*Submitted to the Annals of Applied Statistics*

# NONPARAMETRIC CAUSAL DECOMPOSITION OF GROUP DISPARITIES

BY ANG YU[1,a] AND FELIX ELWERT[1,b]

[1]*Department of Sociology, University of Wisconsin-Madison,* [a]*ayu33@wisc.edu;* [b]*elwert@wisc.edu*

We propose a nonparametric framework that decomposes the causal contributions of a treatment variable to an outcome disparity between two groups. We decompose the causal contributions of treatment into group differences in 1) treatment prevalence, 2) average treatment effects, and 3) selection into treatment based on individual-level treatment effects. Our framework reformulates the classic Kitagawa-Blinder-Oaxaca decomposition nonparametrically in causal terms, complements causal mediation analysis by explaining group disparities instead of group effects, and distinguishes more mechanisms than recent random equalization decomposition. In contrast to all prior approaches, our framework isolates the causal contribution of differential selection into treatment as a novel mechanism for explaining and ameliorating group disparities. We develop nonparametric estimators based on efficient influence functions that are $\sqrt{n}$-consistent, asymptotically normal, semiparametrically efficient, and multiply robust to misspecification. We apply our framework to decompose the causal contributions of education to the disparity in adult income between parental income groups (intergenerational income persistence). We find that both differential prevalence of, and differential selection into, college graduation significantly contribute to intergenerational income persistence.

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

Prior research has addressed such questions using three approaches, none of which is fully appropriate for the task of causally decomposing descriptive disparities. First, popular Kitagawa-Blinder-Oaxaca (KBO) decompositions (Kitagawa, 1955; Blinder, 1973; Oaxaca, 1973) are defined in terms of parametric regression coefficients and do not answer any causal question by design (Fortin, Lemieux and Firpo, 2011, p.13; Lundberg, Johnson and Stewart, 2021, p.542). Second, causal mediation analysis (CMA) (VanderWeele, 2015), though formulated in causal terms, decomposes the causal effects of group membership rather than observed group disparities. Third, the recently developed random-equalization decomposition (VanderWeele and Robinson, 2014; Jackson and VanderWeele, 2018; Jackson, 2021; Lundberg, 2022), conflates distinct disparity-generating mechanisms, which limits its interpretability and usefulness for policy analysis. Importantly, all three prior approaches neglect that disparities in outcomes can in part be explained by group-differential selection of individuals into treatment as a function of their treatment effects.

In this article, we propose a nonparametric causal decomposition approach that remedies these limitations. In contrast to KBO, our decompositions are formulated as model-free

*Keywords and phrases:* social inequality, mediation analysis, double machine learning, income mobility.

causal estimands, with immediate interventional interpretations. In contrast to CMA, we decompose observed group disparities rather than possibly ill-defined causal effects of group membership. In contrast to random equalization decomposition, we separate more explanatory mechanisms. In contrast to all three prior approaches, we reveal a novel mechanism, termed the "selection component," that has not previously been incorporated in decomposition analysis.

Our framework distinguishes three causal mechanisms through which an intermediate treatment variable can contribute to a group disparity in an outcome. First, mean outcomes may differ across groups because the groups receive treatment at different rates (differential prevalence). Second, even if all groups have the same treatment prevalence, outcomes may differ across groups because the treatment has different average effects across groups (differential effects). Third, even if all groups have the same treatment prevalence and the same average effects, outcomes will nonetheless differ across groups when members of one group select into treatment more strongly on their treatment effects than members of another group (differential selection); for example, if treatment is randomly distributed to members of one group, but specifically given to only those members who will most benefit from treatment in the other group. Conceptually, prior decompositions were limited to considering differential prevalence and differential effects, leading to the belief that these two are the only possible mechanisms (Ward et al., 2019; Diderichsen, Hallqvist and Whitehead, 2019). Isolating differential selection into treatment as a source of group disparities is the central conceptual contribution of our approach.

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

We develop nonparametric estimators for our decompositions using efficient influence functions (EIF) under the assumption of conditional ignorability of the treatment. Our estimators can be implemented via data-adaptive methods such as machine learning (ML) and accommodate high-dimensional confounders. We derive the conditions under which the estimators are $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient. The estimators are also doubly or even quadruply robust to misspecification. The estimators are implemented in the R package cdgd (Yu, 2023), available from CRAN.

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

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

## 2. Estimands.

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

Assuming only the stable unit treatment value assumption (SUTVA) (Rubin, 1980),

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

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

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

First, the "baseline" component reflects the difference in mean baseline potential outcomes, $Y^0$, between groups, i.e., the outcome disparity in the complete absence of treatment.[1] In our application, the baseline component is the part of the income disparity in adulthood that is not attributable to college graduation in any way. Second the "prevalence" component indicates how much of the group disparity is due to differential prevalence of treatment. For example, it indicates the extent to which the difference in college graduation rates across parental income groups contributes to the outcome disparity. Third, the "effect" component reflects the difference in average treatment effects (ATE) across groups. Thus, it reveals the contribution of group-differential ATEs of college graduation to the adult income disparity.

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

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

To further explicate our novel selection component, we provide two interpretations for the covariance between treatment and the treatment effect, $\mathrm{Cov}(D,\tau)$. When the receipt of treatment is mainly based on self-selection (e.g, college graduation), the covariance may

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

indicate the extent to which the choice to take up the treatment is rational with respect to returns to the treatment. This interpretation has been the focus of research in economics and sociology (Heckman and Vytlacil, 2005; Brand and Xie, 2010; Heckman, Humphries and Veramendi, 2018), where a quantity closely related to the covariance has been called "sorting on gains" (see Appendix E). On the other hand, if treatment assignment is mainly administered by external decision-makers (e.g., drug prescription), the covariance indicates how effectively treatments are assigned to individuals. Thus, the selection component could also be called the contribution of sorting on gains or the effectiveness component.

2.2. *Interventional interpretation.* Our unconditional decomposition is formulated in counterfactual terms, and hence is prescriptive for future interventions. The decomposition reveals three policy levers for affecting group disparities. First, policy makers could manipulate the prevalence of the treatment in each of the two groups. Second, they could influence within-group selection into treatment based on group members' treatment effects. Third, they might even be able to manipulate the average treatment effect of each group.[2]

Different interventions will have different impacts on the prevalence, selection and effect components. Here, we explicate a two-step intervention that maps exactly onto the prevalence and selection components of our unconditional decomposition. We express this two-step intervention using randomized intervention notation (Didelez, Dawid and Geneletti, 2006; Geneletti, 2007), where $R(D \mid g)$ represents a randomly drawn value of treatment $D$ from group $g$. Then, $\mathrm{E}_g\left(Y^{R(D|g')}\right)$ denotes the post-intervention mean potential outcome for group $g$ after each member of group $g$ has received a random draw of treatment D from group $g'$. When $g = g'$, the intervention amounts to a random redistribution of treatments within the group. Using its definition, we can rewrite the post-intervention mean potential outcome:

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

It follows that our unconditional decomposition components can be re-written as follows:

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

This represents the two-step intervention. The first step internally randomizes the treatment in both groups without changing its prevalence. In this step, the pre-intervention disparity is $\mathrm{E}_a(Y) - \mathrm{E}_b(Y)$, and the post-intervention disparity is $\mathrm{E}_a\left(Y^{R(D|a)}\right) - \mathrm{E}_b\left(Y^{R(D|b)}\right)$. Since randomizing treatment within each group sets $\mathrm{Cov}_a(D,\tau) = \mathrm{Cov}_b(D,\tau) = 0$ and removes differential selection between groups, the change in disparity resulting from this intervention equals our selection component.

The second step equalizes treatment prevalence without changing selection by giving members of group $b$ random draws of treatment from group $a$. In this step, the pre-intervention disparity is $\mathrm{E}_a\left(Y^{R(D|a)}\right) - \mathrm{E}_b\left(Y^{R(D|b)}\right)$, and the post-intervention disparity is

[2]It is more conventional to discuss interventions on treatments than interventions on effects, which cannot be represented using the potential outcome notation. However, the notion of interventions on effects or structural relations have appeared in multiple literatures (Malinsky, 2018; Diderichsen, Hallqvist and Whitehead, 2019; Brady, Finnigan and Hübgen, 2017; Canen and Song, 2023).

$\mathrm{E}_a\left(Y^{R(D|a)}\right) - \mathrm{E}_b\left(Y^{R(D|a)}\right)$. Therefore, the prevalence component is the change in disparity resulting from this equalization intervention.[3] At the end of the two-step intervention, the remaining disparity is the sum of the baseline and the effect components. Using the fact that $\mathrm{Cov}_g(D,\tau) = \mathrm{E}_g(Y) - \mathrm{E}_g\left(Y^{R(D|g)}\right)$, we also present a visualization of our unconditional decomposition in Appendix B.

Isolating the distinct contributions of the selection and the prevalence components is useful in practice and enables policy makers to implement only the first step of the two-step intervention. Such choice is appealing under at least two scenarios. First, it is possible that the selection component is positive but the prevalence component is negative. Second, intervening on prevalence may be normatively undesired or practically impossible due to budget constraints.

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

2.3.1. *Comparison with the KBO decomposition.* Disparities research in the social and biomedical sciences traditionally employs KBO decompositions. The form of the KBO decomposition most closely resembling our approach decomposes the outcome disparity between groups into four components with respect to treatment $D$ and pre-treatment covariates $\boldsymbol{X}$:

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

where $\alpha_g$, $\beta_g$, and $\boldsymbol{\gamma}_g$ are the coefficients from group-specific linear regressions:

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

This decomposition attains a causal interpretation under (i) the causal assumption of conditional ignorability of the treatment, $Y^d \perp\!\!\!\perp D \mid g, \boldsymbol{x}, \forall d, g, \boldsymbol{x}$, and (ii) the parametric assumption that the group-specific linear regressions are correctly specified. If and only if both assumptions are satisfied, the endowment and slope components in the KBO decomposition are equivalent to our prevalence and effect components, respectively; and the sum of the intercept and residual components equals our baseline component.[4]

Our decomposition in equation (1) differs from the KBO decomposition in three respects. First, our decomposition is inherently causal, because it is directly formulated as estimands in potential outcomes notation. In contrast, the KBO decomposition, as an estimation framework, requires additional assumptions to support a causal interpretation. Such assumptions are rarely stated in practice. Second, our decomposition is nonparametric, whereas KBO decompositions are model-based and hence rely on a particular functional form. Third, KBO

[3]This equalization intervention justifies the scaling factors on the prevalence and effect components of equation (1). Intuitively, the prevalence component is scaled by $\mathrm{E}_b(\tau)$, because randomly changing treatment prevalence in group $b$ affects the outcome disparity only to the extent that treatment has an effect in group $b$ on average. The scaling factor on the effect component follows algebraically to complete the decomposition. Different interventions would lead to different scaling factors. For example, intervening to give group $a$ the treatment prevalence of group $b$, the prevalence component would be scaled by $\mathrm{E}_b(\tau)$.

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

decompositions do not contain a selection component, because the assumed functional form imposes effect homogeneity within each group. By contrast, our nonparametric decomposition does not impose effect homogeneity at any level and hence contains a selection component as a distinctive conceptual contribution.[5]

2.3.2. *Comparison with causal mediation analysis.* CMA decomposes a total effect of an exposure on an outcome into components in terms of an intermediate variable (mediator). The CMA literature is vast and contains many different decompositions (VanderWeele, 2015). We focus on a three-way CMA decomposition (VanderWeele and Tchetgen Tchetgen, 2014; VanderWeele, 2014) that is useful for illustrating similarities and differences with our decomposition in equation (1):

(3)

$$\underbrace{\mathrm{E}\left(Y^{a}\right)-\mathrm{E}(Y^{b})}_{\text{Total Effect}}$$
$$=\underbrace{\mathrm{E}\left(Y^{a,0}\right)-\mathrm{E}\left(Y^{b,0}\right)}_{\text{Controlled Direct Effect}}+\underbrace{\mathrm{E}\left[\left(Y^{b,1}-Y^{b,0}\right)\left(D^{a}-D^{b}\right)\right]}_{\text{Pure Indirect Effect}}+\underbrace{\mathrm{E}\left[D^{a}\left(Y^{a,1}-Y^{a,0}-Y^{b,1}+Y^{b,0}\right)\right]}_{\text{Portion Attributable to Interaction}},$$

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

Certain equivalences between equation (3) and our unconditional decomposition can be established under very strong assumptions: two unconditional ignorability assumptions for $G$, i.e., $Y^{g,d} \perp\!\!\!\perp G, \forall d, g$, and $D^g \perp\!\!\!\perp G, \forall g$; two SUTVA-type assumptions, $\mathrm{E}_g\left(Y^{g,d}\right)=\mathrm{E}_g\left(Y^{d}\right)$ and $\mathrm{E}_g\left(D^g\right)=\mathrm{E}_g(D), \forall d, g$; and a cross-world independence assumption (Pearl, 2001), $Y^{g,d} \perp\!\!\!\perp D^{g'}, \forall d, g, g'$. Under these assumptions, the conditional direct effect (CDE) equals our baseline component; the pure indirect effect (PIE) equals our prevalence component; and the portion attributable to interaction (PAI) equals our effect component. These equivalences are intuitive: both the CDE and the baseline component capture a group-based outcome difference when the intermediate variable, $D$, is held at $0$; both the PIE and the prevalence component address the role of the prevalence of $D$ in the relationship between the group and the outcome; finally, the PAI and the effect component both reflect how the effect of $D$ interacts with group membership. Importantly, the assumptions needed for establishing equivalence above are stronger than the corresponding assumptions needed for identifying equation (3).

However, all extant CMA decompositions, to our knowledge, differ from our unconditional decomposition in three crucial respects. First, CMA decomposes a different quantity: a total effect of group membership, rather than the descriptive group disparity. Our focus on decomposing descriptive disparities is useful, because descriptive disparities between groups are often the object of interest in their own right in the social and health sciences and the focus of popular and policy concerns. For example, income disparities in adulthood between the children of rich and poor parents are often viewed as concerning, regardless of whether these disparities originate from the causal effect of parental income or from confounding

[5]Prior work has offered alternative causal interpretations for KBO decompositions under various assumptions. For example, a prominent literature shows that KBO decompositions can estimate the average treatment effect on the treated (ATT) (Fortin, Lemieux and Firpo, 2011; Kline, 2011; Yamaguchi, 2015). Similarly, Chernozhukov, Fernández-Val and Luo (2018) show that a KBO decomposition can estimate the partial treatment effect. And Huber (2015) discusses using a KBO decomposition for estimating natural indirect effects (Pearl, 2001). None of these interpretations accommodates a descriptive group variable.

factors such as parents' education and race. Furthermore, some group variables, such as race and gender, may be immutable attributes on which it is hard to define an intervention (Rubin, 1974; Holland, 1986). As a result, CMA estimands, which encode a hypothetical intervention on group membership, may not even be well-defined.

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

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

2.3.3. *Comparison with unconditional random equalization decomposition.* The newest entry into decomposition methodology is random equalization decomposition (VanderWeele and Robinson, 2014; Jackson and VanderWeele, 2018; Sudharsanan and Bijlsma, 2021; Lundberg, 2022; Park et al., 2023), which exist in unconditional (URED) and conditional versions (CRED).

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

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

Thus, the outcome disparity is decomposed into the change in disparity resulting from the intervention and the remaining disparity after the intervention.

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

However, URED also differs from our unconditional decomposition in important respects. First, as a two-way decomposition, URED contains less information than our four-way decomposition. Second, when selecction is present, URED does not isolate any of the mechanisms identified in our decomposition in equation (1).

Specifically, URED does not have a component isolating the contribution of differential treatment prevalence across groups. Formally, using equation (2), one can show that URED's change in disparity equals the difference between our prevalence component and selection into treatment in the disadvantaged group, $\mathrm{E}_b(\tau)[\mathrm{E}_a(D) - \mathrm{E}_b(D)] - \mathrm{Cov}_b(D, \tau)$. Intuitively, by assigning the disadvantaged group random draws of treatment, URED's one-step intervention simultaneously equalizes treatment prevalence across groups and randomizes treatment in the disadvantaged group. By contrast, as noted in Section 2.2, the two-step intervention corresponding to our unconditional decomposition neatly separates equalization from randomization. Consequently, unlike our prevalence component, URED's change in disparity is generally non-zero even if there is no difference in treatment prevalence, $\mathrm{E}_a(D) = \mathrm{E}_b(D)$.

By the same token, URED does not contain a selection component, because the contribution of selection to the outcome disparity is split between URED's two components. The change-in-disparity component contains the negative of selection into treatment in the disadvantaged group, and the remaining disparity contains selection into treatment in the advantaged group alongside the sum of our baseline and effect components, $\mathrm{E}_a(Y^0) - \mathrm{E}_b(Y^0) + \mathrm{E}_a(D)[\mathrm{E}_a(\tau) - \mathrm{E}_b(\tau)] + \mathrm{Cov}_a(D,\tau)$.[6]

2.4. *Conditional decomposition.* Our conditional decomposition evaluates the contributions of treatment to the outcome disparity conditional on a vector of pre-treatment covariates $\boldsymbol{Q}$. Thus, the conditional decomposition quantifies how the disparity would change if differences in treatment prevalence, effect, and selection were removed only between members of the advantaged and disadvantaged groups who shared the same values of $\boldsymbol{Q}$.

In our application, $\boldsymbol{Q}$ is academic achievement in high school. The conditional decomposition thus informs interventions on college graduation that are conducted only within levels of prior achievement, preserving the relationship between prior achievement and educational attainment. Conditional decompositions are useful in practice because the corresponding conditional interventions may be normatively more desirable or practically more feasible than their unconditional counterparts (see Jackson, 2021).

In order to introduce the conditional decomposition, we assume common support on $\boldsymbol{Q}$.

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

Under Assumptions 1 and 2, we then obtain the following conditional decomposition:

(4)

$$\begin{aligned}\mathrm{E}_a(Y) - \mathrm{E}_b(Y) = &\underbrace{\mathrm{E}_a\left(Y^0\right) - \mathrm{E}_b\left(Y^0\right)}_{\text{Baseline}} + \underbrace{\int [\mathrm{E}_a(D \mid \boldsymbol{q}) - \mathrm{E}_b(D \mid \boldsymbol{q})]\, \mathrm{E}_b(\tau \mid \boldsymbol{q}) f_b(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}}_{\text{Conditional prevalence}} \\ &+ \underbrace{\int [\mathrm{E}_a(\tau \mid \boldsymbol{q}) - \mathrm{E}_b(\tau \mid \boldsymbol{q})]\, \mathrm{E}_a(D \mid \boldsymbol{q}) f_a(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}}_{\text{Conditional effect}} + \underbrace{\mathrm{E}_a[\mathrm{Cov}_a(D,\tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D,\tau \mid \boldsymbol{Q})]}_{\text{Conditional selection}} \\ &+ \underbrace{\int \mathrm{E}_a(D \mid \boldsymbol{q})\, \mathrm{E}_b(\tau \mid \boldsymbol{q}) [f_a(\boldsymbol{q}) - f_b(\boldsymbol{q})] \mathrm{d}\boldsymbol{q}}_{\boldsymbol{Q}\text{-distribution}}.\end{aligned}$$

The baseline component in the conditional decomposition equals the baseline component in the unconditional decomposition; it is the disparity that would obtain if nobody received treatment. The conditional prevalence component gives the part of the disparity that is due to differences in treatment prevalence across groups within levels of $\boldsymbol{Q}$. In our example it indicates the extent to which the difference in college graduation rates between advantaged and disadvantaged students who shared the same academic achievement in high school contribute to the income disparity in adulthood. The conditional effect component reflects the contribution of the difference in the group-specific conditional average treatment effect (CATE) given $\boldsymbol{Q}$. Thus, it reveals the contribution of group-differential CATEs of college graduation given prior academic achievement. The conditional selection component is the contribution of group-differential selection into treatment net of $\boldsymbol{Q}$, i.e., the contribution of differential

[6] Lundberg (2022) introduces a variant of URED whose one-step intervention equalizes treatment prevalences across groups by assigning random draws from the marginal treatment distribution in the population, $R(D)$, to both the advantaged and the disadvantaged groups. The resulting change in disparity is $\mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \left[\mathrm{E}_a\left(Y^{R(D)}\right) - \mathrm{E}_b\left(Y^{R(D)}\right)\right]$. Rewriting this as $\mathrm{E}(\tau)[\mathrm{E}_a(D) - \mathrm{E}_b(D)] - [\Pr(G=a) - \Pr(G=b)][\mathrm{E}_a(D) - \mathrm{E}_b(D)][\mathrm{E}_a(\tau) - \mathrm{E}_b(\tau)] + \mathrm{Cov}_a(D,\tau) - \mathrm{Cov}_b(D,\tau)$ shows that this URED variant also fails to separate the contributions of differential prevalence and differential selection.

sorting into college graduation among equally achieving advantaged and disadvantaged students.

The $\boldsymbol{Q}$-distribution component, which appears only in the conditional decomposition, equals the difference between the conditional and unconditional contributions of treatment to the outcomes disparity.[7] Furthermore, the $\boldsymbol{Q}$-distribution component identifies the amount of the disparity associated with (i) the relationship between $G$ and $\boldsymbol{Q}$ and (ii) the relationship between $\{D, \tau\}$ and $\boldsymbol{Q}$. Clearly, if $G \perp\!\!\!\perp \boldsymbol{Q}$ or $\{D, \tau\} \perp\!\!\!\perp \boldsymbol{Q}$, the $\boldsymbol{Q}$-distribution component will be zero. In our application, if there was no group difference in prior test scores or if prior test scores were not associated with college graduation or the effect of college graduation on adult income, then the $\boldsymbol{Q}$-distribution component would be zero.

2.4.1. *Interventional interpretation.* Analogous to the unconditional case, our conditional decomposition has an interventional interpretation that maps a two-step intervention to the conditional prevalence and selection components. Let $\mathrm{E}_g\left(Y^{R(D|g',\boldsymbol{Q})}\right)$ be the mean potential outcome of group $g$ when each of its members $i$ is given a randomly drawn treatment from those members of group $g'$ who have the same values of $\boldsymbol{Q}$ as $i$. We can rewrite this mean potential outcome as:

$$\mathrm{E}_g\left(Y^{R(D|g',\boldsymbol{Q})}\right) = \mathrm{E}_g\left(Y^0\right) + \int \mathrm{E}_g(\tau \mid \boldsymbol{q})\,\mathrm{E}_{g'}(D \mid \boldsymbol{q}) f_g(\boldsymbol{q})\mathrm{d}\boldsymbol{q} \tag{5}$$

and express the components of the conditional decomposition as:

$$\begin{aligned}
\mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \left[\mathrm{E}_a\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|b,\boldsymbol{Q})}\right)\right] &= \text{conditional selection} \\
\mathrm{E}_b\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|b,\boldsymbol{Q})}\right) &= \text{conditional prevalence} \\
\mathrm{E}_a\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|a,\boldsymbol{Q})}\right) &= \text{baseline} + \text{conditional effect} \\
&\quad + \boldsymbol{Q}\text{-distribution}.
\end{aligned} \tag{6}$$

Therefore, the first step of the intervention randomizes the treatment within groups and within $\boldsymbol{Q}$ levels without changing treatment prevalence. In this step, the pre-treatment disparity is $\mathrm{E}_a(Y) - \mathrm{E}_b(Y)$, and the post-intervention disparity is $\mathrm{E}_a\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|b,\boldsymbol{Q})}\right)$. The resulting change in disparity is thus the conditional selection component. The second step equalizes treatment prevalence across groups and within $\boldsymbol{Q}$ levels without changing selection. In this step, the pre-intervention disparity is $\mathrm{E}_a\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|b,\boldsymbol{Q})}\right)$, the post-intervention disparity is $\mathrm{E}_a\left(Y^{R(D|a,\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|a,\boldsymbol{Q})}\right)$, and the change in disparity is the conditional prevalence component. The disparity remaining after this two-step intervention equals the sum of the baseline, conditional effect, and the $\boldsymbol{Q}$-distribution components.[8]

[7]Here, we define the unconditional contribution of treatment as the sum of the prevalence, effect, and selection components in equation (1), and the conditional contribution of treatment as the sum of the corresponding conditional terms given $\boldsymbol{Q}$ in equation (4).

[8]For equation (5) to hold, we only require $\mathrm{supp}_g(\boldsymbol{Q}) \subseteq \mathrm{supp}_{g'}(\boldsymbol{Q})$. This implies that the two-step intervention is well-defined as long as $\mathrm{supp}_b(\boldsymbol{Q}) \subseteq \mathrm{supp}_a(\boldsymbol{Q})$. Intuitively, at each level of $\boldsymbol{Q}$ in group $b$, we must be able to find members of group $a$ with the same $\boldsymbol{Q}$ values in order to conduct the equalization intervention. Note that $\mathrm{supp}_b(\boldsymbol{Q}) \subseteq \mathrm{supp}_a(\boldsymbol{Q})$ is a weaker condition than Assumption 2. However, under the weaker condition, only the pre- and post-intervention disparities are well-defined, not all components in equation (4).

2.4.2. *Comparison with conditional random equalization decomposition.* We limit the comparison of our conditional decomposition to CRED (Jackson, 2021). Analogous to URED in Section 2.3.3, CRED is defined in terms of a one-step intervention that equalizes treatment prevalence within levels of $\boldsymbol{Q}$ by assigning members of the disadvantaged group who have $\boldsymbol{Q}=\boldsymbol{q}$ to random draws of treatment from advantaged group members who have the same $\boldsymbol{q}$. CRED thus decomposes the outcome disparity into the change in disparity resulting from this intervention and a remaining disparity:

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

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

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

Intuitively, within $\boldsymbol{Q}$ levels, CRED's intervention simultaneously equalizes treatment across groups and randomizes treatment in the disadvantaged group. Consequently, unlike our conditional prevalence component, URED's change in disparity is generally non-zero even if there is no difference in treatment prevalence given any $\boldsymbol{Q}$ value, $\mathrm{E}_a(D\mid\boldsymbol{q})=\mathrm{E}_b(D\mid\boldsymbol{q}),\forall\boldsymbol{q}$. CRED does not contain a conditional selection component either, because CRED's intervention splits the contribution of differential conditional selection across its two components.[9]

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

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

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

We develop nonparametric and efficient estimators for our decompositions. These estimators are "one-step" estimators based on the EIFs of the decomposition components, which remove the bias from naive substitution estimators (Bickel et al., 1998; Van der Vaart, 2000; Hines et al., 2022). The estimators contain some nuisance functions, which can be estimated using flexible ML methods coupled with cross-fitting. Under 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 or quadruple robustness properties.

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

[9] Lundberg (2022) also proposes a variant of CRED, whose change in disparity can be written as $\mathrm{E}_a[\mathrm{Cov}_a(D,\tau\mid\boldsymbol{Q})]-\mathrm{E}_b[\mathrm{Cov}_b(D,\tau\mid\boldsymbol{Q})]+[\mathrm{E}_a(D\mid\boldsymbol{q})-\mathrm{E}_b(D\mid\boldsymbol{q})][\mathrm{E}_a(\tau\mid\boldsymbol{q})f_a(\boldsymbol{q})p_b(\boldsymbol{q})+\mathrm{E}_b(\tau\mid\boldsymbol{q})f_b(\boldsymbol{q})p_a(\boldsymbol{q})]\mathrm{d}\boldsymbol{q}$, where $p_g(\boldsymbol{q})=\Pr(G=g\mid\boldsymbol{q})$. Hence, this CRED variant also does not separate the contributions of differential conditional prevalence and differential conditional selection.

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

$$\begin{aligned}
\text{Baseline} &= \xi_{0a} - \xi_{0b} \\
\text{Prevalence} &= \xi_{1ba} - \xi_{0ba} - \xi_{1bb} + \xi_{0bb} \\
\text{Effect} &= \xi_{1aa} - \xi_{0aa} - \xi_{1ba} + \xi_{0ba} \\
\text{Selection} &= \mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \xi_{0a} + \xi_{0b} - \xi_{1aa} + \xi_{0aa} + \xi_{1bb} - \xi_{0bb}.
\end{aligned}$$

Hence, the EIFs and one-step estimators for the decomposition components directly follow from those for $\xi_{dg}$ and $\xi_{dgg'}$.[10][11] Under Assumptions 1, 3, and 4, $\xi_{dg}$ and $\xi_{dgg'}$ can be identified as the following quantities:

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

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

THEOREM 1 (EIF, unconditional decomposition). *Under Assumptions 1, 3, and 4, the EIF of $\xi_{dg}$ is*

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

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

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

*In Appendix C, we also derive the general EIF with survey weights.*[12]

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

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

Each estimator contains two nuisance functions, $\pi(d, \boldsymbol{X}, g)$ and $\mu(d, \boldsymbol{X}, g)$. The estimators are consistent as long as either one of the two nuisance functions is consistently estimated.

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

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

[12] Related EIFs have appeared in prior work. Park and Kang (2023) give the EIF for $\xi_{1g} - \xi_{0g}$. The EIF for $\xi_{0,g} + \xi_{1gg'} - \xi_{0gg'}$ coincides with the EIF for the quantity denoted as $\theta_c$ in Díaz et al. (2021), when pretreatment confounders in the latter are omitted. Neither of these prior works accommodate survey weights.

THEOREM 2 (Double robustness in consistency, unconditional decomposition). *Under Assumptions 1, 3, and 4, either consistent estimation of $\mu(d, \boldsymbol{X}, g)$ or of $\pi(d, \boldsymbol{X}, g)$ is sufficient for the consistency of $\hat{\xi}_{dg}$ and $\hat{\xi}_{dgg'}$.*[13]

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

To study the asymptotic distribution of the one-step estimators for the unconditional decomposition, we invoke three additional assumptions about the nuisance functions, which are the same as the assumptions required for the double ML estimator of the ATE (Kennedy, 2022; Chernozhukov et al., 2018). The consistent estimation of $p_g$ and $\mathrm{E}(D \mid g), \forall g$, is left implicit. We let $\| \cdot \|$ denote the $L_2$-norm.

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

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

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

THEOREM 3 (Asymptotic distributions, unconditional decomposition). *Under Assumptions 1, 3, 4, 5a, 5b, and 5c, the cross-fitted one-step estimators for $\hat{\xi}_{dg}$ and $\hat{\xi}_{dgg'}$ are $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient, i.e., $\sqrt{n}\left(\hat{\xi}_{dg} - \xi_{dg}\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{dg}\right)$, and $\sqrt{n}\left(\hat{\xi}_{dgg'} - \xi_{dgg'}\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{dgg'}\right)$, where $\sigma^2_{dg} := \mathrm{E}[\phi_{dg}(Y, D, \boldsymbol{X}, G)^2]$ and $\sigma^2_{dgg'} := \mathrm{E}[\phi_{dgg'}(Y, D, \boldsymbol{X}, G)^2]$ are the respective semiparametric efficiency bounds.*

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

3.2. *Conditional decomposition.* Relative to the unconditional decomposition, inference for the conditional decomposition requires consideration of one additional generic function:

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

where $(d, g, g', g'')$ denotes any one of eight combinations of treatment status and group memberships. The relationships between components of the conditional decomposition and the

[13]The estimator of our unconditional decomposition is doubly robust with respect to the same two nuisance functions as the classic augmented-inverse-probability-of-treatment-weighting (AIPW) estimator of the ATE (Robins, Rotnitzky and Zhao, 1994). The proof for Theorem 2 is omitted, as it is similar to the proof of the double robustness of the AIPW estimator.

generic functions are as follows:

$$\begin{aligned}
\text{Baseline} &= \xi_{0a} - \xi_{0b} \\
\text{Conditional Prevalence} &= \xi_{1bab} - \xi_{0bab} - \xi_{1bbb} + \xi_{0bbb} \\
\text{Conditional Effect} &= \xi_{1aaa} - \xi_{0aaa} - \xi_{1baa} + \xi_{0baa} \\
\boldsymbol{Q} \text{ Distribution} &= \xi_{1baa} - \xi_{0baa} - \xi_{1bab} + \xi_{0bab} \\
\text{Conditional Selection} &= \mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \xi_{0a} + \xi_{0b} - \xi_{1aaa} + \xi_{0aaa} + \xi_{1bbb} - \xi_{0bbb}.
\end{aligned}$$

Since estimation of $\xi_{dg}$ was discussed in the previous subsection, we now focus on $\xi_{dgg'g''}$. The EIFs, one-step estimators, and their asymptotic distributions for the components of the conditional decomposition will then follow.[14] Under Assumptions 1, 2, 3, and 4, we identify $\xi_{dgg'g''}$ as

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

THEOREM 4 (EIF, conditional decomposition). *Under Assumptions 1, 2, 3, and 4, the EIF of $\xi_{dgg'g''}$ is*

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

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

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

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

This estimator contains five nuisance functions: $p_g(\boldsymbol{Q})$, $\pi(d,\boldsymbol{X},g)$, $\mu(d,\boldsymbol{X},g)$, $\mathrm{E}(D \mid \boldsymbol{Q},g)$, and $\omega(d,\boldsymbol{Q},g)$. As is the case for the unconditional decomposition, consistent estimation of the conditional decomposition does not require that all nuisance functions be consistently estimated. This leads to its quadruple robustness against misspecification.

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

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

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

TABLE 1

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

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

Laan, 2006; Semenova and Chernozhukov, 2021), where the pseudo outcome for each $d$ is defined as

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

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

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

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

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

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

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

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

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

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

## 4. Application.

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

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

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

Speaking to our effect component, there is an active literature on heterogeneity in the effect of college graduation on adult income, although results vary. Brand et al. (2021) and Cheng et al. (2021) found larger effects of college graduation on the incomes of people from disadvantaged backgrounds. By contrast, Zhou (2019), Fiel (2020), and Yu et al. (2021) found no statistically significant heterogeneity in the effect of college completion on income across parental income groups. We note that none of these works evaluated the extent to which income disparities can be attributed to groupwise differential effects of college.

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

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

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

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

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

4.3. *Results.* Figure 1 presents our main results for the components of the unconditional and the conditional decompositions across different models for the nuisance functions (see Appendix Tables A3 and A4 for details). To aid interpretation, Appendix Table A2 additionally reports estimated group-specific means of baseline potential outcomes, treatment proportions, ATEs, and covariances between treatment and treatment effect, as well as group differences in these quantities.

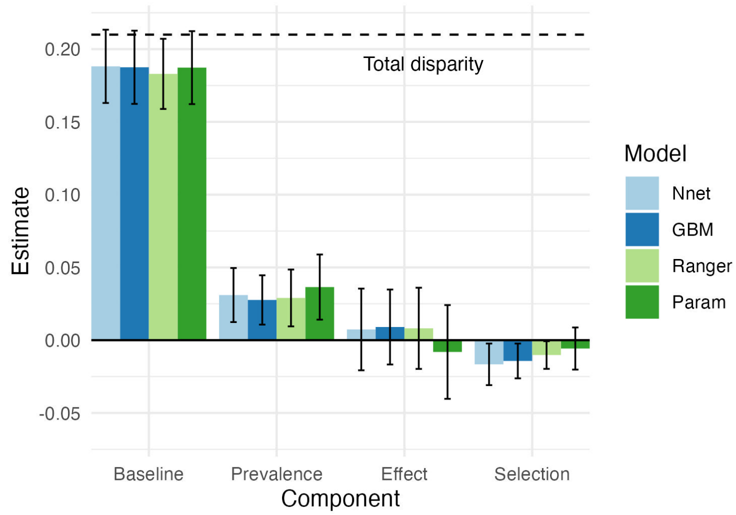
Descriptively, we find that individuals from lower income origin on average achieve 21 percentiles lower incomes in their 30s and 40s than individuals from higher income origin. This confirms the existence of integenerational income persistence and represents the total disparity that we aim to decompose.

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

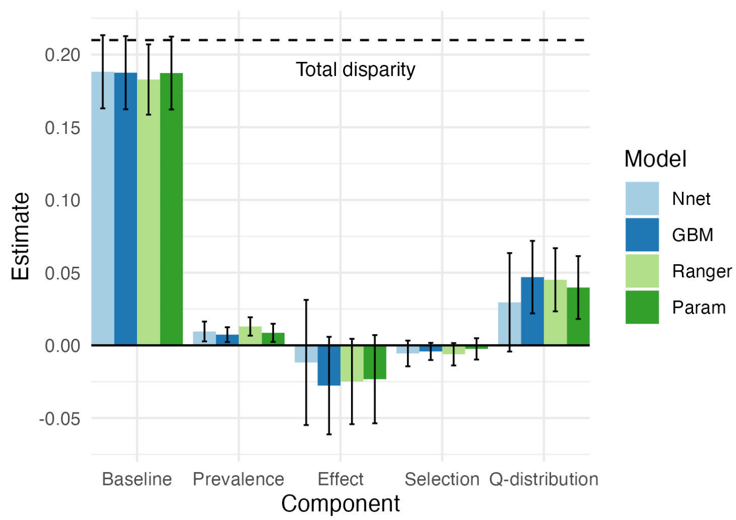
The prevalence component in the unconditional decomposition is positive, substantively large, and statistically significant across all models for the nuisance functions, accounting for about 15% of the total disparity in adult income. Consequently, an intervention to equalize college graduation rates across income-origin groups without changing selection into graduation would reduce the total disparity in adult income by about 15%. In other words, such intervention would decease intergenerational income persistence and increase income mobility. Underlying the prevalence component is the striking inequality in college graduation rates by parental income, as 34% of respondents in the higher-income group, but only 9% of the lower-income group, obtained a college degree by age 29 (see Appendix Table A2).

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

Our estimates for the selection component in the unconditional decomposition are consistently negative. These estimates are statistically significant for the three ML models but not for the parametric models of the nuisance functions. Consequently, randomizing college graduation within each income-origin group to remove selection without changing the prevalence of college graduation would increase the total disparity in adult income by around 7%. Thus, this intervention would increase intergenerational income persistence and decrease income




(a) Unconditional decomposition




(b) Conditional decomposition

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

mobility. Selection into college graduation decreases intergenerational income persistence because selection is positive in the disadvantaged group and negative in the advantaged group (Appendix Table A2). This lends support to stipulations from earlier research that obtaining a college degree is more of a rational decision in pursuit of economic returns among disadvantaged individuals, and more of an adherence to the social norm of college-attendance among advantaged individuals (Mare, 1980; Hout, 2012). This selection component is the central conceptual contribution of our approach and also the most novel finding of our empirical application.

In sum, our unconditional decomposition reveals that college graduation in the United States plays two contradictory roles for intergenerational income persistence. On one hand, higher college graduation rates among high-income origin individuals increase intergenerational income persistence because college graduation increases adult income. On the other hand, part of this prevalence component is offset by our newly identified selection component.

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

In the conditional decomposition, the conditional prevalence component is positive and statistically significant, although it is much smaller than in the unconditional decomposition and only accounts for 3% to 6% of the total disparity. Hence, equalizing chances of graduating college within levels of prior achievement would still somewhat decrease intergenerational income persistence. All estimates for the conditional effect and conditional selection components are negative and statistically insignificant.

The $\boldsymbol{Q}$-distribution component is positive and statistically significant. It reflects the strong associations between AFQT scores ($Q$) on one hand and parental income ($G$), college graduation ($D$), and its effects on income ($\tau$) on the other. The positive $Q$-distribution component also shows that, net of prior test scores, college graduation plays a more limited role in generating intergenerational income persistence.

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

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

In future work, we plan to extend our approach in multiple directions. First, we will develop analogous decompositions for non-binary treatment and group variables, multiple temporally ordered treatments, non-continuous outcomes, and time-to-event outcomes, under the umbrella of a shared two-step intervention. Second, one could employ targeted learning (Van der Laan and Rose, 2011) for improved finite-sample performance for bounded outcomes. Third, it will be important to develop sensitivity analyses and explore alternative identifying assumptions, such as instrumental variables, for applications in which conditional ignorability is unlikely to hold.

**Acknowledgments.** We thank Paul Bauer, Xavier d'Haultfoeuille, Laurent Davezies, Eric Grodsky, Aleksei Opacic, Guanghui Pan, Chan Park, Ben Rosche, Jiwei Zhao, and especially Xiang Zhou for helpful suggestions. We are also grateful for the insightful and thought-provoking comments of AOAS reviewers. Earlier versions of this paper have been presented at GESIS, PAA, ACIC, RC28, JSM, ICLC, CREST, and the Universities of Wisconsin, Heidelberg, and Chicago.

**Funding.** The authors gratefully acknowledge core grants to the Center for Demography and Ecology (P2CHD047873) and to the Center for Demography of Health and Aging (P30AG017266) at UW-Madison, a Romnes Fellowship to Felix Elwert at UW-Madison, and a Wisconsin Partnership Program grant from the UW-Madison School of Medicine and Public Health to Christie Bartels and Felix Elwert.

## SUPPLEMENTARY MATERIAL

### Code for the empirical application

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

### Supplementary appendices

Proofs are contained in Appendices A through E, except for Appendix B. In Appendix B, we present a visualization of the unconditional decomposition. In Appendix F, we discuss the relationship between various concepts of "selection" in the social science literature. Appendix G presents supplemental tables for the empirical application.

## REFERENCES

BAILEY, M. and DYNARSKI, S. (2011). Gains and Gaps: Changing Inequality in U.S. College Entry and Completion Technical Report No. w17633, National Bureau of Economic Research, Cambridge, MA. https://doi.org/10.3386/w17633

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York, NY.

BLAU, P. M. and DUNCAN, O. D. (1967). *The American Occupational Structure*. Free Press.

BLINDER, A. S. (1973). Wage Discrimination: Reduced Form and Structural Estimates. *The Journal of Human Resources* **8** 436. https://doi.org/10.2307/144855

BLOOME, D., DYER, S. and ZHOU, X. (2018). Educational Inequality, Educational Expansion, and Intergenerational Income Persistence in the United States. *American Sociological Review* **83** 1215–1253. https://doi.org/10.1177/0003122418809374

BRADY, D., FINNIGAN, R. M. and HÜBGEN, S. (2017). Rethinking the Risks of Poverty: A Framework for Analyzing Prevalences and Penalties. *American Journal of Sociology* **123** 740–786. https://doi.org/10.1086/693678

BRAND, J. E. and XIE, Y. (2010). Who Benefits Most from College?: Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education. *American Sociological Review* **75** 273–302. https://doi.org/10.1177/0003122410363567

BRAND, J. E., XU, J., KOCH, B. and GERALDO, P. (2021). Uncovering Sociological Effect Heterogeneity Using Tree-Based Machine Learning. *Sociological Methodology* **51** 189–223. https://doi.org/10.1177/0081175021993503

CANEN, N. and SONG, K. (2023). Synthetic Decomposition for Counterfactual Predictions. arXiv:2307.05122 [econ].

CHA, Y., WEEDEN, K. A. and SCHNABEL, L. (2023). Is the Gender Wage Gap Really a Family Wage Gap in Disguise? *American Sociological Review* **88** 972–1001. https://doi.org/10.1177/00031224231212464

CHENG, S., BRAND, J. E., ZHOU, X., XIE, Y. and HOUT, M. (2021). Heterogeneous returns to college over the life course. *Science Advances* **7** eabg7641. https://doi.org/10.1126/sciadv.abg7641

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and LUO, Y. (2018). The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages. *Econometrica* **86** 1911–1938. https://doi.org/10.3982/ECTA14415

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21** C1–C68. https://doi.org/10.1111/ectj.12097

DIDELEZ, V., DAWID, A. P. and GENELETTI, S. (2006). Direct and Indirect Effects of Sequential Treatments. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (R. DECHTER and T. S. RICHARDSON, eds.) 138–146. AUAI Press, Arlington, Virginia.

DIDERICHSEN, F., HALLQVIST, J. and WHITEHEAD, M. (2019). Differential vulnerability and susceptibility: how to make use of recent development in our understanding of mediation and interaction to tackle health inequalities. *International Journal of Epidemiology* **48** 268–274. https://doi.org/10.1093/ije/dyy167

DÍAZ, I., HEJAZI, N. S., RUDOLPH, K. E. and VAN DER LAAN, M. J. (2021). Nonparametric efficient causal mediation with intermediate confounders. *Biometrika* **108** 627–641. https://doi.org/10.1093/biomet/asaa085

FEATHERMAN, D. L. and HAUSER, R. M. (1978). *Opportunity and Change*. *Studies in population*. Academic Press.

FIEL, J. E. (2020). Great Equalizer or Great Selector? Reconsidering Education as a Moderator of Intergenerational Transmissions. *Sociology of Education* 003804072092788. https://doi.org/10.1177/0038040720927886

FORTIN, N., LEMIEUX, T. and FIRPO, S. (2011). Decomposition Methods in Economics. In *Handbook of Labor Economics*, **4** 1–102. Elsevier. https://doi.org/10.1016/S0169-7218(11)00407-2

GENELETTI, S. (2007). Identifying Direct and Indirect Effects in a Non-Counterfactual Framework. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **69** 199–215. https://doi.org/10.1111/j.1467-9868.2007.00584.x

HECKMAN, J. J., HUMPHRIES, J. E. and VERAMENDI, G. (2018). Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking. *Journal of Political Economy* **126** 50.

HECKMAN, J. J., STIXRUD, J. and URZUA, S. (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics* **24** 411–482. https://doi.org/10.1086/504455

HECKMAN, J. J. and VYTLACIL, E. (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica* **73** 669–738. https://doi.org/10.1111/j.1468-0262.2005.00594.x

HINES, O., DUKES, O., DIAZ-ORDAZ, K. and VANSTEELANDT, S. (2022). Demystifying Statistical Learning Based on Efficient Influence Functions. *The American Statistician* 1–13. https://doi.org/10.1080/00031305.2021.2021984

HOLLAND, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association* **81** 945–960.

HOUT, M. (2012). Social and Economic Returns to College Education in the United States. *Annual Review of Sociology* **38** 379–400. https://doi.org/10.1146/annurev.soc.012809.102503

HOWE, C. J., NAPRAVNIK, S., COLE, S. R., KAUFMAN, J. S., ADIMORA, A. A., ELSTON, B., ERON, J. J. and MUGAVERO, M. J. (2014). African American Race and HIV Virological Suppression: Beyond Disparities in Clinic Attendance. *American Journal of Epidemiology* **179** 1484–1492. https://doi.org/10.1093/aje/kwu069

HUBER, M. (2015). Causal Pitfalls in the Decomposition of Wage Gaps. *Journal of Business & Economic Statistics* **33** 179–191. https://doi.org/10.1080/07350015.2014.937437

ISHIDA, H., MULLER, W. and RIDGE, J. M. (1995). Class Origin, Class Destination, and Education: A Cross-National Study of Ten Industrial Nations. *American Journal of Sociology* **101** 145–193. https://doi.org/10.1086/230701

JACKSON, J. W. (2021). Meaningful Causal Decompositions in Health Equity Research: Definition, Identification, and Estimation Through a Weighting Framework. *Epidemiology* **32** 282–290. https://doi.org/10.1097/EDE.0000000000001319

JACKSON, J. W. and VANDERWEELE, T. J. (2018). Decomposition Analysis to Identify Intervention Targets for Reducing Disparities:. *Epidemiology* **29** 825–835. https://doi.org/10.1097/EDE.0000000000000901

KENNEDY, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. Number: arXiv:2203.06469 arXiv:2203.06469 [stat].

KITAGAWA, E. M. (1955). Components of a Difference Between Two Rates. *Journal of the American Statistical Association* **50** 1168–1194.

KLINE, P. (2011). Oaxaca-Blinder as a Reweighting Estimator. *American Economic Review* **101** 532–537. https://doi.org/10.1257/aer.101.3.532

LUNDBERG, I. (2022). The Gap-Closing Estimand: A Causal Approach to Study Interventions That Close Disparities Across Social Categories. *Sociological Methods & Research* 004912412110557. https://doi.org/10.1177/00491241211055769

LUNDBERG, I., JOHNSON, R. and STEWART, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review* **86** 532–565.

MALINSKY, D. (2018). Intervening on structure. *Synthese* **195** 2295–2312. https://doi.org/10.1007/s11229-017-1341-z

MARE, R. D. (1980). Social Background and School Continuation Decisions. *Journal of the American Statistical Association* **75** 295–305.

NAIMI, A. I., SCHNITZER, M. E., MOODIE, E. E. M. and BODNAR, L. M. (2016). Mediation Analysis for Health Disparities Research. *American Journal of Epidemiology* **184** 315–324. https://doi.org/10.1093/aje/kwv329

OAXACA, R. (1973). Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review* **14** 693. https://doi.org/10.2307/2525981

PARK, C. and KANG, H. (2023). A groupwise approach for inferring heterogeneous treatment effects in causal inference. *Journal of the Royal Statistical Society Series A: Statistics in Society* qnad125. https://doi.org/10.1093/jrsssa/qnad125

PARK, S., KANG, S., LEE, C. and MA, S. (2023). Sensitivity analysis for causal decomposition analysis: Assessing robustness toward omitted variable bias. *Journal of Causal Inference* **11** 20220031. https://doi.org/10.1515/jci-2022-0031

PEARL, J. (2001). Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intel ligence* 411–20. Morgan Kaufmann, San Francisco, CA.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American statistical Association* **89** 846–866.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66** 688–701. https://doi.org/10.1037/h0037350

RUBIN, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association* **75** 591–593.

SEMENOVA, V. and CHERNOZHUKOV, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal* **24** 264–289. https://doi.org/10.1093/ectj/utaa027

SUDHARSANAN, N. and BIJLSMA, M. J. (2021). Educational note: causal decomposition of population health differences using Monte Carlo integration and the g-formula. *International Journal of Epidemiology* dyab090. https://doi.org/10.1093/ije/dyab090

VAN DER LAAN, M. J. (2006). Statistical Inference for Variable Importance. *The International Journal of Biostatistics* **2**. https://doi.org/10.2202/1557-4679.1008

VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted learning: Causal Inference for Observational and Experimental Data*. *Springer Series in Statistics*. Springer, New York.

VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. *Cambridge Series on Statistical and Probablistic Mathematics*. Cambridge University Press.

VANDERWEELE, T. J. (2014). A Unification of Mediation and Interaction: A 4-Way Decomposition. *Epidemiology* **25** 749–761. https://doi.org/10.1097/EDE.0000000000000121

VANDERWEELE, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

VANDERWEELE, T. J. and ROBINSON, W. R. (2014). On the Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables. *Epidemiology* **25** 473–484. https://doi.org/10.1097/EDE.0000000000000105

VANDERWEELE, T. J. and TCHETGEN TCHETGEN, E. J. (2014). Attributing Effects to Interactions. *Epidemiology* **25** 711–722. https://doi.org/10.1097/EDE.0000000000000096

WARD, J. B., GARTNER, D. R., KEYES, K. M., FLISS, M. D., MCCLURE, E. S. and ROBINSON, W. R. (2019). How do we assess a racial disparity in health? Distribution, interaction, and interpretation in epidemiological studies. *Annals of Epidemiology* **29** 1–7. https://doi.org/10.1016/j.annepidem.2018.09.007

YAMAGUCHI, K. (2015). Decomposition of Gender or Racial Inequality with Endogenous Intervening Covariates: An Extension of the DiNardo-Fortin-Lemieux Method. *Sociological Methodology* **45** 388–428. https://doi.org/10.1177/0081175015583985

YU, A. (2023). cdgd: Causal Decomposition of Group Disparities R package version 0.3.1.

YU, A., PARK, C., KANG, H. and FLETCHER, J. (2021). Leveraging Machine Learning to Estimate Effect Modification preprint, SocArXiv. https://doi.org/10.31235/osf.io/8ewup

ZHOU, X. (2019). Equalization or Selection? Reassessing the "Meritocratic Power" of a College Degree in Intergenerational Income Mobility. *American Sociological Review* **84** 459–485. https://doi.org/10.1177/0003122419844992

ZIOL-GUEST, K. M. and LEE, K. T. H. (2016). Parent Income–Based Gaps in Schooling: Cross-Cohort Trends in the NLSYs and the PSID. *AERA Open* **2** 233285841664583. https://doi.org/10.1177/2332858416645834

# SUPPLEMENTARY APPENDICES TO NONPARAMETRIC CAUSAL DECOMPOSITION OF GROUP DISPARITIES

BY ANG YU[1,a] AND FELIX ELWERT[1,b]

[1]Department of Sociology, University of Wisconsin-Madison, [a]ayu33@wisc.edu; [b]elwert@wisc.edu

## Appendix A: Proofs for Section 2.

*A.1. Equation (2).* Note that $R(D \mid g')$ denotes a randomly drawn value of treatment $D$ from group $g'$.

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

*A.2. Equivalence Results in Subsection 2.3.2.* For CDE,

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

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

For PIE,

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

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

*A.3. Footnote 6.* We first note that the outcome disparity can also be decomposed as such:

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

where $\mathrm{E}(\tau)$ and $\mathrm{E}(D)$ are the overall ATE and treatment prevalence, and $p_g$ is the proportion of the population in group $g$. And the remaining disparity in Lundberg's (2022) unconditional decomposition can be easily rewritten as $\mathrm{E}_a(Y^0) - \mathrm{E}_b(Y^0) + \mathrm{E}(D)[\mathrm{E}_a(\tau) - \mathrm{E}_b(\tau)]$. It then follows that the change in disparity is what appears in footnote 6.

*A.4. Equation (4).*

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

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

*A.5. Equation (5).*

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

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

*A.6. Equation (6).*

$$
\begin{aligned}
&\mathrm{E}_a(Y) - \mathrm{E}_b(Y) - [\mathrm{E}_a(Y^{R(D \mid a, \boldsymbol{Q})}) - \mathrm{E}_b(Y^{R(D \mid b, \boldsymbol{Q})})] \\
&= \mathrm{E}_a(Y^0) - \mathrm{E}_b(Y^0) + \mathrm{E}_a(D\tau) - \mathrm{E}_b(D\tau) - [\mathrm{E}_a(Y^{R(D \mid a, \boldsymbol{Q})}) - \mathrm{E}_b(Y^{R(D \mid b, \boldsymbol{Q})})] \\
&= \mathrm{E}_a[\mathrm{E}_a(D\tau \mid \boldsymbol{Q})] - \mathrm{E}_a[\mathrm{E}_a(\tau \mid \boldsymbol{Q})\,\mathrm{E}_a(\tau \mid \boldsymbol{Q})] - \{\mathrm{E}_b[\mathrm{E}_b(D\tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{E}_b(\tau \mid \boldsymbol{Q})\,\mathrm{E}_b(\tau \mid \boldsymbol{Q})]\} \\
&= \mathrm{E}_a[\mathrm{Cov}_a(D, \tau \mid \boldsymbol{Q})] - \mathrm{E}_b[\mathrm{Cov}_b(D, \tau \mid \boldsymbol{Q})].
\end{aligned}
$$

Other results in equation (6) follow directly from equation (5).

*A.7. Footnote 9.* The change in disparity of the CRED of Lundberg (2022) is

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

where $Y^{R(D|\boldsymbol{Q})}$ is the potential outcome of an individual when they were given a random draw of $D$ from those in the pooled population who share the same $\boldsymbol{Q}$ value with them.

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

*A.8. Equation (8).* In the main text, we state that the change in disparity in Lundberg's (2022) CRED is $\mathrm{E}_a(Y) - \mathrm{E}_b(Y) - \left[\mathrm{E}_a\left(Y^{R(D|\boldsymbol{Q})}\right) - \mathrm{E}_b\left(Y^{R(D|\boldsymbol{Q})}\right)\right]$. Below, we start from its original form that appears in Lundberg (2022) (see his equation (2)), which is equivalent to the form expressed in randomized intervention notation in our main text.

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

**Appendix B.** With the randomized intervention notation, we present a graphical representation of our unconditional decomposition in Figure A1. The graph visualizes the distinct ways the four components contribute to the observed group disparity in outcome, i.e., one can vary the four components on the graph and obtain different outcome disparities. From this graph, it is clear that the selection component represents the contribution of differential effectiveness of treatment assignment across groups, where effectiveness is defined relative to the random assignment of treatment. Imai and Li (2023) introduce a single-group version of Figure A1. But they do not note the covariance representation of the difference between a treatment assignment rule and the corresponding random assignment with the same treatment prevalence.

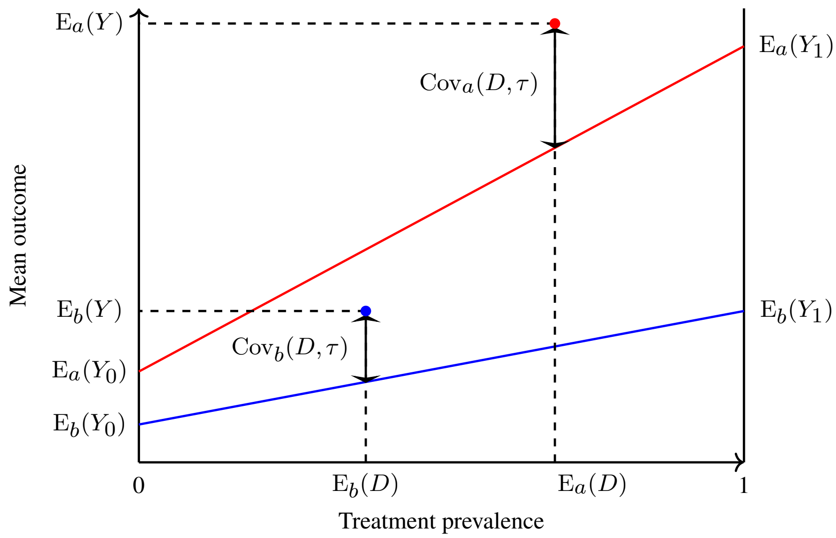



Figure A1. Illustration of the unconditional decomposition. The heights of the red dot and the blue dot respectively indicate the observed mean outcomes in group $a$ and $b$. On the x axis, the position of each dot corresponds to the observed treatment prevalence for the corresponding group. The red line and the blue line respectively represent the mean potential outcomes in group $a$ and $b$ under hypothetically random assignment of the treatment, which vary by the prevalence of the assigned treatment. The lines are straight due to the hypothetical assignment being random. Naturally, when nobody receives the treatment, the height of each line is $\mathrm{E}_g\left(Y_0\right)$; and when everyone receives the treatment, the height of the line is $\mathrm{E}_g(Y_1)$. The slope of each line is hence $\mathrm{E}_g(\tau)$. For a group $g$, the vertical distance between the dot and the line is $\mathrm{E}_g(Y)-\mathrm{E}_g\left(Y^{R(D|g)}\right)=\mathrm{Cov}_g(D,\tau)$.

**Appendix C: 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, $\tilde{o}$. Let subscript $\mathcal{P}_t$ indicate a regular parametric submodel indexed by $t$. The subscript is omitted for the true model. By construction, $f_{\mathcal{P}_t}(o)=t\mathbb{1}_{\tilde{o}}(o)+(1-t)f(o)$, i.e., the submodel is the true model perturbed in the direction of a single observation $\tilde{o}$. Under this construction, the EIF of an estimand, $\xi$, 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, Imbens and Ridder (2003), we assume the survey weight is a known function of the covariates. When no survey weights are needed, $w(\boldsymbol{X},G)$ reduces to 1 for every individual.

In this derivation, we also use the following definitions:

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

*C.1. EIFs for the unconditional decomposition.* First, note that we only need to derive EIFs for two generic functions, $\xi_{dg}:=\mathrm{E}\left(Y^d\frac{w(\boldsymbol{X},g)}{h_g}\mid g\right)$ for an arbitrary group $g$; and

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

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

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

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

We start with the EIF of $\xi_{dg}$.

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

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

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

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

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

Since

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

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

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

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

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

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

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

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

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

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

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

which, without survey weights, becomes

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

Finally, the EIFs for the unconditional decomposition components are

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

*C.2. EIFs for the conditional decomposition.* Similarly to the unconditional case, we focus on the generic function

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

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

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

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

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

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

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

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

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

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

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

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

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

And

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

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

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

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

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

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

Hence,

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

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

which, in the absence of survey weights, simplifies to

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

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

**Appendix D. Asymptotic distribution.** 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 $\xi$. Formally, for all $d, g$ and $\xi$, 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 $\hat{\mu}(d,\boldsymbol{X},g), \hat{\pi}(d,\boldsymbol{X},g), \hat{p}_g(\boldsymbol{Q}), \hat{\omega}(d,\boldsymbol{Q},g)$, and $\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g)$ are fitted using data not in the current subsample, which we implicitly condition on throughout.

*D.1. Inference for the unconditional decomposition.* First, for $\xi_{dg}$, the remainder term is

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

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

The first term is a lower order term. For the second term,

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

where the second inequality uses the Cauchy–Schwarz inequality.

For the empirical process term,

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

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

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

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

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

Also,

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

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

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

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

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

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

For the empirical process term,

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

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

*D.2.1 When $g=g''$.* Note that for the conditional prevalence component, all $\xi$ terms satisfy $g=g''$.

$$
\begin{aligned}
&R_{2,dgg'g''}\\
=&\,\hat{\xi}_{dgg'g}-\xi_{dgg'g}\\
&+\mathbb{P}\left\{-\frac{\mathbb{1}(G=g)}{\hat{p}_g}\hat{\xi}_{dgg'g}\right\}
\end{aligned}
$$

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

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

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

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

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

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

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

*D.2.2 When $g' = g''$.* Note that all $\xi$ terms satisfy $g' = g''$ for the conditional effect component.

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

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

*D.2.3 Summary of convergence rate conditions.* When $g = g' = g''$, which holds for conditional selection, $\|\hat{\omega}(d,\boldsymbol{Q},g)-\omega(d,\boldsymbol{Q},g)\|\left\|\hat{\mathrm{E}}(D\mid\boldsymbol{Q},g)-\mathrm{E}(D\mid\boldsymbol{Q},g)\right\| = o_p(n^{-1/2})$ is sufficient. Hence, we obtain a form of rate double robustness with respect to $\omega(d,\boldsymbol{Q},g)$ and $\mathrm{E}(D\mid\boldsymbol{Q},g)$. Second, when $g = g'' \neq g'$, which holds for conditional prevalence, the following condition is sufficient[1]: $\left\|\hat{\mathrm{E}}(D\mid\boldsymbol{Q},g)-\mathrm{E}(D\mid\boldsymbol{Q},g)\right\| = o_p(n^{-1/2})$. Third, when $g' = g'' \neq g$, a sufficient condition is: $\|\hat{\omega}(d,\boldsymbol{Q},g)-\omega(d,\boldsymbol{Q},g)\| = o_p(n^{-1/2})$. Therefore, the assumption is weaker for conditional selection than other components, which is also the case for Theorem 5.

**Appendix E. Multiple robustness.** Below, we prove the multiple robustness of $\hat{\xi}_{dgg'g''}$, which is stated in Theorem 5. We use a tilde to denote the probability limit of a nuisance estimator, 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 $\hat{\xi}_{dgg'g''}$ converges in probability to

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

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

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

$$=\mathrm{E}\left\{\frac{\mathbb{1}(G=g'')}{p_{g''}}\tilde{\omega}(d,\boldsymbol{Q},g)\tilde{\mathrm{E}}\left(D\mid\boldsymbol{Q},g'\right)\right\}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**Appendix F. Relation between selection concepts.** Various concepts of "selection", "sorting", and "selectivity" have appeared in the social science literature on the effects of education on later-life attainments. Below, we clarify the relationship between our selection concept and those appearing in prior 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 economics (Heckman and Vytlacil, 2005; Heckman, Humphries and Veramendi, 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.

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

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

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

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

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

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

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

## Appendix G. Supplemental tables.

Table A1. Relationship between Baseline Covariates and Outcome Missingness

| Covaraite | | Missing percentage or correlation coefficient | P value |
|---|---|---|---|
| Parental income | | -0.079# | $< 0.001$ |
| Race | White | 25.27 | |
| | Black | 14.45 | $< 0.001$ |
| | Hispanic | 12.38 | |
| Gender | Male | 21.09 | 0.385 |
| | Female | 19.63 | |
| Mother's year of schooling | | 0.003# | 0.870 |
| Parental presence | Yes | 19.31 | 0.089 |
| | No | 22.21 | |
| Number of siblings | | 0.002# | 0.900 |
| Urban residence | Yes | 19.65 | 0.108 |
| | No | 22.86 | |
| Expecting bachelor degree or higher | Yes | 19.28 | 0.250 |
| | No | 21.20 | |
| AFQT score | | 0.021# | 0.294 |
| Age | | -0.015# | 0.453 |
| Friends expecting bachelor degree or higher | Yes | 19.15 | 0.190 |
| | No | 21.31 | |
| Rotter score | | -0.002# | 0.914 |
| Rosenberg score | | -0.013# | 0.517 |
| School satisfaction | | 0.003 | 0.888 |
| Speak foreign language at home | Yes | 18.07 | 0.174 |
| | No | 20.91 | |
| Metropolitan Statistical Area category | Not in SMSA | 22.73 | 0.057 |
| | In SMSA, not central city | 18.00 | |
| | In SMSA, in central city | 22.32 | |
| | In SMSA, central city unknown | 18.59 | |
| Separate from mother | Yes | 25.83 | 0.158 |
| | No | 20.08 | |
| Mother working | Yes | 18.27 | 0.004 |
| | No | 23.00 | |
| SMSA | Northeast | 23.21 | 0.031 |
| | North central | 22.42 | |
| | South | 18.75 | |
| | West | 16.95 | |

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

Table A2. Group-specific Estimates for the Unconditional Decomposition

| | Top 40% income | Bottom 40% income | Top-Bottom |
|---|---|---|---|
| Treatment proportion | 0.337 | 0.086 | 0.251 |
| | (0.310, 0.364) | (0.067, 0.105) | (0.218, 0.284) |
| | | Gradient Boosted Machine | |
| Baseline outcome | 0.599 | 0.411 | 0.188 |
| | (0.581, 0.616) | (0.393, 0.429) | (0.162, 0.213) |
| ATE | 0.137 | 0.110 | 0.027 |
| | (0.100, 0.174) | (0.044, 0.177) | (-0.049, 0.103) |
| $\text{Cov}(D, \tau)$ | -0.011 | 0.004 | -0.014 |
| | (-0.021, 0.000) | (-0.002, 0.009) | (-0.026,-0.002) |
| | | Neural Networks | |
| Baseline outcome | 0.599 | 0.411 | 0.188 |
| | (0.582, 0.617) | (0.393, 0.430) | (0.163,0.213) |
| ATE | 0.145 | 0.123 | 0.022 |
| | (0.103, 0.188) | (0.052, 0.195) | (-0.061, 0.105) |
| $\text{Cov}(D, \tau)$ | -0.014 | 0.002 | -0.017 |
| | (-0.027, -0.002) | (-0.004, 0.009) | (-0.031,-0.002) |
| | | Random Forests | |
| Baseline outcome | 0.593 | 0.410 | 0.183 |
| | (0.578,0.609) | (0.392,0.429) | (0.159,0.207) |
| ATE | 0.140 | 0.116 | 0.024 |
| | (0.107, 0.172) | (0.040, 0.192) | (-0.058, 0.107) |
| $\text{Cov}(D, \tau)$ | -0.006 | 0.004 | -0.010 |
| | (-0.014, 0.001) | (-0.002, 0.010) | (-0.020,-0.001) |
| | | Parametric Regressions | |
| Baseline outcome | 0.598 | 0.410 | 0.187 |
| | (0.580, 0.615) | (0.393, 0.428) | (0.162,0.212) |
| ATE | 0.121 | 0.145 | -0.024 |
| | (0.080, 0.162) | (0.059, 0.232) | (-0.120, 0.072) |
| $\text{Cov}(D, \tau)$ | -0.005 | 0.001 | -0.006 |
| | (-0.017, 0.008) | (-0.006, 0.008) | (-0.020,0.009) |

Note: 95% confidence intervals are in parentheses and are computed according to Theorem 3. Weight stabilization is used. For ML models, cross-fitting is used.

Table A3. Unconditional Decomposition Estimates

| | GBM | Neural Networks | Random Forests | Parametric |
|---|---|---|---|---|
| Total | 0.210 | 0.210 | 0.210 | 0.210 |
| | (0.188,0.232) | (0.188,0.232) | (0.188,0.232) | (0.188,0.232) |
| Baseline | 0.188 | 0.188 | 0.183 | 0.187 |
| | (0.162,0.213) | (0.163,0.213) | (0.159,0.207) | (0.162,0.212) |
| Prevalence | 0.028 | 0.031 | 0.029 | 0.036 |
| | (0.011,0.045) | (0.012,0.050) | (0.009,0.049) | (0.014,0.059) |
| Effect | 0.009 | 0.007 | 0.008 | -0.008 |
| | (-0.017,0.035) | (-0.021,0.035) | (-0.020,0.036) | (-0.040,0.024) |
| Selection | -0.014 | -0.017 | -0.010 | -0.006 |
| | (-0.026,-0.002) | (-0.031,-0.002) | (-0.020,-0.001) | (-0.020,0.009) |
| Change in disparity | 0.024 | 0.029 | 0.025 | 0.035 |
| | (0.005,0.043) | (0.007,0.050) | (0.002,0.048) | (0.008,0.063) |

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

Table A4: Conditional Decomposition Estimates

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

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

In addition, in Table A5, we present a set of estimates for the conditional decomposition as a robustness check. This is motivated by Assumption 6c, which requires parametric convergence rate for either $\hat{\mathrm{E}}(D \mid \boldsymbol{Q}, g)$ or $\hat{\omega}(d, \boldsymbol{Q}, g)$ for the asymptotics of the conditional prevalence and conditional effect components. Thus, to make asymptotic inference more exact, we implement the following procedure. We estimate $\mu(d, \boldsymbol{X}, g)$, $\pi(d, \boldsymbol{X}, g)$, and $p_g(\boldsymbol{Q})$ using ML with cross-fitting. Then we estimate $\mathrm{E}(D \mid \boldsymbol{Q}, g)$ and $\omega(d, \boldsymbol{Q}, g)$ parametrically using linear or logistic regressions without cross-fitting. For the parametric models, we include the group indicator, the AFQT score, the squared AFQT score, and the interactions between the group indicator and the AFQT variables. For $\omega(d, \boldsymbol{Q}, g)$, we apply the pseudo-outcome approach detailed in the main text, but with a different cross-fitting procedure. Here, we use cross-fitted estimates of $\mu(d, \boldsymbol{X}, g)$ and $\pi(d, \boldsymbol{X}, g)$ to construct the pseudo-outcomes, then we estimate $\omega(d, \boldsymbol{Q}, g)$ without cross-fitting. The findings are quantitatively similar and qualitatively identical across Tables A4 and A5.

Table A5: Conditional Decomposition Estimates with Mixed Nonparametric and Parametric Models

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

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

## REFERENCES

BRAND, J. E. and XIE, Y. (2010). Who Benefits Most from College?: Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education. *American Sociological Review* **75** 273–302. https://doi.org/10.1177/0003122410363567

BRAND, J. E., XU, J., KOCH, B. and GERALDO, P. (2021). Uncovering Sociological Effect Heterogeneity Using Tree-Based Machine Learning. *Sociological Methodology* **51** 189–223. https://doi.org/10.1177/0081175021993503

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C. and NEWEY, W. (2017). Double/Debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review* **107** 261–265. https://doi.org/10.1257/aer.p20171038

FIEL, J. E. (2020). Great Equalizer or Great Selector? Reconsidering Education as a Moderator of Intergenerational Transmissions. *Sociology of Education* 003804072092788. https://doi.org/10.1177/0038040720927886

FISHER, A. and KENNEDY, E. H. (2021). Visually Communicating and Teaching Intuition for Influence Functions. *The American Statistician* **75** 162–172. https://doi.org/10.1080/00031305.2020.1717620

HAHN, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica* **66** 315. https://doi.org/10.2307/2998560

HECKMAN, J. J., HUMPHRIES, J. E. and VERAMENDI, G. (2018). Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking. *Journal of Political Economy* **126** 50.

HECKMAN, J. J. and VYTLACIL, E. (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica* **73** 669–738. https://doi.org/10.1111/j.1468-0262.2005.00594.x

HINES, O., DUKES, O., DIAZ-ORDAZ, K. and VANSTEELANDT, S. (2022). Demystifying Statistical Learning Based on Efficient Influence Functions. *The American Statistician* 1–13. https://doi.org/10.1080/00031305.2021.2021984

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica* **71** 1161–1189.

ICHIMURA, H. and NEWEY, W. K. (2022). The influence function of semiparametric estimators. *Quantitative Economics* **13** 29–61. https://doi.org/10.3982/QE826

IMAI, K. and LI, M. L. (2023). Experimental Evaluation of Individualized Treatment Rules. *Journal of the American Statistical Association* **118** 242–256. https://doi.org/10.1080/01621459.2021.1923511

JACKSON, J. W. (2021). Meaningful Causal Decompositions in Health Equity Research: Definition, Identification, and Estimation Through a Weighting Framework. *Epidemiology* **32** 282–290. https://doi.org/10.1097/EDE.0000000000001319

JACKSON, J. W. and VANDERWEELE, T. J. (2018). Decomposition Analysis to Identify Intervention Targets for Reducing Disparities:. *Epidemiology* **29** 825–835. https://doi.org/10.1097/EDE.0000000000000901

KARLSON, K. B. (2019). College as equalizer? Testing the selectivity hypothesis. *Social Science Research* **80** 216–229. https://doi.org/10.1016/j.ssresearch.2018.12.001

KENNEDY, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. Number: arXiv:2203.06469 arXiv:2203.06469 [stat].

LUNDBERG, I. (2022). The Gap-Closing Estimand: A Causal Approach to Study Interventions That Close Disparities Across Social Categories. *Sociological Methods & Research* 004912412110557. https://doi.org/10.1177/00491241211055769

MORGAN, S. L. and WINSHIP, C. (2014). *Counterfactuals and Causal Inference: Methods And Principles For Social Research*, 2nd edition ed. *Analytical Methods for Social Research*. Cambridge University Press, Cambridge, UK.

XIE, Y., BRAND, J. E. and JANN, B. (2012). Estimating Heterogeneous Treatment Effects with Observational Data. *Sociological Methodology* **42** 314–347. https://doi.org/10.1177/0081175012452652

ZHOU, X. (2019). Equalization or Selection? Reassessing the "Meritocratic Power" of a College Degree in Intergenerational Income Mobility. *American Sociological Review* **84** 459–485. https://doi.org/10.1177/0003122419844992