Nonparametric Causal Decompositions of Group Disparities*

Ang Yu† Felix Elwert‡

April 26, 2023

Abstract

We propose a nonparametric method for decomposing group disparities in terms of an intermediate treatment variable. Our decomposition contains four components capturing the contributions of group differences in baseline characteristics, treatment prevalence, average treatment effect, and selection into treatment. These components readily inform policy interventions on the treatment and its effect. Our main contribution is to neatly separate the roles of between-group and within-group effect heterogeneity in explaining group disparities, revealing a new lever for reducing disparities. This decomposition reformulates the classic Kitagawa-Blinder-Oaxaca decomposition in causal and model-free terms, supplements causal mediation analysis by targeting group disparities instead of group effects, and resolves conceptual difficulties of recent random equalization decompositions. We also provide a conditional decomposition that allows researchers to incorporates pre-treatment covariates in defining the estimands and the hypothetical interventions. Using efficient influence functions and cross-fitting, we propose nonparametric estimators that are $\sqrt{n}$-consistent, asymptotically normal, semiparametrically efficient, and doubly robust. We apply our framework to study the causal role of education in intergenerational income mobility.

*We thank Paul Bauer, Eric Grodsky, Aleksei Opacic, Guanghui Pan, Chan Park, Ben Rosche, Jiwei Zhao, and especially Xiang Zhou for helpful comments and suggestions. Earlier versions of this paper has been presented at Causality in the Social Sciences Workshop at GESIS in 2021, Population Association of America annual meeting in 2022, and American Causal Inference Conference in 2022. We thank participants at these conferences for constructive discussions.

†Department of Sociology, University of Wisconsin-Madison. Email: ayu33@wisc.edu

‡Department of Sociology and Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison

1 Introduction

Improving the explanations of inter-group disparities is an urgent task for the social and health sciences. For example, what is the contribution of educational attainment to racial disparities in adult health? How much of the gender wage gap can be explained by occupational sorting? And what is the role of college completion in the relationship between parental income and income during adulthood? The structure of such question is always the same. Researchers compare the outcomes of two groups whose membership is typically established early in life (e.g., race, gender, parental background, birth cohort, country) and ask how differences in outcomes can be attributed to some intermediate treatment variable that occurs between the establishment of group membership and the outcome.

Questions about explanations, contributions, and attribution are inherently causal. First, researchers often are interested in the *causes* of disparities and ask counterfactual questions about the impact of changing one or more intermediate treatment variables, which is necessary for informing real-world interventions that could alleviate the disparities. It is one thing to observe that racial differences in college completion coexist with wage gaps by college completion status; it is another to establish that differences in college completion are responsible for racial wages gaps, such that intervening to equalize college graduation rates would remediate wage inequality. Second, researchers often also posit counterfactual scenarios where the effects of a treatment variable in a group were set to different values than they are in reality. At a minimum, the discussion of effects also presumes causal identification.

Three prior approaches are concerned with questions of attribution and explanation. First, the Kitagawa-Blinder-Oaxaca (KBO) decomposition (Kitagawa, 1955; Blinder, 1973; Oaxaca, 1973) is a classic tool used for explaining group disparities and remains popular to this day (e.g., Barber et al., 2015; Mize, 2016; Laurison and Friedman, 2016; Storer et al., 2020; Zajacova et al., 2021). The KBO decomposition is based on group-specific linear regression models and decomposes the group disparity into an explained component and an unexplained component. The explained component reflects the differential averages of an intermediate variable by group, and the unexplained component captures differential slopes and intercepts. The KBO decomposition is thus defined in terms of regression coefficients and not nonparametrically formulated in causal terms. This renders the KBO decomposition inappropriate for explanation and attribution questions. In fact, sociologists and economists alike have criticized the KBO decomposition for failing to answer any counterfactual question by design (Fortin et al., 2011,

p.13; Lundberg et al., 2021, p.542).

Second, causal mediation analysis is also used to attribute the relationship between two variables to an intermediate variable (VanderWeele, 2015). Causal mediation estimands are clearly defined in counterfactual terms, however, they are not designed for explaining group disparities. This is because what they decompose are summary measures of treatment effects, such as the average treatment effect (ATE). However, the discussion of treatment effects is conceptually challenging when it comes to ascriptive variables such as race and gender, as it is not plausible to impose any intervention on these variables (Rubin, 1974). Some group variables, such as early-life socioeconomic status, can theoretically be intervened upon, but not for people who are already later in their life course. Indeed, in numerous applications in social and health sciences, descriptive group disparities, rather than average effects, is what researchers seek to understand and explain. Causal attribution of group disparities thereby requires analytical frameworks more tailored to it than mediation analysis.

Third, there have recently emerged variants of a causal decomposition of group disparities, which we call the random equalization decomposition (VanderWeele and Robinson, 2014; Jackson and VanderWeele, 2018; Lundberg, 2022). For this decomposition, the extent to which the group disparity can be attributed to the differential prevalence of an intermediate treatment variable is defined as the reduction in disparity brought about by a hypothetical intervention that randomly reassigns values of the treatment such that its prevalence is equalized across groups. The random equalization decomposition marks a important advancement towards establishing an appropriate framework for causal attribution of group disparities to an intermediate treatment variable. However, as we will show, the random equalization decomposition does not purely capture the contribution of differential prevalence of the treatment. Instead, it also incorporates the influence of differential patterns of selection into treatment across groups, which complicates its interpretation.

In this article, we develop a novel decomposition approach for group disparities in terms of an intermediate treatment variable. Our decompositions are defined counterfactually with respect to the treatment variable and descriptively with respect to the group variable. This approach is consistent with common practices in the social sciences and focuses on more modifiable factors in terms of policy intervention. Our unconditional decomposition contains four components, respectively capturing the contributions of baseline differences unrelated to the treatment, differential treatment prevalence, differential ATE, and differential selection into

the treatment. Compared with the KBO decomposition, our decomposition enables causal attribution and nonparametric estimation. It supplements causal mediation analysis by focusing on explaining group disparities instead of "effects" of the group variables. It also improves on the random equalization decomposition by providing components with unambiguous interpretations. The component in our decomposition that captures differential selection is a centerpiece in our contributions. First, this component reveals a novel policy lever for reducing disparities, and it did not appear in any of the previous decompositions. Second, by separating it from the component representing differential prevalence, we clarify the random equalization decomposition and provide a more clearly interpretable decomposition.

Our paper proceeds as follows. Section 2 introduces our unconditional decomposition and its interventional interpretations. We also formally relate our decomposition to the classic KBO decomposition, a causal mediation estimand, and the random equalization decomposition, through which we explicate the contributions of our framework. We also extend the unconditional decomposition to a conditional version, which allows researchers to take pre-treatment covariates into account in defining the estimands and the hypothetical interventions. In section 3, under the conditional ignorability assumption, we develop nonparametric estimators for the unconditional and conditional decompositions based on their efficient influence functions. These estimators can be implemented via data-adaptive methods such as machine learning (ML) models and accommodate high-dimensional confounders. We also derive the conditions under which the estimators are $\sqrt{n}$-consistent, asymptotically normal, and semi-parametrically efficient. In addition, they have double-robustness properties. In section 4, we apply our decompositions to study the causal role of college graduation in intergenerational income mobility. Section 5 concludes with possible future extensions.

2 Estimands

2.1 Unconditional decomposition

For each individual i, consider a binary treatment variable $D_i \in \{0, 1\}$. Let Y_{i0} and Y_{i1} be the potential outcomes (Rubin, 1974) of Y_i under the hypothetical intervention to set $D_i = 0$ and $D_i = 1$, respectively. Let $\tau_i := Y_{i1} - Y_{i0}$ denote the individual-level treatment effect. Henceforth, we largely suppress individual level subscripts on all variables to unburden notation. Suppose that the population contains two disjoint groups, $g \in \{a, b\}$, where a denotes the advantaged

group and b denotes the disadvantaged group. We use subscripts g to indicate group-specific quantities, for example, $\mathrm{E}_g(Y) \coloneqq \mathrm{E}(Y \mid G = g)$.

We now only assume the stable unit treatment value assumption (SUTVA) (Rubin, 1980),

Assumption 1 (SUTVA). $Y = DY_1 + (1 - D)Y_0$.

Then, the observed outcomes disparity between group a and b can be decomposed into four components:

$$\begin{aligned}
&\mathrm{E}_a(Y) - \mathrm{E}_b(Y) \\
&= \underbrace{\mathrm{E}_a(Y_0) - \mathrm{E}_b(Y_0)}_{\text{baseline}} + \underbrace{\mathrm{E}_b(\tau)[\mathrm{E}_a(D) - \mathrm{E}_b(D)]}_{\text{prevalence}} \\
&\quad + \underbrace{\mathrm{E}_a(D)[\mathrm{E}_a(\tau) - \mathrm{E}_b(\tau)]}_{\text{effect}} + \underbrace{\mathrm{Cov}_a(D, \tau) - \mathrm{Cov}_b(D, \tau)}_{\text{selection}}. \qquad (1)
\end{aligned}$$

The "baseline" component reflects the difference in average baseline potential outcome Y_0 between groups, i.e., the difference in outcomes if nobody received treatment. The "prevalence" component indicates how much of the group disparity is due to differential prevalence of treatment. The "effect" component reflects unequal average treatment effects (ATE) across groups. Finally, the "selection" component captures the contribution of the differential selection into treatment based on the treatment effect, τ, to the group disparity. For each group, if group members who would gain more from treatment are more likely to receive the treatment than those who would gain less, there will be a positive covariance between D and τ. To our knowledge, none of prior decompositions of group disparities includes a selection component.[1] Both the effect and the selection components account for the relationship between heterogeneous effects and group disparities. But the "effect" component captures the contribution of *between*-group effect heterogeneity; and the "selection" component captures the contribution of *within*-group effect heterogeneity. In summary, a group will be more advantaged in the outcome if its members have a higher average baseline potential outcome, a higher prevalence of treatment (given a positive ATE), a higher ATE, or a higher level of selection into treatment based on treatment effect.

[1] Zhou (2022) recently developed a causal decomposition method for a special case of mediation analysis, which contains a similar covariance component capturing selection into a mediator.

2.2 Interventional interpretations

Our decomposition is formulated in counterfactual terms, hence is prescriptive for future interventions. The decomposition reveals three levers by which a policy may affect inter-group disparities through the treatment. First, policy makers could manipulate the prevalence of the treatment in each of the two groups. Second, they could manipulate who within each group receives treatment. For example, they could facilitate the matching of high-return individuals to treatment receipt in the disadvantaged group. Third, they might even be able to manipulate the average effect of the treatment for each group.[2] Specifically, each of our decomposition component quantifies the part of the outcome disparity that could be eliminated if the two groups were equalized with respect to that component while other characteristics were held unchanged.

To further fix ideas, we may more exactly pin down the equalization intervention implicated by each decomposition component. The prevalence component indicates the part of the disparity that would be eliminated by giving group b the treatment prevalence of group a, which can be seen by replacing $\mathrm{E}_b(D)$ with $\mathrm{E}_a(D)$ in equation (1). In the same vein, the effect component is the part of the disparity that would disappear if group a was given the ATE of group b. Finally, the disparity would be reduced by the amount of the selection component if the two groups were made to have the same selection level, regardless of the specific level at which their selection would be fixed. However, given that the prevalence, effect, and selection components are determined only by two variables, the treatment and the treatment effect, it is not always possible to intervene on a single component without changing the values of the other components. Nevertheless, we may alternatively conceive of a whole-package intervention. A whole-package intervention transfers the joint distribution of D and τ from one group to the other. Our prevalence, effect, and selection components then add to the reduction in disparity of this whole-package intervention, and each of them can be interpreted as a distinct aspect of the intervention.

Even if one is not willing to consider interventions on treatment effects, the decomposition components can still be interpreted in terms of a two-step intervention. Under the randomized intervention notation proposed by Didelez et al. (2006), $R(D \mid G = g')$ is defined to

[2] It is more unconventional to discuss interventions on treatment effects than on treatments. However, interventions on effects have appeared in both methodological (Malinsky, 2018) and empirical (Brady et al., 2017) literatures. Our notion of interventions on effects is a generalization of Malinsky's (2018) interventions on "structural features", as we allow for heterogeneous effects.

be a randomly drawn value of treatment from group g'. Then, $\mathrm{E}_g\left(Y_{R(D|G=g')}\right)$ denotes the mean outcome for group g after each member of group g is counterfactually given a treatment value randomly drawn from group g'. When $g = g'$, the intervention amounts to a random permutation of treatment status within the group. Using its definition, we can rewrite the post-intervention mean as follows.

$$\mathrm{E}_g\left(Y_{R(D|G=g')}\right) = \mathrm{E}_g(Y_0) + \mathrm{E}_{g'}(D)\,\mathrm{E}_g(\tau). \tag{2}$$

Then it follows that our decomposition components can be re-written in the randomized intervention notation.

$$\begin{aligned}
\mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \left[\mathrm{E}_a\left(Y_{R(D|G=a)}\right) - \mathrm{E}_b\left(Y_{R(D|G=b)}\right)\right] &= \text{selection} \\
\mathrm{E}_b\left(Y_{R(D|G=a)}\right) - \mathrm{E}_b\left(Y_{R(D|G=b)}\right) &= \text{prevalence} \\
\mathrm{E}_a\left(Y_{R(D|G=a)}\right) - \mathrm{E}_b\left(Y_{R(D|G=a)}\right) &= \text{baseline} + \text{effect}
\end{aligned}$$

This points to a two-step intervention. In the first step, the pre-intervention disparity is $\mathrm{E}_a(Y) - \mathrm{E}_b(Y)$, and the post-intervention disparity is $\mathrm{E}_a\left(Y_{R(D|G=a)}\right) - \mathrm{E}_b\left(Y_{R(D|G=b)}\right)$. Hence, this step internally randomizes the treatment in both groups without changing its prevalence, and the selection component represents the reduction in disparity resulting from this randomization intervention. In the second step, the pre-intervention disparity is $\mathrm{E}_a\left(Y_{R(D|G=a)}\right) - \mathrm{E}_b\left(Y_{R(D|G=b)}\right)$, while the post-intervention disparity is $\mathrm{E}_a\left(Y_{R(D|G=a)}\right) - \mathrm{E}_b\left(Y_{R(D|G=a)}\right)$. This step equalizes treatment prevalence by giving group b the treatment prevalence of group a. Therefore, the prevalence component is the reduction in disparity of an equalization intervention. Importantly, this equalization intervention is independent of the randomization intervention in the first step. Finally, the remaining disparity at the end of the two-step intervention is the sum of the baseline and the effect components.

2.3 Relation to the KBO decomposition

Our decomposition expands on the classic KBO decomposition in two ways. First, our decomposition is presented explicitly as a causal estimand, which can be estimated nonparametrically and interpreted interventionally. By contrast, the KBO decomposition is defined in terms of linear regressions and coincides with our decomposition only under a very strong assumption.

Second, our decomposition contains an additional selection component as a novel disparity-producing mechanism, which is assumed away in the KBO decomposition.

Consider a form of the KBO decomposition of the outcome disparity between groups a and b with respect to a treatment variable, D,

$$\begin{aligned} &\overline{Y}_a - \overline{Y}_b \\ &= \underbrace{\hat{\alpha}_a - \hat{\alpha}_b}_{\text{intercept}} + \underbrace{\hat{\beta}_b[\overline{D}_a - \overline{D}_b]}_{\text{characteristic}} + \underbrace{\overline{D}_a[\hat{\beta}_a - \hat{\beta}_b]}_{\text{slope}}, \end{aligned}$$

which is based on estimates of the group-specific linear regressions:

$$Y = \alpha_g + \beta_g D + \epsilon.$$

Under a strong ignorability assumption, $Y_d \perp\!\!\!\perp D \mid G = g, \forall d, g$, in addition to assumption 1, the intercept, characteristic, and slope components in the KBO decomposition are unbiased estimators for our baseline, prevalence, and effect components, respectively. However, the strong ignorability assumption requires random assignment of treatment conditional only on group membership. It also directly rules out the selection component in group disparities. Hence, our decomposition is both more expansive and interventionally interpretable under far less stringent assumptions. Although components of the KBO decomposition have commonly been given counterfactual and intervenional interpretations (e.g., Jann, 2008), our decomposition should be preferred for its more plausibly causal interpretability.

Before the current article, several studies have constructed or inferred causal estimands for the KBO decomposition. A methodological literature centers around interpreting the unexplained component (the sum of the intercept and the slope components) as an estimator for the treatment effect on the treated (ATT) (Fortin et al., 2011; Kline, 2011; Yamaguchi, 2015).[3] This literature addresses how an outcome difference associated with a treatment variable can be accounted for by *pre*-treatment confounders. By contrast, our decomposition focuses on a very different question, i.e., the extent to which an outcome difference between groups is explained by a *post*-group variable. Huber (2015) proposes an interpretation of the KBO components as causal mediation estimands, where the characteristic component is interpreted as the natural indirect effect (Pearl, 2001) of the group indicator via an intermediate variable

[3] Also see Chernozhukov et al. (2018b), which interprets a version of the unexplained component as an average partial effect.

and the unexplained component as the corresponding natural direct effect. The next section contrasts our decomposition with causal mediation analysis.

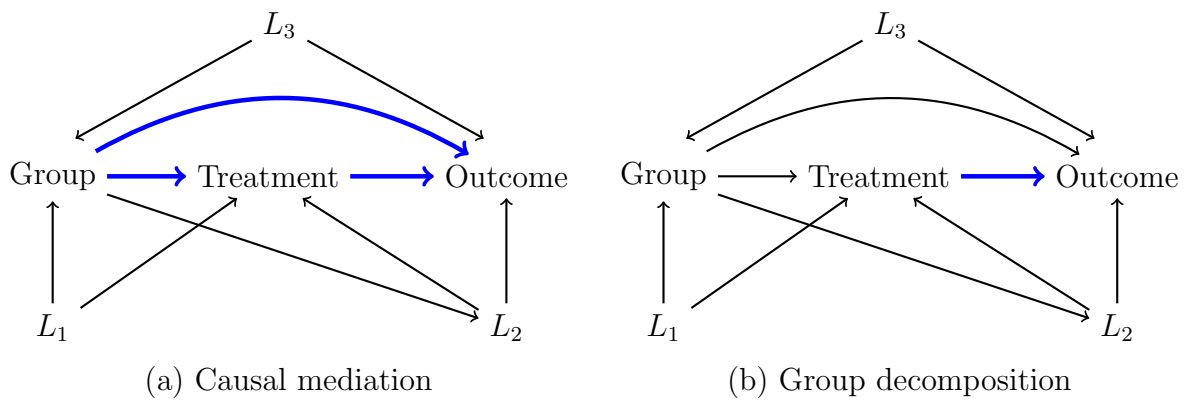
2.4 Relation to randomized interventional analogues of mediation estimands

Among the many causal mediation estimands, a four-way decomposition based on randomized interventional analogues (RIAs) of mediation estimands (VanderWeele, 2015, p.619-21; see also VanderWeele and Tchetgen Tchetgen, 2017; VanderWeele et al., 2014) is most akin to our decomposition. By showing how our decomposition is related to but differs from this decomposition, we highlight the unique advantages of our decomposition compared with all causal mediation estimands. A form of this decomposition can be written as

$$
\begin{aligned}
&\underbrace{\mathrm{E}\left(Y_{G=a,R(D_{G=a})}\right) - \mathrm{E}\left(Y_{G=b,R(D_{G=b})}\right)}_{\text{RIA of total effect}} \\
&= \underbrace{\mathrm{E}(Y_{G=a,D=0}) - \mathrm{E}(Y_{G=b,D=0})}_{\text{conditional direct effect (CDE)}} + \underbrace{\mathrm{E}\left(Y_{G=b,R(D_{G=a})}\right) - \mathrm{E}\left(Y_{G=b,R(D_{G=b})}\right)}_{\text{RIA of pure indirect effect (PIE)}} \\
&+ \underbrace{\mathrm{E}(Y_{G=a,D=1} - Y_{G=a,D=0} - Y_{G=b,D=1} + Y_{G=b,D=0})\,\mathrm{E}(D_{G=b})}_{\text{RIA of reference interaction effect (RIE)}} \\
&+ \underbrace{\mathrm{E}(Y_{G=a,D=1} - Y_{G=a,D=0} - Y_{G=b,D=1} + Y_{G=b,D=0})\,(\mathrm{E}(D_{G=a}) - \mathrm{E}(D_{G=b}))}_{\text{RIA of mediated interaction effect (MIE)}},
\end{aligned}
$$

where $\mathrm{E}\left(Y_{G=g,R(D_{G=g'})}\right)$ is the mean potential outcome of Y under assigning group g and a randomly drawn value of D from the population when assigned group g', and $\mathrm{E}(D_{G=g})$ is the the mean potential outcome of D under assigning group g.

We first note the connections between our decomposition and the RIA-based decomposition. Under unconditional ignorability of G, i.e., $Y_{G=g,D=d} \perp\!\!\!\perp G$, $\forall d, g$, and $D_{G=g} \perp\!\!\!\perp G$, $\forall g$, and two SUTVA-type assumptions, $\mathrm{E}(Y_{G=g,D=d} \mid G = g) = \mathrm{E}(Y_{D=d} \mid G = g)$ and $\mathrm{E}(D_{G=g} \mid G = g) = \mathrm{E}(D \mid G = g), \forall d, g$, the CDE equals our baseline component, the RIA of PIE equals our prevalence component, and the sum of RIAs of RIE and MIE equals our effect component. These connections are intuitive: both the CDE and the baseline component capture a group-based outcome difference when the intermediate treatment variable is held at 0; both the RIA of PIE and the prevalence component address the role of the treatment in the relationship between the group and the outcome; also, the RIA of RIE, the RIA of MIE, and the effect



(a) Causal mediation (b) Group decomposition

Figure 2. Causal graphs illustrating the difference in identification goals between causal mediation analysis and causal decomposition of group disparities. The highlighted edges are the paths each estimand seeks to identify. Causal mediation analysis aims to identify both the effects of group and treatment, hence requiring proper handling of all three types of confounders: L_1, L_2, and L_3. Our causal decomposition only needs to identify the effect of treatment, so only treatment-outcome confounders, i.e., L_2, are of concern.

component all, at least partly, reflect the heterogeneous effects of the treatment by group membership. Thus, as tools of causal explanation, our decomposition and causal mediation analysis are akin to each other.[4]

However, our decomposition and causal mediation analysis differ in two important ways, which makes our decomposition advantageous for studying group disparities. First, we decompose a descriptive group disparity rather than a total effect of group membership. This does not only cater to the goal of studying group disparities but also requires fewer identifying assumptions than causal mediation analysis. We illustrate the difference in identifying assumptions in Figure 1. Consequently, although the components of the RIA-based decomposition can be forcibly equated to the components of our decomposition, it would require the impossible scenario where the group membership is randomly assigned. Second, there is no decomposition in the causal mediation literature that contains a selection component. In this regard, our decomposition makes a novel contribution.

2.5 Relation to unconditional random equalization

In recent years, there have emerged two variants of what we call the unconditional random equalization decomposition, which decomposes the observed disparity into two components.

[4]These intuitions also carry over to VanderWeele's (2014) four-way decomposition that is not based on RIAs. However, in order to establish a similar connection between the conventional four-way decomposition and our decomposition, a cross-world independence assumption needs to be additionally imposed.

One is the reduction in disparity that could be brought about by randomly reassigning the treatment such that the treatment prevalence is equalized across groups. The other is correspondingly a component of remaining disparity (VanderWeele and Robinson, 2014; Jackson and VanderWeele, 2018; Sudharsanan and Bijlsma, 2021; Lundberg, 2022). Like our approach, the random equalization decomposition seeks to identify the causal effect of the treatment, not that of group membership. However, the random equalization decomposition is a two-way decomposition that contains less information than our four-way decomposition. Moreover, the reduction in disparity in the random equalization decomposition combines the selection component and the prevalence component of our decomposition, impeding a clear interpretation.

The first variant of the random equalization decomposition is defined in terms of a random equalization intervention that randomly assigns treatment values of the advantaged group to the disadvantaged group. Using the randomized intervention notation introduced above, Jackson and VanderWeele (2018) decompose the observed group disparity into two components:

$$\mathrm{E}_a(Y) - \mathrm{E}_b(Y) = \underbrace{\mathrm{E}_b\left(Y_{R(D|G=a)}\right) - \mathrm{E}_b(Y)}_{\text{reduction in disparity}} + \underbrace{\mathrm{E}_a(Y) - \mathrm{E}_b\left(Y_{R(D|G=a)}\right)}_{\text{remaining disparity}}.$$

The reduction in disparity is intended to quantify the contribution of the differential prevalence of treatment across groups. However, the random equalization intervention not only equalizes treatment prevalence across groups, but also randomizes the treatment within the disadvantaged group, making any selection into treatment disappear in that group. As a consequence, the reduction in disparity equals the combination of the prevalence component and the group b part of the selection component, i.e., $\mathrm{E}_b(\tau)[\mathrm{E}_a(D) - \mathrm{E}_b(D)] - \mathrm{Cov}_b(D, \tau)$, which follows from equation (2). This leads to an interpretational difficulty that is reflected in an undesirable result: If group a is an exact replicate of group b, the random equalization decomposition would still contain a non-zero reduction in disparity, as long as $\mathrm{Cov}_b(D, \tau) \neq 0$.

The second variant of the random equalization decomposition takes a somewhat different form (Lundberg, 2022), whose hypothetical intervention assigns each individual in both groups a treatment value randomly drawn from the pooled population. Hence, this random equalization intervention changes the treatment values in both groups instead of only in the disadvantaged group as in the first variant. The reduction in disparity of this variant mixes the prevalence component with the selection component, too. To show this, we first note that

the outcome disparity can also be decomposed as such:

$$
\begin{aligned}
&\mathrm{E}_a(Y) - \mathrm{E}_b(Y) \\
&= \mathrm{E}_a(Y_0) - \mathrm{E}_b(Y_0) + \mathrm{E}(\tau)[\mathrm{E}_a(D) - \mathrm{E}_b(D)] \\
&\quad + \mathrm{E}(D)[\mathrm{E}_a(\tau) - \mathrm{E}_b(\tau)] + \mathrm{Cov}_a(D,\tau) - \mathrm{Cov}_b(D,\tau) \\
&\quad - [P_a - P_b][\mathrm{E}_a(D) - \mathrm{E}_b(D)][\mathrm{E}_a(\tau) - \mathrm{E}_b(\tau)],
\end{aligned}
$$

where $\mathrm{E}(\tau)$ and $\mathrm{E}(D)$ are the overall ATE and treatment prevalence, P_a and P_b are the proportions of the population in group a and group b. And the remaining disparity in Lundberg's (2022) unconditional decomposition equals $\mathrm{E}_a(Y_0) - \mathrm{E}_b(Y_0) + \mathrm{E}(D)[\mathrm{E}_a(\tau) - \mathrm{E}_b(\tau)]$. It then follows that the reduction in disparity in this case contains the selection component, $\mathrm{Cov}_a(D,\tau) - \mathrm{Cov}_b(D,\tau)$, which does not involve the difference in treatment prevalence. The intuition is that in the random equalization intervention, both group a and group b would get random draws of D, thus selection into treatment is eliminated in both groups, making $\mathrm{Cov}_a(D,\tau) = \mathrm{Cov}_b(D,\tau) = 0$.

To conclude, the reduction in disparity in both variants of the random equalization decomposition is a mixture of the prevalent component and the selection component of group disparities. This is because its underlying intervention both equalizes and randomizes the treatment. In contrast, as explained in Section 2.1, the prevalence component in our decomposition is solely concerned with the disparity-reducing impact of an equalization intervention and avoids the influence of a randomization intervention.[5] If the goal is to capture the contribution of the unequal treatment prevalence to group disparities, one should prefer using the prevalence component in our decomposition.

2.6 Conditional decomposition

In this subsection, we develop a decomposition method whose corresponding interventions are conditional on a vector of variables $\boldsymbol{Q}$. The conditional decomposition is useful as it can shed light on the impact of normatively more desirable or realistically more feasible interventions, when the unconditional intervention appears less so (Jackson, 2021). For example, it may be

[5]This is also why the RIA of PIE can be translated to the prevalence component, but not to the reduction in disparity component in random equalization decompositions. Both the prevalence component and the RIA of PIE are a difference between two quantities that both contain randomized D values, so both estimands can represent the reduction in disparity resulting from a pure equalization intervention *after* a randomization intervention.

meaningful to consider equalizing college admissions conditional on test scores or equalizing drug prescriptions conditional on certain comorbilidities. Assuming that the the support of $\boldsymbol{Q}$ is the same in the two groups and still under Assumption 1,

$$\begin{aligned}
\mathrm{E}_a(Y) - \mathrm{E}_b(Y) = & \underbrace{\mathrm{E}_a(Y_0) - \mathrm{E}_b(Y_0)}_{\text{baseline}} \\
& + \underbrace{\int [\mathrm{E}_a(D \mid \boldsymbol{Q} = \boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{Q} = \boldsymbol{q})]\, \mathrm{E}_b(\tau \mid \boldsymbol{Q} = \boldsymbol{q}) f_b(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}}_{\text{conditional prevalence}} \\
& + \underbrace{\int [\mathrm{E}_a(\tau \mid \boldsymbol{Q} = \boldsymbol{q}) - \mathrm{E}_b(\tau \mid \boldsymbol{Q} = \boldsymbol{q})]\, \mathrm{E}_a(D \mid \boldsymbol{Q} = \boldsymbol{q}) f_a(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}}_{\text{conditional effect}} \\
& + \underbrace{\mathrm{E}_a[\mathrm{Cov}_a(D, \tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D, \tau \mid \boldsymbol{Q})]}_{\text{conditional selection}} \\
& + \underbrace{\int \mathrm{E}_a(D \mid \boldsymbol{Q} = \boldsymbol{q})\, \mathrm{E}_b(\tau \mid \boldsymbol{Q} = \boldsymbol{q}) [f_a(\boldsymbol{q}) - f_b(\boldsymbol{q})] \mathrm{d}\boldsymbol{q}}_{\boldsymbol{Q} \text{ distribution}} . \qquad (3)
\end{aligned}$$

The conditional prevalence, effect, and selection components indicate the disparity-reducing impacts of equalizing the corresponding group characteristics within levels of $\boldsymbol{Q}$. Thus, the conditional decomposition allows researchers to incorporate prior information about the subjects when designing a re-distribution of treatments or their effects. As policies, the conditional equalization interventions are often more feasible but less influential than the marginal equalizations encoded in equation (1). In addition, the "$\boldsymbol{Q}$ distribution" component captures the between-$\boldsymbol{Q}$ disparity, as opposed to within-$\boldsymbol{Q}$ disparities reflected in other components, but it should not be interpreted causally, as we do not make any identifying assumptions for $\boldsymbol{Q}$.

Similar to the unconditional case, we can also interventionally interpret the conditional decomposition using the randomized intervention notation. Let $\mathrm{E}_g(Y_{R(D|G=g',\boldsymbol{Q})})$ be the mean potential outcome of group g when its members were given treatment values randomly drawn from members of group g' who share the same $\boldsymbol{Q}$ values with them. We can rewrite this mean potential outcome as follows.

$$\mathrm{E}_g(Y_{R(D|G=g',\boldsymbol{Q})}) = \mathrm{E}_g(Y_0) + \int \mathrm{E}_g(\tau \mid \boldsymbol{Q} = \boldsymbol{q})\, \mathrm{E}_{g'}(D \mid \boldsymbol{Q} = \boldsymbol{q}) f_g(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}. \qquad (4)$$

It then follows that components of the conditional decomposition can be written in the ran-

domized intervention notation:

$$\begin{aligned}
\mathrm{E}_a(Y) - \mathrm{E}_b(Y) - [\mathrm{E}_a(Y_{R(D|G=a,\boldsymbol{Q})}) - \mathrm{E}_b(Y_{R(D|G=b,\boldsymbol{Q})})] &= \text{conditional selection} \\
\mathrm{E}_b(Y_{R(D|G=a,\boldsymbol{Q})}) - \mathrm{E}_b(Y_{R(D|G=b,\boldsymbol{Q})}) &= \text{conditional prevalence} \\
\mathrm{E}_a(Y_{R(D|G=a,\boldsymbol{Q})}) - \mathrm{E}_b(Y_{R(D|G=a,\boldsymbol{Q})}) &= \text{conditional effect} + \text{conditional selection} \\
&\quad + \boldsymbol{Q} \text{ distribution.} \qquad (5)
\end{aligned}$$

Therefore, a two-step intervention again underlies our conditional decomposition. In the first step, the treatment is randomized within each $\boldsymbol{Q}$ level for members of both groups. The reduction in disparity resulting from this step is then the conditional selection component. The second step is an equalization intervention independent of the previous randomization step. For the second step, the pre-intervention disparity is $\mathrm{E}_a(Y_{R(D|G=a,\boldsymbol{Q})}) - \mathrm{E}_b(Y_{R(D|G=b,\boldsymbol{Q})})$, and the post-intervention disparity is $\mathrm{E}_a(Y_{R(D|G=a,\boldsymbol{Q})}) - \mathrm{E}_b(Y_{R(D|G=a,\boldsymbol{Q})})$. Hence, the conditional prevalence component is the reduction in disparity brought about by giving group b random draws of treatment from group a within $\boldsymbol{Q}$ levels. What would remain after this two-step intervention is the sum of the conditional effect, the conditional selection, and the $\boldsymbol{Q}$ distribution components.[6]

Similar to the unconditional random equalization decompositions, Jackson (2021) and Lundberg (2022) proposed two variants of what can be called the conditional random equalization decomposition (CRED), which decomposes outcome disparities into a reduction in disparity component and a remaining disparity component. The hypothetical intervention underlying the CRED is a random re-distribution of treatment values among people with the same $\boldsymbol{Q}$ values such that the treatment prevalence would be equalized within $\boldsymbol{Q}$ levels. Analogous to the unconditional case, we show that the reduction in disparity component in both variants of CRED conditionally mixes the prevalence component with the selection component, thus the conditional prevalence component in our decomposition should be preferred for its clearer interpretation.

In Jackson's (2021) variant of the CRED, the reduction in disparity is $\mathrm{E}_b(Y_{R(D|G=a,\boldsymbol{Q})}) -$

[6]For equation (4) to hold, we only require $\mathrm{supp}_g(\boldsymbol{Q}) \subseteq \mathrm{supp}_{g'}(\boldsymbol{Q})$, not necessarily $\mathrm{supp}_g(\boldsymbol{Q}) = \mathrm{supp}_{g'}(\boldsymbol{Q})$ required for the decomposition (3) itself. This implies that the interventional interpretations for conditional selection and conditional prevalence in equation (5) hold 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.

$\mathrm{E}_b(Y)$, which we can rewrite as

$$\int [\mathrm{E}_a(D \mid \boldsymbol{Q} = \boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{Q} = \boldsymbol{q})]\, \mathrm{E}_b(\tau \mid \boldsymbol{Q} = \boldsymbol{q}) f_b(\boldsymbol{q}) \mathrm{d}\boldsymbol{q} - \mathrm{E}_b[\mathrm{Cov}_b(D, \tau \mid \boldsymbol{Q})]. \tag{6}$$

Therefore, the reduction in disparity is a mixture of the conditional prevalence component and part of the conditional selection component. Namely, it captures not only the differential treatment prevalence within $\boldsymbol{Q}$ levels, but also the extent of conditional selection into treatment in group b. Intuitively, this is because the intervention of this CRED involves both a conditional equalization and a conditional randomization, and the randomization part eliminates selection into treatment conditional on $\boldsymbol{Q}$ in group b.

A similar case can be shown for the reduction in disparity in the CRED of Lundberg (2022).

$$\begin{aligned}
&\mathrm{E}_a(Y) - \mathrm{E}_b(Y) \\
&\{\mathrm{E}_a[\Pr(D = 0 \mid \boldsymbol{Q})Y_0 + \Pr(D = 1 \mid \boldsymbol{Q})Y_1] - \mathrm{E}_b[\Pr(D = 0 \mid \boldsymbol{Q})Y_0 + \Pr(D = 1 \mid \boldsymbol{Q})Y_1]\} \\
= {} & \mathrm{E}_a[\mathrm{Cov}_a(D, \tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D, \tau \mid \boldsymbol{Q})] + \int [\mathrm{E}_a(D \mid \boldsymbol{Q} = \boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{Q} = \boldsymbol{q})] \\
&[\mathrm{E}_a(\tau \mid \boldsymbol{Q} = \boldsymbol{q}) f_a(\boldsymbol{q}) \Pr(G = b \mid \boldsymbol{Q} = \boldsymbol{q}) + \mathrm{E}_b(\tau \mid \boldsymbol{Q} = \boldsymbol{q}) f_b(\boldsymbol{q}) \Pr(G = a \mid \boldsymbol{Q} = \boldsymbol{q})] \mathrm{d}\boldsymbol{q},
\end{aligned} \tag{7}$$

where the left hand side is the reduction in disparity in Lundberg's (2022) CRED. Hence, the reduction in disparity again involves both a term for the differential treatment prevalence within $\boldsymbol{Q}$ levels and the conditional selection component. The CRED in this case corresponds to the combination of a conditional equalization and a conditional randomization in both groups. Consequently, both the differential treatment prevalence and the differential selection into treatment would be reduced to zero conditional on $\boldsymbol{Q}$. Distinct from the reduction in disparity in both variants of the CRED, the conditional prevalence component in our decomposition solely measures the impact of a conditional equalization intervention. By implication, only our conditional prevalence component can be interpreted as a pure measure of the contribution of the differential treatment prevalence within $\boldsymbol{Q}$ levels to outcome disparities.

3 Estimation and Inference

We identify the components of our decompositions using two standard assumptions as follows.[7]

[7] Our unconditional decomposition can also be identified using the marginal treatment effect approach based on an instrumental variable (Heckman and Vytlacil, 2005; Zhou and Xie, 2020), as identifying group-specific

Assumption 2 (Conditional ignorability). $Y_d \perp\!\!\!\perp D \mid \boldsymbol{X} = \boldsymbol{x}, G = g, \forall d, \boldsymbol{x}, g$;

Assumption 3 (Overlap). $0 < \mathrm{E}(D \mid \boldsymbol{X} = \boldsymbol{x}, G = g) < 1, \forall \boldsymbol{x}, g$.

Without loss of generality, we assume $\boldsymbol{Q} \in \boldsymbol{X}$. We develop nonparametric and efficient estimators for components of our unconditional and conditional decompositions.[8] 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 specified conditions, 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 robustness properties.

To introduce the EIFs, we define the following functions:

$$\begin{aligned}
\mu(d, \boldsymbol{X}, g) &= \mathrm{E}(Y \mid D, \boldsymbol{X}, G = g) \\
\pi(d, \boldsymbol{X}, g) &= \Pr(D = d \mid \boldsymbol{X}, G = g) \\
p_g &= \Pr(G = g) \\
p_g(\boldsymbol{Q}) &= \Pr(G = g \mid \boldsymbol{Q}).
\end{aligned}$$

Below, we use ξ to denote the estimands (decomposition components), ϕ to denote the EIFs, and Ψ to denote the resulting one-step estimators.

3.1 The unconditional decomposition

All components of the unconditional decomposition are simple linear combinations of the total disparity and two generic functions evaluated at appropriate values of d, g, and g': $\xi_{dg} := \mathrm{E}(Y_d \mid G = g)$ and $\xi_{dgg'} := \mathrm{E}(Y_d \mid G = g)\,\mathrm{E}(D \mid G = g')$. The relationship between components of the unconditional decomposition and the generic functions are as follows:

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

ATTs and ATEs is sufficient (see Appendix E).

[8] Various regression, weighting, or matching estimators can also be used. In particular, under constant treatment effect, the prevalence component in the unconditional decomposition can be estimated using the traditional "product method" for the indirect effect of G through D on Y (Baron and Kenny, 1986). This again explicates the kinship between our decomposition and mediation analysis.

$$
\begin{aligned}
\text{Prevalence} &= \xi_{1ba} - \xi_{0ba} - \xi_{1bb} + \xi_{0bb} \\
\text{Effect} &= \xi_{1aa} - \xi_{0aa} - \xi_{1ba} + \xi_{0ba} \\
\text{Selection} &= \text{Total} - \text{Baseline} - \text{Prevalence} - \text{Effect}.
\end{aligned}
$$

Hence, the EIFs and one-step estimators for the decomposition components directly follow from those of ξ_{dg} and $\xi_{dgg'}$. In addition, these generic functions also provide a basis for estimating Jackson's (2018) version of the unconditional equalization estimand, since its reduction in disparity component can be represented as $\xi_{0b} + \xi_{1ba} - \xi_{0ba} - \mathrm{E}_b(Y)$. Under assumptions 1, 2, and 3, ξ_{dg} and $\xi_{dgg'}$ can be identified as the following quantities:

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

These identification results then enable us to derive the EIFs for ξ_{dg} and $\xi_{dgg'}$.

Theorem 1 (EIF)**.** *Under assumptions 1, 2, and 3, the EIF of ξ_{dg} is*

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

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

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

In Appendix B, we derive the EIFs for the general case with survey weights for both unconditional and conditional decompositions.

It follows that the one-step estimators are

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

There are two nuisance functions, $\pi(d, \boldsymbol{X}, g)$ and $\mu(d, \boldsymbol{X}, g)$. In order to relax the conditions required of nuisance function models (Kennedy, 2022; Chernozhukov et al., 2018a), we use cross-

fitting to estimate all nuisance functions. In particular, we first divide the sample randomly into two subsamples. Then we fit the nuisance functions using each subsample. Finally, we evaluate the fitted nuisance functions using data not in the subsample used to fit the functions. In practice, to improve the finite-sample performance of the estimator, we may stabilize the weights by dividing $\frac{\mathbb{1}(D=d)}{\pi(d,\boldsymbol{X},g)}$ by its sample average.

To study the asymptotic behavior of the one-step estimators for the unconditional decomposition, we invoke three additional assumptions,[9] which are consistent with assumptions required for the double ML estimator of ATE (Kennedy, 2022; Chernozhukov et al., 2018a). We let $\|\cdot\|$ denote the L_2-norm.

Assumption 4a (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 5a (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 6a (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 2 (Asymptotic distribution)**.** *Under assumptions 1, 2, 3, 4a, 5a, and 6a, the one step estimators for components of the unconditional decomposition are asymptotically normal and semiparametrically efficient, i.e.,* $\sqrt{n}(\Psi_{dg} - \xi_{dg}) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{dg})$*, and* $\sqrt{n}(\Psi_{dgg'} - \xi_{dgg'}) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{dgg'})$*, where* $\sigma^2_{dg} = \mathrm{E}[\phi_{dg}(Y,\boldsymbol{X})^2]$ *and* $\sigma^2_{dgg'} = \mathrm{E}[\phi_{dgg'}(Y,\boldsymbol{X})^2]$ *are the respective semiparametric efficiency bounds.*

Assumptions 4a through 6a are mild in the sense that they are often satisfied by popular ML models. The desirable asymptotic properties also carry over from Ψ_{dg} and $\Psi_{dgg'}$ to the final estimators of the decomposition components. We consistently estimate σ^2_{dg} and $\sigma^2_{dgg'}$ using the averages of squared estimated EIFs. The asymptotic distributions can then be used to construct hypothesis testing and confidence intervals. For example, a Wald-style $1-\alpha$ confidence interval for the prevalence component is

$$\Psi_{1ba} - \Psi_{1bb} - \Psi_{0ba} + \Psi_{0bb} \pm z_{1-\alpha/2} \cdot \frac{1}{n}\left[\sum\left(\hat{\phi}_{1ba} - \hat{\phi}_{1bb} - \hat{\phi}_{0ba} + \hat{\phi}_{0bb}\right)^2\right]^{1/2}.$$

Furthermore, our estimators also have a double robustness property.

Theorem 3 (Double robustness)**.** *Either consistent estimation of* $\mu(d,\boldsymbol{X},g)$ *or* $\pi(d,\boldsymbol{X},g)$ *for*

[9] Throughout this paper, the consistent estimation of p_g and $\mathrm{E}(D \mid G = g), \forall g$, are left implicit.

all d and g is sufficient for the consistency of Ψ_{dg} and $\Psi_{dgg'}$.

Hence, our unconditional decomposition estimator only requires the same conditions as the classic augmented inverse probability weighting estimator for ATE (Robins et al., 1994; Hirano et al., 2003) to be consistent.

3.2 The conditional decomposition

Relative to the unconditional case, we only need to additionally consider one generic function:

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

where (d, g, g', g'') is any combination of treatment status and group memberships out of 8 possible combinations. The relationship between components of the conditional decomposition and the generic functions are as follows:

$$\begin{aligned}
\text{Baseline} &= \xi_{0a} - \xi_{0b} \\
\text{Conditional Prevalence} &= \xi_{1bab} - \xi_{0bab} - \xi_{1bbb} + \xi_{0bbb} \\
\text{Conditional Effect} &= \xi_{1aaa} - \xi_{0aaa} - \xi_{1baa} + \xi_{0baa} \\
\boldsymbol{Q}\ \text{Distribution} &= \xi_{1baa} - \xi_{0baa} - \xi_{1bab} + \xi_{0bab} \\
\text{Conditional Selection} &= \text{Total} - \text{Baseline} \\
&\quad - \text{Conditional Prevalence} - \text{Conditional Effect} - \boldsymbol{Q}\ \text{Distribution}.
\end{aligned}$$

Hence, in this subsection, we focus on the inference of $\xi_{dgg'g''}$. The EIFs, one-step estimators, and their asymptotic distributions for components of the conditional decomposition will then straightforwardly follow. Moreover, we thereby also provide nonparametric inference for the reduction in disparity component in the CRED of Jackson (2021), which can be represented as $\xi_{0b} + \xi_{1bab} - \xi_{0bab} - \mathrm{E}_b(Y)$. Under assumptions 1, 2, 3, we identify $\xi_{dgg'g''}$ as

$$\mathrm{E}\left\{\mathrm{E}\left[\mathrm{E}(Y \mid d, \boldsymbol{X}, g) \mid \boldsymbol{Q}, g\right]\mathrm{E}\left(D \mid \boldsymbol{Q}, g'\right) \mid G = g''\right\}.$$

Theorem 4 (EIF)**.** *Under assumptions 1, 2, 3, the EIF of $\xi_{dgg'g''}$ is*

$$\phi_{dgg'g''}(Y, \boldsymbol{X}, \boldsymbol{Q})$$

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

We again construct the one-step estimators based on the EIF.

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

For this estimator, there are 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 $\mathrm{E}(Y_d \mid \boldsymbol{Q},g)$. We again cross-fit these nuisance functions using nonparametric methods. This estimator can be weight-stabilized by dividing $\frac{\mathbb{1}(D=d)}{\pi(d,\boldsymbol{X},g)}$, $\frac{\mathbb{1}(G=g')p_{g''}(\boldsymbol{Q})}{p_{g'}(\boldsymbol{Q})p_{g''}}$, and $\frac{\mathbb{1}(G=g)p_{g''}(\boldsymbol{Q})}{p_g(\boldsymbol{Q})p_{g''}}$ by their respective sample averages. The estimation of $\mathrm{E}(Y_d \mid \boldsymbol{Q},g) = \mathrm{E}[\mu(d,\boldsymbol{X},g) \mid \boldsymbol{Q},g]$ requires a more involved procedure. We adopt a pseudo-outcome approach (e.g., van der Laan, 2006; Semenova and Chernozhukov, 2021), where the pseudo outcome for each d is defined as the decentered EIF for $\mathrm{E}(Y_d)$, i.e.,

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

We first estimate $\pi(d,\boldsymbol{X},G)$ and $\mu(d,\boldsymbol{X},G)$ within two random subsamples of the data *without* cross-fitting and obtain the estimated pseudo outcome $\hat{\delta}_d\left(Y,\boldsymbol{X},G\right)$. Then, using $\hat{\delta}_d\left(Y,\boldsymbol{X},G\right)$ as a proxy of Y_d, we estimate $\mathrm{E}\left[Y_d \mid \boldsymbol{Q},g\right]$ *with* cross-fitting, i.e., we fit $\mathrm{E}\left[\hat{\delta}_d\left(Y,\boldsymbol{X},G\right) \mid \boldsymbol{Q},g\right]$ in one subsample and plug in values of $\boldsymbol{Q}$ in the other subsample. Using this procedure, we ensure that the fitting of $\mathrm{E}(Y_d \mid \boldsymbol{Q},g)$, which relies on estimating the pseudo outcome, is done separately in each subsample.

The asymptotic theory of the one-step estimator for $\xi_{dgg'g''}$ is less general than the theories for ξ_{dg} and $\xi_{dgg'}$, in that some conditions differ by the specific configuration of g, g' and g''. Note that $g = g''$ for the conditional prevalence component, and $g' = g''$ for the conditional effect component. For $\sqrt{n}$-consistency, asymptotic normality, and efficiency, we need three assumptions.

Assumption 4b (Boundedness). With probability 1, $\hat{\pi}(d, \boldsymbol{X}, g) \geq \eta$, $\pi(d, \boldsymbol{X}, g) \geq \eta$, $\hat{p}_g(\boldsymbol{Q}) \geq \eta$, $p_g(\boldsymbol{Q}) \geq \eta$, $|Y - \hat{\mu}(d, \boldsymbol{X}, g)| \leq \zeta$, $|Y - \mu(d, \boldsymbol{X}, g)| \leq \zeta$, $|\mu(d, \boldsymbol{X}, g)| \leq \zeta$, $|\operatorname{E}(Y_d \mid \boldsymbol{Q}, g)| \leq \zeta$, 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 5b (Consistency). $\|\hat{\mu}(d, \boldsymbol{X}, g) - \mu(d, \boldsymbol{X}, g)\| = o_p(1)$, $\|\hat{\pi}(d, \boldsymbol{X}, g) - \pi(d, \boldsymbol{X}, g)\| = o_p(1)$, $\left\|\hat{\operatorname{E}}(Y_d \mid \boldsymbol{Q}, g) - \operatorname{E}(Y_d \mid \boldsymbol{Q}, g)\right\| = o_p(1)$, $\left\|\hat{\operatorname{E}}(D \mid \boldsymbol{Q}, g) - \operatorname{E}(D \mid \boldsymbol{Q}, g)\right\| = o_p(1)$, and $\|\hat{p}_g(\boldsymbol{Q}) - p_g(\boldsymbol{Q})\| = o_p(1)$, $\forall d, g$.

Assumption 6b (Convergence rate). $\|\hat{\pi}(d, \boldsymbol{X}, g) - \pi(d, \boldsymbol{X}, g)\| \, \|\hat{\mu}(d, \boldsymbol{X}, g) - \mu(d, \boldsymbol{X}, g)\| = o_p(n^{-1/2})$, $\left\|\mathbb{1}(G = g) \operatorname{E}(Y_d \mid \boldsymbol{Q}, g) - \mathbb{1}(G = g') \frac{\hat{p}_g(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})} \hat{\operatorname{E}}(Y_d \mid \boldsymbol{Q}, g)\right\| \left\|\hat{\operatorname{E}}(D \mid \boldsymbol{Q}, g') - \operatorname{E}(D \mid \boldsymbol{Q}, g')\right\| = o_p(n^{-1/2})$, and $\left\|\mathbb{1}(G = g') \operatorname{E}(D \mid \boldsymbol{Q}, g') - \mathbb{1}(G = g) \frac{\hat{p}_{g'}(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q})} \hat{\operatorname{E}}(D \mid \boldsymbol{Q}, g')\right\| \left\|\hat{\operatorname{E}}(Y_d \mid \boldsymbol{Q}, g) - \operatorname{E}(Y_d \mid \boldsymbol{Q}, g)\right\| = o_p(n^{-1/2})$, $\forall d, g, g'$.

Theorem 5 (Asymptotic distribution)**.** *Under assumptions 1, 2, 3, 4b, 5b, 6b, the one step estimator for components of the conditional decomposition is asymptotically normal and semiparametrically efficient, i.e.,* $\sqrt{n}\,(\Psi_{dgg'g''} - \xi_{dgg'g''}) \to \mathcal{N}(0, \sigma^2_{dgg'g''})$*, where* $\sigma^2_{dgg'g''} = \operatorname{E}[\phi_{dgg'g''}(Y, \boldsymbol{X}, \boldsymbol{Q})^2]$ *is the semiparametric efficiency bound.*

Some further insights can be gained about the second and third convergence rate assumptions. For a specific $\xi_{dgg'g''}$, first, when $g = g' = g''$, $\left\|\hat{\operatorname{E}}\left(Y_d \mid \boldsymbol{Q}, g\right) - \operatorname{E}(Y_d \mid \boldsymbol{Q}, g)\right\| \left\|\hat{\operatorname{E}}(D \mid \boldsymbol{Q}, g) - \operatorname{E}(D \mid \boldsymbol{Q},\right.$ $o_p(n^{-1/2})$ is sufficient. Hence, we obtain a form of rate double robustness with respect to $\operatorname{E}(Y_d \mid \boldsymbol{Q}, g)$ and $\operatorname{E}(D \mid \boldsymbol{Q}, g)$. Second, when $g = g'' \neq g'$, the following set of conditions is sufficient: for a constant $\zeta > 0$, $\frac{\hat{p}_g(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})} \leq \zeta$ with probability 1, $\left\|\hat{\operatorname{E}}(D \mid \boldsymbol{Q}, g') - \operatorname{E}(D \mid \boldsymbol{Q}, g')\right\| = o_p(n^{-1/2})$, and $\left\|\hat{\operatorname{E}}\left(Y_d \mid \boldsymbol{Q}, g\right) - \operatorname{E}(Y_d \mid \boldsymbol{Q}, g)\right\| = o_p(1)$. Third, when $g' = g'' \neq g$, a sufficient set is: for a constant $\zeta > 0$, $\frac{\hat{p}_{g'}(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q})} \leq \zeta$ with probability 1, $\left\|\hat{\operatorname{E}}(D \mid \boldsymbol{Q}, g') - \operatorname{E}(D \mid \boldsymbol{Q}, g')\right\| = o_p(1)$, and $\left\|\hat{\operatorname{E}}\left(Y_d \mid \boldsymbol{Q}, g\right) - \operatorname{E}(Y_d \mid \boldsymbol{Q}, g)\right\| = o_p(n^{-1/2})$. Therefore, the assumption is weaker for the inference of the conditional prevalence component, which is also the case for Theorem 6 below.

Theorem 6 (Double robustness)**.** *We assume consistent estimation of* $\operatorname{E}(D \mid \boldsymbol{Q}, g), \forall g$*. For a* $\xi_{dgg'g''}$*, when* $g = g''$*, consistent estimation of either* $\mu(d, \boldsymbol{X}, g)$ *or* $\pi(d, \boldsymbol{X}, g)$ *is sufficient for the consistency of* $\Psi_{dgg'g''}$*; when* $g \neq g''$*, if the estimation of* $\operatorname{E}(Y_d \mid \boldsymbol{Q}, g)$ *is doubly robust with respect to* $\mu(d, \boldsymbol{X}, g)$ *and* $\pi(d, \boldsymbol{X}, g)$*, then consistent estimation of either* $\mu(d, \boldsymbol{X}, g)$ *or* $\pi(d, \boldsymbol{X}, g)$ *is still sufficient for the consistency of* $\Psi_{dgg'g''}$*.*

Note that consistent estimation of $\hat{p}_g(\boldsymbol{Q})$ is not required for the consistency of $\Psi_{dgg'g''}$. The double robustness condition for $\operatorname{E}(Y_d \mid \boldsymbol{Q}, g)$ in the case of $g \neq g''$ motivates our pseudo-outcome estimator of $\operatorname{E}(Y_d \mid \boldsymbol{Q}, g)$, which is built on both $\mu(d, \boldsymbol{X}, g)$ and $\pi(d, \boldsymbol{X}, g)$.

Both the unconditional and conditional decompositions are implemented in the R (R Core Team, 2023) package cdgd (Yu, 2023), available from CRAN. The package offers various options for nuisance function estimation, including various ML methods, a parametric option with linear and logistic regressions, and an option that allows users to manually plug in nuisance function estimates.

4 Application

4.1 Overview

We apply our causal decomposition to investigate the role of college graduation in intergenerational income mobility in the United States. In this empirical setting, groups are defined by parental income, the outcome is adult income, and the treatment is college graduation. Previous research has touched upon the components of our unconditional decomposition to various degrees, but this is the first study to comprehensively evaluate the three distinct ways in which college graduation might contribute to the intergenerational transmission of income advantages.

The baseline component represents the part of the income attainment disparity that is unaccounted for by college graduation. Apart from college completion, parental income is associated with a variety of pre-college characteristics that have bearings on income attainment, such as academic achievement, cognitive skills, and noncognitive traits in adolescence (Reardon, 2011; Heckman et al., 2006; Farkas, 2003). Moreover, in adulthood, people from more privileged backgrounds may continue to benefit from various resources passed on within their families regardless of formal educational attainment. Any intervention imposed on the post-secondary credentialing process cannot eliminate channels of income persistence that are independent of college graduation.

In a way conceptually consistent with the prevalence component in our decomposition, social scientists have construed education as a mediator of the intergenerational reproduction of SES inequalities (Blau and Duncan, 1967, chapter 4 & 5; Featherman and Hauser, 1978, p.255-9; Ishida et al., 1995; Breen, 2010). In particular, researchers have documented a large and increasing income gap in college graduation (Ziol-Guest and Lee, 2016; Bailey and Dynarski, 2011). Moreover, using a simulation based on descriptive models, Bloome et al. (2018) concluded that rising educational inequality strengthened intergenerational income persistence

over time.

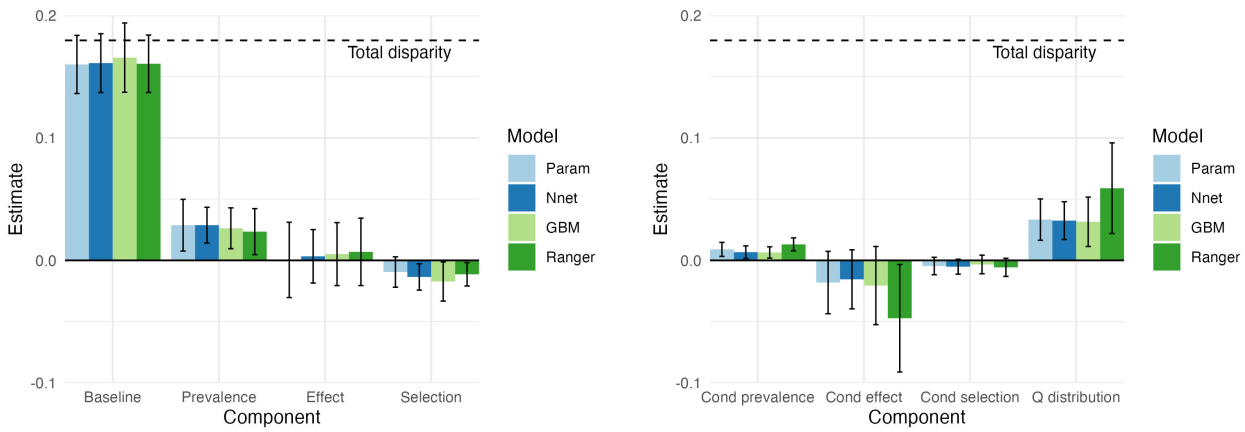
There is also an active literature on the heterogeneity of the college effect on SES attainment by SES origin, which underpins our effect component. Brand et al. (2021) and Cheng et al. (2021) found that the college premiums on wage and income outcomes are larger for people who are economically disadvantaged in adolescence. However, Zhou (2019) and Fiel (2020) found only insignificant heterogeneity in the effect of college completion on income by parental income. However, these works did not evaluate the extent to which income disparities can be attributed to groupwise differential effects of college.

Finally, the literature on college effects on later SES status has yielded mixed findings regarding whether selection on the treatment effect is positive or negative overall in the US. Brand and Xie (2010) and Brand et al. (2021) concluded that there is negative selection into college. However, Breen et al. (2015) warned that Brand and Xie's (2010) result may be biased from unobserved confounding. The instrumental variable analysis of Heckman et al. (2018), on the other hand, reveals positive selection. This line of inquiry does not estimate group-specific selection patterns, thereby missing the linkage between group-differentiated selection and outcome disparities. In Appendix E, we clarify and synthesize the differences and connections between various concepts of "selection into college" in the social science literature.

4.2 Data, variables and estimation details

We analyze the National Longitudinal Survey of Youth 1979 (NLSY79), which is a nationally representative dataset for a cohort of young men and women who were born between 1957 and 1964 and were living in the United States at the baseline survey in 1979. We limit the analysis to respondents who were between 14 to 17 years old in 1979 to ensure that income origin is measured before the respondents had a chance to graduate from college.

We contrast income-origin groups defined as the top 60% and bottom 40% of family incomes averaged over the first three waves of the survey (1979, 1980, and 1981), divided by the square root of the family size. The respondents were 14 to 20 years old during that time. The treatment variable is a binary indicator of whether the respondent graduated from college by the age of 31. We measure the outcome by the percentile rank of adult income, averaged over five survey waves between age 35 and 44, again divided by the square root of family size. For the conditional decomposition, we define $\boldsymbol{Q}$ to be the Armed Forces Qualification Test (AFQT) score measured in 1980, which reflects academic achievement in high school.



(c) Unconditional decomposition

(d) Conditional decomposition

Figure 2. Decomposition estimates. Cond=conditional, Para=parametric models, Nnet=neural networks, GBM=generalized boosting machine, Ranger=random forests. Error bars indicate 95% confidence intervals. Baseline component is omitted in panel (b) as the baseline component estimates for the conditional decomposition are, by definition, the same as those for the unconditional decomposition.

Covariates used for causal identification include 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 locus of control scale score, the Rosenberg esteem item response score, language spoken at home, urban/rural residence, presence of mother, school satisfaction, region of residence, and mother's working status.

We fit the nuisance functions using three alternative ML methods, gradient boosting machine (GBM), neural networks and random forests. For comparison, we also fit the nuisance functions using parametric regressions. The parametric models are set up as follows. For models predicting Y, we use linear regressions with all two-way interactions between the treatment and baseline variables including the group indicator, along with their main effects. For propensity score models and models predicting G, the covariates are entered linearly. For other models, where only G and $\boldsymbol{Q}$ are predictors, we include all main effects and two-way interactions between G and $\boldsymbol{Q}$.[10]

4.3 Results

We illustrate estimates of the decomposition components in Figure 1, while the numeric estimates are presented in Appendix Table A2 and A3. To aid the interpretation of the decomposition components, in Appendix Table A1, we additionally report estimates of group-specific means of baseline potential outcome, treatment proportions, ATEs, and covariances between treatment and treatment effects, as well as group differences in these quantities.

To begin with, according to the estimated total disparity, people from lower income origin on average achieve 18 percentiles lower incomes in their 30s and 40s. This confirms the well-established pattern of integenerational income persistence. Our decompositions reveal aspects of the formation of this persistence. Across models for the nuisance functions, the baseline component constitutes around 90% of the total disparity. This suggests that income advantages are transmitted between generations via many channels other than inequalities related to a college degree. The baseline component hence demarcates an upper limit for interventions aiming at post-secondary education.

Nevertheless, differential prevalence of college degrees across income origin groups plays a substantial and statistically significant role in explaining the income disparity in adulthood. In the unconditional decomposition, the prevalence component causally accounts for about 15% of the total disparity in adult income. Therefore, if an intervention could equalize college graduation rates across income origins, the outcome disparity would be reduced by 15%. What underlies the prevalence component is the vast inequality in college graduation by parental income: 29% in the higher-income group obtained a college degree by age 31 while only 9% in the lower-income group did so (see Appendix Table A1). The conditional prevalence component in the conditional decomposition remains statistically significant in all models, although it has much smaller magnitudes and only accounts for 3% to 7% of the total disparity. Therefore, equalizing chances of graduating college within levels of prior achievement would still make a difference, even small.

The contribution of the effect components in both unconditional and conditional decompositions is small in magnitude and generally not statistically significant. This is due to minimal between-group effect heterogeneity, either unconditionally or conditionally. In Appendix A1, we show that across models, the group-specific ATEs range from 12 to 15 percentiles and are all statistically significant. However, the groupwise difference is always insignificant. This is

[10] Code for this analysis is available at https://github.com/ang-yu/causal_decomposition_case_study.

consistent with findings of Zhou (2019) and Fiel (2020), which are also based on NLSY79.

All three ML models are in agreement that the selection component in the unconditional decomposition is significantly negative, offsetting 48% to 65% of the outcome disparity caused by differential prevalence. Had the selection component been zero, the total disparity would be around 6% higher. In fact, selection into college is positive in the disadvantaged group and negative in the advantaged group, although not significantly. This echos the view that obtaining a college degree is more of a rational decision based on expected returns and less of a cultural norm among the disadvantaged than among the advantaged (Mare, 1980; Hout, 2012). On the other hand, the conditional selection component does not notably contribute to the disparity in income attainment.

Descriptively, the $\boldsymbol{Q}$ distribution component also plays an important role in the generation of the outcome disparity, because of a large income gap in academic achievement in high school. Finally, we also estimate the reduction in disparity components in the unconditional and conditional random equalization decompositions (Jackson and VanderWeele, 2018; Jackson, 2021) and present the results in Appendix Table A2 and A3. In the unconditional case, relative to our prevalence component, the random equalization decomposition underestimates the extent to which differential college graduation rates contribute to the outcome disparity. This is because selection into college is positive in the lower-income group. By randomizing college degrees within the lower-income group, the intervention underlying the random equalization decomposition would be less powerful in reducing the outcome disparity than the pure equalization intervention corresponding to our prevalence component.

5 Discussion

We provide a new decomposition approach for quantifying the ways in which a treatment variable affects outcome disparities between ascriptive groups. Our decompositions improve on previous approaches of explaining group disparities and reveal novel policy levers for ameliorating disparities. Our decompositions are expressed as model-free estimands. We introduced nonparametric estimators that are efficient, asymptotically normal, and doubly robust. We applied our decompositions to study the role of college education in integenerational income persistence. We highlight that differential prevalence of the college degree significantly contributes to the income disparity, even conditional on academic achievement in high school.

Conversely, differential selection into college graduation plays a significantly compensatory role and reduces the income disparity.

Our approach can be extended in multiple directions. First, it is possible to develop analogous decompositions for non-binary treatments. For multi-valued categorical treatments, the basic form of equation (1) can be retained, with prevalence, effect, and selection components replaced with sums of category-specific contrasts. For continuous treatments, the randomization and equalization interventions, as well as their conditional versions, remain well-defined and may serve as bases of decompositions. Second, our framework can be extended to accommodate multiple temporally-ordered treatments. In the case of two treatments, D_1 and D_2, such extension can be based on the following decomposition of the outcome:

$$Y = Y(0,0)+D_1[Y(1,0)-Y(0,0)]+D_2[Y(0,1)-Y(0,0)]+D_1D_2[Y(1,1)-Y(1,0)-Y(0,1)+Y(0,0)],$$

where $Y(d_1, d_2)$ denotes the potential outcome of Y under the assignment of $D_1 = d_1$ and $D_2 = d_2$. Third, as an alternative to the double ML-style estimators in this paper, targeted learning (Van der Laan and Rose, 2011) could also be employed for estimation, which may have better finite-sample performance when the outcome is bounded.

Appendices

Appendix A: Proofs for Section 2

A.1. Equation (2)

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

$$\begin{aligned}
&\mathrm{E}_g\left(Y_{R(D|G=g')}\right) \\
&= \mathrm{E}_g(Y_1 \mid R(D \mid G = g') = 1)\,\mathrm{Pr}_g(R(D \mid G = g') = 1) + \mathrm{E}_g(Y_0 \mid R(D \mid G = g') = 0)\,\mathrm{Pr}_g(R(D \mid G = g') = \\
&= \mathrm{E}_g(Y_1)\,\mathrm{E}_{g'}(D) + \mathrm{E}_g(Y_0)(1 - \mathrm{E}_{g'}(D)) \\
&= \mathrm{E}_g(Y_0) + \mathrm{E}_{g'}(D)\,\mathrm{E}_g(\tau).
\end{aligned}$$

A.2. Equivalence results in subsection 2.3

For the CDE,

$$
\begin{aligned}
&\mathrm{E}(Y_{G=a,D=0}) - \mathrm{E}(Y_{G=b,D=0}) \\
&= \mathrm{E}(Y_{G=a,D=0} \mid G=a) - \mathrm{E}(Y_{G=b,D=0} \mid G=b) \\
&= \mathrm{E}_a(Y_{D=0}) - \mathrm{E}_b(Y_{D=0}),
\end{aligned}
$$

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

For the RIA of PIE,

$$
\begin{aligned}
&\mathrm{E}\left(Y_{G=b,R(D_{G=a})}\right) - \mathrm{E}\left(Y_{G=b,R(D_{G=b})}\right) \\
&= \mathrm{E}\left(Y_{G=b,D=1} \mid R(D_{G=a})=1\right)\Pr(R(D_{G=a})=1) + \mathrm{E}\left(Y_{G=b,D=0} \mid R(D_{G=a})=0\right)\Pr(R(D_{G=a})=0) \\
&\quad - \mathrm{E}\left(Y_{G=b,D=1} \mid R(D_{G=b})=1\right)\Pr(R(D_{G=b})=1) - \mathrm{E}\left(Y_{G=b,D=0} \mid R(D_{G=b})=0\right)\Pr(R(D_{G=b})=0) \\
&= \mathrm{E}\left(Y_{G=b,D=1}\right)\Pr(D_{G=a}=1) + \mathrm{E}\left(Y_{G=b,D=0}\right)\Pr(D_{G=a}=0) \\
&\quad - \mathrm{E}\left(Y_{G=b,D=1}\right)\Pr(D_{G=b}=1) - \mathrm{E}\left(Y_{G=b,D=0}\right)\Pr(D_{G=b}=0) \\
&= \mathrm{E}\left(Y_{G=b,D=1} \mid G=b\right)\Pr(D_{G=a}=1 \mid G=a) + \mathrm{E}\left(Y_{G=b,D=0} \mid G=b\right)\Pr(D_{G=a}=0 \mid G=a) \\
&\quad - \mathrm{E}\left(Y_{G=b,D=1} \mid G=b\right)\Pr(D_{G=b}=1 \mid G=b) - \mathrm{E}\left(Y_{G=b,D=0} \mid G=b\right)\Pr(D_{G=b}=0 \mid G=b) \\
&= \mathrm{E}\left(Y_{D=1} \mid G=b\right)\Pr(D=1 \mid G=a) + \mathrm{E}\left(Y_{D=0} \mid G=b\right)\Pr(D=0 \mid G=a) \\
&\quad - \mathrm{E}\left(Y_{D=1} \mid G=b\right)\Pr(D=1 \mid G=b) - \mathrm{E}\left(Y_{D=0} \mid G=b\right)\Pr(D=0 \mid G=b) \\
&= \mathrm{E}_b(\tau)[\mathrm{E}_a(D) - \mathrm{E}_b(D)],
\end{aligned}
$$

where the third equality holds by the unconditional ignorability of G, and the fourth holds by the consistency assumptions.

For the sum of the RIA of RIE and the RIA of the MIE,

$$
\begin{aligned}
&\mathrm{E}(Y_{G=a,D=1} - Y_{G=a,D=0} - Y_{G=b,D=1} + Y_{G=b,D=0})\,\mathrm{E}(D_{G=a}) \\
&= [\mathrm{E}(Y_{G=a,D=1} \mid G=a) - \mathrm{E}(Y_{G=a,D=0} \mid G=a) - \mathrm{E}(Y_{G=b,D=1} \mid G=b) + \mathrm{E}(Y_{G=b,D=0} \mid G=b)]\,\mathrm{E}(D_{G=a} \mid G \\
&= [\mathrm{E}(Y_{D=1} \mid G=a) - \mathrm{E}(Y_{D=0} \mid G=a) - \mathrm{E}(Y_{D=1} \mid G=b) + \mathrm{E}(Y_{D=0} \mid G=b)]\,\mathrm{E}(D \mid G=a) \\
&= [\mathrm{E}_a(\tau) - \mathrm{E}_b(\tau)]E_a(D),
\end{aligned}
$$

where the first equality holds by unconditional ignorability of G and the second by consistency.

A.3. Equation (3)

$$
\begin{aligned}
\mathrm{E}_a(Y) - \mathrm{E}_b(Y) &= \mathrm{E}_a(Y_0) + \mathrm{E}_b(Y_0) \\
&\quad + \mathrm{E}_a(D\tau) - \mathrm{E}_b(D\tau) - \mathrm{E}_a[\mathrm{E}_a(D \mid \boldsymbol{Q})\,\mathrm{E}_a(\tau \mid \boldsymbol{Q})] + \mathrm{E}_b[\mathrm{E}_b(D \mid \boldsymbol{Q})\,\mathrm{E}_b(\tau \mid \boldsymbol{Q})] \\
&\quad + \mathrm{E}_a[\mathrm{E}_a(D \mid \boldsymbol{Q})\,\mathrm{E}_a(\tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{E}_b(D \mid \boldsymbol{Q})\,\mathrm{E}_b(\tau \mid \boldsymbol{Q})] \\
&= \mathrm{E}_a(Y_0) + \mathrm{E}_b(Y_0) + \mathrm{E}_a[\mathrm{Cov}_a(D, \tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D, \tau \mid \boldsymbol{Q})] \\
&\quad + \int \mathrm{E}_a(D \mid \boldsymbol{Q} = \boldsymbol{q})\,\mathrm{E}_a(\tau \mid \boldsymbol{Q} = \boldsymbol{q}) f_a(\boldsymbol{q})\mathrm{d}\boldsymbol{q} - \int \mathrm{E}_b(D \mid \boldsymbol{Q} = \boldsymbol{q})\,\mathrm{E}_b(\tau \mid \boldsymbol{Q} = \boldsymbol{q}) f_b(\boldsymbol{q})\mathrm{d}\boldsymbol{q} \\
&= \mathrm{E}_a(Y_0) - \mathrm{E}_b(Y_0) \\
&\quad + \int [\mathrm{E}_a(D \mid \boldsymbol{Q} = \boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{Q} = \boldsymbol{q})]\,\mathrm{E}_b(\tau \mid \boldsymbol{Q} = \boldsymbol{q}) f_b(\boldsymbol{q})\mathrm{d}\boldsymbol{q} \\
&\quad + \int \mathrm{E}_a(D \mid \boldsymbol{Q} = \boldsymbol{q})\,\mathrm{E}_b(\tau \mid \boldsymbol{Q} = \boldsymbol{q})[f_a(\boldsymbol{q}) - f_b(\boldsymbol{q})]\mathrm{d}\boldsymbol{q} \\
&\quad + \int [\mathrm{E}_a(\tau \mid \boldsymbol{Q} = \boldsymbol{q}) - \mathrm{E}_b(\tau \mid \boldsymbol{Q} = \boldsymbol{q})]\,\mathrm{E}_a(D \mid \boldsymbol{Q} = \boldsymbol{q}) f_a(\boldsymbol{q})\mathrm{d}\boldsymbol{q} \\
&\quad + \mathrm{E}_a[\mathrm{Cov}_a(D, \tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D, \tau \mid \boldsymbol{Q})].
\end{aligned}
$$

Note that the last equality uses the assumption that $\mathrm{supp}_a(\boldsymbol{Q}) = \mathrm{supp}_b(\boldsymbol{Q})$.

A.4. Equation (4)

$$
\begin{aligned}
&\mathrm{E}_g(Y_{R(D|G=g',\boldsymbol{Q})}) \\
&= \int \mathrm{E}_g(Y_{R(D|G=g',\boldsymbol{Q})} \mid \boldsymbol{Q} = \boldsymbol{q}) f_g(\boldsymbol{q})\mathrm{d}\boldsymbol{q} \\
&= \int \mathrm{E}_g(Y_1 \mid \boldsymbol{Q} = \boldsymbol{q}, R(D \mid G = g', \boldsymbol{Q} = \boldsymbol{q}) = 1)\,\mathrm{Pr}_g(R(D \mid G = g', \boldsymbol{Q} = \boldsymbol{q}) = 1 \mid \boldsymbol{Q} = \boldsymbol{q}) f_g(\boldsymbol{q})\mathrm{d}\boldsymbol{q} \\
&\quad + \int \mathrm{E}_g(Y_0 \mid \boldsymbol{Q} = \boldsymbol{q}, R(D \mid G = g', \boldsymbol{Q} = \boldsymbol{q}) = 0)\,\mathrm{Pr}_g(R(D \mid G = g', \boldsymbol{Q} = \boldsymbol{q}) = 0 \mid \boldsymbol{Q} = \boldsymbol{q}) f_g(\boldsymbol{q})\mathrm{d}\boldsymbol{q} \\
&= \int [\mathrm{E}_g(Y_1 \mid \boldsymbol{Q} = \boldsymbol{q})\,\mathrm{E}_{g'}(D \mid \boldsymbol{Q} = \boldsymbol{q}) + \mathrm{E}_g(Y_0 \mid \boldsymbol{Q} = \boldsymbol{q})(1 - \mathrm{E}_{g'}(D \mid \boldsymbol{Q} = \boldsymbol{q}))] f_g(\boldsymbol{q})\mathrm{d}\boldsymbol{q} \\
&= \mathrm{E}_g(Y_0) + \int \mathrm{E}_g(\tau \mid \boldsymbol{Q} = \boldsymbol{q})\,\mathrm{E}_{g'}(D \mid \boldsymbol{Q} = \boldsymbol{q}) f_g(\boldsymbol{q})\mathrm{d}\boldsymbol{q}.
\end{aligned}
$$

All expectations are taken over $\boldsymbol{q} \in \mathrm{supp}_g(\boldsymbol{Q})$. For $\mathrm{E}_g(Y_{R(D|G=g',\boldsymbol{Q})})$ to be well-defined, we require $\mathrm{supp}_g(\boldsymbol{Q}) \subseteq \mathrm{supp}_{g'}(\boldsymbol{Q})$.

A.5. Equation (5)

$$
\begin{aligned}
&\mathrm{E}_a(Y) - \mathrm{E}_b(Y) - [\mathrm{E}_a(Y_{R(D|G=a,\boldsymbol{Q})}) - \mathrm{E}_b(Y_{R(D|G=b,\boldsymbol{Q})})] \\
&= \mathrm{E}_a(Y_0) - \mathrm{E}_b(Y_0) + \mathrm{E}_a(D\tau) - \mathrm{E}_b(D\tau) - [\mathrm{E}_a(Y_{R(D|G=a,\boldsymbol{Q})}) - \mathrm{E}_b(Y_{R(D|G=b,\boldsymbol{Q})})] \\
&= \mathrm{E}_a[\mathrm{E}_a(D\tau \mid \boldsymbol{Q}=\boldsymbol{q})] - \mathrm{E}_a[\mathrm{E}_a(\tau \mid \boldsymbol{Q}=\boldsymbol{q})\,\mathrm{E}_a(\tau \mid \boldsymbol{Q}=\boldsymbol{q})] \\
&\quad - \{\mathrm{E}_b[\mathrm{E}_b(D\tau \mid \boldsymbol{Q}=\boldsymbol{q})] - \mathrm{E}_b[\mathrm{E}_b(\tau \mid \boldsymbol{Q}=\boldsymbol{q})\,\mathrm{E}_b(\tau \mid \boldsymbol{Q}=\boldsymbol{q})]\} \\
&= \mathrm{E}_a[\mathrm{Cov}_a(D,\tau \mid \boldsymbol{Q}=\boldsymbol{q})] - \mathrm{E}_b[\mathrm{Cov}_b(D,\tau \mid \boldsymbol{Q}=\boldsymbol{q})].
\end{aligned}
$$

Other results in Equation (5) follow directly from Equation (4).

A.6. Equation (6)

$$
\begin{aligned}
&\mathrm{E}_b(Y_{R(D|G=a,\boldsymbol{Q})}) - \mathrm{E}_b(Y) \\
&= \mathrm{E}_b(Y_0) + \int \mathrm{E}_b(\tau \mid \boldsymbol{Q}=\boldsymbol{q})\,\mathrm{E}_a(D \mid \boldsymbol{Q}=\boldsymbol{q}) f_b(\boldsymbol{q})\mathrm{d}\boldsymbol{q} - \mathrm{E}_b(Y_0) - \mathrm{E}_b(D\tau) \\
&= \int [\mathrm{E}_a(D \mid \boldsymbol{Q}=\boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{Q}=\boldsymbol{q})]\,\mathrm{E}_b(\tau \mid \boldsymbol{Q}=\boldsymbol{q}) f_b(\boldsymbol{q})\mathrm{d}\boldsymbol{q} \\
&\quad + \mathrm{E}_b[\mathrm{E}_b(D \mid \boldsymbol{Q})\,\mathrm{E}_b(\tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{E}_b(D\tau \mid \boldsymbol{Q})] \\
&= \int [\mathrm{E}_a(D \mid \boldsymbol{Q}=\boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{Q}=\boldsymbol{q})]\,\mathrm{E}_b(\tau \mid \boldsymbol{Q}=\boldsymbol{q}) f_b(\boldsymbol{q})\mathrm{d}\boldsymbol{q} - \mathrm{E}_b[\mathrm{Cov}_b(D,\tau \mid \boldsymbol{Q}=\boldsymbol{q})].
\end{aligned}
$$

A.7. Equation (7)

$$
\begin{aligned}
&\mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \\
&\{\mathrm{E}_a[\Pr(D=0 \mid \boldsymbol{Q})Y_0 + \Pr(D=1 \mid \boldsymbol{Q})Y_1] - \mathrm{E}_b[\Pr(D=0 \mid \boldsymbol{Q})Y_0 + \Pr(D=1 \mid \boldsymbol{Q})Y_1]\} \\
&= \mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \{\mathrm{E}_a(Y_0) + \mathrm{E}_a[\mathrm{E}(D \mid \boldsymbol{Q})\tau] - \mathrm{E}_b(Y_0) - \mathrm{E}_b[\mathrm{E}(D \mid \boldsymbol{Q})\tau]\} \\
&= \mathrm{E}_a(D\tau) - \mathrm{E}_a[\mathrm{E}(D \mid \boldsymbol{Q})\,\mathrm{E}_a(\tau \mid \boldsymbol{Q})] - \{\mathrm{E}_b(D\tau) - \mathrm{E}_b[\mathrm{E}(D \mid \boldsymbol{Q})\,\mathrm{E}_b(\tau \mid \boldsymbol{Q})]\} \\
&= \int [\mathrm{E}_a(D\tau \mid \boldsymbol{Q}=\boldsymbol{q}) - \mathrm{E}(D \mid \boldsymbol{Q}=\boldsymbol{q})\,\mathrm{E}_a(\tau \mid \boldsymbol{Q}=\boldsymbol{q})] f_a(\boldsymbol{q})\mathrm{d}\boldsymbol{q} \\
&\quad - \int [\mathrm{E}_b(D\tau \mid \boldsymbol{Q}=\boldsymbol{q}) - \mathrm{E}(D \mid \boldsymbol{Q}=\boldsymbol{q})\,\mathrm{E}_b(\tau \mid \boldsymbol{Q}=\boldsymbol{q})] f_b(\boldsymbol{q})\mathrm{d}\boldsymbol{q} \\
&= \mathrm{E}_a[\mathrm{Cov}_a(D,\tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D,\tau \mid \boldsymbol{Q})] \\
&\quad + \int [\mathrm{E}_a(D \mid \boldsymbol{Q}=\boldsymbol{q}) - \mathrm{E}(D \mid \boldsymbol{Q}=\boldsymbol{q})]\,\mathrm{E}_a(\tau \mid \boldsymbol{Q}=\boldsymbol{q}) f_a(\boldsymbol{q})\mathrm{d}\boldsymbol{q} \\
&\quad - \int [\mathrm{E}_b(D \mid \boldsymbol{Q}=\boldsymbol{q}) - \mathrm{E}(D \mid \boldsymbol{Q}=\boldsymbol{q})]\,\mathrm{E}_b(\tau \mid \boldsymbol{Q}=\boldsymbol{q}) f_b(\boldsymbol{q})\mathrm{d}\boldsymbol{q} \\
&= \mathrm{E}_a[\mathrm{Cov}_a(D,\tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D,\tau \mid \boldsymbol{Q})]
\end{aligned}
$$

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

Appendix B: Efficient Influence Functions

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

Let $\mathbb{1}_{\tilde{o}}(o)$ be the point mass density at a single empirical observation, $\tilde{o}$. We also use $\mathbb{1}(\cdot)$ as the indicator function. Let subscript $\mathcal{P}_t$ indicate a regular parametric submodel indexed by t. When the true model is invoked, the subscript is omitted. 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 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:

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

B.1. EIFs for the unconditional decomposition

First of all, note that we only need to derive EIFs for two generic functions, $\xi_{dg} := \mathrm{E}\left(Y_d \frac{w(\boldsymbol{X},g)}{h_g} \mid G = g\right)$ for an arbitrary group g; and $\xi_{dgg'} := \mathrm{E}\left(Y_d \frac{w(\boldsymbol{X},g)}{h_g} \mid G = g\right) \mathrm{E}\left(D \frac{w(\boldsymbol{X},g')}{h_{g'}} \mid G = g'\right)$ 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 and SUTVA

assumptions, these estimands can be written as the following functionals:

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

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

We start with 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 = g)} \mid G = 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 = g\right] + \mathrm{E}[\mu(d, \boldsymbol{X}, g)w(\boldsymbol{X}, g) \mid G = g]\partial\frac{1}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g) \mid G = 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 = g]\} \\
&\quad + \frac{1}{h_g}\mathrm{E}[\partial\mu_{\mathcal{P}_t}(d, \boldsymbol{X}, g)w(\boldsymbol{X}, g) \mid G = 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 = 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 = 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 = 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 $\xi_{dgg'}$,

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

Since

$$\begin{aligned}
&\phi\left[\mathrm{E}\left(D\frac{w(\boldsymbol{X}, g')}{h_{g'}} \mid G = g'\right)\right] \\
&= \partial\, \mathrm{E}_{\mathcal{P}_t}\left(D\frac{w(\boldsymbol{X}, g')}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g') \mid G = g')} \mid G = g'\right)
\end{aligned}$$

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

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

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

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

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

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

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

which, without survey weights, becomes

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

Finally, the EIFs for the unconditional decomposition components are

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

B.2. EIFs for the conditional decomposition

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

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

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

$$\begin{aligned}
&\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. \\
&\quad \left. w(\boldsymbol{X}, g'') \mid G = 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)\right. \\
&\quad \left. w(\boldsymbol{X}, g'') \mid G = g''\right] \frac{1}{h_{g''}} \\
&\quad + \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 = g''\right] \\
&\quad \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''} \\
&\quad + \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)\right. \\
&\quad \left. w(\boldsymbol{X}, g'') \mid G = 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] \\
&\quad + \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)\right. \\
&\quad \left. w(\boldsymbol{X}, g'') \mid G = g''\right] \frac{1}{h_{g''}}.
\end{aligned}$$

And

$$\begin{aligned}
&\mathrm{E}\left[\partial \, \mathrm{E}_{\mathcal{P}_t}\left(Y_d \frac{w(\boldsymbol{X}, g)}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g) \mid \boldsymbol{Q}, g)} \mid \boldsymbol{Q}, g\right) \mathrm{E}_{\mathcal{P}_t}\left(D \frac{w(\boldsymbol{X}, g')}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g') \mid \boldsymbol{Q}, g')} \mid \boldsymbol{Q}, g'\right) w(\boldsymbol{X}, g'') \mid G = g''\right] \frac{1}{h_{g''}} \\
&= \mathrm{E}\left[\partial \, \mathrm{E}_{\mathcal{P}_t}\left(Y_d \frac{w(\boldsymbol{X}, g)}{\mathrm{E}_{\mathcal{P}_t}(w(\boldsymbol{X}, g) \mid \boldsymbol{Q}, g)} \mid \boldsymbol{Q}, g\right) \mathrm{E}\left(D \frac{w(\boldsymbol{X}, g')}{h_{g'}(\boldsymbol{Q})} \mid \boldsymbol{Q}, g'\right) w(\boldsymbol{X}, g'') \mid G = g''\right] \frac{1}{h_{g''}}
\end{aligned}$$

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

Hence,

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

which, in the case of no survey weights, simplifies to

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

Appendix C. Asymptotic inference

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 decomposition to attain this convergence result. By cross-fitting, all of $\hat{\mu}(d, \boldsymbol{X}, g), \hat{\pi}(d, \boldsymbol{X}, g), \hat{p}_g(\boldsymbol{Q})$, $\hat{\mathrm{E}}(Y_d \mid \boldsymbol{Q}, g)$, and $\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g)$ are fitted using data not in the current subsample, which we implicitly condition on throughout.

C.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. For the second term,

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

where the second inequality uses the Cauchy–Schwarz inequality.

For the empirical process term,

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

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

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 $\xi_{dgg'}$,

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

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,

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

$$
\begin{aligned}
&+ \frac{\mathbb{1}(G=g')}{p_{g'}} \mathrm{E}(Y_d \mid G=g) \left[\hat{\mathrm{E}}(D \mid G=g') - \mathrm{E}(D \mid G=g')\right] \\
&+ \frac{\mathbb{1}(G=g')}{\hat{p}_{g'}} \left[\hat{\mathrm{E}}(Y_d \mid G=g) - \mathrm{E}(Y_d \mid G=g)\right] \hat{\mathrm{E}}(D \mid G=g') \\
&+ \mathbb{1}(G=g') \frac{p_{g'} - \hat{p}_{g'}}{\hat{p}_{g'} p_{g'}} \mathrm{E}(Y_d \mid G=g) \hat{\mathrm{E}}(D \mid G=g') \\
&+ \mathbb{1}(G=g) \frac{\hat{p}_g - p_g}{\hat{p}_g p_g} \xi_{dgg'} \\
&+ \frac{\mathbb{1}(G=g)}{\hat{p}_g} (\xi_{dgg'} - \hat{\xi}_{dgg'}).
\end{aligned}
$$

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

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

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

C.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, $(\mathbb{P}_n - \mathbb{P}) \left[\hat{\phi}_{dgg'g''}(Y, \boldsymbol{X}, \boldsymbol{Q}) - \phi_{dgg'g''}(Y, \boldsymbol{X}, \boldsymbol{Q})\right]$, 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,

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

$$
\begin{aligned}
&+ \frac{\mathbb{1}(G=g)\hat{p}_{g''}(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q})\hat{p}_{g''}} \mathrm{E}(Y_d \mid \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[\mathrm{E}(Y_d \mid \boldsymbol{Q}, g) - \hat{\mathrm{E}}(Y_d \mid \boldsymbol{Q}, g)\right] \\
&+ \mathbb{1}(G=g) \frac{p_{g''}(\boldsymbol{Q})}{p_g(\boldsymbol{Q})} \frac{\hat{p}_{g''} - p_{g''}}{\hat{p}_{g''} p_{g''}} \mathrm{E}(Y_d \mid \boldsymbol{Q}, g) \mathrm{E}(D \mid \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}(Y_d \mid \boldsymbol{Q}, g) \mathrm{E}(D \mid \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}(Y_d \mid \boldsymbol{Q}, g) \mathrm{E}(D \mid \boldsymbol{Q}, g') \\
&+ \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] \\
&+ \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] \\
&+ \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\} \\
&+ \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\} \\
&+ \frac{\mathbb{1}(G=g')\hat{p}_{g''}(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})\hat{p}_{g''}} D \left[\hat{\mathrm{E}}(Y_d \mid \boldsymbol{Q}, g) - \mathrm{E}(Y_d \mid \boldsymbol{Q}, g)\right] \\
&+ \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 \, \mathrm{E}(Y_d \mid \boldsymbol{Q}, g) \\
&+ \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 \, \mathrm{E}(Y_d \mid \boldsymbol{Q}, g) \\
&+ \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 \, \mathrm{E}(Y_d \mid \boldsymbol{Q}, g) \\
&+ \frac{\mathbb{1}(G=g')\hat{p}_{g''}(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})\hat{p}_{g''}(\boldsymbol{Q})} \mathrm{E}(Y_d \mid \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''}(\boldsymbol{Q})} \hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') \left[\mathrm{E}(Y_d \mid \boldsymbol{Q}, g) - \hat{\mathrm{E}}(Y_d \mid \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') \mathrm{E}(Y_d \mid \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') \mathrm{E}(Y_d \mid \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') \mathrm{E}(Y_d \mid \boldsymbol{Q}, g).
\end{aligned}
$$

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.

C.2.1 When $g = g''$

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

$$
\begin{aligned}
& 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{\mathrm{E}}\left(Y_d \mid \boldsymbol{Q}, g\right) \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} \mathrm{E}(Y_d \mid \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{\mathrm{E}}\left(Y_d \mid \boldsymbol{Q}, g\right) \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} \mathrm{E}(Y_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}}\left(Y_d \mid \boldsymbol{Q}, g\right) \right] \left[\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') - \mathrm{E}(D \mid \boldsymbol{Q}, g') \right] \right\}.
\end{aligned}
$$

Then it follows from similar arguments as above that $R_{2,dgg'g''} = o_p(n^{-1/2})$ under stated conditions.

To gain some intuition about the second convergence rate condition,

$$
\left\| \mathbb{1}(G=g)\, \mathrm{E}(Y_d \mid \boldsymbol{Q}, g) - \mathbb{1}(G=g') \frac{\hat{p}_g(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})} \hat{\mathrm{E}}\left(Y_d \mid \boldsymbol{Q}, g\right) \right\| \left\| \hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') - \mathrm{E}(D \mid \boldsymbol{Q}, g') \right\| = o_p(n^{-1/2}),
$$

note that if $g = g' = g''$,

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

so $\left\|\mathrm{E}(Y_d \mid \boldsymbol{Q}, g) - \hat{\mathrm{E}}\left(Y_d \mid \boldsymbol{Q}, g\right)\right\| \left\|\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') - \mathrm{E}(D \mid \boldsymbol{Q}, g')\right\| = o_p(n^{-1/2})$ is sufficient to replace the second convergence rate condition.

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

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

so the following conditions are sufficient to replace the second convergence rate condition: for a constant $\zeta > 0$, $\left|\frac{\hat{p}_g(\boldsymbol{Q})}{\hat{p}_{g'}(\boldsymbol{Q})}\right| \leq \zeta$ 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 $\left\|\mathrm{E}(Y_d \mid \boldsymbol{Q}, g) - \hat{\mathrm{E}}\left(Y_d \mid \boldsymbol{Q}, g\right)\right\| = o_p(1)$.

C.2.2 When $g' = g''$

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

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

Under stated conditions, $R_{2,dgg'g''} = o_p(n^{-1/2})$.

For the second convergence rate condition, i.e.,

$$
\left\|\frac{\mathbb{1}(G=g')}{\hat{p}_{g'}}\mathrm{E}(D \mid \boldsymbol{Q}, g') - \frac{\mathbb{1}(G=g)\hat{p}_{g'}(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q})\hat{p}_{g'}}\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g')\right\| \left\|\hat{\mathrm{E}}(Y_d \mid \boldsymbol{Q}, g) - \mathrm{E}(Y_d \mid \boldsymbol{Q}, g)\right\|,
$$

note that when $g' = g'' \neq g$, the following conditions are sufficient: for a constant $\zeta > 0$, $\left|\frac{\hat{p}_{g'}(\boldsymbol{Q})}{\hat{p}_g(\boldsymbol{Q})}\right| \leq \zeta$ with probability 1, $\left\|\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g') - \mathrm{E}(D \mid \boldsymbol{Q}, g')\right\| = o_p(1)$, and $\left\|\mathrm{E}(Y_d \mid \boldsymbol{Q}, g) - \hat{\mathrm{E}}\left(Y_d \mid \boldsymbol{Q}, g\right)\right\| = o_p(n^{-1/2})$.

Appendix D. Double robustness

The proof for Theorem 3, i.e., the double robustness of Ψ_{dg} and $\Psi_{dgg'}$, is omitted. It is similar to the proof of the double robustness of the AIPW estimator of ATE (Robins et al., 1994; Tsiatis, 2006).

Below, we prove the double robustness of $\Psi_{dgg'g''}$, which is stated in Theorem 6. We use a tilde to denote the probability limit of a nuisance estimator, i.e., for example, $\hat{\mu}(d, \boldsymbol{X}, g)$ converges to $\tilde{\mu}(d, \boldsymbol{X}, g)$. Under consistent estimation of p_g and $\mathrm{E}(D \mid \boldsymbol{Q}, g)$, the one-step estimator $\Psi_{dgg'g''}$ converges in probability to

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

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

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

Hence, in this case, the double-robustness is achieved without requiring consistent estimation of $p_g(\boldsymbol{Q})$ and $\mathrm{E}(Y_d \mid \boldsymbol{Q}, g)$ for any d and q. When $g \neq g''$, further assuming that $\tilde{\mathrm{E}}(Y_d \mid \boldsymbol{Q}, g) = \mathrm{E}(Y_d \mid \boldsymbol{Q}, g)$ if either $\tilde{\pi}(d, \boldsymbol{X}, g) = \pi(d, \boldsymbol{X}, g)$ or $\tilde{\mu}(d, \boldsymbol{X}, g) = \mu(d, \boldsymbol{X}, g)$, then the probability limit of $\Psi_{dgg'g''}$ will still be $\xi_{dgg'g''}$. Note that consistent estimation of $p_g(\boldsymbol{Q})$ is still not required in either case.

Appendix E. 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 works. In our framework, group-specific selectivity 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" concept in Heckman et al. (2018), which is defined as the difference between ATT and ATE, i.e.,

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

Since $\mathrm{E}_g(D)$ is always positive, the sign of our group-specific selectivity is always the same as that of sorting on gains in Heckman et al. (2018).

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 selection 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 selection term.

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

$$- \underbrace{\{\mathrm{E}_a(Y_1) - \mathrm{E}_b(Y_1) - [\mathrm{E}_a(Y_0) - \mathrm{E}_b(Y_0)]\}}_{\text{The causal test of the great equalizer thesis}}$$

If the descriptive test returns a smaller 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.

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 outcome. Therefore, the selectivity term in the great equalizer literature captures selection on baseline outcome, on the top of selection on treatment effect captured in sorting on gains and our selectivity term (See Morgan and Winship, 2014, p.58-9).

Third, a literature on the 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 2,

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

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 F. Supplemental tables

Table A1. Group-specific Estimates for the Unconditional Decomposition

	Top 60% income	Bottom 40% income	Top-Bottom
Treatment proportion	0.290	0.086	0.204
	(0.269, 0.312)	(0.066, 0.107)	(0.174, 0.234)
		Gradient Boosted Machine	
Baseline outcome	0.575	0.410	0.165
	(0.553, 0.597)	(0.392, 0.427)	(0.137, 0.194)
ATE	0.146	0.128	0.018
	(0.108, 0.184)	(0.048, 0.208)	(-0.071, 0.106)
Cov(D, τ)	-0.013	0.004	-0.017
	(-0.028, 0.001)	(-0.003, 0.011)	(-0.033, -0.001)
		Neural Networks	
Baseline outcome	0.569	0.408	0.161
	(0.555, 0.584)	(0.389, 0.427)	(0.137, 0.185)
ATE	0.152	0.141	0.011
	(0.118, 0.187)	(0.074, 0.208)	(-0.064, 0.087)
Cov(D, τ)	-0.009	0.004	-0.014
	(-0.018, 0.000)	(-0.002, 0.010)	(-0.024, -0.003)
		Random Forests	
Baseline outcome	0.570	0.410	0.161
	(0.557, 0.584)	(0.390, 0.429)	(0.137, 0.184)
ATE	0.139	0.115	0.024
	(0.109, 0.169)	(0.025, 0.205)	(-0.071, 0.119)
Cov(D, τ)	-0.006	0.005	-0.011
	(-0.013, 0.001)	(-0.001, 0.012)	(-0.021, -0.002)
		Parametric Regressions	
Baseline outcome	0.570	0.410	0.160
	(0.556, 0.584)	(0.391, 0.429)	(0.136, 0.184)
ATE	0.142	0.141	0.001
	(0.107, 0.177)	(0.041, 0.241)	(-0.105, 0.107)
Cov(D, τ)	-0.007	0.003	-0.009
	(-0.016, 0.003)	(-0.005, 0.011)	(-0.022, 0.003)

Note: 95% confidence intervals are in the parentheses. Weight stabilization is used. For ML models, cross-fitting is used.

Table A2. Unconditional Decomposition Estimates

	GBM	Neural Networks	Random Forests	Parametric
Total	0.180	0.180	0.180	0.180
	(0.158, 0.201)	(0.158, 0.201)	(0.158, 0.201)	(0.158, 0.201)
Baseline	0.165	0.161	0.161	0.160
	(0.137, 0.194)	(0.137, 0.185)	(0.137, 0.184)	(0.136, 0.184)
Prevalence	0.026	0.029	0.023	0.029
	(0.010, 0.043)	(0.014, 0.043)	(0.005, 0.042)	(0.008, 0.050)
Effect	0.005	0.003	0.007	0.000
	(-0.021, 0.031)	(-0.018, 0.025)	(-0.021, 0.034)	(-0.030, 0.031)
Selection	-0.017	-0.014	-0.011	-0.009
	(-0.033, -0.001)	(-0.024, -0.003)	(-0.021, -0.002)	(-0.022, 0.003)
Jackson reduction	0.022	0.025	0.018	0.026
	(0.001, 0.043)	(0.007, 0.042)	(-0.006, 0.042)	(-0.001, 0.054)

Note: 95% confidence intervals are in the parentheses. Weight stabilization is used. For ML models, cross-fitting is used.

Table A3: Conditional Decomposition Estimates

	GBM	Neural Networks	Random Forests	Parametric
Total	0.180	0.180	0.180	0.180
	(0.158,0.201)	(0.158,0.201)	(0.158,0.201)	(0.158,0.201)
Baseline	0.165	0.161	0.161	0.160
	(0.137,0.194)	(0.137,0.185)	(0.137,0.184)	(0.136,0.184)
Conditional prevalence	0.006	0.007	0.013	0.009
	(0.002,0.011)	(0.002,0.012)	(0.008,0.018)	(0.003,0.015)
Conditional effect	-0.021	-0.015	-0.047	-0.018
	(-0.052,0.011)	(-0.040,0.009)	(-0.091,-0.003)	(-0.044,0.007)
Conditional selection	-0.003	-0.005	-0.006	-0.005
	(-0.011,0.004)	(-0.011,0.001)	(-0.013,0.002)	(-0.012,0.003)
Q distribution	0.032	0.032	0.059	0.033
	(0.011,0.052)	(0.017,0.048)	(0.022,0.096)	(0.016,0.050)
Conditional Jackson reduction	0.007	0.006	0.011	0.007
	(-0.001,0.014)	(0.000,0.013)	(0.003,0.020)	(-0.003,0.016)

Note: 95% confidence intervals are in the parentheses. Weight stabilization is used. For ML models, cross-fitting is used.

References

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.

Barber, Jennifer S., Jennifer Eckerman Yarger, and Heather H. Gatny. 2015. "Black-White Differences in Attitudes Related to Pregnancy Among Young Women." *Demography* 52:751–786.

Baron, Reuben M and David A Kenny. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51.

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.

Breen, R. 2010. "Educational Expansion and Social Mobility in the 20th Century." *Social Forces* 89:365–388.

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

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.

Farkas, George. 2003. "Cognitive Skills and Noncognitive Traits and Behaviors in Stratification Processes." *Annual Review of Sociology* 29:541–562.

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* p. 003804072092788.

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.

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

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 Evaluation1." *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* pp. 1–13.

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

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

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.

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.

Jann, Ben. 2008. "The Blinder–Oaxaca Decomposition for Linear Regression Models." *The Stata Journal: Promoting communications on statistics and Stata* 8:453–479.

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 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.

Laurison, Daniel and Sam Friedman. 2016. "The Class Pay Gap in Higher Professional and Managerial Occupations." *American Sociological Review* 81:668–695.

Lundberg, Ian. 2022. "The Gap-Closing Estimand: A Causal Approach to Study Interventions That Close Disparities Across Social Categories." *Sociological Methods & Research* p. 004912412110557.

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.

Mize, Trenton D. 2016. "Sexual Orientation in the Labor Market." *American Sociological Review* 81:1132–1160.

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.

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

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

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reardon, Sean F. 2011. "The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations." In *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*, edited by Greg J. Duncan and Richard J. Murnane, pp. 91–115. Russell Sage Foundation.

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.

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

Storer, Adam, Daniel Schneider, and Kristen Harknett. 2020. "What Explains Racial/Ethnic Inequality in Job Quality in the Service Sector?" *American Sociological Review* 85:537–572.

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* p. dyab090.

Tsiatis, Anastasios A. 2006. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. New York, NY: Springer.

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. 2017. "Mediation analysis with time varying exposures and mediators." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79:917–938.

VanderWeele, Tyler J., Stijn Vansteelandt, and James M. Robins. 2014. "Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder:." *Epidemiology* 25:300–306.

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. 2023. *cdgd: Causal Decomposition of Group Disparities*. R package version 0.3.0.

Zajacova, Anna, Hanna Grol-Prokopczyk, and Zachary Zimmer. 2021. "Pain Trends Among American Adults, 2002–2018: Patterns, Disparities, and Correlates." *Demography* 58:711–738.

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. 2022. "Attendance, Completion, and Heterogeneous Returns to College: A Causal Mediation Approach." *Sociological Methods & Research* p. 004912412211138.

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

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.