevalence, 2) average treatment effects, and 3) selection into treatment based on individual-level treatment effects. Our approach reformulates classic Kitagawa-Blinder-Oaxaca decompositions in causal and nonparametric terms, complements causal mediation analysis by explaining group disparities instead of group effects, and isolates conceptually distinct mechanisms conflated in recent random equalization decompositions. In contrast to all prior approaches, our framework uniquely identifies differential selection into treatment as a novel disparity-generating mechanism. Our approach can be used for both the retrospective causal explanation of disparities and the prospective planning of interventions to change disparities. We present both an unconditional and a conditional decomposition, where the latter quantifies the contributions of the treatment within levels of certain covariates. We develop nonparametric estimators that are $\sqrt{n}$-consistent, asymptotically normal, semiparametrically efficient, and multiply robust. We apply our approach to analyze the mechanisms by which college graduation causally contributes to intergenerational income persistence (the disparity in income attainment between parental income groups).

Keywords— social inequality, mediation analysis, double machine learning, income mobility

---

[*]Department of Sociology, University of Wisconsin-Madison. Email: ayu33@wisc.edu

[†]Department of Sociology, Department of Biostatistics and Medical Informatics, and Department of Population Health Sciences, University of Wisconsin-Madison

1   Introduction

Social and health scientists often seek to decompose an outcome disparity between groups in terms of the contributions of an intermediate treatment variable. For example, how much, and in what ways, do racial differences in medical care contribute to racial disparities in health (Howe et al., 2014)? How does childbearing contribute to the gender wage gap (Cha et al., 2023)? And what are the roles of education in the association between parents' and their children's social class (social mobility) (Ishida et al., 1995)? The common structure of these questions is that they seek to quantify the mechanisms by which a treatment variable *causally* explains an observed, *descriptive* group disparity.

Prior research has addressed such questions using three approaches, none of which is fully appropriate for the task of causally decomposing descriptive disparities. First, popular Kitagawa-Blinder-Oaxaca (KBO) decompositions (Kitagawa, 1955; Blinder, 1973; Oaxaca, 1973) are defined in terms of parametric regression coefficients and do not answer any causal question by design (Fortin et al., 2011, p.13; Lundberg et al., 2021, p.542). Second, causal mediation analysis (CMA) (VanderWeele, 2015), though formulated in causal terms, decomposes the causal effects of group membership rather than observed group disparities. Third, the recently developed random-equalization decompositions (VanderWeele and Robinson, 2014; Jackson and Vander-Weele, 2018; Jackson, 2021; Lundberg, 2024) conflate distinct disparity-generating mechanisms, which limits their interpretability and usefulness for policy analysis. Importantly, all three prior approaches neglect that outcome disparities can in part be explained by group-differential selection of individuals into treatment as a function of their individual treatment effects.

In this article, we propose a nonparametric causal decomposition approach that remedies these limitations. In contrast to KBO decompositions, our decompositions are formulated as model-free causal estimands with interventional interpretations. In contrast to CMA, we decompose observed group disparities rather than possibly ill-defined causal effects of group membership. In contrast to random equalization decompositions, our approach divides the outcome disparity into conceptually unambiguous components. In contrast to all three prior approaches, we reveal a novel mechanism, termed the "selection component," that has not previously been incorporated in decomposition analysis.

Our framework distinguishes three causal mechanisms through which an intermediate treatment variable can contribute to a group disparity in an outcome. First, mean outcomes may differ across groups because the groups receive the treatment at different rates (differential prevalence). Second, even if all groups have the same treatment prevalence, outcomes may differ across groups because the treatment has different average effects across groups (differential

effects). Third, even if all groups have the same treatment prevalence and the same average treatment effects, outcomes will nonetheless differ across groups if members of one group select into treatment more strongly on their treatment effects than members of another group (differential selection); for example, if the treatment is randomly distributed to members of one group, but specifically given to only those members who will benefit the most from the treatment in the other group. Conceptually, prior decompositions were limited to considering differential prevalence and differential effects, leading to the belief that these two are the only possible mechanisms (e.g., Ward et al., 2019; Diderichsen et al., 2019). Isolating differential selection into treatment as a source of group disparities is the central conceptual contribution of our approach.

We introduce an unconditional and a conditional decomposition. The unconditional decomposition is useful for investigating the overall (marginal) contributions of a treatment to an outcome disparity. The conditional decomposition is useful for investigating the contributions of the treatment within levels of one or more pre-treatment covariates. For example, in the social sciences, the unconditional decomposition might quantify the contributions of racial differences in incarceration rates to racial income disparities; and the conditional decomposition could identify the contributions of racial differences in mortgage receipt to racial disparities in home ownership conditional on credit scores, i.e., considering only the part of the racial differences in mortgage receipt that remains after accounting for racial differences in credit scores. In medical research, the unconditional decomposition might inform the contributions of vaccinations to racial health disparities; and the conditional decomposition can inform the contributions of receiving a medical treatment to health disparities conditional on indications for that treatment.

We develop nonparametric estimators for our causal decompositions using efficient influence functions (EIF) under the assumption of conditional ignorability of the treatment. Our estimators can be implemented via data-adaptive methods such as machine learning (ML) and accommodate high-dimensional confounders. We derive the conditions under which the estimators are $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient. The estimators are also doubly or even quadruply robust in terms of consistency. The estimators are implemented in the R package cdgd (Yu, 2024), available from CRAN.

As an empirical application, we study the causal contributions of college graduation to intergenerational income persistence, defined as the disparity in income attainment across parental income groups. This application contributes to multiple literatures in sociology and economics. Policy-wise, it provides insights into how interventions on college education may

alter intergenerational income persistence. R Code for the empirical application is available at https://github.com/ang-yu/causal_decomposition_case_study.

Our paper proceeds as follows. Section 2 introduces our causal decompositions and their sequential interventional interpretations. We explicate the contributions of our framework by formally contrasting it with KBO decompositions, CMA, and random equalization decompositions. In section 3, we introduce the estimators and their asymptotic theory. Section 4 presents the empirical application. Section 5 concludes with extensions. All proofs are collected in the supplementary appendices.

2 Estimands

2.1 Unconditional decomposition

We consider a binary treatment variable $D_i \in \{0, 1\}$ for each individual i. Let Y_i^0 and Y_i^1 be the potential outcomes (Rubin, 1974) of Y_i under the hypothetical intervention to set $D_i = 0$ and $D_i = 1$, respectively. Let $\tau_i := Y_i^1 - Y_i^0$ denote the individual-level treatment effect. For expositional convenience, we assume that higher values of Y_i are more favorable. Suppose that the population contains two disjoint groups, $G_i = g \in \{a, b\}$, where a denotes the advantaged group and b denotes the disadvantaged group. We use subscript g to indicate group-specific quantities, for example, $\mathrm{E}_g(Y_i) := \mathrm{E}(Y_i \mid g)$. Henceforth, we suppress subscript i to ease notation. In our empirical application, G is parental income group, Y is income attainment in adulthood, and D is college graduation.

Assumption 1 (SUTVA). $Y = DY^1 + (1 - D)Y^0$.

Assuming only the stable unit treatment value assumption (SUTVA) on the relationship between the treatment and the outcome (Rubin, 1980, 1986), the observed outcome disparity between group a and b can be decomposed into four components:

$$\mathrm{E}_a(Y) - \mathrm{E}_b(Y) \tag{1}$$
$$= \underbrace{\mathrm{E}_a\left(Y^0\right) - \mathrm{E}_b\left(Y^0\right)}_{\text{Baseline}} + \underbrace{\mathrm{E}_b(\tau)[\mathrm{E}_a(D) - \mathrm{E}_b(D)]}_{\text{Prevalence}} + \underbrace{\mathrm{E}_a(D)[\mathrm{E}_a(\tau) - \mathrm{E}_b(\tau)]}_{\text{Effect}} + \underbrace{\mathrm{Cov}_a(D, \tau) - \mathrm{Cov}_b(D, \tau)}_{\text{Selection}}.$$

First, the "baseline" component reflects the difference in mean baseline potential outcome, Y^0, between groups, i.e., the outcome disparity in the complete absence of the treatment.[1] In our

---

[1] The baseline component is identical to the "counterfactual disparity measure" proposed by Naimi et al. (2016). In contrast to the two-way decomposition of Naimi et al. (2016), our four-way decomposition distinguishes more mechanisms.

application, the baseline component is the part of the income disparity in adulthood that is not attributable to college graduation in any way. Second, the "prevalence" component indicates how much of the group disparity is due to differential prevalence of the treatment. For example, it indicates the extent to which the difference in college graduation rates between parental income groups contributes to the outcome disparity. Third, the "effect" component reflects the difference in average treatment effects (ATE) between groups. Thus, it reveals the contribution of group-differential ATEs of college graduation to the adult income disparity.

Fourth, the "selection" component captures the contribution of differential selection into treatment based on the individual-level treatment effects. Selection into treatment within each group is captured by $\text{Cov}_g(D, \tau)$. This covariance is positive if group members who would benefit more from the treatment are more likely to receive the treatment. In our example, differential selection will increase the income disparity in adulthood if selection into college graduation is more positive in the higher parental income group than in the lower parental income group.

Both the effect component and the selection component account for the contribution of effect heterogeneity to group disparities. Whereas the effect component captures the contribution of *between*-group effect heterogeneity, the selection component captures the contribution of *within*-group effect heterogeneity. To our knowledge, no prior decomposition has captured the role of differential selection into treatment in the generation of group disparities.

To further explicate our novel selection component, we provide two interpretations for the covariance between the treatment and the treatment effect, $\text{Cov}(D, \tau)$. When the receipt of treatment is mainly based on self-selection (e.g, college graduation), the covariance may indicate the extent to which the choice to take up the treatment is rational with respect to returns to the treatment. This interpretation has been extensively studied in economics and sociology (Heckman and Vytlacil, 2005; Brand and Xie, 2010; Heckman et al., 2018), where a quantity closely related to $\text{Cov}(D, \tau)$ has been labeled "sorting on gains". We formally explicate the connection to "sorting on gains" in Appendix F. On the other hand, if treatment assignment is mainly administered by external decision-makers (e.g., drug prescription), the covariance indicates how effectively the treatment is assigned to individuals (see Appendix B). Thus, the selection component could also be called the sorting component or the effectiveness component.

Finally, the sum of the prevalence, effect, and selection components can be thought of as the unconditional total contribution of the treatment. Thus, equation (1) decomposes the total contribution of a treatment into three distinct causal mechanisms.

2.2 Interventional interpretation

We expect that most researchers will use our decomposition for retrospective causal explanation of existing disparities. However, since our decomposition is formulated in counterfactual terms, it is also prescriptive for future interventions to change disparities. Here, we explicate a three-step sequential intervention on the treatment, which successively eliminates the selection, prevalence, and effect components, in this order. The baseline component is the remaining disparity after the intervention.

We express this three-step sequential intervention using the randomized intervention notation (Didelez et al., 2006; Geneletti, 2007), where $R(D \mid g)$ represents a randomly drawn value of treatment D from group g. Then, $\mathrm{E}_g\left(Y^{R(D \mid g')}\right)$ denotes the post-intervention mean potential outcome for group g after each member of group g has received a random draw of treatment from group g'. When $g = g'$, the intervention amounts to a random redistribution of treatments within the group. Using its definition, we can rewrite the post-intervention mean potential outcome:

$$\mathrm{E}_g\left(Y^{R(D \mid g')}\right) = \mathrm{E}_g(Y^0) + \mathrm{E}_{g'}(D)\,\mathrm{E}_g(\tau). \tag{2}$$

Then the components of the unconditional decomposition can be re-written as follows:

$$\text{Selection} = \underbrace{\mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \left[\mathrm{E}_a\left(Y^{R(D \mid a)}\right) - \mathrm{E}_b\left(Y^{R(D \mid b)}\right)\right]}_{\text{Change in disparity in the first step}}$$

$$\text{Prevalence} = \underbrace{\mathrm{E}_a\left(Y^{R(D \mid a)}\right) - \mathrm{E}_b\left(Y^{R(D \mid b)}\right) - \left[\mathrm{E}_a\left(Y^{R(D \mid a)}\right) - \mathrm{E}_b\left(Y^{R(D \mid a)}\right)\right]}_{\text{Change in disparity in the second step}}$$

$$\text{Effect} = \underbrace{\mathrm{E}_a\left(Y^{R(D \mid a)}\right) - \mathrm{E}_b\left(Y^{R(D \mid a)}\right) - \left[\mathrm{E}_a\left(Y^0\right) - \mathrm{E}_b\left(Y^0\right)\right]}_{\text{Change in disparity in the third step}}$$

$$\text{Baseline} = \underbrace{\mathrm{E}_a\left(Y^0\right) - \mathrm{E}_b\left(Y^0\right)}_{\text{Remaining disparity}}.$$

The first step of the intervention internally randomizes the treatment within each group. In this step, the pre-intervention disparity is $\mathrm{E}_a(Y) - \mathrm{E}_b(Y)$, and the post-intervention disparity is $\mathrm{E}_a\left(Y^{R(D \mid a)}\right) - \mathrm{E}_b\left(Y^{R(D \mid b)}\right)$. Since randomizing the treatment within each group sets $\mathrm{Cov}_a(D, \tau) = \mathrm{Cov}_b(D, \tau) = 0$ and removes differential selection between groups, the change in disparity resulting from this step equals the selection component.

The second step of the intervention equalizes treatment prevalence by giving members of

group b random draws of treatment from group a. In this step, the pre-intervention disparity is $\mathrm{E}_a\left(Y^{R(D|a)}\right) - \mathrm{E}_b\left(Y^{R(D|b)}\right)$, and the post-intervention disparity is $\mathrm{E}_a\left(Y^{R(D|a)}\right) - \mathrm{E}_b\left(Y^{R(D|a)}\right)$. Thus, the prevalence component is the change in disparity resulting from this equalization step.[2]

The third step of the intervention sets the treatment to zero for all individuals in both groups. In the third step, the pre-intervention disparity is $\mathrm{E}_a\left(Y^{R(D|a)}\right) - \mathrm{E}_b\left(Y^{R(D|a)}\right)$, and the post-intervention disparity is $\mathrm{E}_a\left(Y^0\right) - \mathrm{E}_b\left(Y^0\right)$. When nobody receives the treatment, the influence of differential ATEs across groups is deactivated, which is why the change in disparity in this step equals the effect component.[3]

Finally, at the end of the three-step sequential intervention, the remaining disparity is the baseline component, which is the part of the disparity not attributable to the treatment and cannot be affected by the intervention on the treatment.[4] We present a visualization of the unconditional decomposition in Appendix B.

Importantly, each step of the sequential intervention removes one single component without affecting any other components. Conversely, each component of the decomposition isolates a distinct policy lever that can be sequentially intervened upon. Therefore, estimates of our unconditional decomposition directly enable policy makers to choose between implementing just the first step, the first two steps, or all three steps together. Such freedom of choice is useful, for example, when the first step is effective for reducing the disparity, but the later two steps would either empirically increase the disparity or be infeasible to implement (e.g., due to cost concerns). In our empirical example, if the selection component is positive, then a low-cost information intervention might ameliorate the income disparity by preventing suboptimal selection into college among lower-income students, even if the prevalence and effect components do not provide viable disparity-reducing policies.

---

[2]This step also justifies the scaling factors in the prevalence and effect components. Intuitively, the prevalence component is scaled by $\mathrm{E}_b(\tau)$, because randomly changing treatment prevalence in group b affects the outcome disparity only if the treatment has an effect in group b on average. The scaling factor in the effect component follows algebraically to complete the decomposition. Different interventions would lead to different scaling factors. For example, if group a is instead intervened to have the treatment prevalence of group b, the prevalence component would be scaled by $\mathrm{E}_a(\tau)$.

[3]The effect component also corresponds to the change in disparity brought about by giving group a the ATE of group b. Although the notion of interventions on effects or structural relations often appears in the literature (e.g., Malinsky, 2018; Diderichsen et al., 2019; Brady et al., 2017), these interventions cannot be expressed in the potential outcomes notation. In contrast, the third step of our sequential intervention fits in the potential outcomes framework, as it eliminates the effect component by intervening on the treatment variable.

[4]The three-step interventional interpretation also elucidates the role of the treatment coding scheme. If the labels of treatment ($D = 1$) and control ($D = 0$) are swapped, the selection and prevalence components will remain unchanged, as their corresponding intervention steps are agnostic to treatment labels. However, the effect and baseline components will change, because the third intervention step assigns everybody to what is labelled as control. Therefore, the choice of treatment labels is a substantive decision to the extent that the researcher seeks to interpret the baseline and effect components.

2.3 Comparison with prior work

We compare our new approach to three prior decomposition frameworks. We highlight that no prior decomposition contains a selection component; furthermore, prior approaches require strong assumptions to recover any of our other components.

2.3.1 Comparison with the KBO decomposition

Disparities research in the social and health sciences traditionally employs KBO decompositions. The KBO decomposition that most closely resembles our approach decomposes the outcome disparity between groups into four components with respect to treatment D and pre-treatment covariates $\boldsymbol{X}$:[5]

$$\mathrm{E}_a(Y) - \mathrm{E}_b(Y) \tag{3}$$

$$= \underbrace{\alpha_a - \alpha_b}_{\text{Intercept}} + \underbrace{\beta_b[\mathrm{E}_a(D) - \mathrm{E}_b(D)]}_{\text{Endowment}} + \underbrace{\mathrm{E}_a(D)[\beta_a - \beta_b]}_{\text{Slope}} + \underbrace{\boldsymbol{\gamma}_b^{\mathsf{T}}[\mathrm{E}_a(\boldsymbol{X}) - \mathrm{E}_b(\boldsymbol{X})] + \mathrm{E}_a(\boldsymbol{X})^{\mathsf{T}}(\boldsymbol{\gamma}_a - \boldsymbol{\gamma}_b)}_{\text{Residual}},$$

where α_g, β_g, and $\boldsymbol{\gamma}_g$ are coefficients from the group-specific linear regression that contains only main effects:

$$Y = \alpha_g + \beta_g D + \boldsymbol{\gamma}_g^{\mathsf{T}} \boldsymbol{X} + \epsilon.$$

This decomposition attains a causal interpretation under (i) the causal assumption of conditional ignorability of the treatment, $Y^d \perp\!\!\!\perp D \mid g, \boldsymbol{x}, \forall d, g, \boldsymbol{x}$, and (ii) the parametric assumption that the group-specific linear regressions are correctly specified. If and only if both assumptions are satisfied, the endowment and slope components in the KBO decomposition are equivalent to our prevalence and effect components, respectively; and the sum of the intercept and residual components equals our baseline component.[6]

Our decomposition (1) differs from the KBO decomposition in three respects. First, our decomposition is inherently causal, because it is directly formulated as estimands in potential outcomes notation. In contrast, the KBO decomposition, as an estimation framework, requires additional assumptions to support a causal interpretation. Such assumptions are rarely stated

---

[5]In practice, many KBO decompositions are farther removed from our approach. In particular, research often does not separate D from $\boldsymbol{X}$ or heed the temporal order of variables.

[6]Prior work has offered alternative causal interpretations for KBO decompositions under various assumptions. For example, a prominent literature shows that KBO decompositions can estimate the average treatment effect on the treated (Fortin et al., 2011; Kline, 2011; Yamaguchi, 2015). Similarly, Chernozhukov et al. (2018b) show that a KBO decomposition can estimate the partial treatment effect. And Huber (2015) discusses using a KBO decomposition for estimating the natural indirect effect (Pearl, 2001). None of these interpretations accommodates a descriptive group variable, with one important exception: Jackson and VanderWeele (2018) establish a connection between the KBO decomposition (3) and the unconditional random equalization decomposition that is discussed in Section 2.3.3.

in practice. Second, our decomposition is nonparametric, whereas the KBO decomposition is model-based and hence relies on a particular functional form. Third, the KBO decomposition does not contain a selection component, because the assumed functional form imposes effect homogeneity within each group. By contrast, our nonparametric decomposition does not impose any effect homogeneity hence contains a selection component as a distinctive conceptual contribution.

2.3.2 Comparison with causal mediation analysis

CMA decomposes a total effect of an exposure on an outcome into components in terms of an intermediate mediator. When CMA is used to understand a group disparity, our group variable is treated as the exposure, and our treatment variable becomes the mediator. The CMA literature is vast and contains many different decompositions (VanderWeele, 2015). We focus on a three-way CMA decomposition (VanderWeele and Tchetgen Tchetgen, 2014; VanderWeele, 2014) that is useful for illustrating similarities and differences with our decomposition (1):

$$
\underbrace{\mathrm{E}\left(Y^a\right) - \mathrm{E}(Y^b)}_{\text{Total effect of exposure } G} \tag{4}
$$

$$
= \underbrace{\mathrm{E}\left(Y^{a,0} - Y^{b,0}\right)}_{\text{Controlled direct effect}} + \underbrace{\mathrm{E}\left[\left(Y^{b,1} - Y^{b,0}\right)\left(D^a - D^b\right)\right]}_{\text{Pure indirect effect}} + \underbrace{\mathrm{E}\left[D^a\left(Y^{a,1} - Y^{a,0} - Y^{b,1} + Y^{b,0}\right)\right]}_{\text{Portion attributable to interaction}},
$$

where Y^g and D^g are, respectively, the potential outcomes of Y and D when assigned group g, and $Y^{g,d}$ is the potential outcome of Y when jointly assigned both group g and treatment d.

Certain equivalences between equation (4) and our unconditional decomposition can be established under very strong assumptions: two unconditional ignorability assumptions for G, i.e., $Y^{g,d} \perp\!\!\!\perp G$, $\forall d, g$, and $D^g \perp\!\!\!\perp G$, $\forall g$; two SUTVA-type assumptions, $\mathrm{E}_g\left(Y^{g,d}\right) = \mathrm{E}_g\left(Y^d\right)$ and $\mathrm{E}_g\left(D^g\right) = \mathrm{E}_g(D), \forall d, g$; and a cross-world independence assumption (Pearl, 2001), $Y^{g,d} \perp\!\!\!\perp D^{g'}, \forall d, g, g'$. Under these assumptions, the conditional direct effect (CDE) equals our baseline component; the pure indirect effect (PIE) equals our prevalence component; and the portion attributable to interaction (PAI) equals our effect component. These equivalences are intuitive: both the CDE and the baseline component capture a group-based outcome difference when the intermediate variable, D, is held at 0; both the PIE and the prevalence component address the role of the prevalence of D in the relationship between the group and the outcome; finally, the PAI and the effect component both reflect how the effect of D interacts with group membership.

However, all existing CMA decompositions, to our knowledge, differ from our unconditional decomposition in three crucial respects. First, CMA decomposes a different quantity: a total

effect of group membership, rather than the descriptive group disparity. Our focus on decomposing descriptive disparities is useful, because descriptive disparities between groups are often the object of interest in their own right in the social and health sciences and the focus of popular and policy concerns. For example, income disparities in adulthood between the children of rich and poor parents are often viewed as concerning, regardless of whether these disparities originate from the causal effect of parental income or from confounding factors such as parents' education and race. Furthermore, some group variables, such as race and gender, may be immutable attributes on which it is hard to define an intervention (Rubin, 1974; Holland, 1986). As a result, CMA estimands, when encoding a hypothetical intervention on group membership, may not even be well-defined.

Second, the identification assumptions of CMA are much stronger. Whereas, as we show in Section 3, our approach requires only causal assumptions on D, CMA requires assumptions on both D and G (VanderWeele, 2015). This is why, above, we need assumptions on G to establish equivalences between equation (4) and our unconditional decomposition.

Third, there is no selection component in CMA. Although it is possible to define an analogous selection component for CMA in terms of exposure-induced selection into the mediator based on the net effect of the mediator on the outcome, i.e., $\mathrm{Cov}\left(D^a, Y^{a,1} - Y^{a,0}\right) - \mathrm{Cov}\left(D^b, Y^{b,1} - Y^{b,0}\right)$, such a component, to our knowledge, has not appeared in existing CMA decompositions.[7]

2.3.3 Comparison with the unconditional random equalization decomposition

The newest entry into decomposition methodology is the random equalization decompositions (VanderWeele and Robinson, 2014; Jackson and VanderWeele, 2018; Sudharsanan and Bijlsma, 2021; Lundberg, 2024), which exist in unconditional (URED) and conditional versions (CRED).

The URED is defined in terms of a one-step intervention that equalizes treatment prevalence by assigning random draws of the treatment from the advantaged group, $R(D \mid a)$, to the disadvantaged group (Jackson and VanderWeele, 2018):

$$\mathrm{E}_a(Y) - \mathrm{E}_b(Y) = \underbrace{\mathrm{E}_b\left(Y^{R(D|a)}\right) - \mathrm{E}_b(Y)}_{\text{Change in disparity}} + \underbrace{\mathrm{E}_a(Y) - \mathrm{E}_b\left(Y^{R(D|a)}\right)}_{\text{Remaining disparity}}.$$

Thus, the outcome disparity is decomposed into the change in disparity resulting from the

---

[7]In a companion paper, we discuss in more detail the analogous selection component for CMA in relation to the difference between the total effect and its randomized interventional analogue (Yu et al., 2024). Also, Zhou (2024) presents a decomposition that contains the first half of CMA's selection component, $\mathrm{Cov}(D^a, Y^{a,1} - Y^{a,0})$.

intervention and the remaining disparity after the intervention.

The URED shares several similarities with our approach. It decomposes the observed disparity, and it is formulated as causal estimands with interventional interpretations. Furthermore, when there is no selection into treatment, the change in disparity of the URED equals our prevalence component, and the remaining disparity equals the sum of our baseline and effect components.

However, the URED also differs from our unconditional decomposition in important respects. First, as a two-way decomposition, the URED contains less information than our four-way decomposition. Correspondingly, the URED only informs a one-step intervention, which is less flexible than the three-step sequential intervention our decomposition provides.

Second, when selection is present, the URED does not isolate any of the mechanisms identified in our decomposition (1). Specifically, the URED does not have a component isolating the contribution of differential treatment prevalence across groups. Formally, the change in disparity of the URED equals $E_b(\tau)[E_a(D) - E_b(D)] - \text{Cov}_b(D, \tau)$, which mixes our prevalence component with selection into treatment in the disadvantaged group. Intuitively, by assigning the disadvantaged group random draws of the treatment, the URED's one-step intervention simultaneously equalizes treatment prevalence across groups and randomizes treatment assignment in the disadvantaged group. By contrast, as noted in Section 2.2, the three-step sequential intervention corresponding to our unconditional decomposition neatly separates equalization from randomization. Consequently, unlike our prevalence component, the URED's change in disparity is generally non-zero even if there is no difference in treatment prevalence, $E_a(D) = E_b(D)$.

By the same token, the URED does not contain a selection component, because the contribution of differential selection to the disparity is split between the URED's two components. The change in disparity contains the negative of selection into treatment in the disadvantaged group, and the remaining disparity contains selection into treatment in the advantaged group alongside our baseline and effect components, $E_a(Y^0) - E_b(Y^0) + E_a(D)[E_a(\tau) - E_b(\tau)] + \text{Cov}_a(D, \tau)$.[8]

2.4 Conditional decomposition

Our conditional decomposition evaluates the contributions of the treatment to the outcome disparity conditional on a vector of pre-treatment covariates $\boldsymbol{Q}$. Thus, the conditional decomposi-

---

[8]Lundberg (2024) introduces a variant of the URED whose one-step intervention equalizes treatment prevalence across groups by assigning random draws from the marginal treatment distribution in the population, $R(D)$, to both the advantaged and the disadvantaged groups. The resulting change in disparity is $E_a(Y) - E_b(Y) - \left[E_a\left(Y^{R(D)}\right) - E_b\left(Y^{R(D)}\right)\right]$. Rewriting this as $E(\tau)[E_a(D) - E_b(D)] + \text{Cov}_a(D, \tau) - \text{Cov}_b(D, \tau) - [\Pr(G = a) - \Pr(G = b)][E_a(D) - E_b(D)][E_a(\tau) - E_b(\tau)]$ shows that this URED variant also fails to separate the contributions of differential prevalence and differential selection.

tion quantifies how the disparity would change if differences in treatment selection, prevalence, and effect were removed only between members of the advantaged and disadvantaged groups who share the same values of $\boldsymbol{Q}$. The conditional decomposition is useful in practice because its corresponding interventions may be normatively more desirable or practically more feasible than their unconditional counterparts (see Jackson, 2021).

In our application, $\boldsymbol{Q}$ is academic achievement in high school. The conditional decomposition thus informs interventions on college graduation that are conducted only within levels of prior achievement, preserving the relationship between prior achievement and degree attainment.

Assumption 2 (Common support). $\text{supp}_a(\boldsymbol{Q}) = \text{supp}_b(\boldsymbol{Q})$.

In order to introduce the conditional decomposition, we assume common support on $\boldsymbol{Q}$. Under Assumptions 1 and 2, we then obtain the following conditional decomposition:

$$\mathrm{E}_a(Y) - \mathrm{E}_b(Y) = \underbrace{\mathrm{E}_a\left(Y^0\right) - \mathrm{E}_b\left(Y^0\right)}_{\text{Baseline}} + \underbrace{\int\left[\mathrm{E}_a(D \mid \boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{q})\right]\mathrm{E}_b(\tau \mid \boldsymbol{q})f_b(\boldsymbol{q})\mathrm{d}\boldsymbol{q}}_{\text{Conditional prevalence}} \tag{5}$$

$$+ \underbrace{\int\left[\mathrm{E}_a(\tau \mid \boldsymbol{q}) - \mathrm{E}_b(\tau \mid \boldsymbol{q})\right]\mathrm{E}_a(D \mid \boldsymbol{q})f_a(\boldsymbol{q})\mathrm{d}\boldsymbol{q}}_{\text{Conditional effect}} + \underbrace{\mathrm{E}_a[\mathrm{Cov}_a(D, \tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D, \tau \mid \boldsymbol{Q})]}_{\text{Conditional selection}}$$

$$+ \underbrace{\int \mathrm{E}_a(D \mid \boldsymbol{q})\,\mathrm{E}_b(\tau \mid \boldsymbol{q})[f_a(\boldsymbol{q}) - f_b(\boldsymbol{q})]\mathrm{d}\boldsymbol{q}}_{\boldsymbol{Q}\text{-distribution}}.$$

The baseline component in the conditional decomposition is the same as the baseline component in the unconditional decomposition; it is the level of the disparity had nobody received the treatment. The conditional prevalence component represents the part of the disparity that is due to differences in treatment prevalence across groups within levels of $\boldsymbol{Q}$. In our application, it indicates the extent to which the difference in college graduation rates between equally achieving advantaged and disadvantaged students contributes to the income disparity in adulthood. The conditional effect component reflects the contribution of the difference in the group-specific conditional average treatment effect (CATE) given $\boldsymbol{Q}$. Thus, it reveals the contribution of group-differential CATEs of college graduation given prior academic achievement. The conditional selection component is the contribution of group-differential selection into treatment net of $\boldsymbol{Q}$, for example, the contribution of differential sorting into college graduation after accounting for achievement in high school.

Analogous to the unconditional total contribution of the treatment, the conditional total contribution of the treatment is thus the sum of the conditional prevalence, effect, and selection components. The $\boldsymbol{Q}$-distribution component, which appears only in the conditional decomposi-

tion, equals the difference between the unconditional and conditional total contributions of the treatment to the outcome disparity. Hence, the $\boldsymbol{Q}$-distribution component is the between-$\boldsymbol{Q}$, as opposed to within-$\boldsymbol{Q}$, part of the unconditional total contribution.[9]

2.4.1 Interventional interpretation

Similar to the unconditional case, our conditional decomposition has an interventional interpretation that maps a three-step sequential intervention to the decomposition components. Let $\mathrm{E}_g\left(Y^{R(D|g',\boldsymbol{Q})}\right)$ be the mean potential outcome of group g when its members with certain $\boldsymbol{Q}$ values are given treatments randomly drawn from those members of group g' who have the same $\boldsymbol{Q}$ values. We can rewrite this mean potential outcome as:

$$\mathrm{E}_g\left(Y^{R(D|g',\boldsymbol{Q})}\right) = \mathrm{E}_g\left(Y^0\right) + \int \mathrm{E}_g(\tau \mid \boldsymbol{q})\,\mathrm{E}_{g'}(D \mid \boldsymbol{q})f_g(\boldsymbol{q})\mathrm{d}\boldsymbol{q}. \tag{6}$$

This allows us to express the components of the conditional decomposition as:

$$\text{Conditional selection} = \underbrace{\mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \left[\mathrm{E}_a\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|b,\boldsymbol{Q})}\right)\right]}_{\text{Change in disparity in the first step}} \tag{7}$$

$$\text{Conditional prevalence} = \underbrace{\mathrm{E}_a\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|b,\boldsymbol{Q})}\right) - \left[\mathrm{E}_a\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|a,\boldsymbol{Q})}\right)\right]}_{\text{Change in disparity in the second step}}$$

$$\text{Conditional effect} + \boldsymbol{Q}\text{-distribution} = \underbrace{\mathrm{E}_a\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \left[\mathrm{E}_a\left(Y^0\right) - \mathrm{E}_b\left(Y^0\right)\right]}_{\text{Change in disparity in the third step}}$$

$$\text{Baseline} = \underbrace{\mathrm{E}_a\left(Y^0\right) - \mathrm{E}_b\left(Y^0\right)}_{\text{Remaining disparity}}.$$

Therefore, the first step of the intervention randomizes the treatment within groups and within $\boldsymbol{Q}$ levels. In this step, the pre-intervention disparity is $\mathrm{E}_a(Y) - \mathrm{E}_b(Y)$, the post-intervention disparity is $\mathrm{E}_a\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|b,\boldsymbol{Q})}\right)$, and the resulting change in disparity equals the conditional selection component. The second step equalizes treatment prevalence across groups and within $\boldsymbol{Q}$ levels. In this step, the pre-intervention disparity is $\mathrm{E}_a\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|b,\boldsymbol{Q})}\right)$, the post-intervention disparity is $\mathrm{E}_a\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|a,\boldsymbol{Q})}\right)$, and the change in disparity equals the conditional prevalence component. The third step sets the treat-

---

[9]Furthermore, the $\boldsymbol{Q}$-distribution component identifies the amount of the disparity associated with (i) the relationship between G and $\boldsymbol{Q}$ and (ii) the within-group relationship between $\{D, \tau\}$ and $\boldsymbol{Q}$. Clearly, if $G \perp\!\!\!\perp \boldsymbol{Q}$ or $\{D, \tau\} \perp\!\!\!\perp \boldsymbol{Q} \mid g$, the $\boldsymbol{Q}$-distribution component will be zero. In our application, if there is no group difference in prior achievement or if prior achievement is not associated with college graduation and the effect of college graduation on adult income, then the $\boldsymbol{Q}$-distribution component will be zero.

ment to zero for all individuals. Thus, this final step retains only the baseline component as the post-intervention disparity, while the change in disparity is the sum of the conditional effect and $\boldsymbol{Q}$-distribution components.[10] [11]

2.4.2 Comparison with path-specific effects in causal mediation analysis

Since there is no apparent counterpart of our conditional decomposition in the KBO framework, we proceed to comparison with CMA. As a subfield of CMA, path-specific effects provide fine-grained decompositions of the total effect of an exposure when there is more than one mediator (Avin et al., 2005; VanderWeele, 2015, Section 5.5). Treating the group variable G as the exposure and all variables in $\{\boldsymbol{Q}, D\}$ as mediators, such that $\boldsymbol{Q}$ is temporally after G but before D, we have the following path-specific decomposition of the total effect of the exposure:

$$
\underbrace{\mathrm{E}\left(Y^a\right) - \mathrm{E}(Y^b)}_{\text{Total effect of exposure } G}
$$
$$
= \underbrace{\mathrm{E}\left(Y^{a,\boldsymbol{Q}^a,D^{a,\boldsymbol{Q}^a}} - Y^{b,\boldsymbol{Q}^a,D^{a,\boldsymbol{Q}^a}}\right)}_{\text{Paths involving neither } \boldsymbol{Q} \text{ nor } D} + \underbrace{\mathrm{E}\left(Y^{b,\boldsymbol{Q}^a,D^{a,\boldsymbol{Q}^a}} - Y^{b,\boldsymbol{Q}^b,D^{a,\boldsymbol{Q}^b}}\right)}_{\text{Paths involving } \boldsymbol{Q}} + \underbrace{\left(Y^{b,\boldsymbol{Q}^b,D^{a,\boldsymbol{Q}^b}} - Y^{b,\boldsymbol{Q}^b,D^{b,\boldsymbol{Q}^b}}\right)}_{\text{Paths involving } D \text{ but not } \boldsymbol{Q}},
$$

where $Y^{g,\boldsymbol{Q}^{g'},D^{g''},\boldsymbol{Q}^{g'''}}$ is a potential outcome with "nested counterfactuals" indicating the assignment of G, $\boldsymbol{Q}$, and D variables. The total effect of G is decomposed into the paths involving neither $\boldsymbol{Q}$ nor D (i.e., $A \to Y$), the paths involving $\boldsymbol{Q}$ (i.e., the combination of $G \to \boldsymbol{Q} \to D \to Y$ and $G \to \boldsymbol{Q} \to Y$), and the paths involving D but not $\boldsymbol{Q}$ (i.e., $G \to D \to Y$). Figure A1 in Appendix A.7. illustrates these paths.

The equivalence between the paths involving D but not $\boldsymbol{Q}$ and the conditional prevalence component can be established under some very strong assumptions detailed in Appendix A.7. This equivalence is intuitive, as our conditional prevalence component captures the contribution of differential prevalence of the treatment D after accounting for $\boldsymbol{Q}$. However, none of other components of our conditional decomposition is isolated in the path-specific decomposition

---

[10]Different from the effect component in the unconditional decomposition, there does not exist an obvious interventional step that isolates the conditional effect component itself. Setting $D = 0$ within each level of $\boldsymbol{Q}$ in analogy to the third step of the unconditional decomposition sets $D = 0$ marginally and hence eliminates both the conditional effect and $\boldsymbol{Q}$-distribution components. Equalizing the CATE given $\boldsymbol{Q}$ across groups would eliminate the conditional effects component alone, but, as noted in Footnote 3, interventions on effects are unconventional within the potential outcomes framework.

[11]For equation (6) to hold, we only require $\mathrm{supp}_g(\boldsymbol{Q}) \subseteq \mathrm{supp}_{g'}(\boldsymbol{Q})$. This implies that the three-step intervention is well-defined as long as $\mathrm{supp}_b(\boldsymbol{Q}) \subseteq \mathrm{supp}_a(\boldsymbol{Q})$. Intuitively, at each level of $\boldsymbol{Q}$ in group b, we must be able to find members of group a with the same $\boldsymbol{Q}$ values in order to conduct the equalization intervention. Note that $\mathrm{supp}_b(\boldsymbol{Q}) \subseteq \mathrm{supp}_a(\boldsymbol{Q})$ is a weaker condition than Assumption 2. However, under the weaker condition, only the pre- and post-intervention disparities are well-defined, not all components in equation (5).

under these assumptions. Moreover, unlike the path-specific decomposition, our conditional decomposition does not require any temporal order between G and $\boldsymbol{Q}$.

2.4.3 Comparison with the conditional random equalization decomposition

Analogous to the URED in Section 2.3.3, the CRED (Jackson, 2021) is defined in terms of a one-step intervention that equalizes treatment prevalence within levels of $\boldsymbol{Q}$ by assigning members of the disadvantaged group who have certain $\boldsymbol{Q}$ values to random draws of treatment from advantaged group members who have the same $\boldsymbol{Q}$ values. The CRED thus decomposes the outcome disparity into the change in disparity resulting from this intervention and the corresponding remaining disparity:

$$\mathrm{E}_a(Y) - \mathrm{E}_b(Y) = \underbrace{\mathrm{E}_b\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b(Y)}_{\text{Change in disparity}} + \underbrace{\mathrm{E}_a(Y) - \mathrm{E}_b\left(Y^{R(D|a,\boldsymbol{Q})}\right)}_{\text{Remaining disparity}}.$$

As before, the key difference between our conditional decomposition and the CRED is that the CRED is a two-way decomposition that does not isolate any of the mechanisms identified in our conditional decomposition. Specifically, the CRED does not have a conditional prevalence component. Formally, the CRED's change in disparity equals

$$\int [\mathrm{E}_a(D \mid \boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{q})] \, \mathrm{E}_b(\tau \mid \boldsymbol{q}) f_b(\boldsymbol{q}) \mathrm{d}\boldsymbol{q} - \mathrm{E}_b[\mathrm{Cov}_b(D, \tau \mid \boldsymbol{Q})], \tag{8}$$

which mixes our conditional prevalence component with the disadvantaged group's conditional selection. Intuitively, within $\boldsymbol{Q}$ levels, the CRED's intervention simultaneously equalizes treatment across groups and randomizes treatment in the disadvantaged group. Consequently, unlike our conditional prevalence component, the URED's change in disparity is generally non-zero even if there is no difference in treatment prevalence given any $\boldsymbol{Q}$ value, $\mathrm{E}_a(D \mid \boldsymbol{q}) = \mathrm{E}_b(D \mid \boldsymbol{q}), \forall \boldsymbol{q}$. The CRED does not contain a conditional selection component either, because the CRED splits the contribution of differential conditional selection between its two components.[12]

---

[12]Lundberg (2024) also proposes a variant of the CRED, whose change in disparity can be written as $\mathrm{E}_a[\mathrm{Cov}_a(D, \tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D, \tau \mid \boldsymbol{Q})] + \int [\mathrm{E}_a(D \mid \boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{q})][\mathrm{E}_a(\tau \mid \boldsymbol{q}) f_a(\boldsymbol{q}) p_b(\boldsymbol{q}) + \mathrm{E}_b(\tau \mid \boldsymbol{q}) f_b(\boldsymbol{q}) p_a(\boldsymbol{q})] \mathrm{d}\boldsymbol{q}$, where $p_g(\boldsymbol{q}) = \Pr(G = g \mid \boldsymbol{q})$. Hence, this CRED variant also does not separate the contributions of differential conditional prevalence and differential conditional selection.

3 Identification, estimation, and inference

We identify our unconditional and conditional decompositions using the standard assumptions of conditional ignorability and overlap. Without loss of generality, let $\boldsymbol{Q} \subseteq \boldsymbol{X}$.

Assumption 3 (Conditional ignorability). $Y^d \!\perp\!\!\!\perp D \mid \boldsymbol{x}, g, \forall d, \boldsymbol{x}, g$.

Assumption 4 (Overlap). $0 < \mathrm{E}(D \mid \boldsymbol{x}, g) < 1, \forall \boldsymbol{x}, g$.

We develop nonparametric and efficient estimators for our decompositions. These estimators are "one-step" estimators based on the EIFs of the decomposition components, which remove the bias from naive substitution estimators (Bickel et al., 1998; Van der Vaart, 2000; Hines et al., 2022). The estimators contain some nuisance functions, which can be estimated using flexible ML methods coupled with cross-fitting. Under conditions specified below, our estimators are $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient. Thus, we are able to construct asymptotically accurate Wald-type confidence intervals and hypothesis tests. Our estimators also have double or quadruple robustness properties.

To unburden notation, we define the following functions of the observed data: $\mu(d, \boldsymbol{X}, g) = \mathrm{E}(Y \mid d, \boldsymbol{X}, g)$, $\pi(d, \boldsymbol{X}, g) = \mathrm{Pr}(D = d \mid \boldsymbol{X}, g)$, and $\omega(d, \boldsymbol{Q}, g) = \mathrm{E}[\mu(d, \boldsymbol{X}, g) \mid \boldsymbol{Q}, g]$. Also recall that $p_g = \mathrm{Pr}(G = g)$, and $p_g(\boldsymbol{Q}) = \mathrm{Pr}(G = g \mid \boldsymbol{Q})$. We use circumflexes to denote estimated quantities.

3.1 Unconditional decomposition

All components of the unconditional decomposition can be expressed as linear combinations of the total disparity and two generic functions of the potential outcomes evaluated at appropriate values of d, g, and g': $\xi_{dg} := \mathrm{E}\left(Y^d \mid g\right)$ and $\xi_{dgg'} := \mathrm{E}\left(Y^d \mid g\right) \mathrm{E}\left(D \mid g'\right)$. The relationships between the components of the unconditional decomposition and the generic functions are as follows:

$$\mathrm{Baseline} = \xi_{0a} - \xi_{0b}$$

$$\mathrm{Prevalence} = \xi_{1ba} - \xi_{0ba} - \xi_{1bb} + \xi_{0bb}$$

$$\mathrm{Effect} = \xi_{1aa} - \xi_{0aa} - \xi_{1ba} + \xi_{0ba}$$

$$\mathrm{Selection} = \mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \xi_{0a} + \xi_{0b} - \xi_{1aa} + \xi_{0aa} + \xi_{1bb} - \xi_{0bb}.$$

Hence, the EIFs and one-step estimators for the decomposition components directly follow from those for ξ_{dg} and $\xi_{dgg'}$.[13][14] Under Assumptions 1, 3, and 4, ξ_{dg} and $\xi_{dgg'}$ can be identified as the following nonparametric functionals:

$$\xi_{dg} = \mathrm{E}\left[\mu(d, \boldsymbol{X}, g) \mid g\right]$$

$$\xi_{dgg'} = \mathrm{E}\left[\mu(d, \boldsymbol{X}, g) \mid g\right] \mathrm{E}\left[D \mid g'\right].$$

These identification results then enable the derivation of the EIFs for ξ_{dg} and $\xi_{dgg'}$.[15]

Theorem 1 (EIF, unconditional decomposition). *Under Assumptions 1, 3, and 4, the EIF of* ξ_{dg} *is*

$$\phi_{dg}(Y, D, \boldsymbol{X}, G) := \frac{\mathbb{1}(G = g)}{p_g}\left\{\frac{\mathbb{1}(D = d)}{\pi(d, \boldsymbol{X}, g)}[Y - \mu(d, \boldsymbol{X}, g)] + \mu(d, \boldsymbol{X}, g) - \xi_{dg}\right\},$$

and the EIF of $\xi_{dgg'}$ *is*

$$\phi_{dgg'}(Y, D, \boldsymbol{X}, G) := \frac{\mathbb{1}(G = g)}{p_g}\left\{\frac{\mathbb{1}(D = d)}{\pi(d, \boldsymbol{X}, g)}[Y - \mu(d, \boldsymbol{X}, g)] + \mu(d, \boldsymbol{X}, g)\right\}\mathrm{E}\left(D \mid g'\right)$$
$$+ \frac{\mathbb{1}(G = g')}{p_{g'}}\xi_{dg}\left[D - \mathrm{E}\left(D \mid g'\right)\right] - \frac{\mathbb{1}(G = g)}{p_g}\xi_{dgg'}.$$

In Appendix C, we also derive the general EIFs with survey weights.[16]

---

[13]These generic functions also provide a basis for the estimation of Jackson and VanderWeele's (2018) version of the URED, since its change in disparity can be represented as $\xi_{0b} + \xi_{1ba} - \xi_{0ba} - \mathrm{E}_b(Y)$.

[14]Note that although the selection component is defined in terms of individual-level treatment effects in equation (1), its estimation does not require the estimation of individualized treatment effects.

[15]When the conditional ignorability assumption (3) is not credible, researchers may evaluate the robustness of their causal conclusions by performing a sensitivity analysis, as we do in Appendix H. Alternatively, they may abandon the causal interpretation in favor of a descriptive interpretation. The nonparametric descriptive quantity $\mathrm{E}[\mu(1, \boldsymbol{X}, G) - \mu(0, \boldsymbol{X}, G)]$ has been labelled the average controlled difference (ACD) between treatment and control given $\boldsymbol{X}$ and G (Li et al., 2013, 2018). In the same spirit, one can also descriptively interpret the identified nonparametric functionals for our decomposition components. We define $\mathrm{E}[\mu(1, \boldsymbol{X}, g) - \mu(0, \boldsymbol{X}, g) \mid g]$ as the group specific ACD, $\mu(1, \boldsymbol{X}, G) - \mu(0, \boldsymbol{X}, G)$ as the controlled outcome difference (COD), and $\mathrm{E}(D \mid \boldsymbol{X}, G)$ as the controlled treatment prevalence (CTP). Then, the functional for the prevalence component is the group difference in treatment prevalence scaled by group b's ACD, the functional for the effect component is the group difference in ACD scaled by group a's treatment prevalence, and the functional for the selection component is the group difference in the covariance between the COD and the CTP. Formally,

Prevalence : $\mathrm{E}[\mu(1, \boldsymbol{X}, b) - \mu(0, \boldsymbol{X}, b) \mid b][\mathrm{E}(D \mid a) - \mathrm{E}(D \mid b)]$

Effect : $\mathrm{E}(D \mid a)\{\mathrm{E}[\mu(1, \boldsymbol{X}, a) - \mu(0, \boldsymbol{X}, a) \mid a] - \mathrm{E}[\mu(1, \boldsymbol{X}, b) - \mu(0, \boldsymbol{X}, b) \mid b]\}$

Selection : $\mathrm{Cov}[\mu(1, \boldsymbol{X}, a) - \mu(0, \boldsymbol{X}, a), \mathrm{E}(D \mid \boldsymbol{X}, a) \mid a] - \mathrm{Cov}[\mu(1, \boldsymbol{X}, b) - \mu(0, \boldsymbol{X}, b), \mathrm{E}(D \mid \boldsymbol{X}, b) \mid b]$.

[16]Related EIFs have appeared in prior work. Park and Kang (2024) give the EIF for $\xi_{1g} - \xi_{0g}$. The EIF for $\xi_{0,g} + \xi_{1gg'} - \xi_{0gg'}$ coincides with the EIF for the quantity denoted as θ_c in Díaz et al. (2021), when pre-treatment confounders in the latter are omitted. Neither of these prior works accommodates survey weights.

We use the EIFs as estimating equations, i.e., set their sample averages to zero and solve for ξ_{dg} and $\xi_{dgg'}$. The one-step estimators of ξ_{dg} and $\xi_{dgg'}$ thus are

$$\hat{\xi}_{dg} := \frac{1}{n} \sum \frac{\mathbb{1}(G = g)}{\hat{p}_g} \left\{ \frac{\mathbb{1}(D = d)}{\hat{\pi}(d, \boldsymbol{X}, g)} [Y - \hat{\mu}(d, \boldsymbol{X}, g)] + \hat{\mu}(d, \boldsymbol{X}, g) \right\}$$

$$\hat{\xi}_{dgg'} := \frac{1}{n} \sum \frac{\mathbb{1}(G = g)}{\hat{p}_g} \left\{ \frac{\mathbb{1}(D = d)}{\hat{\pi}(d, \boldsymbol{X}, g)} [Y - \hat{\mu}(d, \boldsymbol{X}, g)] + \hat{\mu}(d, \boldsymbol{X}, g) \right\} \hat{\mathrm{E}} \left(D \mid g' \right).$$

Each estimator contains two nuisance functions, $\pi(d, \boldsymbol{X}, g)$ and $\mu(d, \boldsymbol{X}, g)$. The estimators are consistent as long as either one of the two nuisance functions is consistently estimated. Although p_g and $\mathrm{E}(D \mid g)$ are technically also nuisance functions, their consistent estimation is left implicit hereafter.

Theorem 2 (Double robustness in consistency, unconditional decomposition). *Under Assumptions 1, 3, and 4, either consistent estimation of $\mu(d, \boldsymbol{X}, g)$ or of $\pi(d, \boldsymbol{X}, g)$ is sufficient for the consistency of $\hat{\xi}_{dg}$ and $\hat{\xi}_{dgg'}$.*[17]

The nuisance functions $\pi(d, \boldsymbol{X}, g)$ and $\mu(d, \boldsymbol{X}, g)$ can be estimated using various approaches. To avoid imposing parametric assumptions, we recommend using flexible ML models with cross-fitting (Kennedy, 2022; Chernozhukov et al., 2018a). To improve the finite-sample performance of the estimators, we may stabilize the weight, $\mathbb{1}(D = d)/\hat{\pi}(d, \boldsymbol{X}, g)$, by dividing it by its sample average.

To study the asymptotic distribution of the one-step estimators for the unconditional decomposition, we invoke three additional assumptions about the nuisance functions, which are the same as the assumptions required for the double ML estimator of the ATE (Kennedy, 2022; Chernozhukov et al., 2018a). We let $\| \cdot \|$ denote the L_2-norm.

Assumption 5a (Boundedness). With probability 1, $\hat{\pi}(d, \boldsymbol{X}, g) \geq \eta$, $\pi(d, \boldsymbol{X}, g) \geq \eta$, and $|Y - \hat{\mu}(d, \boldsymbol{X}, g)| \leq \zeta$, for some $\eta > 0$ and some $\zeta < \infty$, $\forall d, g$.

Assumption 5b (Consistency). $\|\hat{\mu}(d, \boldsymbol{X}, g) - \mu(d, \boldsymbol{X}, g)\| = o_p(1)$, and $\|\hat{\pi}(d, \boldsymbol{X}, g) - \pi(d, \boldsymbol{X}, g)\| = o_p(1)$, $\forall d, g$.

Assumption 5c (Convergence rate). $\|\hat{\pi}(d, \boldsymbol{X}, g) - \pi(d, \boldsymbol{X}, g)\| \|\hat{\mu}(d, \boldsymbol{X}, g) - \mu(d, \boldsymbol{X}, g)\| = o_p(n^{-1/2})$, $\forall d, g$.

Theorem 3 (Asymptotic distributions, unconditional decomposition). *Under Assumptions 1, 3, 4, 5a, 5b, and 5c, the cross-fitted one-step estimators for $\hat{\xi}_{dg}$ and $\hat{\xi}_{dgg'}$ are $\sqrt{n}$-consistent,*

---

[17]The estimator of our unconditional decomposition is doubly robust with respect to the same two nuisance functions as the classic augmented-inverse-probability-of-treatment-weighting (AIPW) estimator of the ATE (Robins et al., 1994).

asymptotically normal, and semiparametrically efficient, i.e., $\sqrt{n}\left(\hat{\xi}_{dg} - \xi_{dg}\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_{dg}^2\right)$, *and*
$\sqrt{n}\left(\hat{\xi}_{dgg'} - \xi_{dgg'}\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_{dgg'}^2\right)$, *where* $\sigma_{dg}^2 := \mathrm{E}[\phi_{dg}(Y, D, \boldsymbol{X}, G)^2]$ *and* $\sigma_{dgg'}^2 := \mathrm{E}[\phi_{dgg'}(Y, D, \boldsymbol{X}, G)^2]$
are the respective semiparametric efficiency bounds.

We consistently estimate σ_{dg}^2 and $\sigma_{dgg'}^2$ using the averages of the squared estimated EIFs. The asymptotic distributions can then be used to construct hypothesis tests and confidence intervals. Since the unconditional decomposition components are simple additive functions of the observed disparity, ξ_{dg}, and $\xi_{dgg'}$, all properties established for the estimators of ξ_{dg}, and $\xi_{dgg'}$ (double robustness, $\sqrt{n}$-consistency, asymptotic normality, and semiparametric efficiency) carry over to the final estimators of the decomposition components.

3.2 Conditional decomposition

Relative to the unconditional decomposition, the estimation of the conditional decomposition requires consideration of one additional generic function:

$$\xi_{dgg'g''} := \mathrm{E}\left[\mathrm{E}\left(Y^d \mid \boldsymbol{Q}, g\right) \mathrm{E}\left(D \mid \boldsymbol{Q}, g'\right) \mid g''\right],$$

where (d, g, g', g'') denotes any one of eight combinations of treatment status and group memberships. The relationships between components of the conditional decomposition and the generic functions are as follows:

$$\text{Baseline} = \xi_{0a} - \xi_{0b}$$

$$\text{Conditional prevalence} = \xi_{1bab} - \xi_{0bab} - \xi_{1bbb} + \xi_{0bbb}$$

$$\text{Conditional effect} = \xi_{1aaa} - \xi_{0aaa} - \xi_{1baa} + \xi_{0baa}$$

$$\boldsymbol{Q}\text{-distribution} = \xi_{1baa} - \xi_{0baa} - \xi_{1bab} + \xi_{0bab}$$

$$\text{Conditional selection} = \mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \xi_{0a} + \xi_{0b} - \xi_{1aaa} + \xi_{0aaa} + \xi_{1bbb} - \xi_{0bbb}.$$

Since the estimation of ξ_{dg} was discussed in the previous subsection, we now focus on $\xi_{dgg'g''}$. The EIFs, one-step estimators, and their asymptotic distributions for the components of the conditional decomposition will then follow.[18] Under Assumptions 1, 2, 3, and 4, we identify $\xi_{dgg'g''}$ as

$$\xi_{dgg'g''} = \mathrm{E}\left\{\omega(d, \boldsymbol{Q}, g) \mathrm{E}\left(D \mid \boldsymbol{Q}, g'\right) \mid g''\right\}.$$

Theorem 4 (EIF, conditional decomposition). *Under Assumptions 1, 2, 3, and 4, the EIF of*

---

[18]We thereby also provide efficient and nonparametric estimation for the change in disparity in the CRED of Jackson (2021), which can be represented as $\xi_{0b} + \xi_{1bab} - \xi_{0bab} - \mathrm{E}_b(Y)$.

$\xi_{dgg'g''}$ is

$$
\begin{aligned}
&\phi_{dgg'g''}(Y, D, \boldsymbol{X}, G) \\
&= \frac{\mathbb{1}(G = g'')}{p_{g''}} \left[\omega(d, \boldsymbol{Q}, g) \, \mathrm{E}\left(D \mid \boldsymbol{Q}, g'\right) - \xi_{dgg'g''} \right] + \frac{\mathbb{1}(G = g') p_{g''}(\boldsymbol{Q})}{p_{g'}(\boldsymbol{Q}) p_{g''}} \left[D - \mathrm{E}\left(D \mid \boldsymbol{Q}, g'\right) \right] \omega(d, \boldsymbol{Q}, g) \\
&\quad + \frac{\mathbb{1}(G = g) p_{g''}(\boldsymbol{Q})}{p_g(\boldsymbol{Q}) p_{g''}} \left\{ \frac{\mathbb{1}(D = d)}{\pi(d, \boldsymbol{X}, g)} [Y - \mu(d, \boldsymbol{X}, g)] + \mu(d, \boldsymbol{X}, g) - \omega(d, \boldsymbol{Q}, g) \right\} \mathrm{E}\left(D \mid \boldsymbol{Q}, g'\right).
\end{aligned}
$$

In Appendix C, we also derive the general EIF with survey weights.

We again construct the one-step estimator by using the EIF as an estimating equation.

$$
\begin{aligned}
&\hat{\xi}_{dgg'g''} \\
&= \frac{1}{n} \sum \frac{\mathbb{1}(G = g'')}{\hat{p}_{g''}} \hat{\omega}(d, \boldsymbol{Q}, g) \hat{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g'\right) + \frac{\mathbb{1}(G = g') \hat{p}_{g''}(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q}) \hat{p}_{g''}} \left[D - \hat{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g'\right) \right] \hat{\omega}(d, \boldsymbol{Q}, g) \\
&\quad + \frac{\mathbb{1}(G = g) \hat{p}_{g''}(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q}) \hat{p}_{g''}} \left\{ \frac{\mathbb{1}(D = d)}{\hat{\pi}(d, \boldsymbol{X}, g)} [Y - \hat{\mu}(d, \boldsymbol{X}, g)] + \hat{\mu}(d, \boldsymbol{X}, g) - \hat{\omega}(d, \boldsymbol{Q}, g) \right\} \hat{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g'\right).
\end{aligned}
$$

This estimator contains five nuisance functions: $p_g(\boldsymbol{Q})$, $\pi(d, \boldsymbol{X}, g)$, $\mu(d, \boldsymbol{X}, g)$, $\mathrm{E}(D \mid \boldsymbol{Q}, g)$, and $\omega(d, \boldsymbol{Q}, g)$. As is the case for the unconditional decomposition, consistent estimation of the conditional decomposition does not require all nuisance functions to be consistently estimated. In particular, $\hat{\xi}_{dgg'g''}$ is quadruply robust to inconsistently estimated nuisance functions.

Theorem 5 (Quadruple robustness in consistency, conditional decomposition)**.** *Under Assumptions 1, 2, 3, and 4, $\hat{\xi}_{dgg'g''}$ is consistent if one of four minimal conditions holds, as summarized in Table 1.*

As before, we recommend estimating the nuisance functions nonparametrically using ML and cross-fitting. The estimation of $\omega(d, \boldsymbol{Q}, g)$ deserves particular attention, because it can be doubly robust itself. Specifically, we recommend a pseudo-outcome approach (e.g., van der Laan, 2006; Semenova and Chernozhukov, 2021), where the pseudo outcome for each d is defined as

$$
\delta_d (Y, D, \boldsymbol{X}, G) := \frac{\mathbb{1}(D = d)}{\pi(d, \boldsymbol{X}, G)} [Y - \mu(d, \boldsymbol{X}, G)] + \mu(d, \boldsymbol{X}, G),
$$

which is motivated by the fact that $\omega(d, \boldsymbol{Q}, g) = \mathrm{E}\left[\delta_d (Y, D, \boldsymbol{X}, G) \mid \boldsymbol{Q}, g\right]$. We first randomly draw two disjoint subsamples from the data. Then we estimate $\pi(d, \boldsymbol{X}, G)$ and $\mu(d, \boldsymbol{X}, G)$ in each subsample *without* cross-fitting and obtain the estimated pseudo outcome $\hat{\delta}_d (Y, D, \boldsymbol{X}, G)$. Finally, we obtain estimates of $\omega(d, \boldsymbol{Q}, g)$ using cross-fitting, i.e., we fit $\mathrm{E}\left[\hat{\delta}_d (Y, D, \boldsymbol{X}, G) \mid \boldsymbol{Q}, g\right]$ separately in each subsample and plug in values of $\boldsymbol{Q}$ from the respective other subsample. Using this procedure, we ensure that the fitting of $\omega(d, \boldsymbol{Q}, g)$, which relies on estimating the pseudo

Table 1: Quadruple robustness of $\hat{\xi}_{dgg'g''}$. For each scenario defined by g, g' and g'', $\hat{\xi}_{dgg'g''}$ is consistent if any of four minimal sets of nuisance functions indicated by check marks is consistently estimated. The first three panels concern the conditions for the consistent estimation of the conditional prevalence ($g = g''$), conditional effect ($g' = g''$), and conditional selection ($g = g' = g''$) components, respectively. Since all conditions require that either $\pi(d, \boldsymbol{X}, g)$ or $\mu(d, \boldsymbol{X}, g)$ be consistently estimated, the baseline component is always consistently estimated. The bottom panel shows the four minimal combinations of nuisance functions that must be consistently estimated so that all components of the conditional decomposition are consistently estimated simultaneously.

$\mu(d, \boldsymbol{X}, g)$	$\pi(d, \boldsymbol{X}, g)$	$\omega(d, \boldsymbol{Q}, g)$	$p_g(\boldsymbol{Q})$	$\mathrm{E}(D \mid \boldsymbol{Q}, g)$
		$g = g''$ (conditional prevalence)		
✓		✓	✓	
	✓	✓	✓	
✓				✓
	✓			✓
		$g' = g''$ (conditional effect)		
✓		✓		
	✓	✓		
✓			✓	✓
	✓		✓	✓
		$g = g' = g''$ (conditional selection)		
✓		✓		
	✓	✓		
✓				✓
	✓			✓
		All components		
✓		✓	✓	
	✓	✓	✓	
✓			✓	✓
	✓		✓	✓

outcome, is done separately in each subsample. Provided that $\mathrm{E}\left[\hat{\delta}_d\left(Y, D, \boldsymbol{X}, G\right) \mid \boldsymbol{Q}, g\right]$ can be consistently estimated, this approach enables consistent estimation of $\omega(d, \boldsymbol{Q}, g)$ if either $\mu(d, \boldsymbol{X}, g)$ or $\pi(d, \boldsymbol{X}, g)$ is consistently estimated.

To improve finite-sample performance, $\hat{\xi}_{dgg'g''}$ can be stabilized by dividing $\mathbb{1}(D = d)/\hat{\pi}(d, \boldsymbol{X}, g)$, $\mathbb{1}(G = g')\hat{p}_{g''}(\boldsymbol{Q})/\hat{p}_{g'}(\boldsymbol{Q})\hat{p}_{g''}$, and $\mathbb{1}(G = g)\hat{p}_{g''}(\boldsymbol{Q})/\hat{p}_g(\boldsymbol{Q})p_{g''}$ by their respective sample averages.

To establish the asymptotic distributions of the one-step estimators for the conditional decomposition, we invoke Assumptions 6, which augments Assumptions 5 with respect to the additional nuisance functions needed for the conditional decomposition.

Assumption 6a (Boundedness). With probability 1, $\hat{\pi}(d, \boldsymbol{X}, g) \geq \eta$, $\pi(d, \boldsymbol{X}, g) \geq \eta$, $\hat{p}_g(\boldsymbol{Q}) \geq \eta$, $p_g(\boldsymbol{Q}) \geq \eta$, $|Y - \hat{\mu}(d, \boldsymbol{X}, g)| \leq \zeta$, $|Y - \mu(d, \boldsymbol{X}, g)| \leq \zeta$, $|\mu(d, \boldsymbol{X}, g)| \leq \zeta$, $|\omega(d, \boldsymbol{Q}, g)| \leq \zeta$, and $\left|\frac{\hat{p}_g(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})}\right| \leq \zeta$, for some $\eta > 0$ and $\zeta < \infty$, $\forall d, g, g'$.

Assumption 6b (Consistency). $\|\hat{\mu}(d, \boldsymbol{X}, g) - \mu(d, \boldsymbol{X}, g)\| = o_p(1)$, $\|\hat{\pi}(d, \boldsymbol{X}, g) - \pi(d, \boldsymbol{X}, g)\| = o_p(1)$, $\|\hat{\omega}(d, \boldsymbol{Q}, g) - \omega(d, \boldsymbol{Q}, g)\| = o_p(1)$, $\left\|\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g) - \mathrm{E}(D \mid \boldsymbol{Q}, g)\right\| = o_p(1)$, and $\|\hat{p}_g(\boldsymbol{Q}) - p_g(\boldsymbol{Q})\| =$

$o_p(1), \forall d, g.$

Assumption 6c (Convergence rate). First, we require $\|\hat{\pi}(d, \boldsymbol{X}, g) - \pi(d, \boldsymbol{X}, g)\| \, \|\hat{\mu}(d, \boldsymbol{X}, g) - \mu(d, \boldsymbol{X}, g)\| = o_p(n^{-1/2}), \forall d, g$. Second, depending on the specific combination of g, g', and g'' in a $\xi_{dgg'g''}$, we require $\left\|\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g) - \mathrm{E}(D \mid \boldsymbol{Q}, g)\right\| = o_p(n^{-1/2}), \forall g$, when $g = g''$; $\|\hat{\omega}(d, \boldsymbol{Q}, g) - \omega(d, \boldsymbol{Q}, g)\| = o_p(n^{-1/2}), \forall d, g$, when $g' = g''$; and

$\|\hat{\omega}(d, \boldsymbol{Q}, g) - \omega(d, \boldsymbol{Q}, g)\| \left\|\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g) - \mathrm{E}(D \mid \boldsymbol{Q}, g)\right\| = o_p(n^{-1/2}), \forall d, g$, when $g = g' = g''$.

Theorem 6 (Asymptotic distribution, conditional decomposition). *Under Assumptions 1, 2, 3, 4, 6a, 6b, and 6c, the cross-fitted one-step estimator of $\xi_{dgg'g''}$ is $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient, i.e., $\sqrt{n}\left(\hat{\xi}_{dgg'g''} - \xi_{dgg'g''}\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{dgg'g''}\right)$, where $\sigma^2_{dgg'g''} := \mathrm{E}\left[\phi_{dgg'g''}(Y, D, \boldsymbol{X}, G)^2\right]$ is the semiparametric efficiency bound.*

Since the conditional decomposition components are additive functions of the observed disparity, ξ_{dg}, and $\xi_{dgg'g''}$, double robustness, $\sqrt{n}$-consistency, asymptotic normality, and semiparametric efficiency all carry over to the final estimators of the decomposition components. We conduct hypothesis tests and construct confidence intervals analogously to the unconditional decomposition.

Finally, for ML-based estimation of the conditional decomposition, we note a tension between asymptotic normality and semiparametric efficiency on one hand, and consistency on the other. ML does not typically satisfy the convergence rate conditions for establishing asymptotic normality and semiparametric efficiency for three components of the conditional decomposition: conditional prevalence, conditional effect, and $\boldsymbol{Q}$-distribution. However, we still prefer ML for all components because it does not impose parametric assumptions, which ensures consistent estimation. Note that this tension does not exist for the baseline and conditional selection components, because their convergence rate assumptions can reasonably be achieved with ML.

4   Application

4.1   Overview

Our application decomposes the contributions of college graduation to the perpetuation of income inequality across generations, which is also known as intergenerational income persistence, the complement to income mobility. Groups (G) are defined by parental income; the outcome (Y) is offspring's adult income; and the treatment (D) is college graduation. Although previous research has touched upon all four components of our unconditional decomposition to some extent (even indirectly), this analysis is the first to present a unified decomposition.

The baseline component of our decompositions represents the part of the disparity in adult income that is unaccounted for by college graduation. Prior research enumerates multiple sources of intergenerational income persistence that may operate independently of college graduation. For example, parental income is associated with a variety of pre-college characteristics that may directly influence income attainment, such as cognitive skills and noncognitive traits in adolescence (Heckman et al., 2006). Moreover, people from more privileged backgrounds likely benefit from their parents' human, social, and financial capital regardless of their own formal educational attainment. Interventions on college graduation would not eliminate these channels of income persistence.

Speaking to the prevalence component in our unconditional decomposition, social scientists have long regarded education as a mediator in the intergenerational reproduction of socioeconomic inequalities (Blau and Duncan, 1967, chapter 4 & 5; Featherman and Hauser, 1978, p.255-9; Ishida et al., 1995). Specifically, research documents large differences in college graduation rates across parental income groups (Ziol-Guest and Lee, 2016; Bailey and Dynarski, 2011), and simulations suggest that rising educational inequality has strengthened intergenerational income persistence over time (Bloome et al., 2018).

Related to our effect component, there is an active literature on heterogeneity in the effects of college graduation on adult income, although results vary. Brand et al. (2021) and Cheng et al. (2021) find larger effects of college graduation on the income of people from disadvantaged backgrounds. By contrast, Zhou (2019), Fiel (2020), and Yu et al. (2021) find insignificant heterogeneity in the effects of college completion on income across parental income groups. However, none of these works evaluates the extent to which income disparities can be attributed to groupwise differential effects of college.

Finally, some prior work has addressed selection into college as a function of college effects on income, i.e. $\text{Cov}(D, \tau)$. Here, too, results are mixed. Brand and Xie (2010) and Brand et al. (2021) find negative selection, i.e., those who are least likely to attend college would benefit most from it. By contrast, the instrumental variable analysis of Heckman et al. (2018) finds positive selection into college. However, prior work estimates selection in the pooled population rather than within parental income groups, thereby missing the link between the difference in group-specific selection and the group-based outcome disparity that our approach identifies. (Appendix F clarifies and synthesizes related notions of "selection into college" in the social science literature.)

4.2 Data, variables and estimation

We analyze the National Longitudinal Survey of Youth 1979, a nationally representative U.S. cohort study of individuals born between 1957 and 1964. We restrict the sample to respondents who were between 14 to 17 years old at baseline in 1979 to ensure that income origin is measured prior to respondents' college graduation. We also limit the analysis to respondents who graduated from high school by age 29. The sample size of our complete-case analysis is $N = 2,008$. Missingness mostly occurs in the outcome variable due to loss to follow-up (22%), with less missingness in other variables (¡6%). Appendix Table A1 presents associations between outcome missingness and baseline covariates.

We contrast parental income-origin groups (G), defined as the top 40% and bottom 40% of family income averaged over the first three waves of the survey (1979, 1980, and 1981, when respondents were 14 to 20 years old) and divided by the square root of the family size to adjust for need (e.g., Zhou, 2019). The treatment (D) is a binary indicator of whether the respondent graduated from college by age 29. The outcome is the percentile rank of the respondent's adult income, averaged over five survey waves between age 35 and 44, divided by the square root of family size. For the conditional decomposition, we define $\boldsymbol{Q}$ as the Armed Forces Qualification Test (AFQT) score, measured in 1980. The AFQT score is a widely used measure of academic achievement that predicts college completion.

We measure an extensive set of confounders $(\boldsymbol{X})$ at baseline, including gender, race, parental income percentile, parental education, parental presence, the number of siblings, urban residence, educational expectation, friends' educational expectation, AFQT score, age at baseline survey, the Rotter score of control locus, the Rosenberg self-esteem score, language spoken at home, Metropolitan Statistical Area category, separation from mother, school satisfaction, region of residence, and mother's working status.

We present four estimates each for our unconditional and conditional decompositions, using different models for the nuisance functions to assess robustness: three alternative ML methods (gradient boosting machine [GBM], neural networks, and random forests) and one set of parametric models. Depending on whether the left-hand-side variable is continuous or binary, the parametric models are linear or logit. Specifically, for $\mu(d, \boldsymbol{X}, g)$, we use all two-way interactions between D and $\{\boldsymbol{X}, G\}$, along with their main effects. For $\pi(d, \boldsymbol{X}, g)$ and $p_g(\boldsymbol{Q})$, the logit models contain only main effects. For $\mathrm{E}(D \mid \boldsymbol{Q}, g)$ and $\mathrm{E}[\hat{\delta}_d(Y, D, \boldsymbol{X}, G) \mid \boldsymbol{Q}, g]$, we include all main effects and two-way interactions between G and $\boldsymbol{Q}$. Appendix H contains model diagnostics and a sensitivity analysis for unobserved confounding.

4.3 Results
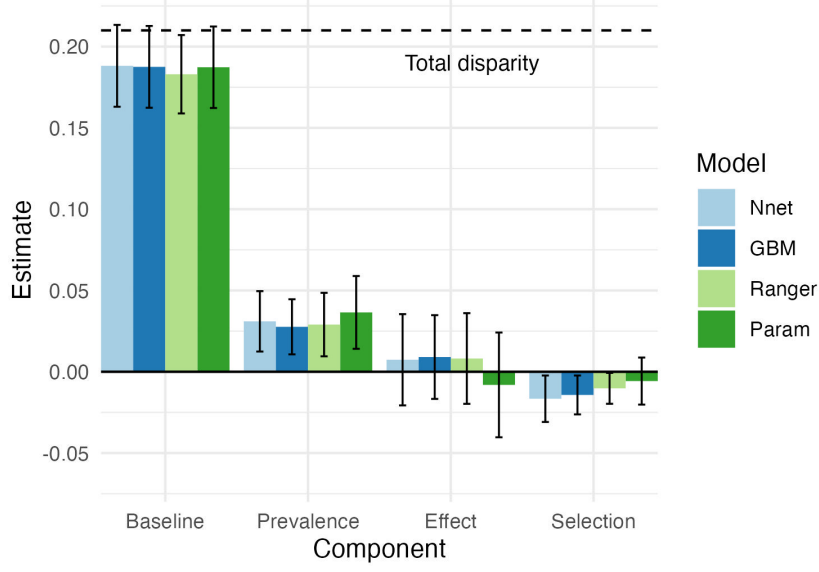
(a) Unconditional decomposition
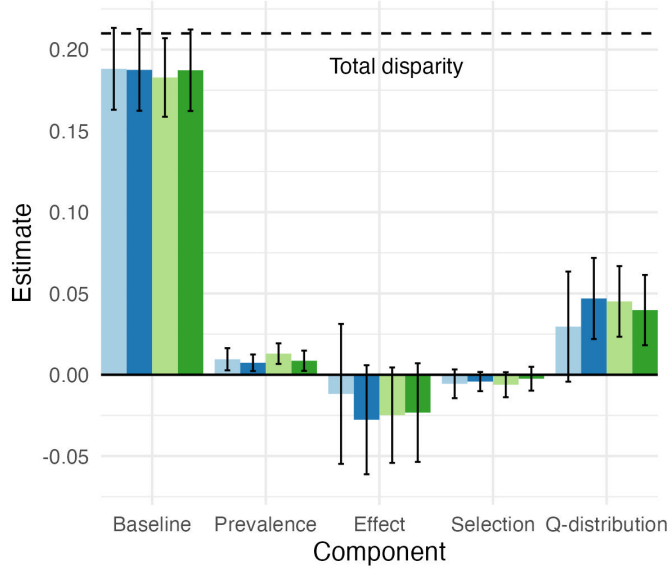
(b) Conditional decomposition

Figure 1: Decomposition estimates. For nuisance function models, Nnet=neural networks, GBM=generalized boosting machines, Ranger=random forests, Para=parametric models. Error bars indicate 95% confidence intervals, computed according to Theorems 3 and 6.

Figure 1 presents our main results for the components of the unconditional and the conditional decompositions across different models for the nuisance functions (see Appendix Tables A3 and A4 for numerical details). To aid interpretation, Appendix Table A2 additionally reports estimated group-specific means of baseline potential outcome, treatment proportions, ATEs,

and covariances between the treatment and the treatment effect, as well as group differences in these quantities. Below, all statistical significance is at the 0.05 level.

Descriptively, we find that individuals from lower-income origins on average achieve income 21 percentiles lower in their 30s and 40s than individuals from higher-income origins. This confirms the existence of intergenerational income persistence and represents the total disparity that we decompose. Below, we present our results in the order of the three-step sequential interventions and start with the unconditional decomposition.

Our estimates for the selection component in the unconditional decomposition are consistently negative. These estimates are statistically significant for all three ML models but not for the parametric models of the nuisance functions. Consequently, randomizing college graduation within each income-origin group to remove selection (the first step of the three-step sequential intervention) would increase the total disparity in adult income by around 7%. Thus, this intervention would increase intergenerational income persistence and decrease income mobility. Differential selection into college graduation reduces intergenerational income persistence because selection is positive in the disadvantaged group and negative in the advantaged group (Appendix Table A2). This lends support to stipulations from earlier research that obtaining a college degree is more of a rational decision in pursuit of economic returns among disadvantaged individuals, and more of an adherence to the social norm of attending college among advantaged individuals (Mare, 1980; Hout, 2012). The selection component is the central conceptual contribution of our approach and also the most novel finding of our empirical application.

The prevalence component in the unconditional decomposition is positive, substantively large, and statistically significant across all models for the nuisance functions, accounting for about 15% of the total disparity in adult income. Consequently, equalizing college graduation rates across income-origin groups as in the second step of the three-step sequential intervention would reduce the total disparity in adult income by about 15%. In other words, such intervention would decease intergenerational income persistence and increase income mobility. Underlying the prevalence component is the striking gap in college graduation rates by parental income, as 34% of respondents in the higher-income group, but only 9% of the lower-income group, obtained a college degree by age 29 (see Appendix Table A2).

The effect component in the unconditional decomposition is statistically insignificant due to minimal between-group effect heterogeneity. In Appendix Table A1, we show that the group-specific ATEs of college graduation on adult income range from 11 to 15 percentiles across models, all of which are statistically significant. However, the group difference in ATEs is always statistically insignificant.

The baseline component constitutes nearly 90% of the total disparity in adult income across all models for the nuisance functions. This demonstrates that most of intergenerational income persistence is due to processes that do not involve college graduation or its effects.

In sum, our unconditional decomposition reveals that college graduation in the United States plays two contradictory roles in intergenerational income persistence. On the one hand, the higher college graduation rate among individuals from higher-income origins increases intergenerational income persistence. On the other hand, some part of the income persistence is offset by the positive selection into college among individuals from lower-income origins.

The conditional decomposition quantifies the contributions of college graduation within levels of AFQT achievement scores. Hence, it informs settings in which policy makers cannot, or do not want to, change the factual relationship between prior achievement and college completion, perhaps due to normative constraints or meritocratic preferences.

In the conditional decomposition, all estimates for the conditional selection component are negative and statistically insignificant. The conditional prevalence component is positive and statistically significant, although it is much smaller than in the unconditional decomposition and only accounts for 3% to 6% of the total disparity. Hence, equalizing chances of graduating college within levels of prior achievement (in the manner of the second step of the three-step sequential intervention in Section 2.4.1) would still somewhat decrease intergenerational income persistence.

Estimates for the conditional effect components are negative and statistically insignificant. The Q-distribution component is positive and statistically significant. It reflects the strong associations between AFQT scores (Q) on one hand and parental income (G), college graduation (D), and its effects on income (τ) on the other (see Footnote 9). The positive Q-distribution component also shows that, net of prior achievement, college graduation plays a much more limited role in producing intergenerational income persistence.[19] Finally, by definition, the baseline component is the same in the unconditional and the conditional decompositions.

In Appendix H, we conduct a sensitivity analysis for the conditional ignorability assumption (3) against unobserved confounding. Leveraging results in Opacic et al. (2023), we derive bias formulas for the unconditional and conditional prevalence components. We find that in order to reduce our point estimates to zero, there must be an unobserved confounder that is about three times as strong as educational expectation, which is likely an especially strong observed

---

[19]In Appendix G, as a robustness check, Table A5 presents estimates of the conditional decomposition based on an estimation procedure combining parametric and ML models for specific nuisance functions. As explained in the last paragraph of Section 3.2 and Appendix G, this theoretically may make the asymptotic inference more exact for the conditional prevalence, conditional effect, and Q-distribution components. In practice, however, the empirical findings are qualitatively unchanged.

confounder. This suggests that our estimates are quite robust to unobserved confounding.

Finally, we also estimate the change-in-disparity components in the URED and CRED of Jackson and VanderWeele (2018) and Jackson (2021). As explained above, these decompositions estimate the impacts of (marginally or conditionally) randomly equalizing the treatment and do not distinguish between group differences in prevalence of and selection into college graduation. Consequently, both the URED and the CRED underestimate the extent to which differential college graduation rates alone, independence of differential selection, contribute to intergenerational income persistence (Appendix Tables A3 and A4).

5   Discussion

The goal of causal decomposition analysis is to enumerate, quantify, and disambiguate the mechanisms by which a treatment variable contributes to an observed outcome disparity between groups. We introduced a new nonparametric decomposition approach that is more appropriate for the causal explanation of descriptive disparities and differentiates more mechanisms than prior approaches. In particular, we identify differential selection into treatment as a previously overlooked mechanism and novel policy lever. We developed ML-based estimators that are semiparametrically efficient, asymptotically normal, and multiply robust. Empirically, we demonstrate that our approach provides new insights by documenting that differential prevalence of, and selection into, college graduation play important but counteracting roles in the production of intergenerational income persistence.

In future work, we plan to extend our approach in multiple directions. First, we will develop analogous decompositions for non-binary treatment and group variables, multiple treatments, non-continuous outcomes, and time-to-event outcomes, under the conceptual umbrella of sequential interventions. Second, as an alternative to the conditional ignorability assumption, it will be valuable to develop an instrumental variable-based identification strategy for the decompositions, possibly via the marginal treatment effect framework (Heckman and Vytlacil, 2005). Third, in social science applications, period or cohort can also play the role of the group variable, so longitudinal changes can also be decomposed. Fourth, in terms of estimation, one could employ targeted learning (Van der Laan and Rose, 2011) for improved finite-sample performance.

Acknowledgments

Funding

Appendices

Appendix A: Proofs for Section 2

A.1. Equation (2)

Note that $R(D \mid g')$ denotes a randomly drawn value of treatment D from group g'.

$$\mathrm{E}_g\left(Y^{R(D \mid g')}\right)$$

$$= \mathrm{E}_g(Y^1 \mid R(D \mid g') = 1)\Pr_g(R(D \mid g') = 1) + \mathrm{E}_g(Y^0 \mid R(D \mid g') = 0)\Pr_g(R(D \mid g') = 0)$$

$$= \mathrm{E}_g(Y^1)\mathrm{E}_{g'}(D) + \mathrm{E}_g(Y^0)(1 - \mathrm{E}_{g'}(D))$$

$$= \mathrm{E}_g(Y^0) + \mathrm{E}_{g'}(D)\mathrm{E}_g(\tau).$$

A.2. Equivalence results in Section 2.3.2

For the CDE,

$$\mathrm{E}(Y^{a,0}) - \mathrm{E}(Y^{b,0})$$

$$= \mathrm{E}(Y^{a,0} \mid G = a) - \mathrm{E}(Y^{b,0} \mid G = b)$$

$$= \mathrm{E}_a(Y^0) - \mathrm{E}_b(Y^0),$$

where the first equality is by the unconditional ignorability of G, and the second equality is by the SUTVA.

For the PIE,

$$
\begin{aligned}
&\mathrm{E}\left[\left(Y^{b,1} - Y^{b,0}\right)\mathrm{E}(D^a)\right] \\
&= \mathrm{E}\left(Y^{b,1} - Y^{b,0}\right)\mathrm{E}(D^a) \\
&= \mathrm{E}\left(Y^{b,1} - Y^{b,0} \mid G = b\right)\mathrm{E}(D^a \mid G = a) \\
&= \mathrm{E}_b\left(Y^1 - Y^0\right)\mathrm{E}_a(D)
\end{aligned}
$$

where the first equality holds by the cross-world independence assumption. The second equality holds by the unconditional ignorability of G. And the third holds by the SUTVA. The equivalence result for the PAI can be similarly proved and is omitted.

A.3. Footnote 8

We first note that the outcome disparity can also be decomposed as such:

$$
\begin{aligned}
&\mathrm{E}_a(Y) - \mathrm{E}_b(Y) \\
&= \mathrm{E}_a(Y^0) - \mathrm{E}_b(Y^0) + \mathrm{E}(D)[\mathrm{E}_a(\tau) - \mathrm{E}_b(\tau)] + \mathrm{E}(\tau)[\mathrm{E}_a(D) - \mathrm{E}_b(D)] \\
&\quad + \mathrm{Cov}_a(D, \tau) - \mathrm{Cov}_b(D, \tau) - [p_a - p_b][\mathrm{E}_a(D) - \mathrm{E}_b(D)][\mathrm{E}_a(\tau) - \mathrm{E}_b(\tau)],
\end{aligned}
$$

where $\mathrm{E}(\tau)$ and $\mathrm{E}(D)$ are the marginal ATE and treatment prevalence, and p_g is the proportion of the population in group g. And the remaining disparity in Lundberg's (2024) unconditional decomposition can be easily rewritten as $\mathrm{E}_a(Y^0) - \mathrm{E}_b(Y^0) + \mathrm{E}(D)[\mathrm{E}_a(\tau) - \mathrm{E}_b(\tau)]$. It then follows that the change in disparity is what appears in Footnote 8.

A.4. Equation (5)

$$
\begin{aligned}
&\mathrm{E}_a(Y) - \mathrm{E}_b(Y) \\
&= \mathrm{E}_a(Y^0) - \mathrm{E}_b(Y^0) + \mathrm{E}_a(D\tau) - \mathrm{E}_b(D\tau) - \mathrm{E}_a[\mathrm{E}_a(D \mid \boldsymbol{Q})\,\mathrm{E}_a(\tau \mid \boldsymbol{Q})] + \mathrm{E}_b[\mathrm{E}_b(D \mid \boldsymbol{Q})\,\mathrm{E}_b(\tau \mid \boldsymbol{Q})] \\
&\quad + \mathrm{E}_a[\mathrm{E}_a(D \mid \boldsymbol{Q})\,\mathrm{E}_a(\tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{E}_b(D \mid \boldsymbol{Q})\,\mathrm{E}_b(\tau \mid \boldsymbol{Q})] \\
&= \mathrm{E}_a(Y^0) - \mathrm{E}_b(Y^0) + \mathrm{E}_a[\mathrm{Cov}_a(D, \tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D, \tau \mid \boldsymbol{Q})] \\
&\quad + \int \mathrm{E}_a(D \mid \boldsymbol{q})\,\mathrm{E}_a(\tau \mid \boldsymbol{q}) f_a(\boldsymbol{q}) \mathrm{d}\boldsymbol{q} - \int \mathrm{E}_b(D \mid \boldsymbol{q})\,\mathrm{E}_b(\tau \mid \boldsymbol{q}) f_b(\boldsymbol{q}) \mathrm{d}\boldsymbol{q} \\
&= \mathrm{E}_a(Y^0) - \mathrm{E}_b(Y^0)
\end{aligned}
$$

$$+ \int [\mathrm{E}_a(D \mid \boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{q})] \, \mathrm{E}_b(\tau \mid \boldsymbol{q}) f_b(\boldsymbol{q}) \mathrm{d}\boldsymbol{q} + \int \mathrm{E}_a(D \mid \boldsymbol{q}) \, \mathrm{E}_b(\tau \mid \boldsymbol{q}) [f_a(\boldsymbol{q}) - f_b(\boldsymbol{q})] \mathrm{d}\boldsymbol{q}$$

$$+ \int [\mathrm{E}_a(\tau \mid \boldsymbol{q}) - \mathrm{E}_b(\tau \mid \boldsymbol{q})] \, \mathrm{E}_a(D \mid \boldsymbol{q}) f_a(\boldsymbol{q}) \mathrm{d}\boldsymbol{q} + \mathrm{E}_a[\mathrm{Cov}_a(D, \tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D, \tau \mid \boldsymbol{Q})].$$

Note that the last equality uses Assumption 2 (common support).

A.5. Equation (6)

$$\mathrm{E}_g \left(Y^{R(D \mid g', \boldsymbol{Q})} \right)$$

$$= \int \mathrm{E}_g \left(Y^{R(D \mid g', \boldsymbol{q})} \mid \boldsymbol{q} \right) f_g(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}$$

$$= \int \mathrm{E}_g(Y^1 \mid \boldsymbol{q}, R(D \mid g', \boldsymbol{q}) = 1) \Pr_g(R(D \mid g', \boldsymbol{q}) = 1 \mid \boldsymbol{q}) f_g(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}$$

$$+ \int \mathrm{E}_g(Y^0 \mid \boldsymbol{q}, R(D \mid g', \boldsymbol{q}) = 0) \Pr_g(R(D \mid g', \boldsymbol{q}) = 0 \mid \boldsymbol{q}) f_g(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}$$

$$= \int [\mathrm{E}_g(Y^1 \mid \boldsymbol{q}) \, \mathrm{E}_{g'}(D \mid \boldsymbol{q}) + \mathrm{E}_g(Y^0 \mid \boldsymbol{q})(1 - \mathrm{E}_{g'}(D \mid \boldsymbol{q}))] f_g(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}$$

$$= \mathrm{E}_g(Y^0) + \int \mathrm{E}_g(\tau \mid \boldsymbol{q}) \, \mathrm{E}_{g'}(D \mid \boldsymbol{q}) f_g(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}.$$

All expectations are taken over $\boldsymbol{q} \in \mathrm{supp}_g(\boldsymbol{Q})$. For $\mathrm{E}_g \left(Y^{R(D \mid g', \boldsymbol{Q})} \right)$ to be well-defined, we require $\mathrm{supp}_g(\boldsymbol{Q}) \subseteq \mathrm{supp}_{g'}(\boldsymbol{Q})$.

A.6. Equation (7)

$$\mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \left[\mathrm{E}_a \left(Y^{R(D \mid a, \boldsymbol{Q})} \right) - \mathrm{E}_b \left(Y^{R(D \mid b, \boldsymbol{Q})} \right) \right]$$

$$= \mathrm{E}_a \left(Y^0 \right) - \mathrm{E}_b \left(Y^0 \right) + \mathrm{E}_a(D\tau) - \mathrm{E}_b(D\tau) - \left[\mathrm{E}_a(Y^{R(D \mid a, \boldsymbol{Q})}) - \mathrm{E}_b(Y^{R(D \mid b, \boldsymbol{Q})}) \right]$$

$$= \mathrm{E}_a[\mathrm{E}_a(D\tau \mid \boldsymbol{Q})] - \mathrm{E}_a[\mathrm{E}_a(\tau \mid \boldsymbol{Q}) \, \mathrm{E}_a(\tau \mid \boldsymbol{Q})] - \{\mathrm{E}_b[\mathrm{E}_b(D\tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{E}_b(\tau \mid \boldsymbol{Q}) \, \mathrm{E}_b(\tau \mid \boldsymbol{Q})]\}$$

$$= \mathrm{E}_a[\mathrm{Cov}_a(D, \tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D, \tau \mid \boldsymbol{Q})].$$

Other results in equation (7) follow directly from equation (6).

A.7. Equivalence result in Section 2.4.2

To establish the equivalence, we invoke the following assumptions:

1. Five ignorability assumptions

   (a) $Y^{b, \boldsymbol{q}, d} \perp\!\!\!\perp G, \forall \boldsymbol{q}, d$

   (b) $Y^{b, \boldsymbol{q}, d} \perp\!\!\!\perp \boldsymbol{Q} \mid b, \forall \boldsymbol{q}, d$

   (c) $D^{g, \boldsymbol{q}} \perp\!\!\!\perp G, \forall g, \boldsymbol{q}$

31

(d) $D^{g,\boldsymbol{q}} \perp\!\!\!\perp \boldsymbol{Q} \mid g, \forall g, \boldsymbol{q}$

(e) $\boldsymbol{Q}^b \perp\!\!\!\perp G$

2. Three SUTVA-type assumptions

    (a) $\mathrm{E}_b\left(Y^{b,\boldsymbol{q},d} \mid \boldsymbol{q}\right) = \mathrm{E}_b(Y^d \mid \boldsymbol{q}), \forall \boldsymbol{q}, d$

    (b) $\mathrm{E}_g(D^{g,\boldsymbol{q}} \mid \boldsymbol{q}) = \mathrm{E}_g(D \mid \boldsymbol{q}), \forall g, \boldsymbol{q}$

    (c) $f_b(\boldsymbol{q}^b) = f_b(\boldsymbol{q}), \forall \boldsymbol{q}$

3. Two cross-world independence-type assumptions

    (a) $Y^{b,\boldsymbol{q},1} \perp\!\!\!\perp \left\{\boldsymbol{Q}^b, D^{a,\boldsymbol{q}}\right\}, \forall \boldsymbol{q}$

    (b) $D^{g,\boldsymbol{q}} \perp\!\!\!\perp \boldsymbol{Q}^b, \forall g, \boldsymbol{q}$.

Note that $f_b(\boldsymbol{q}^b) := f_{\boldsymbol{Q}^b \mid b}(\boldsymbol{q})$ is the density function of $\boldsymbol{Q}^b$ in group b evaluated at $\boldsymbol{q}$, and $f_b(\boldsymbol{q}) := f_{\boldsymbol{Q} \mid b}(\boldsymbol{q})$ is the density function of $\boldsymbol{Q}$ in group b evaluated at $\boldsymbol{q}$.

Under these assumptions,

$$
\begin{aligned}
&\mathrm{E}\left(Y^{b,\boldsymbol{Q}^b,D^{a,\boldsymbol{Q}^b}}\right)\\
&= \int \mathrm{E}\left(Y^{b,\boldsymbol{q},D^{a,\boldsymbol{q}}} \mid \boldsymbol{Q}^b = \boldsymbol{q}\right) f_{\boldsymbol{Q}^b}(\boldsymbol{q})\mathrm{d}\boldsymbol{q}\\
&= \int \mathrm{E}\left(Y^{b,\boldsymbol{q},1} \mid \boldsymbol{Q}^b = \boldsymbol{q}, D^{a,\boldsymbol{q}} = 1\right) \mathrm{E}\left(D^{a,\boldsymbol{q}} \mid \boldsymbol{Q}^b = \boldsymbol{q}\right) f_{\boldsymbol{Q}^b}(\boldsymbol{q})\mathrm{d}\boldsymbol{q}\\
&\quad + \int \mathrm{E}\left(Y^{b,\boldsymbol{q},0} \mid \boldsymbol{Q}^b = \boldsymbol{q}, D^{a,\boldsymbol{q}} = 0\right) \left[1 - \mathrm{E}\left(D^{a,\boldsymbol{q}} \mid \boldsymbol{Q}^b = \boldsymbol{q}\right)\right] f_{\boldsymbol{Q}^b}(\boldsymbol{q})\mathrm{d}\boldsymbol{q}\\
&= \int \mathrm{E}\left(Y^{b,\boldsymbol{q},1}\right) \mathrm{E}\left(D^{a,\boldsymbol{q}}\right) f_{\boldsymbol{Q}^b}(\boldsymbol{q})\mathrm{d}\boldsymbol{q} + \int \mathrm{E}\left(Y^{b,\boldsymbol{q},0}\right) \left[1 - \mathrm{E}\left(D^{a,\boldsymbol{q}}\right)\right] f_{\boldsymbol{Q}^b}(\boldsymbol{q})\mathrm{d}\boldsymbol{q}\\
&= \int \mathrm{E}_b\left(Y^{b,\boldsymbol{q},1} \mid \boldsymbol{q}\right) \mathrm{E}_a\left(D^{a,\boldsymbol{q}} \mid \boldsymbol{q}\right) f_{\boldsymbol{Q}^b \mid b}(\boldsymbol{q})\mathrm{d}\boldsymbol{q} + \int \mathrm{E}_b\left(Y^{b,\boldsymbol{q},0} \mid \boldsymbol{q}\right) \left[1 - \mathrm{E}_a\left(D^{a,\boldsymbol{q}} \mid \boldsymbol{q}\right)\right] f_{\boldsymbol{Q}^b \mid b}(\boldsymbol{q})\mathrm{d}\boldsymbol{q}\\
&= \int \mathrm{E}_b\left(Y^1 \mid \boldsymbol{q}\right) \mathrm{E}_a\left(D \mid \boldsymbol{q}\right) f_{\boldsymbol{Q} \mid b}(\boldsymbol{q})\mathrm{d}\boldsymbol{q} + \int \mathrm{E}_b\left(Y^0 \mid \boldsymbol{q}\right) \left[1 - \mathrm{E}_a\left(D \mid \boldsymbol{q}\right)\right] f_{\boldsymbol{Q} \mid b}(\boldsymbol{q})\mathrm{d}\boldsymbol{q}.
\end{aligned}
$$

Above, the third equality is due to the cross-world independence-type assumptions. The fourth equality holds by the ignorability assumptions. The fifth (last) equality is due to the SUTVA-type assumptions.

Similarly,

$$
\begin{aligned}
&\mathrm{E}\left(Y^{b,\boldsymbol{Q}^b,D^{b,\boldsymbol{Q}^b}}\right)\\
&= \int \mathrm{E}_b\left(Y^1 \mid \boldsymbol{q}\right) \mathrm{E}_b\left(D \mid \boldsymbol{q}\right) f_{\boldsymbol{Q} \mid b}(\boldsymbol{q})\mathrm{d}\boldsymbol{q} + \int \mathrm{E}_b\left(Y^0 \mid \boldsymbol{q}\right) \left[1 - \mathrm{E}_b\left(D \mid \boldsymbol{q}\right)\right] f_{\boldsymbol{Q} \mid b}(\boldsymbol{q})\mathrm{d}\boldsymbol{q}.
\end{aligned}
$$

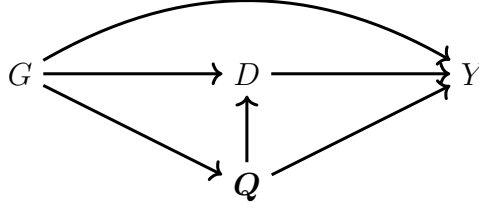The equivalence result then follows. Figure A1 illustrates the effect paths referred to in Section 2.4.2.

Figure A1. Illustration of effect paths in the path-specific decomposition.

A.8. Expression (8)

$$
\mathrm{E}_b\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b(Y)
$$
$$
= \mathrm{E}_b\left(Y^0\right) + \int \mathrm{E}_b(\tau \mid \boldsymbol{q})\,\mathrm{E}_a(D \mid \boldsymbol{q})f_b(\boldsymbol{q})\mathrm{d}\boldsymbol{q} - \mathrm{E}_b\left(Y^0\right) - \mathrm{E}_b(D\tau)
$$
$$
= \int \left[\mathrm{E}_a(D \mid \boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{q})\right]\mathrm{E}_b(\tau \mid \boldsymbol{q})f_b(\boldsymbol{q})\mathrm{d}\boldsymbol{q}
$$
$$
\quad + \mathrm{E}_b[\mathrm{E}_b(D \mid \boldsymbol{Q})\,\mathrm{E}_b(\tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{E}_b(D\tau \mid \boldsymbol{Q})]
$$
$$
= \int \left[\mathrm{E}_a(D \mid \boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{q})\right]\mathrm{E}_b(\tau \mid \boldsymbol{q})f_b(\boldsymbol{q})\mathrm{d}\boldsymbol{q} - \mathrm{E}_b[\mathrm{Cov}_b(D,\tau \mid \boldsymbol{q})].
$$

A.9. Footnote 12

In the randomized intervention notation, the change in disparity of the CRED of Lundberg (2024) is

$$
\mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \left[\mathrm{E}_a\left(Y^{R(D|\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|\boldsymbol{Q})}\right)\right],
$$

where $Y^{R(D|\boldsymbol{Q})}$ is the potential outcome of an individual when they were given a random draw of D from those in the pooled population who share the same $\boldsymbol{Q}$ values with them. Below, we start from the original form that appears in Lundberg (2024) (see his equation [2]), which is equivalent to the form expressed in the randomized intervention notation.

$$
\mathrm{E}_a(Y) - \mathrm{E}_b(Y) -
$$
$$
\left\{\mathrm{E}_a\left[\mathrm{Pr}(D = 0 \mid \boldsymbol{Q})Y^0 + \mathrm{Pr}(D = 1 \mid \boldsymbol{Q})Y^1\right] - \mathrm{E}_b\left[\mathrm{Pr}(D = 0 \mid \boldsymbol{Q})Y^0 + \mathrm{Pr}(D = 1 \mid \boldsymbol{Q})Y^1\right]\right\}
$$
$$
= \mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \left\{\mathrm{E}_a(Y^0) + \mathrm{E}_a[\mathrm{E}(D \mid \boldsymbol{Q})\tau] - \mathrm{E}_b(Y^0) - \mathrm{E}_b[\mathrm{E}(D \mid \boldsymbol{Q})\tau]\right\}
$$
$$
= \mathrm{E}_a(D\tau) - \mathrm{E}_a[\mathrm{E}(D \mid \boldsymbol{Q})\,\mathrm{E}_a(\tau \mid \boldsymbol{Q})] - \left\{\mathrm{E}_b(D\tau) - \mathrm{E}_b[\mathrm{E}(D \mid \boldsymbol{Q})\,\mathrm{E}_b(\tau \mid \boldsymbol{Q})]\right\}
$$
$$
= \int \left[\mathrm{E}_a(D\tau \mid \boldsymbol{q}) - \mathrm{E}(D \mid \boldsymbol{q})\,\mathrm{E}_a(\tau \mid \boldsymbol{q})\right]f_a(\boldsymbol{q})\mathrm{d}\boldsymbol{q} - \int \left[\mathrm{E}_b(D\tau \mid \boldsymbol{q}) - \mathrm{E}(D \mid \boldsymbol{q})\,\mathrm{E}_b(\tau \mid \boldsymbol{q})\right]f_b(\boldsymbol{q})\mathrm{d}\boldsymbol{q}
$$
$$
= \mathrm{E}_a[\mathrm{Cov}_a(D,\tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D,\tau \mid \boldsymbol{Q})]
$$

$$+ \int [\mathrm{E}_a(D \mid \boldsymbol{q}) - \mathrm{E}(D \mid \boldsymbol{q})] \, \mathrm{E}_a(\tau \mid \boldsymbol{q}) f_a(\boldsymbol{q}) \mathrm{d}\boldsymbol{q} - \int [\mathrm{E}_b(D \mid \boldsymbol{q}) - \mathrm{E}(D \mid \boldsymbol{q})] \, \mathrm{E}_b(\tau \mid \boldsymbol{q}) f_b(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}$$

$$= \mathrm{E}_a[\mathrm{Cov}_a(D, \tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D, \tau \mid \boldsymbol{Q})]$$

$$+ \int \mathrm{Pr}(G = b \mid \boldsymbol{q})[\mathrm{E}_a(D \mid \boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{q})] \, \mathrm{E}_a(\tau \mid \boldsymbol{q}) f_a(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}$$

$$- \int \mathrm{Pr}(G = a \mid \boldsymbol{q})[\mathrm{E}_b(D \mid \boldsymbol{q}) - \mathrm{E}_a(D \mid \boldsymbol{q})] \, \mathrm{E}_b(\tau \mid \boldsymbol{q}) f_b(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}$$

$$= \mathrm{E}_a[\mathrm{Cov}_a(D, \tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D, \tau \mid \boldsymbol{Q})]$$

$$+ \int [\mathrm{E}_a(D \mid \boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{q})][\mathrm{E}_a(\tau \mid \boldsymbol{q}) f_a(\boldsymbol{q}) \mathrm{Pr}(G = b \mid \boldsymbol{q}) + \mathrm{E}_b(\tau \mid \boldsymbol{q}) f_b(\boldsymbol{q}) \mathrm{Pr}(G = a \mid \boldsymbol{q})] \mathrm{d}\boldsymbol{q}.$$

Appendix B. Graphical illustration of the unconditional decomposition

Using the randomized intervention notation, we present a graphical representation of our unconditional decomposition in Figure A2. The graph visualizes the distinct ways the four components contribute to the observed group disparity in outcome, i.e., one can vary the four components on the graph and obtain different outcome disparities. From this graph, it is clear that the selection component represents the contribution of differential effectiveness of treatment assignment across groups, where effectiveness is defined relative to random assignment of the treatment. Imai and Li (2023) introduce a single-group version of Figure A2. But they do not note the covariance representation of the difference between an observed treatment assignment rule and the hypothetical random assignment with the same treatment prevalence.

Appendix C. Efficient influence functions

We use the Gateaux derivative approach to derive the EIFs (Ichimura and Newey, 2022), which is more succinct than the approach traditionally used in the semiparametric causal inference literature (e.g., Hahn, 1998). To further simplify the derivation, we leverage some rules of calculating Gateaux derivatives in Hines et al. (2022) and Kennedy (2022).

Let $\mathbb{1}_{\tilde{o}}(o)$ be the point mass density at a single empirical observation, $\tilde{o}$. Let subscript $\mathcal{P}_t$ indicate a regular parametric submodel indexed by t. The subscript is omitted for the true model. By construction, $f_{\mathcal{P}_t}(o) = t \mathbb{1}_{\tilde{o}}(o) + (1 - t) f(o)$, i.e., the submodel is the true model perturbed in the direction of a single observation $\tilde{o}$. Under this construction, the EIF of an estimand, ξ, is the Gateaux derivative at the truth, i.e., $\phi(\xi) = \frac{\partial \xi_{\mathcal{P}_t}}{\partial t}\big|_{t=0}$. For an arbitrary function $g(o)$, we denote $\frac{\partial g_{\mathcal{P}_t}(o)}{\partial t}\big|_{t=0}$ as $\partial g(o)$.

We derive the EIFs for the general case of weighted estimands. Let $w(\boldsymbol{X}, G)$ be the survey
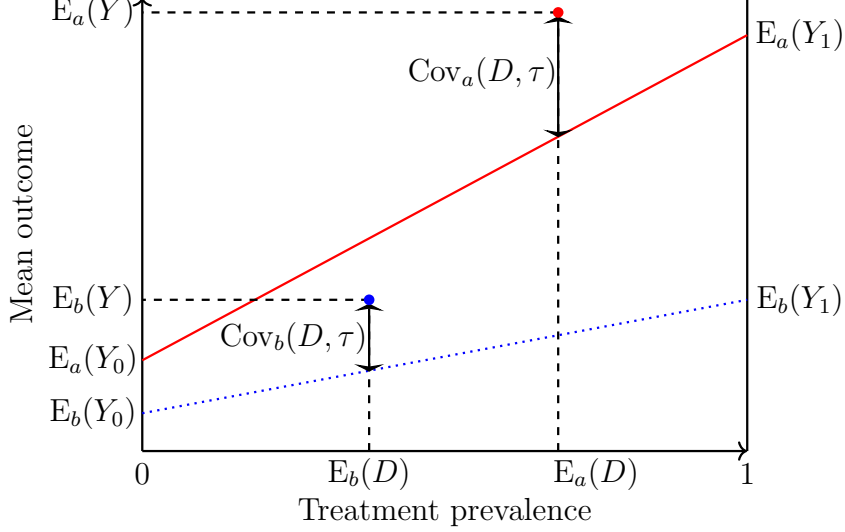
Figure A2. Illustration of the unconditional decomposition. The heights of the red dot and the blue dot respectively indicate the observed mean outcomes in group a and b. The position of each dot on the x axis corresponds to the observed treatment prevalence for the corresponding group. The red (solid) line and the blue (dotted) line respectively represent the mean potential outcomes in group a and group b under random assignment of the treatment, which vary by the prevalence of the assigned treatment. The lines are straight due to the hypothetical assignment being random. When nobody receives the treatment, the height of each line is $\mathrm{E}_g\left(Y^0\right)$; and when everyone receives the treatment, the height of the line is $\mathrm{E}_g\left(Y^1\right)$. The slope of each line is hence $\mathrm{E}_g(\tau)$. For a group g, the vertical distance between the dot and the line is $\mathrm{E}_g(Y) - \mathrm{E}_g\left(Y^{R(D|g)}\right) = \mathrm{Cov}_g(D, \tau)$.

weight. Following Hirano et al. (2003), we assume the survey weight is a known function of the covariates. When no survey weights are needed, $w(\boldsymbol{X}, G)$ reduces to 1 for every individual.

In this derivation, we also use the following definitions:

$$h_g := \mathrm{E}(w(\boldsymbol{X}, G) \mid g)$$

$$h_g(\boldsymbol{Q}) := \mathrm{E}(w(\boldsymbol{X}, G) \mid \boldsymbol{Q}, g).$$

C.1. EIFs for the unconditional decomposition

First, note that we only need to derive EIFs for two generic functions, $\xi_{dg} := \mathrm{E}\left(Y^d \frac{w(\boldsymbol{X}, g)}{h_g} \mid g\right)$ for an arbitrary group g; and $\xi_{dgg'} := \mathrm{E}\left(Y^d \frac{w(\boldsymbol{X}, g)}{h_g} \mid g\right) \mathrm{E}\left(D \frac{w(\boldsymbol{X}, g')}{h_{g'}} \mid g'\right)$ for two arbitrary groups g and g', which may be the same group, and an arbitrary treatment status d. The EIFs for the decomposition components then follow from adding and subtracting these functions evaluated at appropriate g, g', and d values. Under conditional ignorability (Assumption 3), these estimands can be identified as the following functionals:

$$\xi_{dg} = \mathrm{E}\left[\mu(d, \boldsymbol{X}, g) \frac{w(\boldsymbol{X}, g)}{h_g} \mid g\right]$$

35

$$\xi_{dgg'} = \mathrm{E}\left[\mu(d, \boldsymbol{X}, g)\frac{w(\boldsymbol{X}, g)}{h_g} \mid g\right] \mathrm{E}\left[D\frac{w(\boldsymbol{X}, g')}{h_{g'}} \mid g'\right].$$

We will also rely on the overlap assumption (Assumption 4) below, as $\pi(d, \boldsymbol{X}, g)$ will appear in the denominator.

We start with the EIF of ξ_{dg}.

$$
\begin{aligned}
\phi(\xi_{dg}) &= \partial \mathrm{E}_{\mathcal{P}_t}\left[\mu_{\mathcal{P}_t}(d, \boldsymbol{X}, g)\frac{w(\boldsymbol{X}, g)}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g) \mid g)} \mid g\right] \\
&= \frac{1}{h_g}\partial \mathrm{E}_{\mathcal{P}_t}\left[\mu_{\mathcal{P}_t}(d, \boldsymbol{X}, g)w(\boldsymbol{X}, g) \mid g\right] + \mathrm{E}[\mu(d, \boldsymbol{X}, g)w(\boldsymbol{X}, g) \mid g]\partial\frac{1}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g) \mid g)} \\
&= \frac{1}{h_g}\frac{\mathbb{1}_{\tilde{g}}(g)}{p_g}\{\mu(d, \tilde{\boldsymbol{x}}, g)w(\tilde{\boldsymbol{x}}, g) - \mathrm{E}[\mu(d, \boldsymbol{X}, g)w(\boldsymbol{X}, g) \mid g]\} \\
&\quad + \frac{1}{h_g}\mathrm{E}[\partial\mu_{\mathcal{P}_t}(d, \boldsymbol{X}, g)w(\boldsymbol{X}, g) \mid g] - \frac{1}{(h_g)^2}\frac{\mathbb{1}_{\tilde{g}}(g)}{p_g}[w(\tilde{\boldsymbol{x}}, g) - h_g]\mathrm{E}[\mu(d, \boldsymbol{X}, g)w(\boldsymbol{X}, g) \mid g] \\
&= \frac{\mathbb{1}_{\tilde{g}}(g)}{p_g}\frac{w(\tilde{\boldsymbol{x}}, g)}{h_g}[\mu(d, \tilde{\boldsymbol{x}}, g) - \xi_{dg}] + \frac{1}{h_g}\mathrm{E}[\partial\mu_{\mathcal{P}_t}(d, \boldsymbol{X}, g)w(\boldsymbol{X}, g) \mid g] \\
&= \frac{\mathbb{1}_{\tilde{g}}(g)}{p_g}\frac{w(\tilde{\boldsymbol{x}}, g)}{h_g}[\mu(d, \tilde{\boldsymbol{x}}, g) - \xi_{dg}] + \frac{1}{h_g}\mathrm{E}\left\{\frac{\mathbb{1}_{\tilde{d}, \tilde{\boldsymbol{x}}, \tilde{g}}(d, \boldsymbol{X}, g)}{f(d, \boldsymbol{X}, g)}[\tilde{y} - \mu(d, \boldsymbol{X}, g)]w(\boldsymbol{X}, g) \mid g\right\} \\
&= \frac{\mathbb{1}_{\tilde{g}}(g)}{p_g}\frac{w(\tilde{\boldsymbol{x}}, g)}{h_g}[\mu(d, \tilde{\boldsymbol{x}}, g) - \xi_{dg}] + \frac{\mathbb{1}_{\tilde{g}}(g)}{p_g}\frac{w(\tilde{\boldsymbol{x}}, g)}{h_g}\frac{\mathbb{1}_{\tilde{d}}(d)}{\pi(d, \tilde{\boldsymbol{x}}, g)}[\tilde{y} - \mu(d, \tilde{\boldsymbol{x}}, g)] \\
&= \frac{\mathbb{1}(G = g)}{p_g}\frac{w(\boldsymbol{X}, g)}{h_g}\left\{\frac{\mathbb{1}(D = d)}{\pi(d, \boldsymbol{X}, g)}[Y - \mu(d, \boldsymbol{X}, g)] + \mu(d, \boldsymbol{X}, g) - \xi_{dg}.\right\}
\end{aligned}
$$

And without survey weights, $\phi(\xi_{dg})$ simplifies to

$$\frac{\mathbb{1}(G = g)}{p_g}\left\{\frac{\mathbb{1}(D = d)}{\pi(d, \boldsymbol{X}, g)}[Y - \mu(d, \boldsymbol{X}, g)] + \mu(d, \boldsymbol{X}, g) - \xi_{dg}\right\}.$$

Now, for $\xi_{dgg'}$,

$$\phi(\xi_{dgg'}) = \phi(\xi_{dg})\mathrm{E}\left(D\frac{w(\boldsymbol{X}, g')}{h_{g'}} \mid g'\right) + \xi_{dg}\phi\left[\mathrm{E}\left(D\frac{w(\boldsymbol{X}, g')}{h_{g'}} \mid g'\right)\right].$$

Since

$$
\begin{aligned}
&\phi\left[\mathrm{E}\left(D\frac{w(\boldsymbol{X}, g')}{h_{g'}} \mid g'\right)\right] \\
&= \partial \mathrm{E}_{\mathcal{P}_t}\left(D\frac{w(\boldsymbol{X}, g')}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g') \mid g')} \mid g'\right) \\
&= \frac{1}{h_{g'}}\frac{\mathbb{1}_{\tilde{g}}(g')}{p_{g'}}\left[\tilde{d}w(\tilde{\boldsymbol{x}}, g') - \mathrm{E}(Dw(\boldsymbol{X}, g') \mid g')\right] - \frac{1}{(h_{g'})^2}\frac{\mathbb{1}_{\tilde{g}}(g')}{p_{g'}}[w(\tilde{\boldsymbol{x}}, g') - h_{g'}]\mathrm{E}(Dw(\boldsymbol{X}, g') \mid g') \\
&= \frac{\mathbb{1}_{\tilde{g}}(g')}{p_{g'}}\frac{w(\tilde{\boldsymbol{x}}, g')}{h_{g'}}\left[\tilde{d} - \mathrm{E}\left(D\frac{w(\boldsymbol{X}, g')}{h_{g'}} \mid g'\right)\right] \\
&= \frac{\mathbb{1}(G = g')}{p_{g'}}\frac{w(\boldsymbol{X}, g')}{h_{g'}}\left[D - \mathrm{E}\left(D\frac{w(\boldsymbol{X}, g')}{h_{g'}} \mid g'\right)\right],
\end{aligned}
$$

we obtain the EIF for $\xi_{dgg'}$,

$$\phi(\xi_{dgg'}) = \frac{\mathbb{1}(G = g)}{p_g} \frac{w(\boldsymbol{X}, g)}{h_g} \left[\frac{\mathbb{1}(D = d)}{\pi(d, \boldsymbol{X}, g)} (Y - \mu(d, \boldsymbol{X}, g)) + \mu(d, \boldsymbol{X}, g) \right] \mathrm{E} \left(D \frac{w(\boldsymbol{X}, g')}{h_{g'}} \mid g' \right)$$
$$+ \frac{\mathbb{1}(G = g')}{p_{g'}} \frac{w(\boldsymbol{X}, g')}{h_{g'}} \mathrm{E} \left[Y^d \frac{w(\boldsymbol{X}, g)}{h_g} \mid g \right] \left[D - \mathrm{E} \left(D \frac{w(\boldsymbol{X}, g')}{h_{g'}} \mid g' \right) \right]$$
$$- \frac{\mathbb{1}(G = g)}{p_g} \frac{w(\boldsymbol{X}, g)}{h_g} \xi_{dgg'}.$$

Without survey weights, $\phi(\xi_{dgg'})$ simplifies to

$$\frac{\mathbb{1}(G = g)}{p_g} \left[\frac{\mathbb{1}(D = d)}{\pi(d, \boldsymbol{X}, g)} (Y - \mu(d, \boldsymbol{X}, g)) + \mu(d, \boldsymbol{X}, g) \right] \mathrm{E} \left(D \mid g' \right)$$
$$+ \frac{\mathbb{1}(G = g')}{p_{g'}} \mathrm{E} \left[Y^d \mid g \right] \left[D - \mathrm{E} \left(D \mid g' \right) \right] - \frac{\mathbb{1}(G = g)}{p_g} \xi_{dgg'}.$$

In the main text of the paper, we write $\phi(\xi_{dg})$ and $\phi(\xi_{dgg'})$ as $\phi_{dg}(Y, D, \boldsymbol{X}, G)$ and $\phi_{dgg'}(Y, D, \boldsymbol{X}, G)$, respectively, to highlight that they are functions of observed variables.

Also note that the EIF for the total disparity, $\mathrm{E} \left(Y \frac{w(\boldsymbol{X}, a)}{h_a} \mid a \right) - \mathrm{E} \left(Y \frac{w(\boldsymbol{X}, b)}{h_b} \mid b \right)$, is

$$\phi(\text{Total})$$
$$= \frac{\mathbb{1}(G = a)}{p_a} \frac{w(\boldsymbol{X}, a)}{h_a} \left[Y - \mathrm{E} \left(Y \frac{w(\boldsymbol{X}, a)}{h_a} \mid a \right) \right] - \frac{\mathbb{1}(G = b)}{p_b} \frac{w(\boldsymbol{X}, b)}{h_b} \left[Y - \mathrm{E} \left(Y \frac{w(\boldsymbol{X}, b)}{h_b} \mid b \right) \right],$$

which, without survey weights, becomes

$$\frac{\mathbb{1}(G = a)}{p_a} \left[Y - \mathrm{E} \left(Y \mid a \right) \right] - \frac{\mathbb{1}(G = b)}{p_b} \left[Y - \mathrm{E} \left(Y \mid b \right) \right].$$

Finally, the EIFs for the unconditional decomposition components are

$$\phi(\text{Baseline}) = \phi(\xi_{0a}) - \phi(\xi_{0b})$$
$$\phi(\text{Prevalence}) = \phi(\xi_{1ba}) - \phi(\xi_{1bb}) - \phi(\xi_{0ba}) + \phi(\xi_{0bb})$$
$$\phi(\text{Effect}) = \phi(\xi_{1aa}) - \phi(\xi_{0aa}) - \phi(\xi_{1ba}) + \phi(\xi_{0ba})$$
$$\phi(\text{Selection}) = \phi(\text{Total}) - \phi(\text{Baseline}) - \phi(\text{Prevalence}) - \phi(\text{Effect}).$$

C.2. EIFs for the conditional decomposition

Similar to the unconditional case, we focus on the generic function

$$\xi_{dgg'g''} := \mathrm{E} \left[\mathrm{E} \left(Y^d \frac{w(\boldsymbol{X}, g)}{h_g(\boldsymbol{Q})} \mid \boldsymbol{Q}, g \right) \mathrm{E} \left(D \frac{w(\boldsymbol{X}, g')}{h_{g'}(\boldsymbol{Q})} \mid \boldsymbol{Q}, g' \right) \frac{w(\boldsymbol{X}, g'')}{h_{g''}} \mid g'' \right],$$

where (g, g', g'') is an arbitrary combination of group memberships out of the 8 possible combinations. We maintain the conditional ignorability and overlap assumptions.

$$\phi(\xi_{dgg'g''})$$

$$= \partial \mathrm{E}_{\mathcal{P}_t} \left[\mathrm{E}_{\mathcal{P}_t} \left(Y^d \frac{w(\boldsymbol{X}, g)}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g) \mid \boldsymbol{Q}, g)} \mid \boldsymbol{Q}, g \right) \mathrm{E}_{\mathcal{P}_t} \left(D \frac{w(\boldsymbol{X}, g')}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g') \mid \boldsymbol{Q}, g')} \mid \boldsymbol{Q}, g' \right) \right.$$

$$\left. w(\boldsymbol{X}, g'') \mid g'' \right] \frac{1}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g) \mid g'')}$$

$$= \partial \mathrm{E}_{\mathcal{P}_t} \left[\mathrm{E}_{\mathcal{P}_t} \left(Y^d \frac{w(\boldsymbol{X}, g)}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g) \mid \boldsymbol{Q}, g)} \mid \boldsymbol{Q}, g \right) \mathrm{E}_{\mathcal{P}_t} \left(D \frac{w(\boldsymbol{X}, g')}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g') \mid \boldsymbol{Q}, g')} \mid \boldsymbol{Q}, g' \right) w(\boldsymbol{X}, g'') \mid g'' \right] \frac{1}{h_{g''}}$$

$$+ \mathrm{E} \left[\mathrm{E} \left(Y^d \frac{w(\boldsymbol{X}, g)}{\mathrm{E}(w(\boldsymbol{X}, g) \mid \boldsymbol{Q}, g)} \mid \boldsymbol{Q}, g \right) \mathrm{E} \left(D \frac{w(\boldsymbol{X}, g')}{\mathrm{E}(w(\boldsymbol{X}, g') \mid \boldsymbol{Q}, g')} \mid \boldsymbol{Q}, g' \right) w(\boldsymbol{X}, g'') \mid g'' \right]$$

$$\partial \frac{1}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g) \mid g'')}$$

$$= \frac{\mathbb{1}_{\tilde{g}(g'')}}{p_{g''}} \frac{w(\boldsymbol{X}, g'')}{h_{g''}} \mathrm{E} \left(Y^d \frac{w(\boldsymbol{X}, g)}{h_g(\boldsymbol{Q})} \mid \boldsymbol{Q}, g \right) \mathrm{E} \left(D \frac{w(\boldsymbol{X}, g')}{h_{g'}(\boldsymbol{Q})} \mid \boldsymbol{Q}, g' \right) - \frac{\mathbb{1}_{\tilde{g}(g'')}}{p_{g''}} \xi_{dgg'g''}$$

$$+ \mathrm{E} \left[\partial \mathrm{E}_{\mathcal{P}_t} \left(Y^d \frac{w(\boldsymbol{X}, g)}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g) \mid \boldsymbol{Q}, g)} \mid \boldsymbol{Q}, g \right) \mathrm{E}_{\mathcal{P}_t} \left(D \frac{w(\boldsymbol{X}, g')}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g') \mid \boldsymbol{Q}, g')} \mid \boldsymbol{Q}, g' \right) w(\boldsymbol{X}, g'') \mid g'' \right] \frac{1}{h_{g''}}$$

$$- \xi_{dgg'g''} h_{g''} \frac{1}{(h_{g''})^2} \frac{\mathbb{1}_{\tilde{g}(g'')}}{p_{g''}} \left[w(\boldsymbol{X}, g'') - h_{g''} \right]$$

$$= \frac{\mathbb{1}(G = g'')}{p_{g''}} \frac{w(\boldsymbol{X}, g'')}{h_{g''}} \left[\mathrm{E} \left(Y^d \frac{w(\boldsymbol{X}, g)}{h_g(\boldsymbol{Q})} \mid \boldsymbol{Q}, g \right) \mathrm{E} \left(D \frac{w(\boldsymbol{X}, g')}{h_{g'}(\boldsymbol{Q})} \mid \boldsymbol{Q}, g' \right) - \xi_{dgg'g''} \right]$$

$$+ \mathrm{E} \left[\partial \mathrm{E}_{\mathcal{P}_t} \left(Y^d \frac{w(\boldsymbol{X}, g)}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g) \mid \boldsymbol{Q}, g)} \mid \boldsymbol{Q}, g \right) \mathrm{E}_{\mathcal{P}_t} \left(D \frac{w(\boldsymbol{X}, g')}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g') \mid \boldsymbol{Q}, g')} \mid \boldsymbol{Q}, g' \right) w(\boldsymbol{X}, g'') \mid g'' \right] \frac{1}{h_{g''}}.$$

And

$$\mathrm{E} \left[\partial \mathrm{E}_{\mathcal{P}_t} \left(Y^d \frac{w(\boldsymbol{X}, g)}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g) \mid \boldsymbol{Q}, g)} \mid \boldsymbol{Q}, g \right) \mathrm{E}_{\mathcal{P}_t} \left(D \frac{w(\boldsymbol{X}, g')}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g') \mid \boldsymbol{Q}, g')} \mid \boldsymbol{Q}, g' \right) w(\boldsymbol{X}, g'') \mid g'' \right] \frac{1}{h_{g''}}$$

$$= \mathrm{E} \left[\partial \mathrm{E}_{\mathcal{P}_t} \left(Y^d \frac{w(\boldsymbol{X}, g)}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g) \mid \boldsymbol{Q}, g)} \mid \boldsymbol{Q}, g \right) \mathrm{E} \left(D \frac{w(\boldsymbol{X}, g')}{h_{g'}(\boldsymbol{Q})} \mid \boldsymbol{Q}, g' \right) w(\boldsymbol{X}, g'') \mid g'' \right] \frac{1}{h_{g''}}$$

$$+ \mathrm{E} \left[\mathrm{E} \left(Y^d \frac{w(\boldsymbol{X}, g)}{h_g(\boldsymbol{Q})} \mid \boldsymbol{Q}, g \right) \partial \mathrm{E}_{\mathcal{P}_t} \left(D \frac{w(\boldsymbol{X}, g')}{\mathrm{E}_{\mathcal{P}_t}(\boldsymbol{X}, g')} \mid \boldsymbol{Q}, g' \right) w(\boldsymbol{X}, g'') \mid g'' \right] \frac{1}{h_{g''}}$$

$$= \frac{\mathbb{1}(G = g) p_{g''}(\boldsymbol{Q})}{p_g(\boldsymbol{Q}) p_{g''}} \frac{w(\boldsymbol{X}, g)}{h_g(\boldsymbol{Q})} \frac{w(\boldsymbol{X}, g'')}{h_{g''}} \left\{ \frac{\mathbb{1}(D = d)}{\pi(d, \boldsymbol{X}, g)} [Y - \mu(d, \boldsymbol{X}, g)] + \mu(d, \boldsymbol{X}, g) \right.$$

$$\left. - \mathrm{E} \left(Y^d \frac{w(X, Q, g)}{h_g(\boldsymbol{Q})} \mid \boldsymbol{Q}, g \right) \right\} \mathrm{E} \left(D \frac{w(\boldsymbol{X}, Q, g')}{h_{g'}(\boldsymbol{Q})} \mid \boldsymbol{Q}, g' \right)$$

$$+ \frac{\mathbb{1}(G = g') p_{g''}(\boldsymbol{Q})}{p_{g'}(\boldsymbol{Q}) p_{g''}} \frac{w(\boldsymbol{X}, g')}{h_{g'}(\boldsymbol{Q})} \frac{w(\boldsymbol{X}, g'')}{h_{g''}} \left[D - \mathrm{E} \left(D \frac{w(\boldsymbol{X}, g')}{h_{g'}(\boldsymbol{Q})} \mid \boldsymbol{Q}, g' \right) \right] \mathrm{E} \left(Y^d \frac{w(\boldsymbol{X}, g)}{h_g(\boldsymbol{Q})} \mid \boldsymbol{Q}, g \right).$$

Hence,

$$\phi(\xi_{dgg'g''})$$

$$
\begin{aligned}
=\; & \frac{\mathbb{1}(G=g'')}{p_{g''}}\frac{w(\boldsymbol{X},g'')}{h_{g''}}\left[\mathrm{E}\left(Y^d\frac{w(\boldsymbol{X},g)}{h_g(\boldsymbol{Q})}\mid \boldsymbol{Q},g\right)\mathrm{E}\left(D\frac{w(\boldsymbol{X},g')}{h_{g'}(\boldsymbol{Q})}\mid \boldsymbol{Q},g'\right)-\xi_{dgg'g''}\right] \\
& +\frac{w(\boldsymbol{X},g'')}{h_{g''}}\frac{\mathbb{1}(G=g)p_{g''}(\boldsymbol{Q})}{p_g(\boldsymbol{Q})p_{g''}}\frac{w(\boldsymbol{X},g)}{h_g(\boldsymbol{Q})}\left\{\frac{\mathbb{1}(D=d)}{\pi(d,\boldsymbol{X},g)}[Y-\mu(d,\boldsymbol{X},g)]+\mu(d,\boldsymbol{X},g)\right. \\
& \left.-\mathrm{E}\left(Y^d\frac{w(\boldsymbol{X},g)}{h_g(\boldsymbol{Q})}\mid \boldsymbol{Q},g\right)\right\}\mathrm{E}\left(D\frac{w(\boldsymbol{X},g')}{h_{g'}(\boldsymbol{Q})}\mid \boldsymbol{Q},g'\right) \\
& +\frac{w(\boldsymbol{X},g'')}{h_{g''}}\frac{\mathbb{1}(G=g')p_{g''}(\boldsymbol{Q})}{p_{g'}(\boldsymbol{Q})p_{g''}}\frac{w(\boldsymbol{X},g')}{h_{g'}(\boldsymbol{Q})}\left[D-\mathrm{E}\left(D\frac{w(\boldsymbol{X},g')}{h_{g'}(\boldsymbol{Q})}\mid \boldsymbol{Q},g'\right)\right]\mathrm{E}\left(Y^d\frac{w(\boldsymbol{X},g)}{h_g(\boldsymbol{Q})}\mid \boldsymbol{Q},g\right),
\end{aligned}
$$

which, in the absence of survey weights, simplifies to

$$
\begin{aligned}
& \frac{\mathbb{1}(G=g'')}{p_{g''}}\left[\mathrm{E}\left(Y^d\mid \boldsymbol{Q},g\right)\mathrm{E}\left(D\mid \boldsymbol{Q},g'\right)-\xi_{dgg'g''}\right] \\
& +\frac{\mathbb{1}(G=g)p_{g''}(\boldsymbol{Q})}{p_g(\boldsymbol{Q})p_{g''}}\left\{\frac{\mathbb{1}(D=d)}{\pi(d,\boldsymbol{X},g)}[Y-\mu(d,\boldsymbol{X},g)]+\mu(d,\boldsymbol{X},g)-\mathrm{E}\left(Y^d\mid \boldsymbol{Q},g\right)\right\}\mathrm{E}\left(D\mid \boldsymbol{Q},g'\right) \\
& +\frac{\mathbb{1}(G=g')p_{g''}(\boldsymbol{Q})}{p_{g'}(\boldsymbol{Q})p_{g''}}\left[D-\mathrm{E}\left(D\mid \boldsymbol{Q},g'\right)\right]\mathrm{E}\left(Y^d\mid \boldsymbol{Q},g\right).
\end{aligned}
$$

Recall that $\mathrm{E}\left(Y^d\mid \boldsymbol{Q},g\right)$ is identified as $\omega(d,\boldsymbol{Q},g):=\mathrm{E}[\mu(d,\boldsymbol{X},g)\mid \boldsymbol{Q},g]$. Finally, in the main text of the paper, we write $\phi(\xi_{dgg'g''})$ as $\phi_{dgg'g''}(Y,D,\boldsymbol{X},G)$ to highlight that it is a function of observed variables.

Appendix D. Asymptotic distributions

We follow the procedure of using the von Mises expansion to prove asymptotic properties of cross-fitting EIF-based one-step estimators (Hines et al., 2022; Kennedy, 2022; Fisher and Kennedy, 2021). In order for the cross-fitting one-step estimator to be $\sqrt{n}-$consistent, asymptotically normal, and semiparametrically efficient, we just need two conditions to hold. That is, both the empirical process term and the "remainder term" in the von Mises expansion are $o_p(n^{-1/2})$.

We use the notation $\mathbb{P}(f(O)):=\int f(O)d\mathbb{P}(O)$, and $\mathbb{P}_n$ denotes the corresponding sample average. Also let $\|\cdot\|$ denote the L_2-norm, such that $\|f(O)\|^2=\mathbb{P}(f(O)^2)$. And $\hat{\xi}$ is defined to be a substitution estimator for ξ. Formally, for all d,g and ξ, we need $(\mathbb{P}-\mathbb{P}_n)\left[\hat{\phi}(Y,d,\boldsymbol{X},g)-\phi(Y,d,\boldsymbol{X},g)\right]=o_p(n^{-1/2})$, and $\hat{\xi}+\mathbb{P}[\hat{\phi}(Y,d,\boldsymbol{X},g)]-\xi=o_p(n^{-1/2})$. In this appendix, we prove that the assumptions specified in the main text are sufficient for the unconditional and conditional decompositions to attain this convergence result. By cross-fitting, all of $\hat{\mu}(d,\boldsymbol{X},g),\hat{\pi}(d,\boldsymbol{X},g),\hat{p}_g(\boldsymbol{Q}),\hat{\omega}(d,\boldsymbol{Q},g)$, and $\hat{\mathrm{E}}(D\mid \boldsymbol{Q},g)$ are fitted using data not in the current subsample, which we implicitly condition on throughout.

D.1. Inference for the unconditional decomposition

First, for ξ_{dg}, the remainder term is

$$
\begin{aligned}
R_{2,dg} &= \hat{\xi}_{dg} + \mathbb{P}\left\{\frac{\mathbb{1}(G=g)}{\hat{p}_g}\left[\frac{\mathbb{1}(D=d)}{\hat{\pi}(d,\boldsymbol{X},g)}(Y-\hat{\mu}(d,\boldsymbol{X},g)) + \hat{\mu}(d,\boldsymbol{X},g) - \hat{\xi}_{dg}\right]\right\} - \xi_{dg} \\
&= \hat{\xi}_{dg} + \mathbb{P}\left\{\frac{\mathbb{1}(G=g)}{\hat{p}_g}\left[\frac{\mathbb{1}(D=d)}{\hat{\pi}(d,\boldsymbol{X},g)}(Y-\hat{\mu}(d,\boldsymbol{X},g)) + \hat{\mu}(d,\boldsymbol{X},g) - \mu(d,\boldsymbol{X},g)\right]\right\} \\
&\quad + \mathbb{P}\left[\frac{\mathbb{1}(G=g)}{\hat{p}_g}\mu(d,\boldsymbol{X},G)\right] - \mathbb{P}\left[\frac{\mathbb{1}(G=g)}{\hat{p}_g}\hat{\xi}_{dg}\right] - \xi_{dg} \\
&= \hat{\xi}_{dg} + \mathbb{P}\left\{\frac{\mathbb{1}(G=g)}{\hat{p}_g}\left[1 - \frac{\pi(d,\boldsymbol{X},g)}{\hat{\pi}(d,\boldsymbol{X},g)}\right](\hat{\mu}(d,\boldsymbol{X},g) - \mu(d,\boldsymbol{X},g))\right\} + \frac{p_g}{\hat{p}_g}\xi_{dg} - \frac{p_g}{\hat{p}_g}\hat{\xi}_{dg} - \xi_{dg} \\
&= \left(1 - \frac{p_g}{\hat{p}_g}\right)\left(\hat{\xi}_{dg} - \xi_{dg}\right) + \mathbb{P}\left\{\frac{\mathbb{1}(G=g)}{\hat{p}_g}\left[1 - \frac{\pi(d,\boldsymbol{X},g)}{\hat{\pi}(d,\boldsymbol{X},g)}\right][\hat{\mu}(d,\boldsymbol{X},g) - \mu(d,\boldsymbol{X},g)]\right\}.
\end{aligned}
$$

The first term is a lower order term (see the proof of Theorem 1 in Vansteelandt and Dukes, 2022). For the second term,

$$
\begin{aligned}
&\left|\mathbb{P}\left\{\frac{\mathbb{1}(G=g)}{\hat{p}_g}\left[1 - \frac{\pi(d,\boldsymbol{X},g)}{\hat{\pi}(d,\boldsymbol{X},g)}\right][\hat{\mu}(d,\boldsymbol{X},g) - \mu(d,\boldsymbol{X},g)]\right\}\right| \\
&\leq \frac{1}{\eta\hat{p}_g}|\mathbb{P}\left\{[\hat{\pi}(d,\boldsymbol{X},g) - \pi(d,\boldsymbol{X},g)][\hat{\mu}(d,\boldsymbol{X},g) - \mu(d,\boldsymbol{X},g)]\right\}| \\
&\leq \frac{1}{\eta\hat{p}_g}\|\hat{\pi}(d,\boldsymbol{X},g) - \pi(d,\boldsymbol{X},g)\|\|\hat{\mu}(d,\boldsymbol{X},g) - \mu(d,\boldsymbol{X},g)\| \\
&= o_p(n^{-1/2}),
\end{aligned}
$$

where the second inequality uses the Cauchy–Schwarz inequality.

For the empirical process term,

$$
\begin{aligned}
&\hat{\phi}_{dg}(Y,\boldsymbol{X}) - \phi_{dg}(Y,\boldsymbol{X}) \\
&= \frac{\mathbb{1}(G=g)}{\hat{p}_g}\left[1 - \frac{\mathbb{1}(D=d)}{\pi(d,\boldsymbol{X},g)}\right][\hat{\mu}(d,\boldsymbol{X},g) - \mu(d,\boldsymbol{X},g)] \\
&\quad + \frac{\mathbb{1}(G=g)}{\hat{p}_g}\frac{\mathbb{1}(D=d)[Y-\hat{\mu}(d,\boldsymbol{X},g)]}{\hat{\pi}(d,\boldsymbol{X},g)\pi(d,\boldsymbol{X},g)}[\pi(d,\boldsymbol{X},g) - \hat{\pi}(d,\boldsymbol{X},g)] \\
&\quad + \mathbb{1}(G=g)\frac{p_g - \hat{p}_g}{\hat{p}_g p_g}\left\{\frac{\mathbb{1}(D=d)}{\pi(d,\boldsymbol{X},g)}[Y - \mu(d,\boldsymbol{X},g)] + \mu(d,\boldsymbol{X},g)\right\} \\
&\quad + \frac{\mathbb{1}(G=g)}{\hat{p}_g}\left(\xi_{dg} - \hat{\xi}_{dg}\right) + \mathbb{1}(G=g)\frac{\hat{p}_g - p_g}{\hat{p}_g p_g}\xi_{dg}.
\end{aligned}
$$

Note that $(\mathbb{P}_n - \mathbb{P})\left[\frac{\mathbb{1}(G=g)}{\hat{p}_g}\left(\xi_{dg} - \hat{\xi}_{dg}\right)\right]$ is a lower order term. Then, using the Chebyshev's inequality argument commonly used in the double ML literature (Chernozhukov et al., 2017), the empirical process term is $o_p(n^{-1/2})$ under stated conditions.

Second, for $\xi_{dgg'}$,

$$
\begin{aligned}
R_{2,dgg'} &= \hat{\xi}_{dgg'} + \mathbb{P}\left\{\frac{\mathbb{1}(G=g)}{\hat{p}_g}\left[\frac{\mathbb{1}(D=d)}{\hat{\pi}(d,\boldsymbol{X},g)}(Y-\hat{\mu}(d,\boldsymbol{X},g))+\hat{\mu}(d,\boldsymbol{X},g)\right]\hat{\mathrm{E}}\left(D\mid g'\right)\right. \\
&\quad \left. + \frac{\mathbb{1}(G=g')}{\hat{p}_{g'}}\hat{\mathrm{E}}\left(Y^d\mid g\right)\left[D-\hat{\mathrm{E}}\left(D\mid g'\right)\right]-\frac{\mathbb{1}(G=g)}{\hat{p}_g}\hat{\xi}_{dgg'}\right\}-\xi_{dgg'} \\
&= \mathbb{P}\left\{\frac{\mathbb{1}(G=g)}{\hat{p}_g}\left[1-\frac{\pi(d,\boldsymbol{X},g)}{\hat{\pi}(d,\boldsymbol{X},g)}\right][\hat{\mu}(d,\boldsymbol{X},g)-\mu(d,\boldsymbol{X},g)]\right\}\hat{\mathrm{E}}(D\mid g') \\
&\quad + \left[\frac{p_g}{\hat{p}_g}\mathrm{E}(Y^d\mid g)-\frac{p_{g'}}{\hat{p}_{g'}}\hat{\mathrm{E}}(Y^d\mid g)\right]\left[\hat{\mathrm{E}}(D\mid g')-\mathrm{E}(D\mid g')\right]+\left(1-\frac{p_g}{\hat{p}_g}\right)\left(\hat{\xi}_{dgg'}-\xi_{dgg'}\right),
\end{aligned}
$$

where, under stated conditions, the first term is $o_p(n^{-1/2})$, the second term is $o_p(1)O_p(n^{-1/2})=o_p(n^{-1/2})$, and the last term is again a lower order term.

Also,

$$
\begin{aligned}
&\hat{\phi}_{dgg'}(Y,\boldsymbol{X})-\phi_{dgg'}(Y,\boldsymbol{X}) \\
&= \frac{\mathbb{1}(G=g)}{\hat{p}_g}\hat{\mathrm{E}}(D\mid g')\left[1-\frac{\mathbb{1}(D=d)}{\pi(d,\boldsymbol{X},g)}\right][\hat{\mu}(d,\boldsymbol{X},g)-\mu(d,\boldsymbol{X},g)] \\
&\quad + \frac{\mathbb{1}(G=g)}{\hat{p}_g}\hat{\mathrm{E}}(D\mid g')\frac{\mathbb{1}(D=d)[Y-\hat{\mu}(d,\boldsymbol{X},g)]}{\hat{\pi}(d,\boldsymbol{X},g)\pi(d,\boldsymbol{X},g)}[\pi(d,\boldsymbol{X},g)-\hat{\pi}(d,\boldsymbol{X},g)] \\
&\quad + \frac{\mathbb{1}(G=g)}{\hat{p}_g}\left[\hat{\mathrm{E}}(D\mid g')-\mathrm{E}(D\mid g')\right]\left\{\frac{\mathbb{1}(D=d)}{\pi(d,\boldsymbol{X},g)}[Y-\mu(d,\boldsymbol{X},g)]+\mu(d,\boldsymbol{X},g)\right\} \\
&\quad + \mathbb{1}(G=g)\frac{p_g-\hat{p}_g}{\hat{p}_gp_g}\mathrm{E}(D\mid g')\left\{\frac{\mathbb{1}(D=d)}{\pi(d,\boldsymbol{X},g)}[Y-\mu(d,\boldsymbol{X},g)]+\mu(d,\boldsymbol{X},g)\right\} \\
&\quad + \frac{\mathbb{1}(G=g')}{\hat{p}_{g'}}\left[\hat{\mathrm{E}}(Y^d\mid g)-\mathrm{E}(Y^d\mid g)\right]D+\mathbb{1}(G=g')\frac{p_g-\hat{p}_g}{\hat{p}_gp_g}\mathrm{E}(Y^d\mid g)D \\
&\quad + \frac{\mathbb{1}(G=g')}{p_{g'}}\mathrm{E}(Y^d\mid g)\left[\hat{\mathrm{E}}(D\mid g')-\mathrm{E}(D\mid g')\right]+\frac{\mathbb{1}(G=g')}{\hat{p}_{g'}}\left[\hat{\mathrm{E}}(Y^d\mid g)-\mathrm{E}(Y^d\mid g)\right]\hat{\mathrm{E}}(D\mid g') \\
&\quad + \mathbb{1}(G=g')\frac{p_{g'}-\hat{p}_{g'}}{\hat{p}_{g'}p_{g'}}\mathrm{E}(Y^d\mid g)\hat{\mathrm{E}}(D\mid g')+\mathbb{1}(G=g)\frac{\hat{p}_g-p_g}{\hat{p}_gp_g}\xi_{dgg'}+\frac{\mathbb{1}(G=g)}{\hat{p}_g}(\xi_{dgg'}-\hat{\xi}_{dgg'}).
\end{aligned}
$$

Thus, the empirical process term, $(\mathbb{P}_n-\mathbb{P})\left[\hat{\phi}_{dgg'}(Y,\boldsymbol{X})-\phi_{dgg'}(Y,\boldsymbol{X})\right]$, is $o_p(n^{-1/2})$ under stated conditions. Note that

$$
\hat{\mathrm{E}}(Y^d\mid g)=\mathbb{P}_n\left\{\frac{\mathbb{1}(D=d)}{\hat{\pi}(d,\boldsymbol{X},g)}[Y-\hat{\mu}(d,\boldsymbol{X},g)]+\hat{\mu}(d,\boldsymbol{X},g)\right\},
$$

hence consistent estimation of $\pi(d,\boldsymbol{X},g)$ and $\mu(d,\boldsymbol{X},g)$ ensures that $\hat{\mathrm{E}}(Y^d\mid g)-\mathrm{E}(Y^d\mid g)=o_p(1)$.

D.2. Inference for the conditional decomposition

For the components of our conditional decomposition, either $g = g''$ or $g' = g''$. In what follows, we first show that the empirical process term, $(\mathbb{P}_n - \mathbb{P})\left[\hat{\phi}_{dgg'g''}(Y, \boldsymbol{X}) - \phi_{dgg'g''}(Y, \boldsymbol{X})\right]$, is $o_p(n^{-1/2})$. Then, we show that the remainder term is also $o_p(n^{-1/2})$ in both cases relevant to us, i.e., when $g = g''$ and when $g' = g''$.

For the empirical process term,

$$
\begin{aligned}
&\hat{\phi}_{dgg'g''}(Y, \boldsymbol{X}) - \phi_{dgg'g''}(Y, \boldsymbol{X}) \\
&= \frac{\mathbb{1}(G = g'')}{\hat{p}_{g''}}\left(\xi_{dgg'g''} - \hat{\xi}_{dgg'g''}\right) + \mathbb{1}(G = g'')\frac{\hat{p}_{g''} - p_{g''}}{\hat{p}_{g''}p_{g''}}\xi_{dgg'g''} \\
&\quad + \mathbb{1}(G = g'')\frac{p_{g''} - \hat{p}_{g''}}{p_{g''}\hat{p}_{g''}}\omega(d, \boldsymbol{Q}, g)\,\mathrm{E}(D \mid \boldsymbol{Q}, g') \\
&\quad + \frac{\mathbb{1}(G = g'')}{\hat{p}_{g''}}\omega(d, \boldsymbol{Q}, g)\left[\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') - \mathrm{E}(D \mid \boldsymbol{Q}, g')\right] \\
&\quad + \frac{\mathbb{1}(G = g'')}{\hat{p}_{g''}}\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g')\left[\hat{\omega}(d, \boldsymbol{Q}, g) - \omega(d, \boldsymbol{Q}, g)\right] \\
&\quad + \frac{\mathbb{1}(G = g)\hat{p}_{g''}(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q})\hat{p}_{g''}}\omega(d, \boldsymbol{Q}, g)\left[\mathrm{E}(D \mid \boldsymbol{Q}, g') - \hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g')\right] \\
&\quad + \frac{\mathbb{1}(G = g)\hat{p}_{g''}(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q})\hat{p}_{g''}}\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g')\left[\omega(d, \boldsymbol{Q}, g) - \hat{\omega}(d, \boldsymbol{Q}, g)\right] \\
&\quad + \mathbb{1}(G = g)\frac{p_{g''}(\boldsymbol{Q})}{p_g(\boldsymbol{Q})}\frac{\hat{p}_{g''} - p_{g''}}{\hat{p}_{g''}p_{g''}}\omega(d, \boldsymbol{Q}, g)\,\mathrm{E}(D \mid \boldsymbol{Q}, g') \\
&\quad + \frac{\mathbb{1}(G = g)}{\hat{p}_{g''}}\frac{1}{\hat{p}_g(\boldsymbol{Q})}\left[p_{g''}(\boldsymbol{Q}) - \hat{p}_{g''}(\boldsymbol{Q})\right]\omega(d, \boldsymbol{Q}, g)\,\mathrm{E}(D \mid \boldsymbol{Q}, g') \\
&\quad + \frac{\mathbb{1}(G = g)}{\hat{p}_{g''}}\frac{\hat{p}_g(\boldsymbol{Q}) - p_g(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q})p_g(\boldsymbol{Q})}p_{g''}(\boldsymbol{Q})\omega(d, \boldsymbol{Q}, g)\,\mathrm{E}(D \mid \boldsymbol{Q}, g') \\
&\quad + \frac{\mathbb{1}(G = g)}{\hat{p}_{g''}}\frac{\hat{p}_{g''}(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q})}\left[1 - \frac{\mathbb{1}(D = d)}{\pi(d, \boldsymbol{X}, g)}\right]\left[\hat{\mu}(d, \boldsymbol{X}, g) - \mu(d, \boldsymbol{X}, g)\right] \\
&\quad + \frac{\mathbb{1}(G = g)}{\hat{p}_{g''}}\frac{\hat{p}_{g''}(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q})}\frac{\mathbb{1}(D = d)[Y - \hat{\mu}(d, \boldsymbol{X}, g)]}{\hat{\pi}(d, \boldsymbol{X}, g)\pi(d, \boldsymbol{X}, g)}\left[\pi(d, \boldsymbol{X}, g) - \hat{\pi}(d, \boldsymbol{X}, g)\right] \\
&\quad + \mathbb{1}(G = g)\frac{p_{g''} - \hat{p}_{g''}}{p_{g''}\hat{p}_{g''}}\frac{p_{g''}(\boldsymbol{Q})}{p_g(\boldsymbol{Q})}\left\{\frac{\mathbb{1}(D = d)}{\pi(d, \boldsymbol{X}, g)}[Y - \mu(d, \boldsymbol{X}, g)] + \mu(d, \boldsymbol{X}, g)\right\} \\
&\quad + \frac{\mathbb{1}(G = g)}{\hat{p}_{g''}}\frac{p_g(\boldsymbol{Q}) - \hat{p}_g(\boldsymbol{Q})}{p_g(\boldsymbol{Q})\hat{p}_g(\boldsymbol{Q})}p_{g''}(\boldsymbol{Q})\left\{\frac{\mathbb{1}(D = d)}{\pi(d, \boldsymbol{X}, g)}[Y - \mu(d, \boldsymbol{X}, g)] + \mu(d, \boldsymbol{X}, g)\right\} \\
&\quad + \frac{\mathbb{1}(G = g')\hat{p}_{g''}(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})\hat{p}_{g''}}D\left[\hat{\omega}(d, \boldsymbol{Q}, g) - \omega(d, \boldsymbol{Q}, g)\right] \\
&\quad + \mathbb{1}(G = g')\frac{p_{g''} - \hat{p}_{g''}}{p_{g''}\hat{p}_{g''}}\frac{p_{g''}(\boldsymbol{Q})}{p_{g'}(\boldsymbol{Q})}D\omega(d, \boldsymbol{Q}, g) \\
&\quad + \frac{\mathbb{1}(G = g')}{\hat{p}_{g''}}\frac{1}{\hat{p}_{g'}(\boldsymbol{Q})}\left[\hat{p}_{g''}(\boldsymbol{Q}) - p_{g''}(\boldsymbol{Q})\right]D\omega(d, \boldsymbol{Q}, g) \\
&\quad + \frac{\mathbb{1}(G = g')}{\hat{p}_{g''}}\frac{p_{g'}(\boldsymbol{Q}) - \hat{p}_{g'}(\boldsymbol{Q})}{p_{g'}(\boldsymbol{Q})\hat{p}_{g'}(\boldsymbol{Q})}p_{g''}(\boldsymbol{Q})D\omega(d, \boldsymbol{Q}, g)
\end{aligned}
$$

$$+ \frac{\mathbb{1}(G = g')\hat{p}_{g''}(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})\hat{p}_{g''}} \omega(d, \boldsymbol{Q}, g) \left[\mathrm{E}(D \mid \boldsymbol{Q}, g') - \hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') \right]$$

$$+ \frac{\mathbb{1}(G = g')\hat{p}_{g''}(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})\hat{p}_{g''}} \hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') \left[\omega(d, \boldsymbol{Q}, g) - \hat{\omega}(d, \boldsymbol{Q}, g) \right]$$

$$+ \mathbb{1}(G = g') \frac{\hat{p}_{g''} - p_{g''}}{\hat{p}_{g''} p_{g''}} \frac{p_{g''}(\boldsymbol{Q})}{p_{g'}(\boldsymbol{Q})} \mathrm{E}(D \mid \boldsymbol{Q}, g')\omega(d, \boldsymbol{Q}, g)$$

$$+ \frac{\mathbb{1}(G = g')}{\hat{p}_{g''}} \frac{1}{\hat{p}_{g'}(\boldsymbol{Q})} \left[p_{g''}(\boldsymbol{Q}) - \hat{p}_{g''}(\boldsymbol{Q}) \right] \mathrm{E}(D \mid \boldsymbol{Q}, g')\omega(d, \boldsymbol{Q}, g)$$

$$+ \frac{\mathbb{1}(G = g')}{\hat{p}_{g''}} \frac{\hat{p}_{g'}(\boldsymbol{Q}) - p_{g'}(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})p_{g'}(\boldsymbol{Q})} p_{g''}(\boldsymbol{Q}) \mathrm{E}(D \mid \boldsymbol{Q}, g')\omega(d, \boldsymbol{Q}, g).$$

Using arguments similar to above, we can show that the empirical process term is indeed asymptotically negligible under stated conditions. Next, we turn to the remainder term.

D.2.1 When $g = g''$

Note that for the conditional prevalence component, all ξ terms satisfy $g = g''$.

$$R_{2,dgg'g''}$$

$$= \hat{\xi}_{dgg'g} - \xi_{dgg'g}$$

$$\quad + \mathbb{P}\left\{ -\frac{\mathbb{1}(G = g)}{\hat{p}_g} \hat{\xi}_{dgg'g} \right\}$$

$$\quad + \mathbb{P}\left\{ \frac{\mathbb{1}(G = g)}{\hat{p}_g} \left\{ \frac{\mathbb{1}(D = d)}{\hat{\pi}(d, \boldsymbol{X}, g)}[Y - \hat{\mu}(d, \boldsymbol{X}, g)] + \hat{\mu}(d, \boldsymbol{X}, g) \right\} \hat{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g' \right) \right\}$$

$$\quad + \mathbb{P}\left\{ \frac{\mathbb{1}(G = g')\hat{p}_g(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})\hat{p}_g} \left[D - \hat{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g' \right) \right] \hat{\omega}(d, \boldsymbol{Q}, g) \right\}$$

$$= \left(1 - \frac{p_g}{\hat{p}_g} \right) \hat{\xi}_{dgg'g} - \xi_{dgg'g}$$

$$\quad + \mathbb{P}\left\{ \frac{\mathbb{1}(G = g)}{\hat{p}_g} \left[1 - \frac{\pi(d, \boldsymbol{X}, g)}{\hat{\pi}(d, \boldsymbol{X}, g)} \right] [\hat{\mu}(d, \boldsymbol{X}, g) - \mu(d, \boldsymbol{X}, g)]\hat{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g' \right) \right\}$$

$$\quad + \mathbb{P}\left\{ \frac{\mathbb{1}(G = g)}{\hat{p}_g} \omega(d, \boldsymbol{Q}, g)\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') \right\}$$

$$\quad + \mathbb{P}\left\{ \frac{\mathbb{1}(G = g')\hat{p}_g(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})\hat{p}_g} \left[\mathrm{E}(D \mid \boldsymbol{Q}, g') - \hat{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g' \right) \right] \hat{\omega}(d, \boldsymbol{Q}, g) \right\}$$

$$= \left(1 - \frac{p_g}{\hat{p}_g} \right) \left(\hat{\xi}_{dgg'g} - \xi_{dgg'g} \right)$$

$$\quad + \mathbb{P}\left\{ \frac{\mathbb{1}(G = g)}{\hat{p}_g} \left[1 - \frac{\pi(d, \boldsymbol{X}, g)}{\hat{\pi}(d, \boldsymbol{X}, g)} \right] [\hat{\mu}(d, \boldsymbol{X}, g) - \mu(d, \boldsymbol{X}, g)]\hat{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g' \right) \right\}$$

$$\quad + \mathbb{P}\left\{ \left[\frac{\mathbb{1}(G = g)}{\hat{p}_g} \omega(d, \boldsymbol{Q}, g) - \frac{\mathbb{1}(G = g')\hat{p}_g(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})\hat{p}_g} \hat{\omega}(d, \boldsymbol{Q}, g) \right] \left[\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') - \mathrm{E}(D \mid \boldsymbol{Q}, g') \right] \right\}.$$

Then it follows from similar arguments as above that $R_{2,dgg'g''} = o_p(n^{-1/2})$ under stated conditions. In particular, for the last line, note that if $g = g' = g''$,

$$\mathbb{P}\left\{\left[\mathbb{1}(G = g)\omega(d, \boldsymbol{Q}, g) - \mathbb{1}(G = g')\frac{\hat{p}_g(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})}\hat{\omega}(d, \boldsymbol{Q}, g)\right]\left[\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') - \mathrm{E}(D \mid \boldsymbol{Q}, g')\right]\right\}$$

$$= \mathbb{P}\left\{\mathbb{1}(G = g)\left[\omega(d, \boldsymbol{Q}, g) - \hat{\omega}(d, \boldsymbol{Q}, g)\right]\left[\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') - \mathrm{E}(D \mid \boldsymbol{Q}, g')\right]\right\},$$

so $\|\omega(d, \boldsymbol{Q}, g) - \hat{\omega}(d, \boldsymbol{Q}, g)\| \left\|\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g) - \mathrm{E}(D \mid \boldsymbol{Q}, g)\right\| = o_p(n^{-1/2}), \forall d, g$, is sufficient for the last line.

If $g = g'' \neq g'$,

$$\mathbb{P}\left\{\left[\mathbb{1}(G = g)\omega(d, \boldsymbol{Q}, g) - \mathbb{1}(G = g')\frac{\hat{p}_g(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})}\hat{\omega}(d, \boldsymbol{Q}, g)\right]\left[\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') - \mathrm{E}(D \mid \boldsymbol{Q}, g')\right]\right\}$$

$$= \mathbb{P}\left\{\mathbb{1}(G = g')\frac{\hat{p}_g(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})}\left[\omega(d, \boldsymbol{Q}, g) - \hat{\omega}(d, \boldsymbol{Q}, g)\right]\left[\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') - \mathrm{E}(D \mid \boldsymbol{Q}, g')\right]\right\}$$

$$+ \mathbb{P}\left\{\frac{\hat{p}_g(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})}\left[\mathbb{1}(G = g) - \mathbb{1}(G = g')\right]\omega(d, \boldsymbol{Q}, g)\left[\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') - \mathrm{E}(D \mid \boldsymbol{Q}, g')\right]\right\}$$

$$+ \mathbb{P}\left\{\mathbb{1}(G = g)\left[1 - \frac{\hat{p}_g(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})}\right]\omega(d, \boldsymbol{Q}, g)\left[\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') - \mathrm{E}(D \mid \boldsymbol{Q}, g')\right]\right\},$$

so the following conditions are sufficient for the last line: for some $\zeta < \infty$, $\left|\frac{\hat{p}_g(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})}\right| \leq \zeta$ with probability 1, $\left\|\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g) - \mathrm{E}(D \mid \boldsymbol{Q}, g)\right\| = o_p(n^{-1/2})$, and $\|\omega(d, \boldsymbol{Q}, g) - \hat{\omega}(d, \boldsymbol{Q}, g)\| = o_p(1)$, $\forall d, g, g'$.

D.2.2 When $g' = g''$

Note that all ξ terms satisfy $g' = g''$ for the conditional effect component.

$$R_{2,dgg'g''}$$

$$= \hat{\xi}_{dgg'g'} - \xi_{dgg'g'}$$

$$+ \mathbb{P}\left\{-\frac{\mathbb{1}(G = g')}{\hat{p}_{g'}}\hat{\xi}_{dgg'g'}\right\}$$

$$+ \mathbb{P}\left\{\frac{\mathbb{1}(G = g)\hat{p}_{g'}(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q})\hat{p}_{g'}}\left\{\frac{\mathbb{1}(D = d)}{\hat{\pi}(d, \boldsymbol{X}, g)}[Y - \hat{\mu}(d, \boldsymbol{X}, g)] + \hat{\mu}(d, \boldsymbol{X}, g) - \hat{\omega}(d, \boldsymbol{Q}, g)\right\}\hat{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g'\right)\right\}$$

$$+ \mathbb{P}\left\{\frac{\mathbb{1}(G = g')}{\hat{p}_{g'}}D \cdot \hat{\omega}(d, \boldsymbol{Q}, g)\right\}$$

$$= \left(1 - \frac{p_g}{\hat{p}_g}\right)\left(\hat{\xi}_{dgg'g'} - \xi_{dgg'g'}\right)$$

$$+ \mathbb{P}\left\{\frac{\mathbb{1}(G = g)\hat{p}_{g'}(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q})\hat{p}_{g'}}\left[1 - \frac{\pi(d, \boldsymbol{X}, g)}{\hat{\pi}(d, \boldsymbol{X}, g)}\right][\hat{\mu}(d, \boldsymbol{X}, g) - \mu(d, \boldsymbol{X}, g)]\hat{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g'\right)\right\}$$

$$+ \mathbb{P}\left\{\left[\frac{\mathbb{1}(G = g')}{\hat{p}_{g'}}\mathrm{E}(D \mid \boldsymbol{Q}, g') - \frac{\mathbb{1}(G = g)\hat{p}_{g'}(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q})\hat{p}_{g'}}\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g')\right][\hat{\omega}(d, \boldsymbol{Q}, g) - \omega(d, \boldsymbol{Q}, g)]\right\}.$$

Under stated conditions, $R_{2,dgg'g''} = o_p(n^{-1/2})$. In particular, for the last line, the following conditions are sufficient: for some $\zeta < \infty$, $\left|\frac{\hat{p}_{g'}(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q})}\right| \leq \zeta$ with probability 1, $\left\|\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g) - \mathrm{E}(D \mid \boldsymbol{Q}, g)\right\| = o_p(1)$, and $\|\omega(d, \boldsymbol{Q}, g) - \hat{\omega}(d, \boldsymbol{Q}, g)\| = o_p(n^{-1/2})$, $\forall d, g, g'$.

D.2.3 Summary of convergence rate conditions

When $g = g' = g''$, which holds for the conditional selection component, $\|\hat{\omega}(d, \boldsymbol{Q}, g) - \omega(d, \boldsymbol{Q}, g)\| \left\|\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g) - \right.$ $o_p(n^{-1/2})$ is sufficient. Hence, we obtain a form of rate double robustness with respect to $\omega(d, \boldsymbol{Q}, g)$ and $\mathrm{E}(D \mid \boldsymbol{Q}, g)$. Second, when $g = g'' \neq g'$, which holds for the conditional prevalence component, the following condition is sufficient[20]: $\left\|\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g) - \mathrm{E}(D \mid \boldsymbol{Q}, g)\right\| = o_p(n^{-1/2})$. Third, when $g' = g'' \neq g$, which holds for the conditional effect component, a sufficient condition is: $\|\hat{\omega}(d, \boldsymbol{Q}, g) - \omega(d, \boldsymbol{Q}, g)\| = o_p(n^{-1/2})$. Therefore, the assumption is weaker for conditional selection than other components, which is also the case for Theorem 5.

Appendix E. Multiple robustness

The proof for Theorem 2 is omitted, as it is similar to the proof of the double robustness of the AIPW estimator for the ATE (Robins et al., 1994).

Below, we prove the multiple robustness of $\hat{\xi}_{dgg'g''}$, which is stated in Theorem 5. We use a tilde to denote the probability limit of a nuisance estimator. For example, $\hat{\mu}(d, \boldsymbol{X}, g)$ converges to $\tilde{\mu}(d, \boldsymbol{X}, g)$. Under consistent estimation of p_g, the one-step estimator $\hat{\xi}_{dgg'g''}$ converges in probability to

$$
\mathrm{E}\left\{\frac{\mathbb{1}(G = g'')}{p_{g''}}\tilde{\omega}(d, \boldsymbol{Q}, g)\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g'\right)\right\}
$$
$$
+ \mathrm{E}\left\{\frac{\mathbb{1}(G = g)\tilde{p}_{g''}(\boldsymbol{Q})}{\tilde{p}_g(\boldsymbol{Q})p_{g''}}\left\{\frac{\mathbb{1}(D = d)}{\tilde{\pi}(d, \boldsymbol{X}, g)}[Y - \tilde{\mu}(d, \boldsymbol{X}, g)] + \tilde{\mu}(d, \boldsymbol{X}, g) - \tilde{\omega}(d, \boldsymbol{Q}, g)\right\}\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g'\right)\right\}
$$
$$
+ \mathrm{E}\left\{\frac{\mathbb{1}(G = g')\tilde{p}_{g''}(\boldsymbol{Q})}{\tilde{p}_{g'}(\boldsymbol{Q})p_{g''}}\left[D - \tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g'\right)\right]\tilde{\omega}(d, \boldsymbol{Q}, g)\right\}
$$
$$
= \mathrm{E}\left\{\frac{\mathbb{1}(G = g'')}{p_{g''}}\tilde{\omega}(d, \boldsymbol{Q}, g)\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g'\right)\right\}
$$
$$
+ \mathrm{E}\left\{\frac{\mathbb{1}(G = g)\tilde{p}_{g''}(\boldsymbol{Q})}{\tilde{p}_g(\boldsymbol{Q})p_{g''}}\left\{\frac{\pi(d, \boldsymbol{X}, g)}{\tilde{\pi}(d, \boldsymbol{X}, g)}[\mu(d, \boldsymbol{X}, g) - \tilde{\mu}(d, \boldsymbol{X}, g)] + \tilde{\mu}(d, \boldsymbol{X}, g)\right\}\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g'\right)\right\}
$$
$$
- \mathrm{E}\left\{\frac{\mathbb{1}(G = g)\tilde{p}_{g''}(\boldsymbol{Q})}{\tilde{p}_g(\boldsymbol{Q})p_{g''}}\tilde{\omega}(d, \boldsymbol{Q}, g)\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g'\right)\right\}
$$
$$
+ \mathrm{E}\left\{\frac{\mathbb{1}(G = g')\tilde{p}_{g''}(\boldsymbol{Q})}{\tilde{p}_{g'}(\boldsymbol{Q})p_{g''}}\left[\mathrm{E}(D \mid \boldsymbol{Q}, g') - \tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q}, g'\right)\right]\tilde{\omega}(d, \boldsymbol{Q}, g)\right\}
$$

---

[20] Recall that by Assumption 6a, for some $\zeta < \infty$, $\hat{p}_g(\boldsymbol{Q})/\hat{p}_{g'}(\boldsymbol{Q}) \leq \zeta$ with probability 1, $\forall g, g'$. And by Assumption 6b, $\|\hat{\omega}(d, \boldsymbol{Q}, g) - \omega(d, \boldsymbol{Q}, g)\| = o_p(1)$, $\left\|\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g) - \mathrm{E}(D \mid \boldsymbol{Q}, g)\right\| = o_p(1)$, $\forall d, g$.

$$
= \mathrm{E}\left\{\frac{p_{g''}(\boldsymbol{Q})}{p_{g''}}\tilde{\omega}(d,\boldsymbol{Q},g)\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q},g'\right)\right\}
$$

$$
+ \mathrm{E}\left\{\frac{p_g(\boldsymbol{Q})\tilde{p}_{g''}(\boldsymbol{Q})}{\tilde{p}_g(\boldsymbol{Q})p_{g''}}\mathrm{E}\left\{\frac{\pi(d,\boldsymbol{X},g)}{\tilde{\pi}(d,\boldsymbol{X},g)}[\mu(d,\boldsymbol{X},g)-\tilde{\mu}(d,\boldsymbol{X},g)] + \tilde{\mu}(d,\boldsymbol{X},g) \mid \boldsymbol{Q},g\right\}\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q},g'\right)\right\}
$$

$$
- \mathrm{E}\left\{\frac{p_g(\boldsymbol{Q})\tilde{p}_{g''}(\boldsymbol{Q})}{\tilde{p}_g(\boldsymbol{Q})p_{g''}}\tilde{\omega}(d,\boldsymbol{Q},g)\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q},g'\right)\right\}
$$

$$
+ \mathrm{E}\left\{\frac{p_{g'}(\boldsymbol{Q})\tilde{p}_{g''}(\boldsymbol{Q})}{\tilde{p}_{g'}(\boldsymbol{Q})p_{g''}}\left[\mathrm{E}(D \mid \boldsymbol{Q},g')-\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q},g'\right)\right]\tilde{\omega}(d,\boldsymbol{Q},g)\right\}.
$$

Now, when $g = g''$, this probability limit becomes

$$
\mathrm{E}\left\{\frac{p_g(\boldsymbol{Q})}{p_g}\mathrm{E}\left\{\frac{\pi(d,\boldsymbol{X},g)}{\tilde{\pi}(d,\boldsymbol{X},g)}[\mu(d,\boldsymbol{X},g)-\tilde{\mu}(d,\boldsymbol{X},g)] + \tilde{\mu}(d,\boldsymbol{X},g) \mid \boldsymbol{Q},g\right\}\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q},g'\right)\right\}
$$

$$
+ \mathrm{E}\left\{\frac{p_{g'}(\boldsymbol{Q})\tilde{p}_g(\boldsymbol{Q})}{\tilde{p}_{g'}(\boldsymbol{Q})p_g}\left[\mathrm{E}(D \mid \boldsymbol{Q},g')-\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q},g'\right)\right]\tilde{\omega}(d,\boldsymbol{Q},g)\right\}.
$$

Next, when $g' = g''$, the probability limit becomes

$$
\mathrm{E}\left\{\frac{p_g(\boldsymbol{Q})\tilde{p}_{g'}(\boldsymbol{Q})}{\tilde{p}_g(\boldsymbol{Q})p_{g'}}\mathrm{E}\left\{\frac{\pi(d,\boldsymbol{X},g)}{\tilde{\pi}(d,\boldsymbol{X},g)}[\mu(d,\boldsymbol{X},g)-\tilde{\mu}(d,\boldsymbol{X},g)] + \tilde{\mu}(d,\boldsymbol{X},g) \mid \boldsymbol{Q},g\right\}\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q},g'\right)\right\}
$$

$$
- \mathrm{E}\left\{\frac{p_g(\boldsymbol{Q})\tilde{p}_{g'}(\boldsymbol{Q})}{\tilde{p}_g(\boldsymbol{Q})p_{g'}}\tilde{\omega}(d,\boldsymbol{Q},g)\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q},g'\right)\right\} + \mathrm{E}\left\{\frac{p_{g'}(\boldsymbol{Q})}{p_{g'}}\mathrm{E}(D \mid \boldsymbol{Q},g')\tilde{\omega}(d,\boldsymbol{Q},g)\right\}.
$$

Finally, when $g = g' = g''$, it becomes

$$
\mathrm{E}\left\{\frac{p_g(\boldsymbol{Q})}{p_g}\mathrm{E}\left\{\frac{\pi(d,\boldsymbol{X},g)}{\tilde{\pi}(d,\boldsymbol{X},g)}[\mu(d,\boldsymbol{X},g)-\tilde{\mu}(d,\boldsymbol{X},g)] + \tilde{\mu}(d,\boldsymbol{X},g) \mid \boldsymbol{Q},g\right\}\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q},g\right)\right\}
$$

$$
+ \mathrm{E}\left\{\frac{p_g(\boldsymbol{Q})}{p_g}\left[\mathrm{E}(D \mid \boldsymbol{Q},g)-\tilde{\mathrm{E}}\left(D \mid \boldsymbol{Q},g\right)\right]\tilde{\omega}(d,\boldsymbol{Q},g)\right\}.
$$

Then it is relatively easy to verify the results in Theorem 5. Note that the condition for the conditional selection component to be consistently estimated is the weakest compared with other components. In particular, for the conditional selection component, it is not necessary to consistently estimate $p_g(\boldsymbol{Q})$.

Appendix F. Relation between selection concepts

Various concepts of "selection", "sorting", and "selectivity" have appeared in the social science literature on the effects of education on later-life attainments. Below, we clarify the relationship between our selection concept and those appearing in prior work. In our framework, group-specific selection is defined as

$$
\mathrm{Cov}_g(D,\tau) = [\mathrm{E}_g(\tau \mid D = 1) - \mathrm{E}_g(\tau)]\,\mathrm{E}_g(D).
$$

First, our definition is closely related to the "sorting on gains" in economics (Heckman and Vytlacil, 2005; Heckman et al., 2018), which is defined as the difference between the average treatment effect on the treated and the ATE, i.e.,

$$\mathrm{E}_g(\tau \mid D = 1) - \mathrm{E}_g(\tau).$$

Since $\mathrm{E}_g(D)$ is always non-negative, the sign of our group-specific selection is always the same as that of sorting on gains.

Second, akin to our framework, recent works in the "great equalizer" literature (Zhou, 2019; Fiel, 2020; Karlson, 2019) have discussed *differential* selection into college completion by parental income. Implicitly, these authors define group-specific selectivity as

$$\mathrm{E}_g(Y \mid D = 1) - \mathrm{E}_g(Y \mid D = 0) - \mathrm{E}_g(\tau).$$

This implicit definition can be validated by rewriting the group difference in this selectivity term.

$$
\mathrm{E}_a(Y \mid D = 1) - \mathrm{E}_a(Y \mid D = 0) - \mathrm{E}_a(\tau) - [\mathrm{E}_b(Y \mid D = 1) - \mathrm{E}_b(Y \mid D = 0) - \mathrm{E}_b(\tau)]
$$

$$
= \underbrace{\mathrm{E}_a(Y \mid D = 1) - \mathrm{E}_b(Y \mid D = 1) - [\mathrm{E}_a(Y \mid D = 0) - \mathrm{E}_b(Y \mid D = 0)]}_{\text{The descriptive test of the great equalizer thesis}}
$$

$$
- \underbrace{\{\mathrm{E}_a(Y^1) - \mathrm{E}_b(Y^1) - [\mathrm{E}_a(Y^0) - \mathrm{E}_b(Y^0)]\}}_{\text{The causal test of the great equalizer thesis}}
$$

If the descriptive test returns a more negative value than the causal test, these authors will conclude that members of the advantaged group are less selected than their disadvantaged peers in the transition to college completion.[21]

Taking the difference between the selectivity term in the great equalizer literature and sorting on gains, we obtain $\mathrm{E}_g(Y^0 \mid D = 1) - \mathrm{E}_g(Y^0 \mid D = 0)$, which is selection on baseline potential outcome.[22] Therefore, the selectivity term in the great equalizer literature captures selection on baseline outcome, on top of selection on treatment effect captured in sorting on gains and our selection term (see Morgan and Winship, 2014, p.58-9).

---

[21] This conclusion, however, corresponds more closely to $\mathrm{E}_a(Y \mid D = 1) - \mathrm{E}_a(Y^1) - [\mathrm{E}_b(Y \mid D = 1) - \mathrm{E}_b(Y^1)]$ being negative. For example, if $\mathrm{E}_a(Y \mid D = 1) - \mathrm{E}_a(Y^1) = 0$, while $\mathrm{E}_b(Y \mid D = 1) - \mathrm{E}_b(Y^1) > 0$, then the treated individuals in group b (and group b only) have higher Y_1 than the untreated individuals in the same group, which means that the treated individuals in group b (and group b only) are positively selected on Y_1 (see Yu and Zhao, 2024).

[22] In Heckman et al. (1998), Heckman and Vytlacil (2005) and Heckman et al. (2018), selection on baseline outcome is referred to as the "selection bias".

Third, a large literature on effect heterogeneity by propensity score (Brand and Xie, 2010; Xie et al., 2012; Brand et al., 2021) is closely aligned with our framework in its conceptualization of selection into treatment. Originally developed in the context of education effects, this framework has been applied in a wide array of topics. Recall that $\pi(1, \boldsymbol{X}, g)$ denotes the propensity score, then under Assumption 3 (conditional ignorability),

$$
\begin{aligned}
&\mathrm{Cov}_g(D, \tau) \\
&= \mathrm{E}_g[\mathrm{Cov}_g(D, \tau \mid \boldsymbol{X})] + \mathrm{Cov}_g[\mathrm{E}_g(D \mid \boldsymbol{X}), \mathrm{E}_g(\tau \mid \boldsymbol{X})] \\
&= \mathrm{Cov}_g[\pi(1, \boldsymbol{X}, g), \mathrm{E}_g(\tau \mid \boldsymbol{X})] \\
&= \mathrm{E}_g[\pi(1, \boldsymbol{X}, g) \, \mathrm{E}_g(\tau \mid \boldsymbol{X})] - \mathrm{E}_g(D) \, \mathrm{E}_g(\tau) \\
&= \mathrm{E}_g\{\mathrm{E}_g[\pi(1, \boldsymbol{X}, g) \, \mathrm{E}_g(\tau \mid \boldsymbol{X}) \mid \pi(1, \boldsymbol{X}, g)]\} - \mathrm{E}_g(D) \, \mathrm{E}_g(\tau) \\
&= \mathrm{E}_g\{\pi(1, \boldsymbol{X}, g) \, \mathrm{E}_g[\tau \mid \pi(1, \boldsymbol{X}, g)]\} - \mathrm{E}_g[\pi(1, \boldsymbol{X}, g)] \, \mathrm{E}_g\{\mathrm{E}_g[\tau \mid \pi(1, \boldsymbol{X}, g)]\} \\
&= \mathrm{Cov}_g[\pi(1, \boldsymbol{X}, g), \mathrm{E}_g(\tau \mid \pi(1, \boldsymbol{X}, g)].
\end{aligned}
$$

Hence, under conditional ignorability, our selectivity term equals the covariance between the propensity score and the conditional treatment effect given the propensity score. Due to this relationship, the estimators we developed for the generic functions underlying our unconditional decomposition can also be used to nonparametrically estimate effect heterogeneity by propensity score.

Appendix G. Supplemental tables

Table A1. Relationship between baseline covariates and outcome missingness

Covariate		Missing percentage or correlation coefficient	P value
Parental income		-0.079$^{\#}$	< 0.001
Race	White	25.27	
	Black	14.45	< 0.001
	Hispanic	12.38	
Gender	Male	21.09	
	Female	19.63	0.385
Mother's year of schooling		0.003$^{\#}$	0.870
Parental presence	Yes	19.31	
	No	22.21	0.089
Number of siblings		0.002$^{\#}$	0.900

Urban residence	Yes		19.65	0.108
	No		22.86	
Expecting bachelor	Yes		19.28	0.250
degree or higher	No		21.20	
AFQT score			$0.021^{\#}$	0.294
Age			$-0.015^{\#}$	0.453
Friends expecting	Yes		19.15	0.190
bachelor degree or higher	No		21.31	
Rotter score			$-0.002^{\#}$	0.914
Rosenberg score			$-0.013^{\#}$	0.517
School satisfaction			0.003	0.888
Speak foreign language	Yes		18.07	0.174
at home	No		20.91	
	Not in SMSA		22.73	
Metropolitan Statistical	In SMSA, not central city		18.00	0.057
Area category	In SMSA, in central city		22.32	
	In SMSA, central city unknown		18.59	
Separate from mother	Yes		25.83	0.158
	No		20.08	
Mother working	Yes		18.27	0.004
	No		23.00	
	Northeast		23.21	
SMSA	North central		22.42	0.031
	South		18.75	
	West		16.95	

Note: N=2580. The sample is individuals with no missing values in any baseline covariates. In the column for missing percentage or correlation coefficient, values with # are correlation coefficients between covariates and the outcome missingness indicator. P values for missing percentages are based on Chi-squared tests.

Table A2. Group-specific estimates for the unconditional decomposition

	Top 40% income	Bottom 40% income	Top-Bottom
Treatment proportion	0.337	0.086	0.251
	(0.310, 0.364)	(0.067, 0.105)	(0.218, 0.284)
	Neural networks		
Baseline outcome	0.599	0.411	0.188
	(0.582, 0.617)	(0.393, 0.430)	(0.163,0.213)

ATE	0.145	0.123	0.022
	(0.103, 0.188)	(0.052, 0.195)	(-0.061, 0.105)
Cov(D,τ)	-0.014	0.002	-0.017
	(-0.027, -0.002)	(-0.004, 0.009)	(-0.031,-0.002)

Gradient boosted machines

Baseline outcome	0.599	0.411	0.188
	(0.581, 0.616)	(0.393, 0.429)	(0.162, 0.213)
ATE	0.137	0.110	0.027
	(0.100, 0.174)	(0.044, 0.177)	(-0.049, 0.103)
Cov(D,τ)	-0.011	0.004	-0.014
	(-0.021, 0.000)	(-0.002, 0.009)	(-0.026,-0.002)

Random forests

Baseline outcome	0.593	0.410	0.183
	(0.578,0.609)	(0.392,0.429)	(0.159,0.207)
ATE	0.140	0.116	0.024
	(0.107, 0.172)	(0.040, 0.192)	(-0.058, 0.107)
Cov(D,τ)	-0.006	0.004	-0.010
	(-0.014, 0.001)	(-0.002, 0.010)	(-0.020,-0.001)

Parametric models

Baseline outcome	0.598	0.410	0.187
	(0.580, 0.615)	(0.393, 0.428)	(0.162,0.212)
ATE	0.121	0.145	-0.024
	(0.080, 0.162)	(0.059, 0.232)	(-0.120, 0.072)
Cov(D,τ)	-0.005	0.001	-0.006
	(-0.017, 0.008)	(-0.006, 0.008)	(-0.020,0.009)

Note: 95% confidence intervals are in parentheses and are computed according to Theorem 3.

Weight stabilization is used. For ML models, cross-fitting is used.

Table A3. Unconditional decomposition estimates

	Neural networks	GBM	Random forests	Parametric
Total	0.210	0.210	0.210	0.210
	(0.188,0.232)	(0.188,0.232)	(0.188,0.232)	(0.188,0.232)
Baseline	0.188	0.188	0.183	0.187
	(0.163,0.213)	(0.162,0.213)	(0.159,0.207)	(0.162,0.212)
Prevalence	0.031	0.028	0.029	0.036
	(0.012,0.050)	(0.011,0.045)	(0.009,0.049)	(0.014,0.059)
Effect	0.007	0.009	0.008	-0.008

	(-0.021,0.035)	(-0.017,0.035)	(-0.020,0.036)	(-0.040,0.024)
Selection	-0.017	-0.014	-0.010	-0.006
	(-0.031,-0.002)	(-0.026,-0.002)	(-0.020,-0.001)	(-0.020,0.009)
Change in disparity	0.029	0.024	0.025	0.035
	(0.007,0.050)	(0.005,0.043)	(0.002,0.048)	(0.008,0.063)

Note: 95% confidence intervals are in parentheses and are computed according to Theorem 3. Weight stabilization is used. For ML models, cross-fitting is used. The change in disparity uses the definition in Jackson and Vanderweele's (2018) URED.

Table A4. Conditional decomposition estimates

	Neural networks	GBM	Random forests	Parametric
Total	0.210	0.210	0.210	0.210
	(0.188,0.232)	(0.188,0.232)	(0.188,0.232)	(0.188,0.232)
Baseline	0.188	0.188	0.183	0.187
	(0.163,0.213)	(0.162,0.213)	(0.159,0.207)	(0.162,0.212)
Conditional prevalence	0.010	0.007	0.013	0.009
	(0.003,0.016)	(0.002,0.012)	(0.007,0.019)	(0.002,0.015)
Conditional effect	-0.012	-0.028	-0.025	-0.023
	(-0.055,0.031)	(-0.061,0.006)	(-0.054,0.004)	(-0.054,0.007)
Conditional selection	-0.006	-0.004	-0.006	-0.002
	(-0.014,0.003)	(-0.010,0.002)	(-0.014,0.002)	(-0.010,0.005)
$\boldsymbol{Q}$-distribution	0.030	0.047	0.045	0.040
	(-0.004,0.063)	(0.022,0.072)	(0.023,0.067)	(0.018,0.061)
Change in disparity	0.006	0.006	0.007	0.006
	(-0.003,0.016)	(-0.001,0.013)	(-0.002,0.017)	(-0.003,0.016)

Note: 95% confidence intervals are in parentheses and are computed according to Theorem 6. Weight stabilization is used. For ML models, cross-fitting is used. The change in disparity uses the definition in Jackson's (2021) CRED.

In addition, in Table A5, we present a set of estimates for the conditional decomposition as a robustness check. This is motivated by Assumption 6c, which requires the parametric convergence rate for $\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g)$ for the conditional prevalence component, the parametric convergence rate for $\hat{\omega}(d, \boldsymbol{Q}, g)$ for the conditional effect component, and the same rate for both nuisance function models for the $\boldsymbol{Q}$-distribution component. Thus, to make asymptotic inference more exact, we implement the following procedure.

We estimate $\mu(d, \boldsymbol{X}, g)$, $\pi(d, \boldsymbol{X}, g)$, and $p_g(\boldsymbol{Q})$ using ML with cross-fitting. Then we estimate $\mathrm{E}(D \mid \boldsymbol{Q}, g)$ and $\omega(d, \boldsymbol{Q}, g)$ parametrically using linear or logistic regressions without cross-

fitting. For the parametric models, we include the group indicator, the AFQT score, the squared AFQT score, and the interactions between the group indicator and the AFQT variables. For $\omega(d, \boldsymbol{Q}, g)$, we apply the pseudo-outcome approach detailed in the main text, but with a different cross-fitting procedure. Here, we use cross-fitted estimates of $\mu(d, \boldsymbol{X}, g)$ and $\pi(d, \boldsymbol{X}, g)$ to construct the pseudo-outcomes, then we estimate $\omega(d, \boldsymbol{Q}, g)$ without cross-fitting. The findings are quantitatively similar and qualitatively identical across Tables A4 and A5.

Table A5. Conditional decomposition estimates with mixed nonparametric and parametric models

	Neural networks	GBM	Random forests
Total	0.210	0.210	0.210
	(0.188,0.232)	(0.188,0.232)	(0.188,0.232)
Baseline	0.188	0.182	0.183
	(0.163,0.213)	(0.157,0.208)	(0.159,0.207)
Conditional prevalence	0.008	0.008	0.008
	(0.002,0.014)	(0.002,0.013)	(0.002,0.014)
Conditional effect	-0.018	-0.019	-0.024
	(-0.049,0.013)	(-0.048,0.010)	(-0.054,0.006)
Conditional selection	-0.006	-0.001	-0.004
	(-0.014,0.003)	(-0.008,0.005)	(-0.010,0.002)
$\boldsymbol{Q}$-distribution	0.037	0.040	0.047
	(0.016,0.058)	(0.022,0.059)	(0.025,0.069)
Change in disparity	0.006	0.007	0.006
	(-0.002,0.015)	(-0.000,0.014)	(-0.003,0.014)

Note: 95% confidence intervals are in parentheses and are computed according to Theorem 6. Weight stabilization is used. $\mu(d, \boldsymbol{X}, g)$, $\pi(d, \boldsymbol{X}, g)$, and $p_g(\boldsymbol{Q})$ are estimated using ML with cross-fitting. $\mathrm{E}(D \mid \boldsymbol{Q}, g)$ and $\omega(d, \boldsymbol{Q}, g)$ are estimated using parametric models without cross-fitting. The change in disparity uses the definition in Jackson's (2021) CRED.

Appendix H. Diagnostic checks and sensitivity analysis for the empirical application

H.1. Covariate overlap

First, we assess covariate overlap in the data. Our Assumption 4 requires that at any level of the covariates, there are both individuals receiving the treatment and individuals receiving the control. We empirically assess this assumption by plotting the histograms of estimated propensity scores in Figure A3.

Across the four alternative models for the propensity score, the basic pattern is the same. There appear to be a large number of individuals whose probability of completing college is close to 0. However, there is no exact 0 or 1 in the estimated propensity scores, which means that Assumption 4 is minimally satisfied.
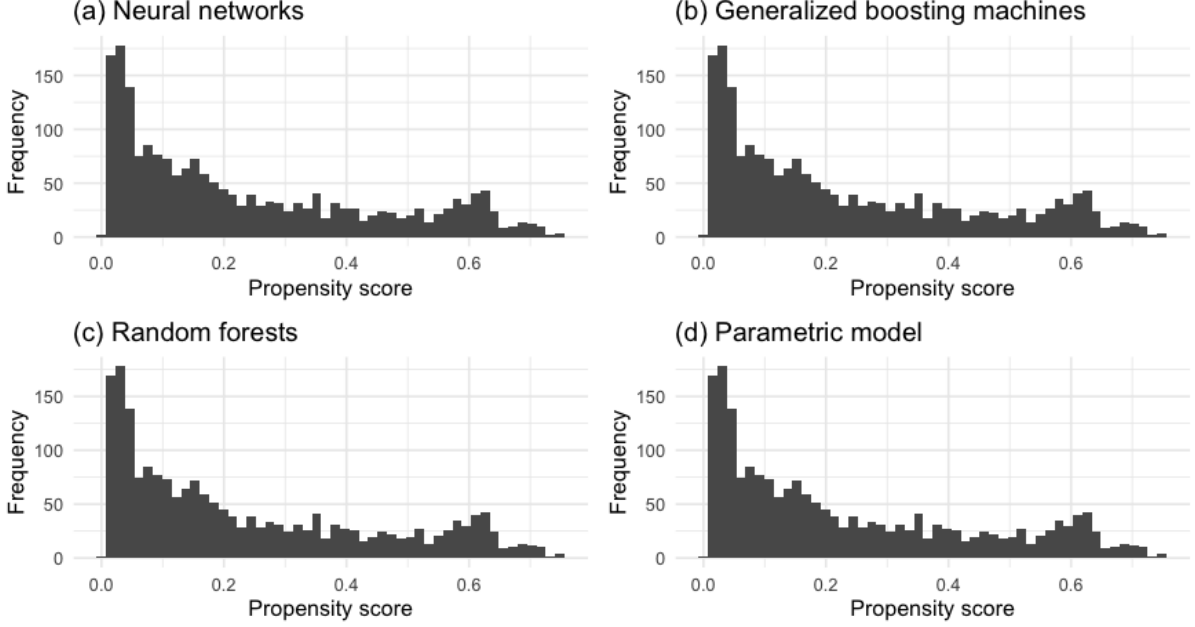

Figure A3. Covariate overlap

H.2. Covariate balance after weighting

We perform a diagnostic check for the propensity score models. The propensity score function is $\pi(d, \boldsymbol{X}, G) := \Pr(D = d \mid \boldsymbol{X}, G)$, which is a nuisance function shared by both the unconditional and conditional decompositions. Our estimators can be viewed as involving a weight based on the propensity score, $\frac{\mathbb{1}(D=d)}{\pi(d,\boldsymbol{X},G)}$. Also recall that $\hat{\pi}(d, \boldsymbol{X}, G)$ is the estimated counterpart of $\pi(d, \boldsymbol{X}, G)$. In this subsection, covariates refer to $\boldsymbol{C} := \{\boldsymbol{X}, G\}$, and C_j is the j- covariate.

Theoretically, all covariates should have the same distribution between treatment and control observations if the true propensity score is used to weight the data (Li et al., 2018). Thus, covariate balance after weighting has been used as a diagnostic check for the estimator of the propensity score. There is a variety of covariate balance measures in the weighting literature (Austin and Stuart, 2015; Li et al., 2018; Cannas and Arpino, 2019). We focus on the absolute standardized difference (ASD) as defined in Austin and Stuart (2015, p.3666):

$$100 \times \left| \frac{\sum C_j \frac{D}{\hat{\pi}(1,\boldsymbol{X},G)}}{\sum \frac{D}{\hat{\pi}(1,\boldsymbol{X},G)}} - \frac{\sum C_j \frac{1-D}{\hat{\pi}(0,\boldsymbol{X},G)}}{\sum \frac{1-D}{\hat{\pi}(0,\boldsymbol{X},G)}} \right| \Bigg/ \sqrt{\left(s_1^2 + s_0^2\right)/2},$$

where

$$s_d^2 := \frac{\sum \frac{\mathbb{1}(D=d)}{\hat{\pi}(d,\boldsymbol{X},G)}}{\left(\sum \frac{\mathbb{1}(D=d)}{\hat{\pi}(d,\boldsymbol{X},G)}\right)^2 - \sum \left(\frac{\mathbb{1}(D=d)}{\hat{\pi}(d,\boldsymbol{X},G)}\right)^2} \sum \frac{\mathbb{1}(D=d)}{\hat{\pi}(d,\boldsymbol{X},G)} \left(C_j - \frac{\sum C_j \frac{\mathbb{1}(D=d)}{\hat{\pi}(d,\boldsymbol{X},G)}}{\sum \frac{\mathbb{1}(D=d)}{\hat{\pi}(d,\boldsymbol{X},G)}}\right)^2.$$

The unweighted ASD is defined by replacing $\hat{\pi}(1,\boldsymbol{X},G)$ and $\hat{\pi}(0,\boldsymbol{X},G)$ with 1 (see Imbens and Rubin, 2015, p.310-311). Note that the way s_d^2 is computed treats $\frac{\mathbb{1}(D=d)}{\hat{\pi}(d,\boldsymbol{X},G)}$ as a "reliability weight" and corrects the weighted sample variance accordingly.
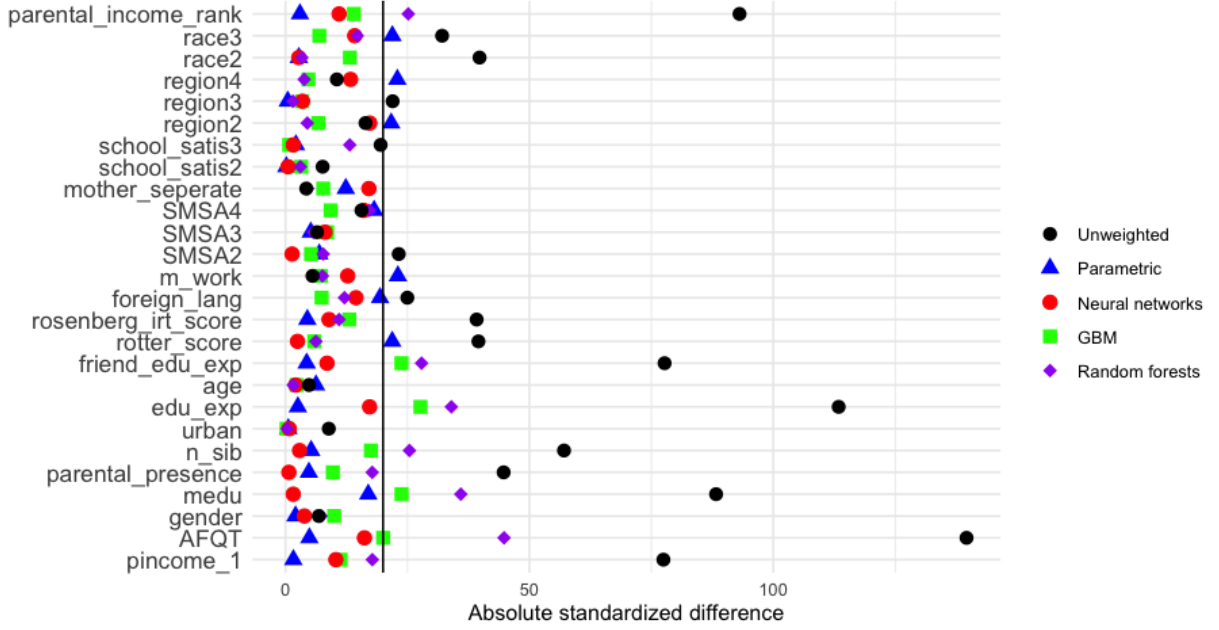
Figure A4. Covariate balance before and after weighting by covariate and propensity score model. The vertical line indicates the rule-of-thumb threshold of 20 (Cannas and Arpino, 2019).

We present the ASDs in Figure A4. Two observations are noteworthy. First, before weighting, several covariates are very imbalanced across treatment and control groups. After weighting, extreme imbalances almost all disappeared, regardless of the model. Using the rule-of-thumb threshold of 20 (Cannas and Arpino, 2019), all models achieve good enough balance for most covariates. In particular, neural networks keep the ASDs for all covariates under 20. On the other hand, random forests leave a few covariates quite unbalanced after weighting. It is also important to recall that our estimators are multiply robust, such that even if the propensity score function is not consistently estimated, the decomposition components may still be consistently estimated.

H.3. Outcome model fit

We also perform a diagnostic check for the outcome regression models. The outcome regression function is $\mu(d, \boldsymbol{X}, g) := \mathrm{E}(Y \mid d, \boldsymbol{X}, g)$, and $\hat{\mu}(d, \boldsymbol{X}, g)$ is its estimated counterpart. It is another nuisance function shared by both the unconditional and conditional decompositions.

Here, we measure model fit using the mean squared error (MSE): $\frac{1}{n} \sum [Y - \hat{\mu}(D, \boldsymbol{X}, G)]^2$. We compare the ML models with the intercept-only model and the parametric model used in the empirical application in terms of the MSE as a sanity check. The ML models should have lower MSEs than the intercept-only model and the parametric model. Otherwise, hyper-parameters may not have been sufficiently tuned. To make comparison across models fair, we use cross-fitting for all models, as opposed to only the ML models like we did in the main text. In other words, the MSEs are all calculated out of sample.

Table A6. Mean squared errors of outcome models

Model	MSE
Intercept model	0.0680
Parametric model	0.0484
Neural networks	0.0480
Generalized boosted machines	0.0474
Random forests	0.0480

Note: To ensure fair comparison, cross-fitting is used for all models.
As specified in Section 4.2 of the main text, the parametric model contains all two-way interactions between D and $\{\boldsymbol{X}, G\}$ as well as the main effects.

Results are shown in Table A6. All models used in the empirical application obtain much lower MSEs than the intercept-only model. Furthermore, all ML models have lower MSEs than the parametric model. These suggest that the hyper-parameters of the ML models are satisfactorily tuned via cross-validation. Again, due to the multiple robustness of our estimators, consistent estimation of the outcome regression function is not necessary for consistent estimation of the decomposition components.

H.4. Sensitivity analysis

We conduct a sensitivity analysis for our decompositions using results in Opacic et al. (2023), who provide bias formulas for a class of estimands, including $\mathrm{E}_g\left(Y^{R(D|g')}\right)$ and $\mathrm{E}_g\left(Y^{R(D|g', \boldsymbol{Q})}\right)$.[23] They do so in a setting where there is a binary unobserved confounder $U \in \{0, 1\}$, and they make two simplifying assumptions. In particular,

Sensitivity assumption 1. $Y^d \perp\!\!\!\perp D \mid \boldsymbol{x}, g, u, \forall d, \boldsymbol{x}, g, u.$

---

[23]There is an emerging literature on sensitivity analysis for random equalization decompositions (Park et al., 2023, 2024; Shen et al., 2024).

Sensitivity assumption 2. $\rho_{1g} := \mathrm{E}(U \mid D = 1, \boldsymbol{x}, g) - \mathrm{E}(U \mid D = 0, \boldsymbol{x}, g)$ does not depend on $\boldsymbol{x}$.

Sensitivity assumption 3. $\rho_{2g} := \mathrm{E}(Y \mid d, \boldsymbol{x}, g, U = 1) - \mathrm{E}(Y \mid d, \boldsymbol{x}, g, U = 0)$ does not depend on d or $\boldsymbol{x}$.

Note that ρ_{1g} is the conditional association between U and the treatment, and ρ_{2g} is the conditional association between U and the outcome. The bias is defined as the identified functional derived under Assumption 3 in the main text minus the one derived under Sensitivity assumption 1. Based on the results of Opacic et al. (2023), it is easy to show that, under Sensitivity Assumptions 1, 2, and 3,

$$\mathrm{bias}\left[\mathrm{E}_g\left(Y^{R(D|g')}\right)\right] = \rho_{1g}\rho_{2g}[\mathrm{E}(D \mid g') - \mathrm{E}(D \mid g)]$$
$$\mathrm{bias}\left[\mathrm{E}_g\left(Y^{R(D|g',\boldsymbol{Q})}\right)\right] = \rho_{1g}\rho_{2g}\{\mathrm{E}[\mathrm{E}(D \mid \boldsymbol{Q}, g') \mid g] - \mathrm{E}(D \mid g)\}.$$

In our decompositions, there are five components that are estimated to be significantly different from zero by all ML models: baseline, prevalence, selection, conditional prevalence, and $\boldsymbol{Q}$-distribution. Baseline and $\boldsymbol{Q}$-distribution components are not part of the unconditional and conditional total contributions of the treatment, and their biases cannot be directly evaluated using the bias formulas above. For the selection component, under Sensitivity Assumptions 1, 2, and 3, the bias is necessarily 0.[24] Therefore, we focus on the possible biases of the prevalence and conditional prevalence components. In particular, using results from Sections 2.2 and 2.4.1 in the main text, we have

$$\mathrm{bias}(\mathrm{prevalence}) = \rho_{1b}\rho_{2b}[\mathrm{E}(D \mid a) - \mathrm{E}(D \mid b)]$$
$$\mathrm{bias}(\mathrm{conditional\ prevalence}) = \rho_{1b}\rho_{2b}\{\mathrm{E}[\mathrm{E}(D \mid \boldsymbol{Q}, a) \mid b] - \mathrm{E}(D \mid b)\}.$$

Therefore, in order to conduct sensitivity analysis for the prevalence and conditional prevalence components, we need to estimate $\mathrm{E}[\mathrm{E}(D \mid \boldsymbol{Q}, a) \mid b]$ and specify $\{\rho_{1b}, \rho_{2b}\}$. For the conditional prevalence estimate based on each nuisance function model, we use the same nuisance function model to estimate $\mathrm{E}[\mathrm{E}(D \mid \boldsymbol{Q}, a) \mid b]$. For ρ_{1b} and ρ_{2b}, we benchmark their values using the observed confounder, educational expectation (binary indicator for "bachelor degree or higher"), which we denote as $\tilde{U}$. In Figure A4, this is the binary baseline covariate that has the strongest imbalance in the unweighted data. Hence, it is likely an exceptionally strong

---

[24]The bias for the conditional selection component is also constrained to be 0 under these assumptions. Echoing Theorems 5 and 6 in the main text, the conditional selection component is thus more robust to inconsistencies and lower convergence rates in nuisance function estimation as well as to unobserved confounding relative to other components of the conditional contribution of the treatment.

confounder. We estimate what ρ_{1b} and ρ_{2b} would be had educational expectation been U by fitting two models in group b with only main effects:

$$\mathrm{E}(\tilde{U} \mid D, \boldsymbol{X} \setminus \tilde{U}, b) = \kappa_{10} + \kappa_{11}D + \boldsymbol{\kappa}_{12}^{\mathsf{T}}\{\boldsymbol{X} \setminus \tilde{U}\}$$

$$\mathrm{E}(Y \mid D, \boldsymbol{X} \setminus \tilde{U}, \tilde{U}, b) = \kappa_{20} + \kappa_{21}D + \kappa_{22}^{\mathsf{T}}\{\boldsymbol{X} \setminus \tilde{U}\} + \kappa_{23}\tilde{U}.$$

Then, κ_{11} and κ_{23} respectively serves as benchmarks for ρ_{1b} and ρ_{2b}. These benchmarks suggest the magnitude of the biases if the unobserved confounder confounds the prevalence and conditional prevalence components as much as the observed educational expectation does. In our application, κ_{11} and κ_{23} are estimated to be 0.27 and 0.05, respectively.
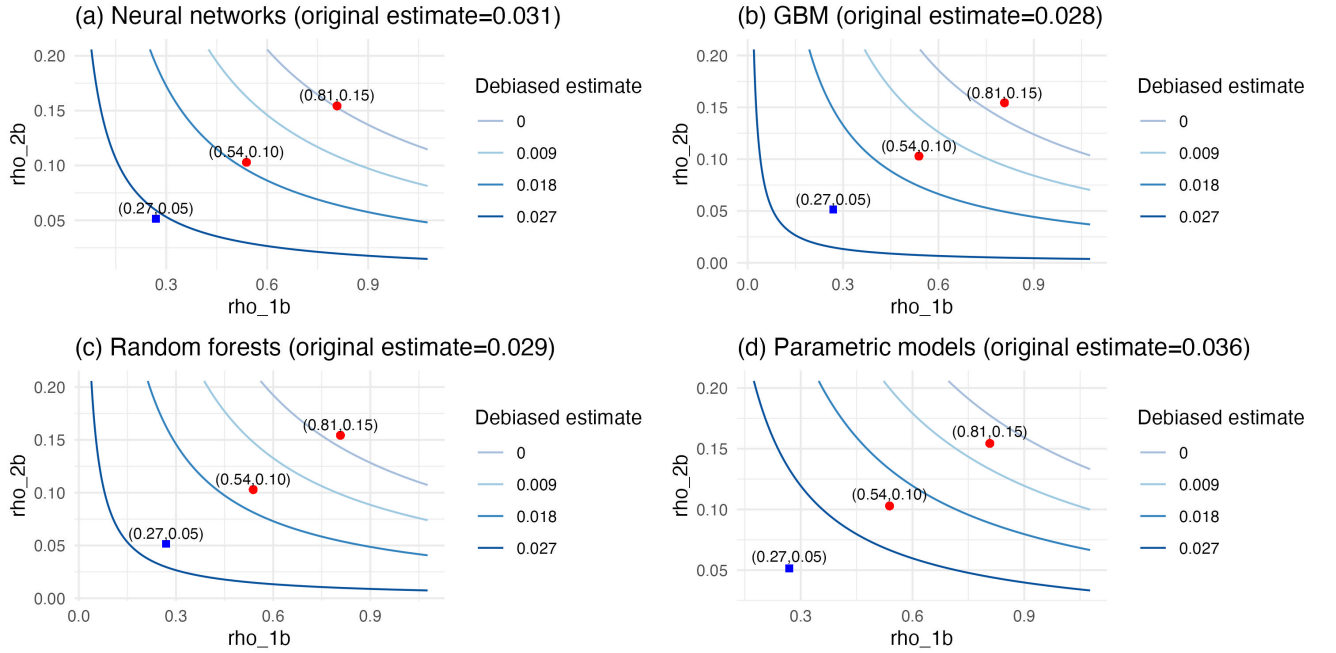
Figure A5. Contour plot for debiased estimates of the prevalence component. The blue (square) dot indicates the benchmark values of ρ_{1b} and ρ_{2b} based on treating educational expectation as an unobserved confounder. The two red (circle) dots show twice and thrice the benchmark values, respectively.

In Figures A5 and A6, we show debiased estimates of the prevalence and the conditional prevalence components, respectively. The debiased estimates are obtained by subtracting the biases under certain ρ_{1b} and ρ_{2b} values from the original estimates. For both unconditional and conditional prevalence components and across nuisance function models, a confounder whose conditional associations with the treatment and the outcome are twice as large as those of educational expectation would not bring the estimates to zero. In fact, to reduce the original estimates to zero, there must be a confounder that is at least around three times as strong as educational expectation. Based on these results, our empirical estimates appear reasonably
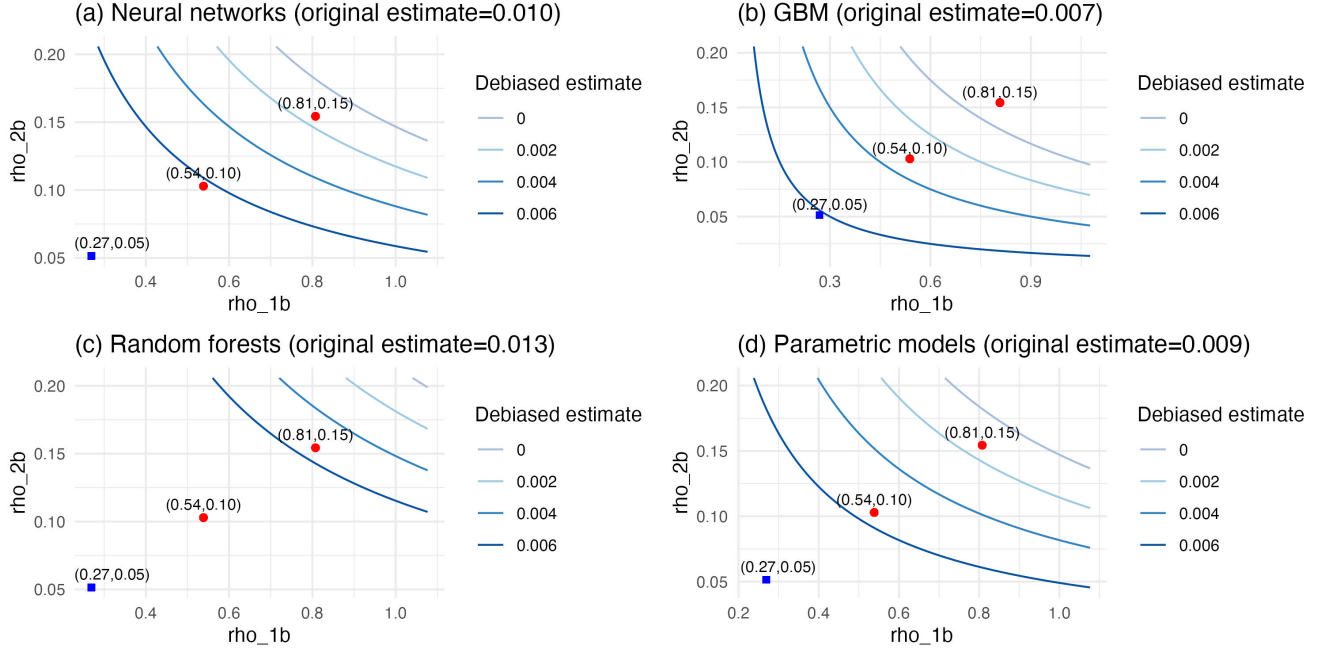
Figure A6. Contour plot for debiased estimates of the conditional prevalence component. The blue (square) dot indicates the benchmark values of ρ_{1b} and ρ_{2b} based on treating educational expectation as an unobserved confounder. The two red (circle) dots show twice and thrice the benchmark values, respectively.

robust to unobserved confounding. We leave the development of sensitivity analysis for other components of our decompositions to future work.

References

Austin, Peter C. and Elizabeth A. Stuart. 2015. "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies." *Statistics in Medicine* 34:3661–3679.

Avin, Chen, Ilya Shpitser, and Judea Pearl. 2005. "Identifiability of Path-Specific Effects." In *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 357–363, Edinburgh, Schotland.

Bailey, Martha and Susan Dynarski. 2011. "Gains and Gaps: Changing Inequality in U.S. College Entry and Completion." Technical Report w17633, National Bureau of Economic Research, Cambridge, MA.

Bickel, Peter J, Chris A.J Klaassen, Ya'acov Ritov, and Jon A Wellner. 1998. *Efficient and Adaptive Estimation for Semiparametric Models*. New York, NY: Springer.

Blau, Peter Michael and Otis Dudley Duncan. 1967. *The American Occupational Structure*. Free Press.

Blinder, Alan S. 1973. "Wage Discrimination: Reduced Form and Structural Estimates." *The Journal of Human Resources* 8:436.

Bloome, Deirdre, Shauna Dyer, and Xiang Zhou. 2018. "Educational Inequality, Educational Expansion, and Intergenerational Income Persistence in the United States." *American Sociological Review* 83:1215–1253.

Brady, David, Ryan M. Finnigan, and Sabine Hübgen. 2017. "Rethinking the Risks of Poverty: A Framework for Analyzing Prevalences and Penalties." *American Journal of Sociology* 123:740–786.

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

Brand, Jennie E., Jiahui Xu, Bernard Koch, and Pablo Geraldo. 2021. "Uncovering Sociological Effect Heterogeneity Using Tree-Based Machine Learning." *Sociological Methodology* 51:189–223.

Cannas, Massimo and Bruno Arpino. 2019. "A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting." *Biometrical Journal* 61:1049–1072.

Cha, Youngjoo, Kim A. Weeden, and Landon Schnabel. 2023. "Is the Gender Wage Gap Really a Family Wage Gap in Disguise?" *American Sociological Review* 88:972–1001.

Cheng, Siwei, Jennie E. Brand, Xiang Zhou, Yu Xie, and Michael Hout. 2021. "Heterogeneous returns to college over the life course." *Science Advances* 7:eabg7641.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. "Double/Debiased/Neyman Machine Learning of Treatment Effects." *American Economic Review* 107:261–265.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018a. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21:C1–C68.

Chernozhukov, Victor, Iván Fernández-Val, and Ye Luo. 2018b. "The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages." *Econometrica* 86:1911–1938.

Didelez, Vanessa, A Philip Dawid, and Sara Geneletti. 2006. "Direct and Indirect Effects of Sequential Treatments." In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, edited by R Dechter and T. S Richardson, pp. 138–146, Arlington, Virginia. AUAI Press.

Diderichsen, Finn, Johan Hallqvist, and Margaret Whitehead. 2019. "Differential vulnerability and susceptibility: how to make use of recent development in our understanding of mediation and interaction to tackle health inequalities." *International Journal of Epidemiology* 48:268–274.

Díaz, I, N S Hejazi, K E Rudolph, and M J Van Der Laan. 2021. "Nonparametric efficient causal mediation with intermediate confounders." *Biometrika* 108:627–641.

Featherman, David L. and Robert M. Hauser. 1978. *Opportunity and Change*. Studies in population. Academic Press.

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

Fisher, Aaron and Edward H. Kennedy. 2021. "Visually Communicating and Teaching Intuition for Influence Functions." *The American Statistician* 75:162–172.

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. 2011. "Decomposition Methods in Economics." In *Handbook of Labor Economics*, volume 4, pp. 1–102. Elsevier.

Geneletti, Sara. 2007. "Identifying Direct and Indirect Effects in a Non-Counterfactual Framework." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 69:199–215.

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66:315.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66:1017.

Heckman, James J, John Eric Humphries, and Gregory Veramendi. 2018. "Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking." *Journal of Political Economy* 126:50.

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

Heckman, James J. and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73:669–738.

Hines, Oliver, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. 2022. "Demystifying Statistical Learning Based on Efficient Influence Functions." *The American Statistician* 76:292–304.

Hirano, Keisuke, Guido W Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71:1161–1189.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.

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

Howe, Chanelle J., Sonia Napravnik, Stephen R. Cole, Jay S. Kaufman, Adaora A. Adimora, Beth Elston, Joseph J. Eron, and Michael J. Mugavero. 2014. "African American Race and HIV Virological Suppression: Beyond Disparities in Clinic Attendance." *American Journal of Epidemiology* 179:1484–1492.

Huber, Martin. 2015. "Causal Pitfalls in the Decomposition of Wage Gaps." *Journal of Business & Economic Statistics* 33:179–191.

Ichimura, Hidehiko and Whitney K. Newey. 2022. "The influence function of semiparametric estimators." *Quantitative Economics* 13:29–61.

Imai, Kosuke and Michael Lingzhi Li. 2023. "Experimental Evaluation of Individualized Treatment Rules." *Journal of the American Statistical Association* 118:242–256.

Imbens, Guido W and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Ishida, Hiroshi, Walter Muller, and John M. Ridge. 1995. "Class Origin, Class Destination, and Education: A Cross-National Study of Ten Industrial Nations." *American Journal of Sociology* 101:145–193.

Jackson, John W. 2021. "Meaningful Causal Decompositions in Health Equity Research: Definition, Identification, and Estimation Through a Weighting Framework." *Epidemiology* 32:282–290.

Jackson, John W. and Tyler J. VanderWeele. 2018. "Decomposition Analysis to Identify Intervention Targets for Reducing Disparities:." *Epidemiology* 29:825–835.

Karlson, Kristian Bernt. 2019. "College as equalizer? Testing the selectivity hypothesis." *Social Science Research* 80:216–229.

Kennedy, Edward H. 2022. "Semiparametric doubly robust targeted double machine learning: a review." Number: arXiv:2203.06469 [stat].

Kitagawa, Evelyn M. 1955. "Components of a Difference Between Two Rates." *Journal of the American Statistical Association* 50:1168–1194.

Kline, Patrick. 2011. "Oaxaca-Blinder as a Reweighting Estimator." *American Economic Review* 101:532–537.

Li, Fan, Kari Lock Morgan, and Alan M. Zaslavsky. 2018. "Balancing Covariates via Propensity Score Weighting." *Journal of the American Statistical Association* 113:390–400.

Li, Fan, Alan M. Zaslavsky, and Mary Beth Landrum. 2013. "Propensity score weighting with multilevel data." *Statistics in Medicine* 32:3373–3387.

Lundberg, Ian. 2024. "The Gap-Closing Estimand: A Causal Approach to Study Interventions That Close Disparities Across Social Categories." *Sociological Methods & Research* 53:507–570.

Lundberg, Ian, Rebecca Johnson, and Brandon M Stewart. 2021. "What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review* 86:532–565.

Malinsky, Daniel. 2018. "Intervening on structure." *Synthese* 195:2295–2312.

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

Morgan, Stephen L. and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Methods And Principles For Social Research*. Analytical Methods for Social Research. Cambridge, UK: Cambridge University Press, 2nd edition edition.

Naimi, Ashley I., Mireille E. Schnitzer, Erica E. M. Moodie, and Lisa M. Bodnar. 2016. "Mediation Analysis for Health Disparities Research." *American Journal of Epidemiology* 184:315–324.

Oaxaca, Ronald. 1973. "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review* 14:693.

Opacic, Aleksei, Lai Wei, and Xiang Zhou. 2023. "Disparity Analysis: A Tale of Two Approaches." *Unpublished manuscript* .

Park, Chan and Hyunseung Kang. 2024. "A groupwise approach for inferring heterogeneous treatment effects in causal inference." *Journal of the Royal Statistical Society Series A: Statistics in Society* 187:374–392.

Park, Soojin, Suyeon Kang, Chioun Lee, and Shujie Ma. 2023. "Sensitivity analysis for causal decomposition analysis: Assessing robustness toward omitted variable bias." *Journal of Causal Inference* 11:20220031.

Park, Soojin, Xu Qin, and Chioun Lee. 2024. "Estimation and sensitivity analysis for causal decomposition in health disparity research." *Sociological Methods & Research* 53:571–602.

Pearl, Judea. 2001. "Direct and Indirect Effects." In *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intel ligence*, pp. 411–20, San Francisco, CA. Morgan Kaufmann.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American statistical Association* 89:846–866.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66:688–701.

Rubin, Donald B. 1980. "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment." *Journal of the American Statistical Association* 75:591–593.

Rubin, Donald B. 1986. "Comment: Which Ifs Have Causal Answers." *Journal of the American Statistical Association* 81:961–962.

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

Shen, Andy, Elina Visoki, Ran Barzilay, and Samuel D. Pimentel. 2024. "A Calibrated Sensitivity Analysis for Weighted Causal Decompositions."

Sudharsanan, Nikkil and Maarten J Bijlsma. 2021. "Educational note: causal decomposition of population health differences using Monte Carlo integration and the g-formula." *International Journal of Epidemiology* 50:2098–2107.

van der Laan, Mark J. 2006. "Statistical Inference for Variable Importance." *The International Journal of Biostatistics* 2.

Van der Laan, Mark J and Sherri Rose. 2011. *Targeted learning: Causal Inference for Observational and Experimental Data*. Springer Series in Statistics. New York: Springer.

Van der Vaart, Aad W. 2000. *Asymptotic Statistics*. Cambridge Series on Statistical and Probablistic Mathematics. Cambridge University Press.

VanderWeele, Tyler. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

VanderWeele, Tyler J. 2014. "A Unification of Mediation and Interaction: A 4-Way Decomposition." *Epidemiology* 25:749–761.

VanderWeele, Tyler J. and Whitney R. Robinson. 2014. "On the Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables." *Epidemiology* 25:473–484.

VanderWeele, Tyler J. and Eric J. Tchetgen Tchetgen. 2014. "Attributing Effects to Interactions." *Epidemiology* 25:711–722.

Vansteelandt, Stijn and Oliver Dukes. 2022. "Assumption-lean Inference for Generalised Linear Model Parameters." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84:657–685.

Ward, Julia B., Danielle R. Gartner, Katherine M. Keyes, Mike D. Fliss, Elizabeth S. McClure, and Whitney R. Robinson. 2019. "How do we assess a racial disparity in health? Distribution, interaction, and interpretation in epidemiological studies." *Annals of Epidemiology* 29:1–7.

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

Yamaguchi, Kazuo. 2015. "Decomposition of Gender or Racial Inequality with Endogenous Intervening Covariates: An Extension of the DiNardo-Fortin-Lemieux Method." *Sociological Methodology* 45:388–428.

Yu, Ang. 2024. *cdgd: Causal Decomposition of Group Disparities*. R package version 0.3.5.

Yu, Ang, Li Ge, and Felix Elwert. 2024. "When Do Natural Mediation Effects Differ from Their Randomized Interventional Analogues: Test and Theory." Number: arXiv:2407.02671 [stat].

Yu, Ang, Chan Park, Hyunseng Kang, and Jason Fletcher. 2021. "Leveraging Machine Learning to Estimate Effect Modification." preprint, SocArXiv.

Yu, Ang and Jiwei Zhao. 2024. "Counterfactual Slopes and Their Applications in Social Stratification." Number: arXiv:2401.07000 [stat].

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

Zhou, Xiang. 2024. "Attendance, Completion, and Heterogeneous Returns to College: A Causal Mediation Approach." *Sociological Methods & Research* 53:1136–1166.

Ziol-Guest, Kathleen M. and Kenneth T. H. Lee. 2016. "Parent Income–Based Gaps in Schooling: Cross-Cohort Trends in the NLSYs and the PSID." *AERA Open* 2:233285841664583.