# Disparity Analysis: A Tale of Two Approaches*

Aleksei Opacic — Harvard University
Lai Wei — Princeton University
Xiang Zhou — Harvard University

June 21, 2023

**Abstract**

To understand the patterns and trends of various forms of inequality, quantitative social science research has typically relied on statistical models linking the conditional mean of an outcome of interest to a range of explanatory factors. A prime example of this approach is the widely used Kitagawa-Oaxaca-Blinder (KOB) method. By fitting two linear models separately for an advantaged group and a disadvantaged group, the KOB method decomposes the between-group outcome disparity into two parts, a part explained by group differences in a set of background characteristics and an unexplained part often dubbed "residual inequality." In this paper, we explicate and contrast two distinct approaches to studying group disparities, which we term the *descriptive* approach, as epitomized by the KOB method and its variants, and the *prescriptive* approach, which focuses on how a disparity of interest would change under a hypothetical intervention to one or more manipulable treatments. For the descriptive approach, we propose a generalized KOB decomposition that considers multiple (sets of) explanatory variables sequentially. For the prescriptive approach, we introduce a variety of stylized interventions, such as lottery-type and affirmative-action-type interventions that close between-group gaps in treatment. We illustrate the two approaches to disparity analysis by assessing the Black-White gap in college completion, how it is statistically explained by racial differences in demographic and socioeconomic background, family structure, ability and behavior, and college selectivity, and the extent to which it would be reduced under hypothetical reallocations of college-goers from different racial/economic backgrounds into different tiers of college --- reallocations that could be targeted by race- or class-conscious admissions policies.

*Direct all correspondence to Xiang Zhou, Department of Sociology, Harvard University, 33 Kirkland Street, Cambridge MA 02138; email: xiang_zhou@fas.harvard.edu. An earlier version of this paper was presented at the 2023 Annual Meeting of the Population Association of America. The authors benefitted from the comments of Geoffrey Wodtke and seminar participants at New York University, Princeton University, the University of Hong Kong, and the Chinese University of Hong Kong.

# 1 Introduction

Social and economic inequalities are a striking indictment of American society and a perennial topic of economic, demographic, and sociological research. To understand the patterns and trends of various forms of inequality, quantitative social science research has typically relied on statistical models that link the conditional mean (and, sometimes, the conditional variance) of an outcome of interest, such as logged hourly wage, to a range of explanatory variables, such as education (Katz and Murphy 1992; Juhn et al. 1993), occupation (e.g., Mouw and Kalleberg 2010), and union membership (e.g., DiNardo et al. 1996; Western and Rosenfeld 2011). A prime example of this approach is the Kitagawa-Oaxaca-Blinder (KOB) decomposition method (Kitagawa 1955; Blinder 1973; Oaxaca 1973; see also Duncan 1968; Winsborough and Dickinson 1971; Althauser and Wigler 1972).[1] By fitting two linear regression models, one for an advantaged group (e.g., Whites) and one for a disadvantaged group (e.g., African Americans), it decomposes the between-group outcome disparity into two parts, a part accounted for by group differences in a set of explanatory variables and an unexplained part often dubbed "residual inequality." While the canonical KOB method decomposes the gap in mean outcomes based on linear models, various extensions of the KOB method have also been proposed, including decomposition based on nonlinear outcome models (Fairlie 2005; Bauer and Sinning 2008), decomposition based on models for group membership (Barsky et al. 2002), and decomposition of differences in distributional statistics such as quantiles (Firpo et al. 2018) and the variance (Lemieux 2006; Western and Bloome 2009). The KOB decomposition and its extensions have been widely used to study "categorical inequalities" (Tilly 1998) such as disparities across gender (e.g., Petersen and Morgan 1995; Smith-Doerr et al. 2019), race (e.g., Mandel and Semyonov 2016; Storer et al. 2020), class (Laurison and Friedman 2016), and sexual orientation (Mize 2016).

While the KOB method and its extensions are useful for quantifying the contributions of various factors to a disparity of interest, they do so in an accounting or descriptive sense. Specifically,

[1]For a long time, the KOB method was (and is still by many) called the Oaxaca-Blinder or Blinder-Oaxaca decomposition (Blinder 1973; Oaxaca 1973). In recent years, it has been increasingly referred to as the Kitagawa-Oaxaca-Blinder or Kitagawa-Blinder-Oaxaca decomposition, in recognition of Kitagawa (1955), who proposed the same method in a nonparametric framework. We follow this emerging convention and use the KOB acronym.

because the KOB method does not explicitly address potential confounding of the relationship between the explanatory variables and the outcome, it does not speak to the causal effects of the explanatory variables on the disparity of interest. As such, they are ill-equipped to inform policy. As an illustration, consider a KOB-based analysis that assesses the extent to which the Black-White gap in college completion rates can be statistically explained by racial differences in socioeconomic background, academic preparation, and college selectivity (Ciocca Eller and DiPrete 2018; Voss et al. 2022). While such an analysis may provide useful indications of possible causal processes leading to the completion gap, it does not target a causal estimand concerning the effect of any of the explanatory variables in question. Consequently, it cannot be used to address policy-relevant questions. From a policy standpoint, we may wish to assess the extent to which the Black-White gap in college completion rates would be reduced if college selectivity were equalized by race, by class, or by both — for example, through race- and/or class-conscious admissions policies. Unfortunately, such questions are not targeted by the KOB method or its current extensions.

To address these and other policy-relevant questions concerning inequality, an emerging line of research has begun to incorporate the language and logic of causal inference into the study of group disparities (VanderWeele and Robinson 2014; Jackson and VanderWeele 2018; Zhou 2019; An and Glynn 2021; Jackson 2021; Lundberg 2022; Yu and Elwert 2022; Zhou and Pan 2023). In this line of research, analysts specify one or more explanatory variables as manipulable treatments and then ask how a disparity of interest would change under a hypothetical intervention to these treatments. For example, using data from the National Longitudinal Surveys, Jackson and VanderWeele (2018) investigate how racial disparities in wages and other outcomes would change under several hypothetical interventions that equalize childhood SES, test score, or both by race. Similarly, to assess the equalizing potential of a college degree, Zhou (2019) considers how the level of intergenerational income mobility would change under a hypothetical intervention that sets everyone's educational level to "bachelor's degree." Generalizing the quantities of interest examined in Jackson and VanderWeele (2018) and Zhou (2019), Lundberg (2022) proposes the concept of "gap-closing estimand," which "quantifies how much a gap (e.g., incomes by race) would close if we intervened to equalize a treatment (e.g., access to college)" (p. 1). More recently, extending the gap-closing estimand to a setting with sequential treatments, Zhou and Pan (2023) estimate how much the Black-White earnings gap would be reduced under a series of stylized interventions

to college attendance and college completion.

In this paper, we explicate and contrast the aforementioned two approaches for studying group disparities, which we term the *descriptive* approach, as epitomized by the KOB decomposition and its variants, and the *prescriptive* approach, which focuses on counterfactual disparities under hypothetical interventions to one or more manipulable treatments. Beyond a systematic review of the two approaches, we also suggest extensions that allow us to address a broader range of research questions. For the descriptive approach, we highlight a generalized KOB decomposition that considers multiple (sets of) explanatory variables sequentially. For the prescriptive approach, we introduce new interventional estimands that are more akin to real-world policies than those proposed in previous work, and introduce a novel sensitivity analysis technique for gauging potential bias induced by unobserved confounding. For all estimands in both the descriptive and prescriptive approaches, we present model-free definitions and nonparametric identification results. The nonparametric definition and identification of estimands improves the clarity and transparency of research goals and assumptions (Lundberg et al. 2021), and as we will see, opens the door to a range of parametric and nonparametric estimation strategies, including those leveraging modern nonparametric theory and machine learning methods. Finally, we illustrate the two approaches to disparity analysis by considering the Black-White gap in college completion rates in the US, how it is statistically explained by racial differences in demographic and socioeconomic background, family structure, ability and behavior, and college selectivity, and the extent to which it would be reduced under hypothetical reallocations of college-goers from different racial/economic backgrounds into different tiers of college --- reallocations that could be targeted by race- or class-conscious admissions policies.

## 2 A Motivating Example

In the United States, Black students are much less likely to complete a bachelor's degree relative to their White peers. Among those who started at a four-year college in 2010, for example, 64% of White students graduated with a bachelor's degree within six years, compared with only 40% of Black students (Jeffrey 2020; see also Snyder et al. 2019). The Black-White disparity in college graduation rates has barely changed over the past several decades (Voss et al. 2022), and,

as documented by previous research, it is shaped by a range of factors, including Black students' disadvantages in socioeconomic background (e.g., Bailey and Dynarski 2011), academic preparation (e.g., Bowen et al. 2009), college selectivity (e.g., Reardon et al. 2012), as well as academic performance and social engagement during college (Tinto 1994; Braxton et al. 1997). In a recent study, Ciocca Eller and DiPrete (2018) examine the relative importance of these factors using data from the Educational Longitudinal Study of 2002. These authors apply a nonlinear variant of the KOB decomposition (Fairlie 2005) as well as counterfactual simulations based on several different models of college dropout to quantify the importance of pre-college resources, institutional characteristics, and college experience. These analyses suggest that racial disparities in academic resources prior to and within college constitute the most significant driver of the BA completion gap, and that racial differences in college quality (due to either over- or under-match of Black students) play a much smaller role.

In this article, we revisit the above example using data from the National Longitudinal Survey of Youth, 1997 cohort (NLSY97). Specifically, we illustrate a generalized KOB decomposition that, in this context, considers four sets of explanatory variables successively: (a) demographic and socioeconomic background; (b) family structure; (c) pre-college ability and behavior; and (d) college selectivity. Unlike the model-based decompositions employed by Ciocca Eller and DiPrete (2018), the generalized KOB decomposition is defined nonparametrically, i.e., without reference to a particular linear or logistic regression model. Thus, the estimands and their interpretations are more generic and transparent. Moreover, we highlight that the "contributions" of different variables quantified by the generalized KOB decomposition have no causal interpretations. As such, they do not inform the effectiveness of different institutional interventions on reducing the Black-White completion gap. For example, they cannot tell us the extent to which an equalization in college selectivity by race or by class (either marginally or conditional on some pre-college characteristics) can help reduce the gap. Such questions motivate what we call the *prescriptive* approach to disparity analysis, which we discuss in Section 4.

# 3 Disparity Analysis for Description

In this section, we discuss the descriptive approach to disparity analysis as characterized by the KOB decomposition and its variants. We first review the canonical KOB method based on linear models and then introduce a nonparametric version of the KOB method. Next, we consider the case of multiple explanatory variables and contrast two alternative ways of quantifying their respective contributions to the disparity of interest: a simultaneous approach and a sequential approach. We highlight a fully nonparametric formulation of the sequential approach, which we call a generalized KOB/Duncan decomposition. Finally, we illustrate the generalized KOB/Duncan decomposition by reanalyzing the Black-White gap in college completion.

## 3.1 The Canonical KOB Method

Let $Y$ denote an outcome of interest, $R$ a binary indicator for group membership, and $X$ an explanatory variable for the group disparity in the mean outcome. To motivate our discussion, let us consider how parental income helps explain the Black-White gap in college completion rates among four-year college goers. In this context, $Y$ denotes college completion, $R$ denotes race, and $X$ denotes parental income. The Black-White gap in college completion rates can be written as $\Delta = \mathbb{E}[Y|R = \text{White}] - \mathbb{E}[Y|R = \text{Black}]$. The goal of the KOB method is to decompose this quantity into two components: a component explained by racial differences in parental income and an unexplained or "residual" component.

Assuming we have a representative sample of the population (with or without the aid of sampling weights), the canonical KOB method can be implemented as follows:

1. Compute completion rates separately for Black and White students in the sample, which we denote by $\bar{Y}_B$ and $\bar{Y}_W$;

2. Fit a linear probability model of college completion ($Y$) on parental income ($X$) among Black students;

3. Compute the average predicted value of the outcome when we substitute White students' parental incomes for Black students' in the above model, which we denote by $\bar{Y}_B^{\text{pred}}$.

Then, the sample analog of the Black-White completion gap can be decomposed as

$$\widehat{\Delta} = \underbrace{(\bar{Y}_B^{\text{pred}} - \bar{Y}_B)}_{\widehat{\Delta}_{\text{Explained}}} + \underbrace{(\bar{Y}_W - \bar{Y}_B^{\text{pred}})}_{\widehat{\Delta}_{\text{Unexplained}}}. \tag{1}$$

The first component, $\widehat{\Delta}_{\text{Explained}}$, gauges how the predicted completion rate (according to the linear probability model in step 2) would change if the parental income distribution of Black students shifted to that of White students. The second component, $\widehat{\Delta}_{\text{Unexplained}}$, gauges the remaining difference between the observed completion rate among White students and the predicted completion rate among Black students under this hypothetical shift of their parental income distribution.

In the above algorithm, we fit a model of college completion among Black students and use it to obtain their predicted completion rate under White students' parental income distribution. Alternatively, we can fit a model of college completion among White students and use this model to obtain White students' predicted completion rate under Black students' parental income distribution. Denoting the latter quantity by $\bar{Y}_W^{\text{pred}}$, we have a different partition of $\widehat{\Delta}$:

$$\widehat{\Delta} = \underbrace{(\bar{Y}_W - \bar{Y}_W^{\text{pred}})}_{\widehat{\Delta}_{\text{Explained}}} + \underbrace{(\bar{Y}_W^{\text{pred}} - \bar{Y}_B)}_{\widehat{\Delta}_{\text{Unexplained}}}. \tag{2}$$

Equations (1) and (2) can also be expressed in terms of regression coefficients. If we use $(\hat{\alpha}_B, \hat{\beta}_B)$ and $(\hat{\alpha}_W, \hat{\beta}_W)$ to denote the estimated intercept-slope pairs of the regression models for Black and White students, equations (1) and (2) can be written as

$$\widehat{\Delta} = \underbrace{\hat{\beta}_B'(\bar{X}_W - \bar{X}_B)}_{\widehat{\Delta}_{\text{Explained}}} + \underbrace{(\hat{\alpha}_W - \hat{\alpha}_B) + (\hat{\beta}_W - \hat{\beta}_B)'\bar{X}_W}_{\widehat{\Delta}_{\text{Unexplained}}} \tag{3}$$

$$= \underbrace{\hat{\beta}_W'(\bar{X}_W - \bar{X}_B)}_{\widehat{\Delta}_{\text{Explained}}} + \underbrace{(\hat{\alpha}_W - \hat{\alpha}_B) + (\hat{\beta}_W - \hat{\beta}_B)'\bar{X}_B}_{\widehat{\Delta}_{\text{Unexplained}}}, \tag{4}$$

where $\bar{X}_B$ and $\bar{X}_W$ denote the sample means of parental income among Black and White students, respectively.[2] From these equations, we can see that the two alternative decompositions differ only

[2]We use $\hat{\beta}_B'$ and $\hat{\beta}_W'$ to denote the transposes of the slope coefficients $\hat{\beta}_B$ and $\hat{\beta}_W$. When there is only one explanatory variable (e.g., parental income), $\hat{\beta}_B$ and $\hat{\beta}_W$ are scalars, so their tranposes are themselves. When there are multiple explanatory variables, $\hat{\beta}_B$ and $\hat{\beta}_W$ are column vectors,

in the "reference groups" used to gauge $\Delta_{\text{Explained}}$: while the first decomposition uses $\beta_B$, the second uses $\beta_W$. This difference has motivated many researchers to interpret equations (1) and (2) in terms of distinct "counterfactuals." Specifically, they would regard $\Delta_{\text{Explained}}$ in equation (1) as addressing the question: "How much would the Black completion rate rise if their parental incomes rose to those of White students?" Similarly, $\Delta_{\text{Explained}}$ in equation (2) would be seen as capturing how much the White completion rate would drop if their parental incomes fell to those of Black students (Barsky et al. 2002).

However, intuitive as they may seem, these "counterfactual" interpretations of the KOB decomposition entail an implicit and potentially implausible assumption that the slope coefficients $\beta_B$ and $\beta_W$ capture the causal effects of the explanatory variable on the outcome. In our example, this assumption means that the slope coefficient in a simple linear regression of college completion on parental income among Black/White students captures the causal effect of parental income on college completion in this group. Since the relationship between parental income and college completion can be confounded by a range of other factors, such as parental education and family structure, this assumption seems unrealistic. In sum, because the KOB decomposition is typically implemented via an ad hoc regression model that includes multiple predictors and does not explicitly address potential confounding of any of the predictor-outcome relationships, the counterfactual interpretations outlined above cannot be justified in most applications.

For this reason, we consider the KOB method a purely descriptive approach to analyzing group disparities. In other words, we regard $\Delta_{\text{Explained}}$ in equation (1) or (2) as addressing largely the same question: "How much of the Black-White gap in completion rates can be *statistically* explained by racial differences in parental income?" In fact, if we detach counterfactual interpretations from the KOB decomposition, we may consider equations (3) and (4) as merely two special cases of a more general decomposition:

$$\widehat{\Delta} = \underbrace{\hat{\beta}_R'(\bar{X}_W - \bar{X}_B)}_{\widehat{\Delta}_{\text{Explained}}} + \underbrace{(\hat{\alpha}_W - \hat{\alpha}_B) + (\hat{\beta}_W - \hat{\beta}_R)'\bar{X}_W + (\hat{\beta}_R - \hat{\beta}_B)'\bar{X}_B}_{\widehat{\Delta}_{\text{Unexplained}}}, \tag{5}$$

where $\hat{\beta}_R$ are called the *reference* coefficients (Jann 2008). It is easy to see that equations (3)

and the transpose operation turns them into row vectors.

and (4) correspond to the cases of $\hat{\beta}_R = \hat{\beta}_B$ and $\hat{\beta}_R = \hat{\beta}_W$, respectively. To make the reference coefficients more "neutral," several other choices have been proposed, such as $\hat{\beta}_R = 0.5\hat{\beta}_B + 0.5\hat{\beta}_W$ (Reimers 1983) and $\hat{\beta}_R = p_B\hat{\beta}_B + p_W\hat{\beta}_W$, where $p_B$ and $p_W$ represent the proportions of the Black and White observations in the sample (Cotton 1988).

### 3.2 A Nonparametric Formulation of the KOB Method

We now provide a nonparametric formulation of the KOB method, highlighting the distinction between population-level estimands and their estimators. Without loss of generality, we focus on a nonparametric version of equation (1) and discuss similar generalizations of (2) and (5) in Appendix A. First, note that equation (1) involves only three quantities: $\bar{Y}_B$, $\bar{Y}_W$, and $\bar{Y}_B^{\text{pred}}$. The population counterparts of $\bar{Y}_B$ and $\bar{Y}_W$ are simply the conditional means of the outcome for Black and White students, i.e., $\mathbb{E}[Y|R = \text{Black}]$ and $\mathbb{E}[Y|R = \text{White}]$. Since $\bar{Y}_B^{\text{pred}}$ represents the average predicted outcome among Black students under the White parental income distribution, its population counterpart can be written as

$$\mu_B^{\text{pred}} = \int \mathbb{E}[Y|R = \text{Black}, x]dP(x|R = \text{White}), \tag{6}$$

where $P(x|R = \text{White})$ denotes the parental income distribution among White students. In words, $\mu_B^{\text{pred}}$ represents the conditional mean of college completion given parental income among Black students marginalized over the parental income distribution of White students. Hence, the explained and unexplained components of the KOB decomposition become

$$\Delta_{\text{Explained}} = \mu_B^{\text{pred}} - \mu_B, \tag{7}$$

$$\Delta_{\text{Unexplained}} = \mu_W - \mu_B^{\text{pred}}. \tag{8}$$

Note that to identify equation (6) nonparametrically, we need to be able to compute $\mathbb{E}[Y|R = \text{Black}, x]$ at all possible values of $X$ among White students. Similarly, to identify a nonparametric version of equation (2) (see Appendix A), we need to be able to compute $\mathbb{E}[Y|R = \text{White}, x]$ at all possible values of $X$ among Black students. These conditions require that the support of $X$ (i.e., the values of $X$ with a nonzero density) be the same for each racial group, or, equivalently, the

conditional probability $\Pr[R = \text{Black}|X = x]$ lie strictly between zero and one. This requirement is similar to the positivity assumption that is often invoked to identify the average causal effect of a binary treatment from observational data (see Section 4). Yet, it has received little attention in the KOB literature, where a linear model is typically imposed to extrapolate the values of $\mathbb{E}[Y|R = \text{Black}, x]$ (or $\mathbb{E}[Y|R = \text{White}, x]$) outside the support of $X$.

In the canonical KOB algorithm, if the linear model for $\mathbb{E}[Y|R = \text{Black}, x]$ is correctly specified, then $\widehat{\Delta}_{\text{Explained}}$ and $\widehat{\Delta}_{\text{Unexplained}}$ shown in equation (3) will be consistent estimates of $\Delta_{\text{Explained}}$ and $\Delta_{\text{Unexplained}}$. However, if $\mathbb{E}[Y|R = \text{Black}, x]$ is a nonlinear function of $x$, the canonical KOB decomposition will target a different set of population-level quantities. Compared with a model-based definition of $\Delta_{\text{Explained}}$ and $\Delta_{\text{Unexplained}}$, our definition (equations 7 and 8) is fully nonparametric, i.e., not depending on any statistical model. In other words, they are still meaningful estimands even if the linear model specified in the canonical algorithm is wrong. Thus, the nonparametric formulation decouples the KOB decomposition from parametric models on which it is routinely based, eliminating the model dependency of our estimands.

In the nonparametric decomposition, $\mu_B$ and $\mu_W$ can be estimated directly by their sample analogs ($\bar{Y}_B$ and $\bar{Y}_W$). To estimate $\mu_B^{\text{pred}}$, several strategies can be used. First, note that equation (6) can be written as an iterated expectation:

$$\mu_B^{\text{pred}} = \mathbb{E}\big[\mathbb{E}[Y|R = \text{Black}, X]|R = \text{White}\big],$$

which suggests a regression-imputation (RI) estimator:

$$\hat{\mu}_B^{\text{pred, RI}} = \hat{\mathbb{E}}\big[\hat{\mathbb{E}}[Y|R = \text{Black}, X]|R = \text{White}\big]. \tag{9}$$

To implement the RI estimator, we first fit a regression model for $\mathbb{E}[Y|R, X]$ and then, for all units, obtain their imputed values of the inner expectation (i.e., $\hat{\mathbb{E}}[Y|R = \text{Black}, X]$) by setting race to be Black. Finally, we estimate the outer expectation by averaging these imputed values only among White students. Here, if the inner expectation $\mathbb{E}[Y|R = \text{Black}, X]$ is estimated through a linear model within the subsample of Black students, $\hat{\mu}_B^{\text{pred, RI}}$ will coincide with $\bar{Y}_B^{\text{pred}}$ in the canonical KOB decomposition. However, the generic form of equation (9) suggests that $\mathbb{E}[Y|R = \text{Black}, X]$

need not be estimated via a linear model; we can use a logit or probit model for binary outcomes, a Poisson or negative binomial model for count outcomes, or even a complex machine learning model if the explanatory variable $X$ is high-dimensional.

Second, equation (6) can also be written as the marginal mean of a reweighted outcome:

$$\mu_B^{\text{pred}} = \mathbb{E}\big[\frac{\mathbb{I}(R = \text{Black})}{\Pr[R = \text{White}]} \frac{\Pr[R = \text{White}|X]}{\Pr[R = \text{Black}|X]} Y\big],$$

where $\mathbb{I}(\cdot)$ is an indicator function. The above equation suggests a weighting estimator for $\mu_B^{\text{pred}}$:

$$\hat{\mu}_B^{\text{pred, W}} = \frac{1}{n} \sum_i W_i Y_i, \text{ where } W_i = \frac{\mathbb{I}[R_i = \text{Black}]}{\widehat{\Pr}[R_i = \text{White}]} \frac{\widehat{\Pr}[R_i = \text{White}|X_i]}{\widehat{\Pr}[R_i = \text{Black}|X_i]}. \tag{10}$$

Here $n$ is the sample size and $\widehat{\Pr}[R_i = \text{White}]$ is the sample proportion of White students. We can see that the weighting estimator does not involve fitting an outcome model; instead, it involves fitting a model for group membership (in this case, race) given $X$, which will be used to obtain $\widehat{\Pr}[R_i = \text{Black}|X_i]$ and $\widehat{\Pr}[R_i = \text{White}|X_i]$. The weighting estimator was first proposed by Barsky et al. (2002) for analyzing the Black-White wealth gap. These authors argue that the RI estimator of $\mu_B^{\text{pred}}$ can be highly sensitive to misspecification of the outcome model, and that the weighting estimator may be more robust because it avoids making functional-form assumptions about the $X$-$Y$ relationship. However, as we can see from equation (10), the weighting estimator relies on a correct specification of the group membership model. When this model is misspecified, it can also lead to biased estimates of $\mu_B^{\text{pred}}$.

Finally, to mitigate potential bias due to model misspecification, we can combine RI and weighting to form a "doubly robust" (DR) estimator of $\mu_B^{\text{pred}}$:

$$\hat{\mu}_B^{\text{pred, DR}} = \frac{1}{n} \sum_{i=1}^{n} \big(W_i(Y_i - \hat{\mathbb{E}}[Y_i|\text{Black}, X_i]) + \frac{\mathbb{I}(R_i = \text{White})}{\widehat{\Pr}[R_i = \text{White}]}(\hat{\mathbb{E}}[Y_i|\text{Black}, X_i] - \hat{\mu}_B^{\text{pred, RI}}) + \hat{\mu}_B^{\text{pred, RI}}\big), \tag{11}$$

where $W_i$ denotes the same weights used in the weighting estimator (10). We can see from equation (11) that the DR estimator involves fitting both an outcome model (to obtain $\hat{\mathbb{E}}[Y_i|\text{Black}, X_i]$ and $\hat{\mu}_B^{\text{pred, RI}}$) and a group membership model (to obtain $W_i$). This estimator is doubly robust in the sense that it is consistent if either the outcome model or the group membership model, but not

necessarily both, is correctly specified. The DR estimator (11) is closely related to DR estimators developed for the natural direct effect (NDE) and the average treatment effect on the treated (ATT) in causal inference settings (Tchetgen Tchetgen and Shpitser 2012; Chernozhukov et al. 2018). Yet, to the best of our knowledge, it has not been proposed in the context of KOB decomposition.[3]

### 3.3 The Case of Multiple Explanatory Variables: A Simultaneous Approach

So far, we have dealt with the case of one explanatory variable $X$. Note that $X$ can be vector-valued, in which case the decomposition given by equations (7-8) does not distinguish the contributions of its individual components. For example, if different components of $X$ are indicators of the same theoretical construct (e.g., socioeconomic status captured by both parental education and parental income), we may want to treat them as a whole. However, the researcher often has multiple explanatory variables capturing distinct concepts. For instance, to account for the Black-White completion gap, we may want to isolate the respective contributions of parental income, family structure, academic preparation, and college selectivity. In this and the next subsection, we discuss two alternative strategies for teasing out covariate-specific contributions: a *simultaneous* approach and a *sequential* approach.

To ease exposition, let us now write $X$ as a column vector $X = (X_1, X_2, \dots X_J)'$, where $X_j$ is the $j$th (possibly vector-valued) explanatory variable and $J$ denotes the total number of explanatory variables. In the canonical KOB decomposition, the conditional expectation function $\mathbb{E}[Y|R = r, X]$ is estimated via a race-specific linear model with no interaction terms among the $X_j$'s. In this case, both the explained and unexplained components can be subdivided into covariate-specific contributions. This can be seen by rewriting equation (3) as

$$\widehat{\Delta} = \underbrace{\sum_{j=1}^{J} \hat{\beta}_{B,j}(\bar{X}_{W,j} - \bar{X}_{B,j})}_{\widehat{\Delta}_{\text{Explained}}} + \underbrace{(\hat{\alpha}_W - \hat{\alpha}_B) + \sum_{j=1}^{J}(\hat{\beta}_{W,j} - \hat{\beta}_{B,j})\bar{X}_{W,j}}_{\widehat{\Delta}_{\text{Unexplained}}}. \tag{12}$$

where $\hat{\beta}_{B,j}$ and $\hat{\beta}_{W,j}$ denote the slope coefficients for Black and White students, respectively, and

[3]Kline (2011) shows that the canonical KOB estimator based on a linear model for $\mathbb{E}[Y|R = \text{Black}, X]$ is also doubly robust in the sense that it is consistent for $\mu_B^{\text{pred}}$ if either the linear model for $\mathbb{E}[Y|R = \text{Black}, X]$ is correctly specified or if $\frac{\Pr[R=\text{White}|X]}{\Pr[R=\text{Black}|X]}$ is a linear function of $X$. This condition for double robustness is much more stringent than that for the DR estimator (11).

$\bar{X}_{B,j}$ and $\bar{X}_{W,j}$ denote the corresponding sample means of the $j$th covariate. In this decomposition, $\widehat{\Delta}_{\text{Explained}}$ consists of $J$ components, each representing the contribution of one covariate. $\widehat{\Delta}_{\text{Unexplained}}$, on the other hand, consists of $J+1$ components, one capturing the baseline gap when all the explanatory variables equal zero and the other $J$ components capturing racial differences in the slope coefficients, i.e., the partial effects of $X_j$ on $Y$. In the KOB literature, equation (12) is called a *detailed decomposition*, and the detailed components of $\widehat{\Delta}_{\text{Explained}}$ and $\widehat{\Delta}_{\text{Unexplained}}$ are known as the endowment effects and the coefficient effects, respectively (e.g., Jann 2008). Because the detailed decomposition quantifies the endowment and coefficient effects of each covariate concurrently, it constitutes a *simultaneous approach* to tackling multiple explanatory variables.

While computationally straightforward, the detailed decomposition has several limitations. First, it is not model-free, as its estimands (the endowment and coefficient effects) hinge on the linear and additive specification of the outcome model. In fact, when the outcome model is either nonlinear (e.g., a logit model) or non-additive (e.g., including interaction terms among the covariates), the overall explained and unexplained components (which can still be estimated via the RI method) cannot be neatly partitioned into covariate-specific contributions in the manner of equation (12). Second, it suffers from the so-called "omitted group" problem (Oaxaca and Ransom 1999). Specifically, the coefficient effects that compose $\widehat{\Delta}_{\text{Unexplained}}$ in equation (12) depend on the omitted group, i.e., the group for which all the explanatory variables equal zero. As Fortin et al. (2011, p. 9) note, "[s]ince the choice of the omitted group is arbitrary, the elements of the detailed decomposition can be viewed as arbitrary as well." In other words, the coefficient effects of the detailed decomposition are not only model-dependent but also indeterminate (i.e., unidentified) even under the same model.

Finally, in the detailed decomposition, even the covariate-specific endowment effects — which, unlike the coefficient effects, do not suffer from the omitted group problem — can be hard to interpret, and they may deviate from the researcher's theoretical estimands (Lundberg et al. 2021). This is especially the case if the covariates are temporally or causally ordered. For example, to explain the Black-White completion gap, we can reasonably expect parental income to be causally prior to academic preparation, and academic preparation causally prior to college selectivity. In this case, the coefficient of parental income in the outcome model reflects its partial effect on college completion net of academic preparation and college selectivity. Given that differences in

academic preparation and college selectivity are likely induced by parental income, the contribution of parental income in the detailed decomposition cannot fully capture its cumulative influence on the Black-White completion gap. At most, it can be interpreted as a "direct contribution" of parental income, net of its indirect pathways via academic preparation and college selectivity. Similarly, the contribution of academic preparation in the detailed decomposition can at most be interpreted as its "direct contribution" net of its indirect pathway via college selectivity. However, in such cases, the researcher may not be as interested in such "direct contributions" as in the cumulative contributions of the upstream variables. For example, we may want to assess the overall contribution of Black-White differences in parental income to the Black-White completion gap, whether it stems from the direct influence of parental income during college or its indirect influence via academic preparation or college quality. Unfortunately, this quantity is not targeted by the detailed KOB decomposition. To capture estimands of this type, we need to quantify the contributions of parental income, academic preparation, and college quality sequentially rather than simultaneously — a strategy to which we now turn.

### 3.4 A Generalized KOB/Duncan Decomposition

In an article titled "Inheritance of Poverty or Inheritance of Race?", Duncan (1968) attempted to assess whether the persistence of racial economic inequality across generations is driven by "class" versus "race" effects. Using linear regression models, Duncan conducted a sequence of KOB decompositions for the Black-White income gap, adding one factor at a time: (a) class origin as measured by parental education and occupation; (b) number of siblings; (c) respondent's education; and (d) respondent's occupation. Then, he quantified the net contribution of each factor by comparing the explained components of every two consecutive models. For example, the difference in $\Delta_{\text{Explained}}$ between the first and second models reflects the net contribution of racial differences in the average number of siblings. Although proposed prior to Blinder (1973) and Oaxaca (1973), Duncan's decomposition can be viewed as a generalization of the KOB decomposition. In what follows, we provide a fully nonparametric formulation of Duncan's sequential approach to decomposing group disparities, which we call a generalized KOB/Duncan decomposition.

Following our earlier notation, let $X_1, X_2, \ldots X_J$ denote $J$ ordered sets of covariates that may help explain a disparity of interest. In our running example, we may consider five ordered sets

of covariates: $X_1$ for demographic and socioeconomic background, $X_2$ for family structure, $X_3$ for pre-college measures of ability and behavior, $X_4$ for contextual factors such as peer and school characteristics, $X_5$ for college selectivity. It might be reasoned that these sets of covariates are causally ordered, although this is not required for purely descriptive interpretations of the sequential decomposition, which we detail below.

Let the calligraphic font denote a "cumulative" set of covariates such that $\mathcal{X}_0 = \varnothing$ and $\mathcal{X}_j = \{X_1, X_2, \dots X_j\}$. Define the following "counterfactual" expectations:

$$\mu_{B,j}^{\text{pred}} = \mathbb{E}[\mathbb{E}[Y|R = \text{Black}, \mathcal{X}_j]|R = \text{White}], \quad 0 \leq j \leq J. \tag{13}$$

In words, $\mu_{B,j}^{\text{pred}}$ represents the conditional mean of college completion given the covariate set $\mathcal{X}_j = \{X_1, X_2, \dots X_j\}$ among Black students marginalized over the distribution of these covariates among White students. Then, the Black-White gap in college completion rates can be decomposed as

$$\Delta = \mu_W - \mu_B = \sum_{j=1}^{J} \underbrace{(\mu_{B,j}^{\text{pred}} - \mu_{B,j-1}^{\text{pred}})}_{\Delta_{\text{Explained},X_j}} + \underbrace{\mu_W - \mu_{B,J}^{\text{pred}}}_{\Delta_{\text{Unexplained}}}. \tag{14}$$

Note that since $\mathcal{X}_0 = \varnothing$, $\mu_{B,0}^{\text{pred}} = \mathbb{E}[\mathbb{E}[Y|R = \text{Black}]|R = \text{White}] = \mathbb{E}[Y|R = \text{Black}] = \mu_B$. Here, $\Delta_{\text{Explained},X_j}$ reflects the net contribution of $X_j$ to the completion gap, i.e., the amount of the Black-White completion gap that can be statistically explained by racial differences in $X_j$ after racial differences in $X_1, X_2, \dots X_{j-1}$ are accounted for.

To implement the generalized KOB/Duncan decomposition, we need only estimates of $\mu_W$, $\mu_B$, and $\mu_{B,j}^{\text{pred}}$ for each $j$, which can then be plugged into equation (14) to obtain estimates of $\Delta_{\text{Explained},X_j}$ and $\Delta_{\text{Unexplained}}$. As in the canonical KOB method, $\mu_W$ and $\mu_B$ can be directly estimated by their sample analogs. Thus, the key is to estimate $\mu_{B,j}^{\text{pred}}$. Similar to $\mu_B^{\text{pred}}$ discussed in Section 3.2, $\mu_{B,j}^{\text{pred}}$ can be estimated via RI, weighting, or a doubly robust estimator that combines RI and weighting. First, an RI estimator of $\mu_{B,j}^{\text{pred}}$ can be written as

$$\hat{\mu}_{B,j}^{\text{pred, RI}} = \hat{\mathbb{E}}\big[\hat{\mathbb{E}}[Y|R = \text{Black}, \mathcal{X}_j]|R = \text{White}\big]. \tag{15}$$

Specifically, we first fit a regression model for $\mathbb{E}[Y|R, \mathcal{X}_j]$, and, for all units, obtain their imputed

values of the inner expectation (i.e., $\hat{\mathbb{E}}[Y|R = \text{Black}, \mathcal{X}_j]$) by setting race to be Black. Then, we estimate the outer expectation by averaging these imputed values only among White students. By iterating the above steps for all $j$ ($1 \leq j \leq J$), we obtain all RI estimates of $\mu^{\text{pred}}_{B,j}$, which can then be substituted into equation (14). Note that when all of the inner expectations $\mathbb{E}[Y|R = \text{Black}, \mathcal{X}_j]$ ($1 \leq j \leq J$) are estimated through a linear model within the subsample of Black students, this procedure coincides with Duncan's decomposition.

Second, a weighting estimator of $\mu^{\text{pred}}_{B,j}$ can be written as

$$\hat{\mu}^{\text{pred, W}}_{B,j} = \frac{1}{n}\sum_i W_{ji} Y_i, \quad \text{where } W_{ji} = \frac{\mathbb{I}[R_i = \text{Black}]}{\widehat{\Pr}[R_i = \text{White}]} \frac{\widehat{\Pr}[R_i = \text{White}|\mathcal{X}_{ji}]}{\widehat{\Pr}[R_i = \text{Black}|\mathcal{X}_{ji}]}. \tag{16}$$

Here, $n$ is the sample size and $\mathcal{X}_{ji}$ denotes the observed covariate values of $\mathcal{X}_j = \{X_1, X_2, \ldots X_j\}$ for individual $i$. We can see that this weighting estimator is very similar to the weighting estimator for $\mu^{\text{pred}}_B$ described in Section 3.2, except that the group membership model is now fit as a function of the covariate set $\mathcal{X}_j$ and that this procedure is repeated for each $j$.

Both the RI and weighting estimators involve fitting $J$ different models — an outcome model for each $\mathbb{E}[Y|R, \mathcal{X}_j]$ or a group membership model for each $\Pr[R = \text{Black}|\mathcal{X}_j]$. When any of the $J$ models is misspecified, the corresponding estimator of the explained and unexplained components can be biased. From this perspective, both the RI and weighting estimators are prone to model misspecification bias. To mitigate potential bias, we can combine RI and weighting to form a doubly robust estimator of $\mu^{\text{pred}}_{B,j}$:

$$\hat{\mu}^{\text{pred, DR}}_{B,j} = \frac{1}{n}\sum_{i=1}^{n} \left\{ W_{ji}(Y_i - \hat{\mathbb{E}}[Y_i|R = \text{Black}, \mathcal{X}_{ji}]) + \frac{\mathbb{I}(R_i = \text{White})}{\widehat{\Pr}[R_i = \text{White}]}(\hat{\mathbb{E}}[Y_i|R = \text{Black}, \mathcal{X}_{ji}] - \hat{\mu}^{\text{pred, RI}}_{B,j}) + \hat{\mu}^{\text{pred, RI}}_{B,j} \right\}, \tag{17}$$

where $W_{ji}$ denotes the same weights used in the weighting estimator (16). We can see that the DR estimator involves fitting both an outcome model for $\mathbb{E}[Y|R, \mathcal{X}_j]$ or a group membership model for $\Pr[R = \text{Black}|\mathcal{X}_j]$. It is doubly robust in the sense that it is consistent if either the outcome model or the group membership model, but not necessarily both, is correctly specified. Moreover, this double robustness of $\hat{\mu}^{\text{pred, DR}}_{B,j}$ holds for each $j \in \{1, 2, \ldots J\}$. Thus, if we use the DR estimator for all of the $\mu^{\text{pred}}_{B,j}$'s, the resulting estimators of $\Delta_{\text{Explained},X_j}$ and $\Delta_{\text{Unexplained}}$ will be consistent if for each $j$, either the outcome model or the group membership model is correctly specified.

In this sense, the doubly robust method for the generalized KOB/Duncan decomposition is $2^J$-robust, as the number of conditions for the resulting estimates of the explained and unexplained components to be consistent is $2^J$ (Zhou 2022). Thus, it offers much greater protection from model misspecification bias than the RI and weighting estimators.

More importantly, the doubly robust estimator $\hat{\mu}_{B,j}^{\text{pred, DR}}$ is particularly amenable to the use of machine learning estimates of the outcome and group membership models, in which case our estimates of $\Delta_{\text{Explained},X_j}$ and $\Delta_{\text{Unexplained}}$ will be not only robust to model misspecification but also relatively efficient, especially compared with the weighting estimator. When machine learning methods are used to estimate the outcome and group membership models, it is advisable to use sample splitting to avoid potential bias due to over-fitting. In particular, Chernozhukov et al. (2018) introduce a procedure called "cross-fitting," which involves alternate use of different subsamples to estimate nuisance functions[4] and target parameters. In our context, cross-fitting involves the following steps: (a) randomly partition the sample $S$ into $K$ folds: $S_1, S_2 \ldots S_K$; (b) for each $k$, obtain a fold-specific estimate of $\hat{\mu}_{B,j}^{\text{pred, DR}}$ using only data from $S_k$ ("main sample"), but with the outcome and group membership models learned from the remainder of the sample (i.e., $S \backslash S_k$; "auxiliary sample"); (c) average these fold-specific estimates to form final estimates of $\mu_{B,j}^{\text{pred}}$. This combination of $\hat{\mu}_{B,j}^{\text{pred, DR}}$, machine learning, and cross-fitting is an instance of what Chernozhukov et al. (2018) call "double/debiased machine learning" (DML). Apart from improved robustness and efficiency, another advantage of DML is that it leads to theoretically justified analytical standard errors, obviating the need for computationally intensive inferential methods such as the nonparametric bootstrap.

### 3.5 Illustration

We now illustrate the generalized KOB/Duncan decomposition via a reanalysis of the Black-White gap in college completion rates in the US. We use data from a sample of Black and White respondents who had attended a four-year college by age 25 in the NLSY97. Our outcome ($Y$) is whether the respondent had completed a BA degree by age 29. We consider five sets of explanatory

[4] A nuisance function is a function that is not of our primary interest but needed to estimate our target parameters. For the doubly robust estimator (17), the nuisance functions are the outcome and group membership models, i.e., $\mathbb{E}[Y|R, \mathcal{X}_j]$ and $\Pr[R = \text{Black}|\mathcal{X}_j]$.

variables: $X_1$ for demographic and socioeconomic background (gender, race, ethnicity, age in 1997, parental education, parental income, parental assets, rural residence, southern residence), $X_2$ for family structure (co-residence with both biological parents, presence of a paternal figure), $X_3$ for pre-college measures of ability and behavior (percentile score on the ASVAB test, high school GPA, an index of substance use, an index of delinquency, whether the respondent had any children by age 18), $X_4$ for peer and school-level characteristics (college expectation among peers and three dummy variables denoting whether the respondent ever had property stolen at school, was ever threatened at school, and was ever in a fight at school), and $X_5$ for college selectivity (whether the student attended one of the "most competitive," "highly competitive," and "very competitive" colleges according to Barron's Profile of American Colleges 2000).

In what follows, we focus on DML estimates of $\mu^{\text{pred}}_{B,j}$, $\Delta_{\text{Explained},X_j}$, and $\Delta_{\text{Unexplained}}$ based on equation (17), cross-fitting, and machine learning estimates of the outcome and group membership models.[5] Specifically, we estimate each of the outcome and group membership models using a super learner (van der Laan et al. 2007) composed of Lasso and random forests. Because random forests allow for nonlinear and interaction effects, potential bias due to model misspecification is minimized. In keeping with Chernozhukov et al. (2018), we use five-fold cross-fitting, meaning that $K = 5$. Moreover, following Cole and Hernán (2008), we stabilize our estimates by censoring all inverse-probability-weights ($W_{ji}$ in equation 17) at their 1st and 99th percentiles. Standard errors are constructed using the sample variances of the estimated influence functions and adjusted for multiple imputation via Rubin's (1987) method.

In Figure 1, the first and last rows show the observed completion rates for Black and White students in our sample, which are 0.51 and 0.69, respectively. The intermediate rows report our DML estimates of the "adjusted" completion rates for Black students, i.e. $\mu^{\text{pred}}_{B,j}$ for $j = 1, 2, \ldots 5$. We find that the first three sets of the explanatory variables, i.e., demographic and socioeconomic background, family structure, and pre-college ability and behavior, account for over 90% of the Black-White gap in college completion. Of them, the first (demographic and socioeconomic background) and the third (ability and behavior) sets appear to have the largest explanatory power.

[5]We also obtained estimates of $\mu^{\text{pred}}_{B,j}$, $\Delta_{\text{Explained},X_j}$, and $\Delta_{\text{Unexplained}}$ from the RI, weighting, and doubly robust estimators (equations 15-17) where the outcome and group membership models were both fit via a logistic regression with linear and additive effects of the explanatory variables. The results are qualitatively similar and detailed in Appendix D.

[Figure 1 here]

How would our findings differ if we were to implement the simultaneous KOB decomposition, as discussed in Section 3.3, rather than our proposed sequential approach? To address this question, Figure 2 contrasts estimates obtained under these two approaches. Here, for the generalized KOB/Duncan decomposition, the specific contributions of the five sets of explanatory variables are disaggregated from the incremental results reported in Figure 1. For the simultaneous KOB decomposition, to improve efficiency, we use the slope coefficients from a pooled regression model that includes the group indicator as a predictor. We can see that compared with the simultaneous approach, the generalized KOB/Duncan decomposition yields a larger estimate of the contribution of demographic and socioeconomic background, and smaller estimates of the contributions of family structure, ability and behavior, contextual factors, and college selectivity. This discrepancy occurs because, for upstream variables, the generalized KOB/Duncan decomposition captures their full influence, while the simultaneous decomposition captures only their direct influence net of the downstream variables. These results demonstrate that the widely used simultaneous KOB decomposition can produce misleading results if the researcher is concerned with the temporal/causal ordering of the explanatory variables in question. In fact, while the simultaneous decomposition suggests that college selectivity can explain a non-trivial portion of racial inequality, results from the generalized KOB/Duncan decomposition imply that college selectivity has no additional explanatory power when the upstream variables are accounted for.

[Figure 2 here]

# 4 Disparity Analysis for Prescription

As noted earlier, unless the effects of the explanatory variables on the outcome are unconfounded, estimates from the KOB method and its variants have no causal interpretations. Therefore, any policy implications drawn from such analyses can be misguided. For example, in our empirical illustration, we found that after accounting for other background characteristics, college selectivity does not (statistically) contribute to the Black-White gap in college completion. Given this finding, it might be supposed that the Black-White completion gap cannot be reduced by interventions on college selectivity. However, this is not necessarily the case, because if college selectivity has a

causal effect on degree completion, an intervention that induces more Black students to attend selective colleges should work toward reducing the gap. To quantify the impacts of different types of hypothetical interventions, we introduce a more policy-relevant approach to disparity analysis, which we call the *prescriptive* approach. In what follows, we first outline the key elements of this approach, including the concept of interventional disparity, its identification assumptions, several estimation strategies, and a method for sensitivity analysis. We then introduce some stylized interventions in the context of the Black-White completion gap and illustrate them with the NLSY97 data.

### 4.1 Interventional Disparity

In the prescriptive approach, we specify one or more variables as manipulable treatments and ask how the disparity of interest would change under a hypothetical intervention to these treatments. Without loss of generality, let us consider a binary point-in-time treatment, which we denote by $A$. In our running example, $A$ denotes whether a student attends a selective college. Let $A^*$ and $Y^*$ denote the treatment and outcome variables, respectively, under an intervention of interest. Then, we can define *interventional disparity* ($\Delta^*$) as the difference in means between Black and White students in their interventional outcome $Y^*$:

$$\Delta^* = \mathbb{E}[Y^*|R = \text{White}] - \mathbb{E}[Y^*|R = \text{Black}].$$

Since $Y^*$ denotes the outcome under a hypothetical intervention, it is unobserved. In fact, because the hypothetical intervention would change the treatment status for some units, $Y^*$ involves potential outcomes under treatment conditions that are not realized under the status quo. Below, we outline a set of assumptions under which $\Delta^*$ can be identified from observational data.

### 4.2 Identification and Estimation

Let $X$ denote a set of pretreatment covariates that may confound the treatment-outcome relationship. In our example, $X$ may include the first four sets of explanatory variables described in Section 3.5 (demographic and socioeconomic background, family structure, pre-college ability and behavior, and peer and school-level characteristics). In addition, let $Y(a)$ denote the potential

outcome associated with treatment status $a$. To identify the interventional disparity, we invoke the following four assumptions:

**Assumption 1.** *Consistency:* $Y = Y(A)$.

**Assumption 2.** *System invariance:* $Y^* = Y(A^*)$.

**Assumption 3.** *Unconfoundedness: For any* $a \in \{0, 1\}$, $Y(a) \perp\!\!\!\perp A|X, R$ *and* $Y(a) \perp\!\!\!\perp A^*|X, R$.

**Assumption 4.** *Positivity:* $0 < \Pr[A = 1|X = x, R = r] < 1$ *for all* $x$ *and* $r$.

Assumption 1 (*consistency*) states that a unit's observed outcome under the status quo equals its potential outcome under the observed treatment status. This assumption implies that for any unit and any $a \in \{0, 1)$, $Y(a)$ is fixed, thus ruling out the possibility that a treatment has multiple versions, each corresponding to a different potential outcome. Similarly, it implies that the potential outcomes for any unit do not depend on the treatment assigned to other units. Thus, there should be no interference between units. The consistency assumption is also known as the stable unit treatment value assumption (SUTVA; Rubin 1986) in the causal inference literature. Assumption 2 (*system invariance*) requires that the potential outcomes $Y(0)$ and $Y(1)$ be unaffected by the intervention. In other words, the intervention is not allowed to change a unit's outcome other than through changing its treatment status. This assumption is violated if the intervention affects a unit's potential outcome either "directly" (i.e., via pathways other than changing $A$) or via other units' treatment status (i.e., interference). For example, when analyzing the interventional effect associated with an expansion in selective college attendance, Assumption 2 might be violated if an increasing prevalence of selective college attendance leads to a dilution of resources that each student receives, which may lower the effect of selective college attendance on degree completion. Figure 3 illustrates the conceptual relationship between Assumptions 1 and 2: Assumption 1 bridges potential outcomes and the observed outcome ($Y$), whereas Assumption 2 bridges potential outcomes and the interventional outcome ($Y^*$). Assumption 3 (*unconfoundedness*) means that among units with the same race and covariate values, treatment assignment is as-if random, i.e., independent of potential outcomes, under both the status quo and the intervention. Finally, Assumption 4 (*positivity*) requires that treatment assignment under the status quo is probabilistic at all possible values of race and the covariates.

[Figure 3 here]

Under Assumptions 1-4, the group-specific mean of the interventional outcome can be identified as:

$$\mathbb{E}[Y^*|R=r] = \mathbb{E}_{X|R=r}\mathbb{E}_{A^*|R=r,X}\mathbb{E}[Y|R=r,X,A=A^*] \tag{18}$$

$$= \mathbb{E}[\frac{p^*(A|R,X)}{p(A|R,X)}Y|R=r], \tag{19}$$

where $p^*(\cdot|r,x)$ denotes the probability mass function (PMF) of the interventional treatment $A^*$ given race and the covariates. Since this PMF is a defining characteristic of an intervention, we assume that it is known (i.e., not an unknown quantity that needs to be estimated from data). A proof of equations (18-19) is given in Appendix B.

Equation (18) suggests that we can estimate $\mathbb{E}[Y^*|R=r]$ using a regression-imputation (RI) estimator:

$$\hat{\mathbb{E}}_{\text{RI}}[Y^*|R=r] = \frac{1}{n_r}\sum_{i:R_i=r} \underbrace{\sum_{a=0,1} p^*(a|R_i,X_i)\hat{\mathbb{E}}[Y|R_i,X_i,A=a]}_{\text{predicted outcome for unit } i \text{ under the intervention}} \quad , \tag{20}$$

where $n_r$ denotes the number of individuals in racial group $r$ in the sample. To implement the RI estimator, we first fit a model for the conditional mean of the outcome given race, the covariates, and the treatment. Then, for each unit, we obtain two predicted outcomes, one under treatment and one under control, and compute a weighted average of them where the weights are the conditional probabilities of treatment and control given race and the covariates under the intervention. This weighted average serves as a predicted outcome for this unit under the intervention. Finally, we average these predicted outcomes for each racial group $r$. In our example, the difference between $\hat{\mathbb{E}}_{\text{RI}}[Y^*|R=\text{White}]$ and $\hat{\mathbb{E}}_{\text{RI}}[Y^*|R=\text{Black}]$ constitutes our RI estimator of the interventional disparity $\Delta^*$.

By contrast, equation (19) suggests a weighting estimator of $\mathbb{E}[Y^*|R=r]$:

$$\hat{\mathbb{E}}_{\text{W}}[Y^*|R=r] = \frac{1}{n_r}\sum_{i:R_i=r} W_i Y_i, \quad \text{where } W_i = \frac{p^*(A_i|R_i,X_i)}{\hat{p}(A_i|R_i,X_i)}. \tag{21}$$

Here, the weight $W_i$ is the ratio of the probability of receiving the observed treatment under the

intervention ($p^*(A_i|R_i, X_i)$) to the estimated probability of receiving the observed treatment under the status quo ($\hat{p}(A_i|R_i, X_i)$). The weighting estimator can be viewed as a generalization of the inverse-probability-weighted (IPW) estimator for the mean of a potential outcome, say $\mathbb{E}[Y(a)]$, where the numerator is an indicator function for whether the observed treatment $A$ equals $a$ (see Hernan and Robins 2023 for an introduction to IPW estimators). To understand this connection, note that under Assumptions 1-4, $\mathbb{E}[Y(a)]$ is simply the mean outcome under a deterministic intervention that sets everyone's treatment status to $a$. Also, note that the weighting estimator (21) is very different from the weighting estimator (10) for the KOB decomposition, where we need a group membership model rather than a treatment model.

We can also combine RI and weighting to form a doubly robust estimator of $\mathbb{E}[Y^*|R = r]$ (Lundberg 2022):

$$\hat{\mathbb{E}}_{\text{DR}}[Y^*|R = r] = \frac{1}{n_r} \sum_{i:R_i=r} \left\{ \hat{Y}_i^* + W_i(Y_i - \hat{\mathbb{E}}[Y|R_i, X_i, A_i]) \right\}, \tag{22}$$

where $W_i$ denotes the same weights used in the weighting estimator (21) and $\hat{Y}_i^* = \sum_{a=0,1} p^*(a|R = r, X_i)\hat{\mathbb{E}}[Y|R_i, X_i, A = a]$ denotes the individual-level predicted outcome under the intervention as used in the RI estimator (20). Assuming that the interventional distribution $p^*(\cdot|r, x)$ is known, this estimator is doubly robust — it is consistent if either the outcome model or the treatment model is correctly specified. This assumption requires some qualification. On the one hand, we rarely have a fully specified intervention a priori, and often consider a stylized intervention where $A^*|R, X$ is a function of the observed distribution of $(R, X, A)$. For example, we may want to consider an intervention that sets the treatment probability for Black students to be the same as that for Whites. Yet, the latter is still a population quantity that needs to be estimated from data. In this sense, the interventional distribution $p^*(\cdot|r, x)$ is also unknown, albeit estimable. On the other hand, real-world policies are rarely calibrated using population data, meaning that their interventional distributions are also approximations of their "population counterparts." Thus, we consider it reasonable to use our sample data to formulate an intervention, and, then, by treating it as fixed, evaluate its impact on our disparity of interest. As with the doubly robust estimator for the generalized KOB decomposition, $\hat{\mathbb{E}}_{\text{DR}}[Y^*|R = r]$ is also amenable to the use of DML. When cross-fitting is used, and when the outcome and treatment models are both estimated with

flexible machine learning methods, $\hat{\mathbb{E}}_{\text{DR}}[Y^*|R=r]$ will be $\sqrt{n}$-consistent, asymptotically normal, and semiparametric efficient under relatively mild technical conditions.

### 4.3 Sensitivity Analysis

In practice, due to either data limitations or measurement error, it is unlikely that a researcher is able to capture all confounders of the treatment-outcome relationship to make Assumption 3 (unconfoundedness) hold. If there exists an unobserved confounder that affects both the treatment and the outcome, then the estimated interventional disparity may be biased. Sensitivity analysis is a popular approach to assess potential bias induced by unobserved confounding. We propose an original technique for sensitivity analysis that links the bias in the estimated interventional disparity ($\Delta^*$) to the associations between the unobserved confounder, the treatment, and the outcome. Let us consider an unobserved binary confounder $U$, such as extracurricular excellence, that affects both selective college attendance and BA completion. Under some simplifying assumptions about the nature of the $U$-$A$ and $U$-$Y$ associations, the bias for the estimated post-intervention mean ($\mathbb{E}[Y^*|R=r]$) can be written as (see Appendix C)

$$\alpha_r \beta_r (\Pr[A^*=1|R=r] - \Pr[A=1|R=r]),$$

where $\alpha_r$ denotes the difference in the prevalence of $U$ between selective- and non-selective college goers given race and the covariates (a measure of $U$-$A$ association), and $\beta_r$ denotes the difference in BA completion probability between students with and without $U$ given race, the covariates, and treatment status (a measure of $U$-$Y$ association).

If we further assume that $\alpha_{\text{White}} = \alpha_{\text{Black}} = \alpha$ and that $\beta_{\text{White}} = \beta_{\text{Black}} = \beta$, the bias of the estimated interventional disparity can be written as

$$\text{bias}(\Delta^*) = \alpha\beta C,$$

where $C = (\Pr[A^*=1|\text{White}] - \Pr[A=1|\text{White}]) - (\Pr[A^*=1|\text{Black}] - \Pr[A=1|\text{Black}])$, i.e., the Black-White difference in how the intervention affects treatment prevalence. Similarly, the bias of the estimated "gap reduced" ($\Delta - \Delta^*$) can be written as

$$\text{bias}(\Delta - \Delta^*) = -\alpha\beta C. \tag{23}$$

Since $C$ is directly estimable from the observed data given a user-specified intervention, it is possible to draw a contour plot showing bias-adjusted estimates of $\Delta - \Delta^*$ (or $\Delta^*$) at different possible values of $(\alpha, \beta)$. In such a plot, we can identify how strong the unobserved confounder would need to be, in terms of its association with the treatment ($\alpha$) and the outcome ($\beta$), to reduce our estimated $\Delta - \Delta^*$ to zero. In addition, we can use observed covariates to suggest plausible values for $\alpha$ and $\beta$. For example, if we have an observed binary confounder $Z \in X$, we can fit a linear model of $Y$ on $X$, $R$, and $A$, whose coefficient on $Z$ will provide a plausible value of $\beta$. In the meantime, we can fit a linear model of $Z$ on $R$, $A$, and other components of $X$, whose coefficient on $A$ will provide a plausible value of $\alpha$. By combining these plausible values of $\alpha$ and $\beta$, we can assess the amount of bias that would result if an unobserved variable "worked exactly like" $Z$ in confounding the treatment-outcome relationship. We illustrate these techniques in our empirical illustration below.

### 4.4 Some Stylized Interventions

We now introduce some stylized interventions that could be used to guide real-world policies for reducing disparities. In keeping with our running example, we consider a family of interventions that would equalize college selectivity across race or class background, either marginally or conditionally on some metric of merit or demonstrated ability. Given the limited availability of seats in selective colleges, we focus on interventions that maintain the overall share of students attending selective colleges. Interventions of this type are also less likely to induce violations of system invariance (Lundberg 2022).

First, we could consider a stochastic intervention under which treatment assignment becomes a random draw from the marginal distribution of treatment. For example, if 20% of all Black and White students currently attend a selective college, then whether a student (Black or White) attends a selective college under this intervention will be determined by a random draw from a Bernoulli distribution with probability 0.2. Such an intervention was first introduced by Jackson and VanderWeele (2018) and considered also in Lundberg (2022). Its associated PMF of treatment

can be written as

$$\Pr[A^* = 1|R, X] = \Pr[A = 1].$$

Under this intervention, the prevalence of treatment will be equalized between Black and White students. However, at the individual level, treatment becomes a lottery, and pre-college characteristics such as academic performance is no longer predictive of whether a student attends a selective college or not. Thus, this intervention might be called "lottery-type equalization." Previous research suggests that such lottery-type interventions would be unpopular with both admissions officers and the general public because they completely dismiss the role of merit (Carnevale and Rose 2013).

Alternatively, we may want to equalize treatment across race without breaking the association between treatment and other pretreatment characteristics. Thus, individuals more likely to attend a selective college under the status quo will still be more likely to attend a selective college under the intervention. This can be achieved, for example, through an intervention that multiplies a person's odds of treatment by a race-specific constant such that the prevalence of treatment will be the same between Black and White students and also the same as its overall prevalence under the status quo. Under this intervention, the PMF of treatment can be written as

$$\Pr[A^* = 1|\text{Black}, x] = \frac{e^{\delta_B} \Pr[A = 1|\text{Black}, x]}{1 - \Pr[A = 1|\text{Black}, x] + e^{\delta_B} \Pr[A = 1|\text{Black}, x]}, \tag{24}$$

$$\Pr[A^* = 1|\text{White}, x] = \frac{e^{\delta_W} \Pr[A = 1|\text{White}, x]}{1 - \Pr[A = 1|\text{White}, x] + e^{\delta_W} \Pr[A = 1|\text{White}, x]}, \tag{25}$$

where $\delta_B$ and $\delta_W$ are race-specific constants chosen such that $\Pr^*[A = 1|\text{Black}] = \Pr^*[A = 1|\text{White}]$=$\Pr[A = 1]$. To see why this intervention does not break the association between $X$ and $A$ (conditional on $R$), we can rewrite the above equations in terms of log-odds:

$$\log \frac{\Pr[A^* = 1|r, x]}{\Pr[A^* = 0|r, x]} = \delta_r + \log \frac{\Pr[A = 1|r, x]}{\Pr[A = 0|r, x]}.$$

Thus, in each racial group, this intervention changes everyone's log-odds of treatment by the same amount ($\delta_r$) — without changing the way it depends on $X$; the association between pre-college characteristics and selective college attendance within each racial group is unchanged by the in-

tervention. The functional form characterizing equations (24) and (25) was referred to as an incremental propensity score intervention (IPSI) by Kennedy (2019), who considered the impact of a uniform change in everyone's log-odds of treatment on the mean outcome (i.e., $\mathbb{E}[Y^*]$). Here, the IPSI is applied differentially to Black and White students to achieve racial parity in treatment (see Zhou and Pan 2023 for a recent application). Since this intervention is race-conscious but does not alter the roles of other pretreatment characteristics, it is more akin to real-world affirmation action policies than lottery-type equalization. For this reason, we might call it "affirmation-action-type (AA-type) equalization."

Due to the contested state of race-conscious admissions policies, some scholars have considered the extent to which racial disparities in college attendance, completion, and later-life outcomes can be reduced by class-based affirmative action — as either a substitute for or supplement to race-based affirmative action (e.g., Carnevale and Rose 2013, Alon 2015). In our framework, "AA-type equalization by class" can be operationalized through the following intervention:

$$\Pr[A^* = 1|r, x] = \frac{e^{\delta_0 + \delta_1 s} \Pr[A = 1|r, x]}{1 - \Pr[A = 1|r, x] + e^{\delta_0 + \delta_1 s} \Pr[A = 1|r, x]},$$

where $S$ is an indicator of class background, such as parental income rank. Here, $\delta_0$ and $\delta_1$ are chosen such that (a) the prevalence of treatment under the intervention is the same as that under the status quo, i.e., $\Pr^*[A = 1] = \Pr[A = 1]$, and (b) treatment is no longer correlated with class background, i.e., $\text{Cor}[A^*, S] = 0$. The first of these conditions ensures that the overall share of students attending selective colleges will stay the same, while the second ensures equalization of treatment across students from different class backgrounds. However, to the extent that Black and White students with the same class background may still differ in other predictors of selective college attendance, such as high school GPA, this intervention does not ensure racial parity in treatment.

To achieve both economic and racial parity in treatment while again maintaining the association between $X$ and $A$, we could consider "AA-type equalization by class and race." This can be achieved through the use of race-specific $\delta_0$ and $\delta_1$ parameters:

$$\Pr[A^* = 1|\text{Black}, x] = \frac{e^{\delta_{0B} + \delta_{1B} s} \Pr[A = 1|\text{Black}, x]}{1 - \Pr[A = 1|\text{Black}, x] + e^{\delta_{0B} + \delta_{1B} s} \Pr[A = 1|\text{Black}, x]},$$

$$\Pr[A^* = 1|\text{White}, x] = \frac{e^{\delta_{0W}+\delta_{1W}s}\Pr[A = 1|\text{White}, x]}{1 - \Pr[A = 1|\text{White}, x] + e^{\delta_{0W}+\delta_{1W}s}\Pr[A = 1|\text{White}, x]},$$

where $\delta_{0B}$, $\delta_{1B}$, $\delta_{0W}$, and $\delta_{1W}$ are chosen such that (a) the prevalence of treatment will be the same between Black and White students and also the same as its overall prevalence under the status quo, i.e., $\Pr[A^* = 1|\text{Black}] = \Pr[A^* = 1|\text{White}] = \Pr[A = 1]$, and (b) treatment is uncorrelated with class background within each racial group, i.e., $\text{Cor}[A^*, S|\text{Black}] = \text{Cor}[A^*, S|\text{White}] = 0$. Since conditions (a) and (b) involve four equality constraints (each of them involves two), the four parameters $\delta_{0B}$, $\delta_{1B}$, $\delta_{0W}$, and $\delta_{1W}$ are uniquely identified, and can be solved for numerically.

So far, we have considered only interventions that equalize treatment *marginally*, either by race, by class, or by both. Since marginal equalization often involves a substantial redistribution of opportunity, it could be difficult to achieve due to political resistance from advantaged groups. In such cases, we could also consider less ambitious interventions that equalize treatment not necessarily marginally, but conditional on one or more prespecified covariates. In the context of selective college attendance, we might consider a "conditional equalization" intervention that targets racial parity among students with the same high school GPA. This can be operationalized, for example, through the following post-intervention PMF:

$$\Pr[A^* = 1|R, X] = w\Pr[A = 1|R = \text{Black}, Z] + (1 - w)\Pr[A = 1|R = \text{White}, Z],$$

where $Z$ denotes high school GPA. We can see that under this intervention, the probability of treatment given $R$ and $X$ is now solely a function of high school GPA, which is defined by a weighted average of the conditional probabilities of treatment given high school GPA for Black students and for White students. Here, to maintain the prevalence of treatment at its pre-intervention level, we can choose $w$ such that $\Pr[A^* = 1] = \Pr[A = 1]$. By doing so, we achieve racial parity in selective college attendance conditional on high school GPA while keeping the overall share of students attending selective colleges unchanged.

### 4.5 Illustration

We now illustrate the prescriptive approach to disparity analysis by assessing how much of the Black-White gap in college completion rates would be reduced under the different stylized inter-

ventions outlined above. We use the same sample of the NLSY97 data as described in Section 3.5. However, we now treat college selectivity as a treatment and all the other explanatory variables (demographic and socioeconomic background, family structure, ability and behavior, and peer and school-level characteristics) as pretreatment covariates. In what follows, we focus on DML estimates of the gap reduced, i.e., $\Delta - \Delta^*$, based on equation (22), cross-fitting, and machine learning estimates of the outcome and treatment models.[6] As in Section 3.5, we use five-fold cross-fitting, estimate each of the outcome and treatment models using a super learner composed of Lasso and random forests, and stabilize our estimates by censoring all inverse-probability-weights ($W_i$ in equation 22) at their 1st and 99th percentiles. Standard errors are constructed using the sample variances of the estimated influence functions and adjusted for multiple imputation via Rubin's (1987) method.

Figure 4 shows the estimated reductions in the Black-White completion gap under the five stylized interventions described in Section 4.4: lottery-type equalization, AA-type equalization, AA-type equalization by class, AA-type equalization by class and race, and conditional equalization (given high school GPA). In AA-type equalization by class and AA-type equalization by class and race, class background is measured by the percentile rank of parental income. We can see that different interventions can reduce the gap to varying degrees. The largest reductions are associated with AA-type equalization by race and AA-type equalization by class and race, in which case the gap would be reduced by about 2.6 percentage points (from a baseline of 18 percentage points to 15.4). Lottery-type equalization might reduce the gap by a similar amount, although its estimated effect is not statistically significant due to a larger standard error. By contrast, AA-type equalization by class only would have a much smaller and statistically insignificant effect on the Black-White completion gap. This finding is consistent with previous research showing that economic affirmation action would not improve racial parity in *selective college attendance* in the first place (Carnevale and Rose 2013; Alon 2015). Finally, the estimated reduction associated with conditional equalization (given high school GPA) is also relatively small and statistically insignificant. Overall, our analyses suggest that interventions that target racial parity in selective

[6]We also obtained estimates of the gap reduced from the RI, weighting, and doubly robust estimators based on equations (20-22) where the outcome and treatment models were both fit via a logistic regression with linear and additive effects of the explanatory variables. The results are qualitatively similar and reported in Appendix E.

college attendance can still lead to appreciable reductions in the Black-White completion gap, despite the fact that little of the observed gap can be statistically explained by racial differences in selective college attendance after other characteristics are accounted for (see Section 3.5).

[Figure 4 here]

In order to assess the robustness of the above findings to potential violations of Assumption 3 (*Unconfoundedness*), we implement our sensitivity analysis as discussed in Section 4.3. For illustrative purposes, we focus on our estimate of the gap reduced ($\Delta - \Delta^*$) under AA-type equalization, which is about 0.026. In this case, because $\Pr[A^* = 1|\text{White}] = \Pr[A^* = 1|\text{Black}]$, the parameter $C$ in the bias formula (23) reduces to $\Pr[A = 1|\text{Black}] - \Pr[A = 1|\text{White}]$, which, according to our data, is about $-0.18$. Hence, the amount of bias associated with given values of $(\alpha, \beta)$ can be approximated as $0.18\alpha\beta$, and the corresponding bias-adjusted estimate can be written as $0.026 - 0.18\alpha\beta$, where $\alpha$ and $\beta$ represent the strengths of the $U$-$A$ and $U$-$Y$ relationships conditional on race and other covariates, as defined earlier. Figure 5 shows contours of the bias-adjusted estimates associated with a range of potential values of $\alpha$ and $\beta$. To benchmark these contours, we also show the values of $(\alpha, \beta)$ that would arise if the unobserved binary confounder $U$ behaved like one of two observed binary confounders of the $A$-$Y$ relationship: *urban residence* and *peers' college expectation.*[7] We can see that the original estimate (0.026) can be explained away by unobserved confounding only when both $\alpha$ and $\beta$ are much larger than the sizes implied by these observed confounders. For example, the bias-adjusted estimate would reduce to zero only when both $\alpha$ and $\beta$ of the unobserved confounder are about 10 times that of *peers' college expectation.*

[Figure 5 here]

## 5 Concluding Remarks

In this article, we have explicated, contrasted, and extended two distinct approaches to disparity analysis. In the descriptive approach, all estimands are a function of *observed data* (i.e., not

[7]Among all observed binary confounders of the $A$-$Y$ relationship in our data, *urban residence* and *peers' college expectation* are the two strongest according to their implied values of $\alpha\beta$. *Peers' college expectation* is a dummy variable denoting whether at least 75% of the respondent's peers expected college in 1997.

involving counterfactuals). As with a scatter plot, a variance-covariance analysis (ANOVA), or a linear regression model, a descriptive analysis of disparity is a useful tool for summarizing our data. In some cases, it may also provide useful indications of possible causal processes that can be explored in more detail. Yet, in general, results from a descriptive analysis, no matter what kinds of statistical models have been used to decompose the disparity, have no causal interpretations. Thus, they are not well-suited for answering policy-relevant questions. By contrast, the prescriptive approach directly interrogates the extent to which a disparity of interest would change under a given intervention. Such an approach fosters more rigorous, and less ad hoc, discussions of the potential impacts of different policies.

For both approaches, we demonstrate that one can define and identify all relevant estimands nonparametrically, i.e., without reference to any statistical model. For example, while the canonical KOB decomposition is routinely implemented via linear regression models, the underlying estimands can be defined in an entirely model-free manner. The model-free definition and identification of estimands greatly improves the clarity of research goals and the transparency of the associated assumptions. Moreover, as we have shown, it opens the door to a range of flexible estimation strategies, including regression-imputation, weighting, and doubly robust methods. The doubly robust estimator, in particular, can be combined with flexible machine learning models to produce estimators with desirable statistical properties. To be sure, the separation of the definition of estimands from their identification and estimation should be a goal not only for disparity analysis but for all quantitative social science research (Lundberg et al. 2021).

For both the descriptive and prescriptive approaches, we have suggested several extensions that allow us to examine a broader range of research questions than those enabled by extant methods. For example, in decomposing a disparity with a set of temporally ordered explanatory variables, researchers may wish to assess the overall contribution of some factor, regardless of whether that contribution is "direct" or "flows through" subsequent explanatory variables. Yet, extant methods, such as the simultaneous KOB decomposition, are not geared toward this goal. Our proposed generalized KOB/Duncan decomposition, by contrast, enables researchers to assess the cumulative contributions of different covariates successively. Moreover, our identification and estimation results for the prescriptive approach are general enough for us to evaluate the impacts of a wide range of potential interventions, among which we have discussed a few stylized examples. Finally, for

interventional disparities, we have also proposed and illustrated an original method for assessing the robustness of our estimates to potential unobserved confounding.

The prescriptive approach we discussed can be extended in several ways. First, while the point-in-time interventions we consider align closely with the way most real-world policies are formulated, researchers may be interested in assessing the impact of a bundle of hypothetical interventions implemented at multiple points of an individual's life course. For example, a policy designed to equalize admission rates in postsecondary education may concur with an intervention designed to equalize rates of college completion. Researchers may then want to assess the impact of this dual set of policies on later-life outcomes, such as labor market earnings (Zhou and Pan 2023). Second, while we have focused on "categorical inequalities" across disparate groups, many axes of inequality may be best characterized as lying on a continuum. A prime example is intergenerational income (im)mobility, where the "group" indicator is parental income or a monotone transformation of it. Extending our proposed framework to assess interventional disparities by a continuous characteristic is an important avenue for future research.

## References


Alon, S. (2015). *Race, Class, and Affirmative action*. Russell Sage Foundation.

Althauser, R. P. and Wigler, M. (1972). Standardization and component analysis. *Sociological Methods & Research*, 1(1):97–135.

An, W. and Glynn, A. N. (2021). Treatment effect deviation as an alternative to blinder–oaxaca decomposition for studying social inequality. *Sociological Methods & Research*, 50(3):1006–1033.

Bailey, M. J. and Dynarski, S. M. (2011). Gains and gaps: Changing inequality in us college entry and completion. Technical report, National Bureau of Economic Research.

Barsky, R., Bound, J., Charles, K. K., and Lupton, J. P. (2002). Accounting for the black–white wealth gap: a nonparametric approach. *Journal of the American Statistical Association*, 97(459):663–673.

Bauer, T. K. and Sinning, M. (2008). An extension of the blinder–oaxaca decomposition to nonlinear models. *AStA Advances in Statistical Analysis*, 92:197–206.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, pages 436–455.

Bowen, W. G., Chingos, M. M., and McPherson, M. S. (2009). *Crossing the Finish Line: Completing College at America's Public Universities*. Princeton University Press, Princeton, NJ.

Braxton, J. M., Hirschy, A. S., and McClendon, S. A. (1997). Understanding and reducing college student departure. In *ASHE-ERIC Higher Education Report*, volume 30. San Francisco, CA: Wiley Periodicals.

Carnevale, A. P. and Rose, S. (2013). Socioeconomic status, race/ethnicity, and selective college admissions.

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(1):C1–C68.

Ciocca Eller, C. and DiPrete, T. A. (2018). The paradox of persistence: Explaining the black-white gap in bachelor's degree completion. *American Sociological Review*, 83(6):1171–1214.

Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664.

Cotton, J. (1988). On the decomposition of wage differentials. *The Review of Economics and Statistics*, pages 236–243.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, 64(5):1001–1044.

Duncan, O. D. (1968). Inheritance of poverty or inheritance of race? *On Understanding Poverty*, pages 85–110.

Fairlie, R. W. (2005). An extension of the blinder-oaxaca decomposition technique to logit and probit models. *Journal of Economic and Social measurement*, 30(4):305–316.

Firpo, S. P., Fortin, N. M., and Lemieux, T. (2018). Decomposing wage distributions using recentered influence function regressions. *Econometrics*, 6(2):28.

Fortin, N., Lemieux, T., and Firpo, S. (2011). Decomposition methods in economics. In *Handbook of Labor Economics*, volume 4, pages 1–102. Elsevier.

Hernan, M. A. and Robins, J. M. (2023). *Causal inference: What If*. Boca Raton: Chapman & Hall/CRC.

Jackson, J. W. (2021). Meaningful causal decompositions in health equity research: Definition, identification, and estimation through a weighting framework. *Epidemiology (Cambridge, Mass.)*, 32(2):282.

Jackson, J. W. and VanderWeele, T. J. (2018). Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology*, 29(6):825.

Jann, B. (2008). The blinder–oaxaca decomposition for linear regression models. *The Stata Journal*, 8(4):453–479.

Jeffrey, W. (2020). Crossing the finish line? a review of college completion inequality in the united states by race and class. *Sociology Compass*, 14(5):e12787.

Juhn, C., Murphy, K. M., and Pierce, B. (1993). Wage inequality and the rise in returns to skill. *Journal of Political Economy*, 101(3):410–442.

Katz, L. F. and Murphy, K. M. (1992). Changes in relative wages, 1963–1987: Supply and demand factors. *The Quarterly Journal of Economics*, 107(1):35–78.

Kennedy, E. H. (2019). Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656.

Kitagawa, E. M. (1955). Components of a difference between two rates. *Journal of the American Statistical Association*, 50(272):1168–1194.

Kline, P. (2011). Oaxaca-blinder as a reweighting estimator. *American Economic Review*, 101(3):532–537.

Laurison, D. and Friedman, S. (2016). The class pay gap in higher professional and managerial occupations. *American Sociological Review*, 81(4):668–695.

Lemieux, T. (2006). Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill? *American Economic Review*, 96(3):461–498.

Lundberg, I. (2022). The gap-closing estimand: A causal approach to study interventions that close disparities across social categories. *Sociological Methods & Research*, 0(0):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(3):532–565.

Mandel, H. and Semyonov, M. (2016). Going back in time? gender differences in trends and sources of the racial pay gap, 1970 to 2010. *American Sociological Review*, 81(5):1039–1068.

Mize, T. D. (2016). Sexual orientation in the labor market. *American Sociological Review*, 81(6):1132–1160.

Mouw, T. and Kalleberg, A. L. (2010). Occupations and the structure of wage inequality in the united states, 1980s to 2000s. *American Sociological Review*, 75(3):402–431.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, pages 693–709.

Oaxaca, R. L. and Ransom, M. R. (1999). Identification in detailed wage decompositions. *Review of Economics and Statistics*, 81(1):154–157.

Petersen, T. and Morgan, L. A. (1995). Separate and unequal: Occupation-establishment sex segregation and the gender wage gap. *American Journal of Sociology*, 101(2):329–365.

Reardon, S. F., Baker, R., and Klasik, D. (2012). Race, income, and enrollment patterns in highly selective colleges, 1982-2004.

Reimers, C. W. (1983). Labor market discrimination against hispanic and black men. *The Review of Economics and Statistics*, pages 570–579.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American statistical association*, 81(396):961–962.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley and Sons.

Smith-Doerr, L., Alegria, S., Husbands Fealing, K., Fitzpatrick, D., and Tomaskovic-Devey, D. (2019). Gender pay gaps in us federal science agencies: An organizational approach. *American Journal of Sociology*, 125(2):534–576.

Snyder, T. D., De Brey, C., and Dillow, S. A. (2019). Digest of education statistics 2017, nces 2018-070. *National Center for Education Statistics*.

Storer, A., Schneider, D., and Harknett, K. (2020). What explains racial/ethnic inequality in job quality in the service sector? *American Sociological Review*, 85(4):537–572.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, 40(3):1816.

Tilly, C. (1998). *Durable Inequality*. University of California Press.

Tinto, V. (1994). *Leaving College: Rethinking the Causes and Cures of Student Attrition*. University of Chicago Press, Chicago, IL, 2 edition.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

VanderWeele, T. J. and Robinson, W. R. (2014). On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology (Cambridge, Mass.)*, 25(4):473.

Voss, K., Hout, M., and George, K. (2022). Persistent inequalities in college completion, 1980–2010. *Social Problems*.

Western, B. and Bloome, D. (2009). Variance function regressions for studying inequality. *Sociological Methodology*, 39(1):293–326.

Western, B. and Rosenfeld, J. (2011). Unions, norms, and the rise in us wage inequality. *American Sociological Review*, 76(4):513–537.

Winsborough, H. and Dickinson, P. (1971). Components of negro-white income differences. *Age*, 25(34):35–44.

Yu, A. and Elwert, F. (2022). Causal decomposition of group disparities: The role of within-group heterogeneity. In *PAA 2022 Annual Meeting*. PAA.

Zhou, X. (2019). Equalization or selection? reassessing the "meritocratic power" of a college degree in intergenerational income mobility. *American Sociological Review*, 84(3):459–485.

Zhou, X. (2022). Semiparametric estimation for causal mediation analysis with multiple causally ordered mediators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):794–821.

Zhou, X. and Pan, G. (2023). Higher education and the black-white earnings gap. *American Sociological Review*, 88(1):154–188.

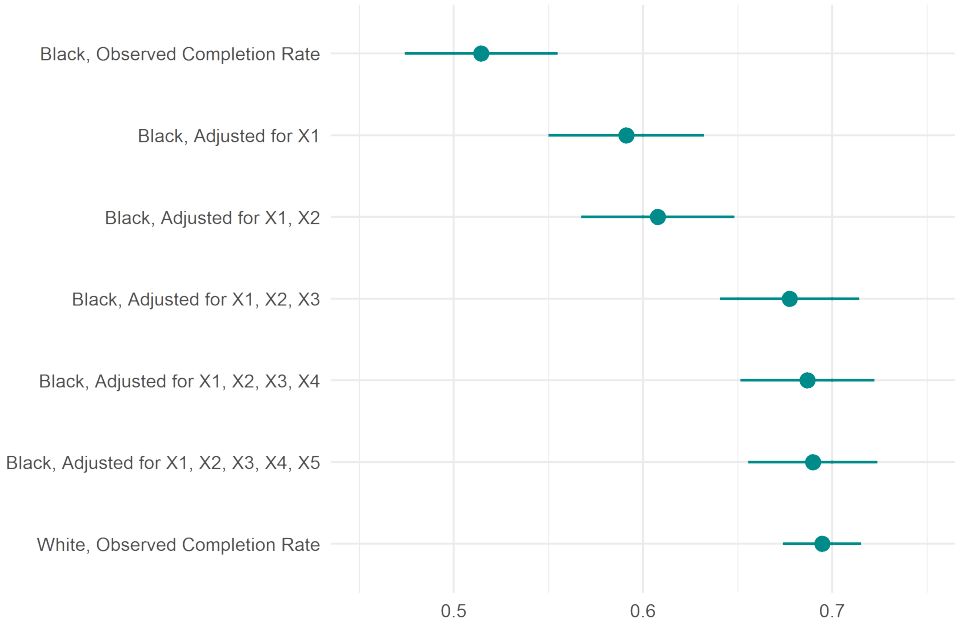


Figure 1: DML estimates of observed and adjusted completion rates.

Note: Data are from NLSY97. $X_1$ denotes demographic and socioeconomic background (gender, race, ethnicity, age in 1997, parental education, parental income, parental assets, rural residence, southern residence), $X_2$ denotes family structure (co-residence with both biological parents, presence of a paternal figure), $X_3$ denotes pre-college measures of ability and behavior (percentile score on the ASVAB test, high school GPA, an index of substance use, an index of delinquency, whether the respondent had any children by age 18), $X_4$ denotes peer and school-level characteristics (college expectation among peers and three dummy variables denoting whether the respondent ever had property stolen at school, was ever threatened at school, and was ever in a fight at school), and $X_5$ denotes college selectivity (whether the student attended one of the "most competitive," "highly competitive," and "very competitive" colleges according to Barron's Profile of American Colleges 2000). In the DML estimation, both the outcome and group membership models are fitted via a super learner composed of Lasso and random forests.

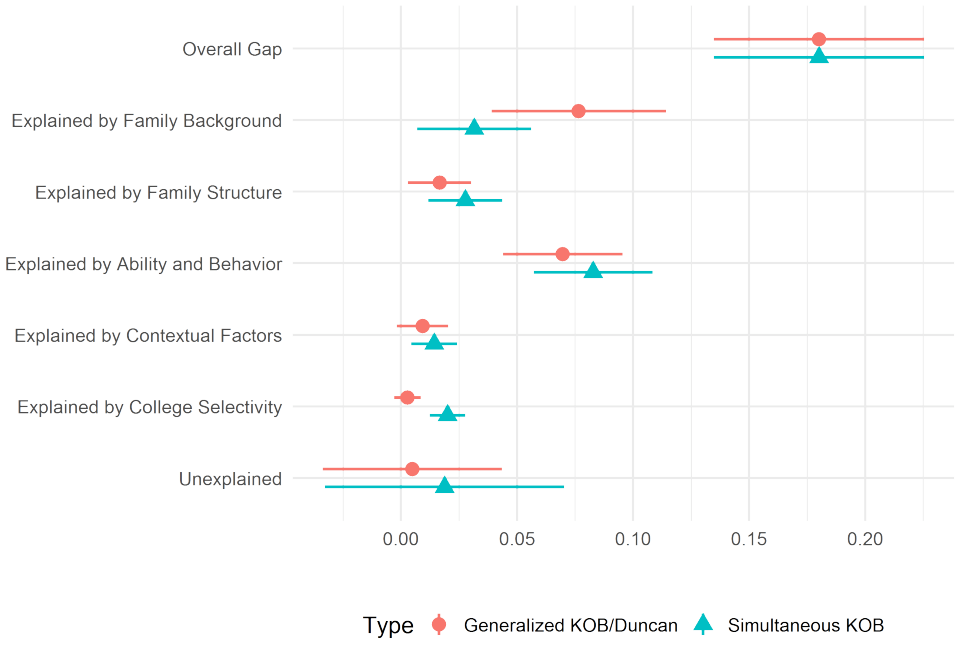


Figure 2: Estimates of the explained and unexplained components of the Black-White gap in college completion under the generalized KOB/Duncan decomposition and the simultaneous KOB decomposition.

Note: Data are from NLSY97. *Family background* denotes a broad set of variables concerning demographic and socioeconomic background, including gender, race, ethnicity, age in 1997, parental education, parental income, parental assets, rural residence, southern residence; *family structure* includes co-residence with both biological parents, presence of a paternal figure; *ability and behavior* includes percentile score on the ASVAB test, high school GPA, an index of substance use, an index of delinquency, whether the respondent had any children by age 18; *contextual factors* include college expectation among peers and three dummy variables denoting whether the respondent ever had property stolen at school, was ever threatened at school, and was ever in a fight at school; *college selectivity* is a dummy variable for whether the student attended one of the "most competitive," "highly competitive," and "very competitive" colleges according to Barron's Profile of American Colleges 2000. In the DML estimation of the generalized KOB/Duncan decomposition, both the outcome and group membership models are fitted via a super learner composed of Lasso and random forests.

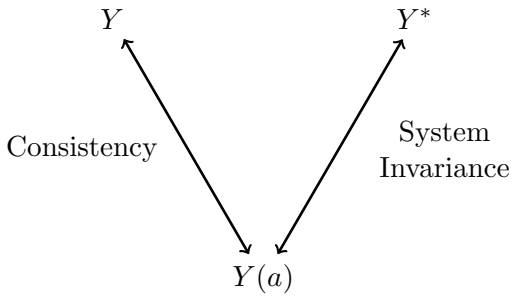


Figure 3: Relationship between Assumption 1 (*Consistency*) and Assumption 2 (*System Invariance*).

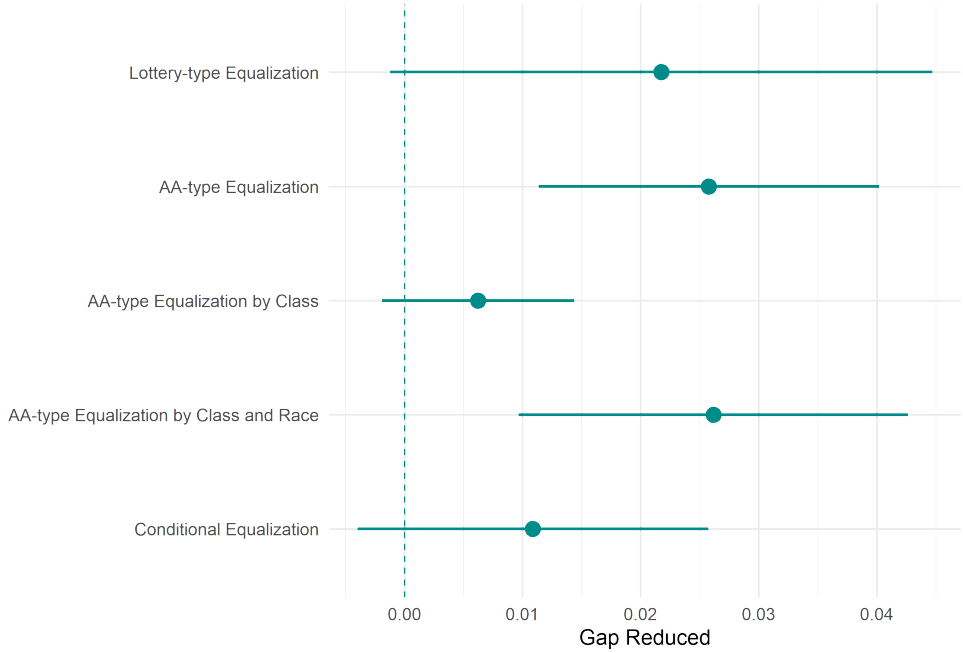


Figure 4: DML estimates of the gap reduced ($\Delta - \Delta^*$) under different stylized interventions to selective college attendance.

Note: Data are from NLSY97. Pretreatment covariates include demographic and socioeconomic background, family structure, ability and behavior, and peer- and school-level characteristics, as detailed in the descriptive analysis. In the DML estimation, both the outcome and treatment models are fitted via a super learner composed of Lasso and random forests. Lottery-type equalization applies the formula $\Pr[A^* = 1|R, X] = \Pr[A = 1]$. AA-type equalization applies the formula $\Pr[A^* = 1|\text{Black}, x] = \frac{e^{\delta_B} \Pr[A=1|\text{Black},x]}{1-\Pr[A=1|\text{Black},x]+e^{\delta_B} \Pr[A=1|\text{Black},x]}$ and $\Pr[A^* = 1|\text{White}, x] = \frac{e^{\delta_W} \Pr[A=1|\text{White},x]}{1-\Pr[A=1|\text{White},x]+e^{\delta_W} \Pr[A=1|\text{White},x]}$, where $\delta_B$ and $\delta_W$ are chosen such that $\Pr[A^* = 1|\text{Black}] = \Pr[A^* = 1|\text{White}] = \Pr[A = 1]$. AA-type equalization by class applies the formula $\Pr[A^* = 1|r, x] = \frac{e^{\delta_0+\delta_1 s} \Pr[A=1|r,x]}{1-\Pr[A=1|r,x]+e^{\delta_0+\delta_1 s} \Pr[A=1|r,x]}$, where $\delta_0$ and $\delta_1$ are chosen such that $\Pr[A^* = 1] = \Pr[A = 1]$ and $\text{Cov}[A^*, S] = 0$. AA-type equalization by class and race applies the formula $\Pr[A^* = 1|\text{Black}, x] = \frac{e^{\delta_{0B}+\delta_{1B} s} \Pr[A=1|\text{Black},x]}{1-\Pr[A=1|\text{Black},x]+e^{\delta_{0B}+\delta_{1B} s} \Pr[A=1|\text{Black},x]}$ and $\Pr[A^* = 1|\text{White}, x] = \frac{e^{\delta_{0W}+\delta_{1W} s} \Pr[A=1|\text{White},x]}{1-\Pr[A=1|\text{White},x]+e^{\delta_{0W}+\delta_{1W} s} \Pr[A=1|\text{White},x]}$, where $\delta_{0B}$, $\delta_{1B}$, $\delta_{0W}$, and $\delta_{1W}$ are chosen such that $\Pr^*[A = 1|\text{Black}] = \Pr^*[A = 1|\text{White}] = \Pr[A = 1]$ and $\text{Cor}[A^*, S|\text{Black}] = \text{Cor}[A^*, S|\text{White}] = 0$. Conditional equalization applies the formula $\Pr[A^* = 1|R, X] = w \Pr[A = 1|R = \text{Black}, Z] + (1 - w) \Pr[A = 1|R = \text{White}, Z]$, where $w$ is chosen such that $\Pr[A^* = 1] = \Pr[A = 1]$.

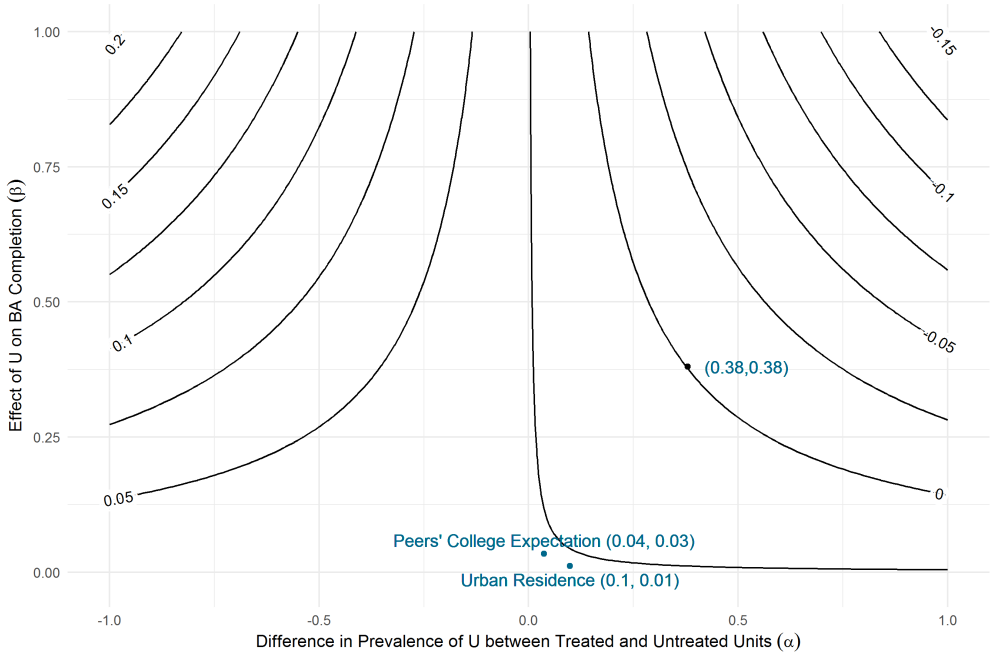


Figure 5: Contour plot showing bias-adjusted estimates of the gap reduced ($\Delta - \Delta^*$) under AA-type equalization as a function of $(\alpha, \beta)$. The two annotated points show values of $(\alpha, \beta)$ if the unobserved confounder $U$ behaved like one of two observed binary confounders in the NLSY97.

# A Nonparametric Formulations of Equations (2) and (5)

The decomposition in the form of equation (2) involves fitting an outcome model among White students, which is then used to obtain the average predicted outcome among White students under the Black covariate distribution, $\bar{Y}_W^{\text{pred}}$. The population counterpart of $\bar{Y}_W^{\text{pred}}$ can be written as

$$\mu_W^{\text{pred}} = \int \mathbb{E}[Y|R = \text{White}, x] dP(x|R = \text{Black}), \tag{26}$$

where $P(x|R = \text{Black})$ denotes the covariate distribution among Black students. Consequently, the explained and unexplained components of the KOB decomposition become

$$\Delta_{\text{Explained}} = \mu_W - \mu_W^{\text{pred}}, \tag{27}$$

$$\Delta_{\text{Unexplained}} = \mu_W^{\text{pred}} - \mu_B. \tag{28}$$

The explained component is the difference between the average observed outcome of White students and the average predicted outcome of White students when their covariate distribution is adjusted to that of Black students. The unexplained component is the difference between the average predicted outcome of White students after the adjustment and the average observed outcome of Black students.

The more general form of the canonical KOB decomposition shown in equation (5) can also be defined nonparametrically. Consider a weighted average of the conditional expectations of the outcome given the covariates for Black students and for White students, $\mu_R(X) = q_B \mathbb{E}[Y|R = \text{Black}, X] + q_W \mathbb{E}[Y|R = \text{White}, X]$, where $q_B + q_W = 1$. The function $\mu_R(X)$ can be seen as a composite conditional expectation function (CEF). The $X$-$Y$ relationship encoded in this composite CEF is analogous to the reference coefficients $\beta_R$ in equation (5). Accordingly, a nonparametric counterpart of equation (5) can be written as

$$\Delta_{\text{Explained}} = \int \mu_R(x) \big( dP(x|R = \text{White}) - dP(x|R = \text{Black}) \big), \tag{29}$$

$$\Delta_{\text{Unexplained}} = \int (\mu_W(x) - \mu_R(x)) dP(x|R = \text{White}) + \int (\mu_R(x) - \mu_W(x)) dP(x|R = \text{Black}). \tag{30}$$

Clearly, this decomposition reduces to equations (7-8) when $q_B = 1$ and equations (27-28) when $q_W = 1$.

## B Identification of $\mathbb{E}[Y^*|R=r]$ under Assumptions 1-4

$$
\begin{aligned}
\mathbb{E}[Y^*|R=r] &= \int\int \mathbb{E}[Y^*|R=r, X=x, A^*=a] p^*(a|R=r,x) p(x|R=r) da dx \\
&\underbrace{=}_{A.2} \int\int \mathbb{E}[Y(a)|R=r, X=x, A^*=a] p^*(a|R=r,x) p(x|R=r) da dx \\
&\underbrace{=}_{A.3} \int\int \mathbb{E}[Y(a)|R=r, X=x] p^*(a|R=r,x) p(x|R=r) da dx \\
&\underbrace{=}_{A.3} \int\int \mathbb{E}[Y(a)|R=r, X=x, A=a] p^*(a|R=r,x) p(x|R=r) da dx \\
&\underbrace{=}_{A.1} \int\int \mathbb{E}[Y|R=r, X=x, A=a] p^*(a|R=r,x) p(x|R=r) da dx \\
&= \mathbb{E}_{X|R=r} \mathbb{E}_{A^*|R=r,X} \mathbb{E}[Y|R=r, X, A=A^*] \\
&= \int\int \frac{p^*(a|R=r,x)}{p(a|R=r,x)} \mathbb{E}[Y|R=r, X=x, A=a] p(a|R=r,x) p(x|R=r) da dx \\
&= \mathbb{E}[\frac{p^*(A|R,X)}{p(A|R,X)} Y|R=r]
\end{aligned}
\tag{31}
$$

## C Derivation of Bias Formulas for Sensitivity Analysis

Assume we have a binary unobserved confounder, $U$, for the treatment-outcome relationship. We further assume that (i) $\alpha_r \stackrel{\Delta}{=} \Pr[U=1|x,r,A=1] - \Pr[U=1|x,r,A=0]$ does not depend on $x$ and that (ii) $\beta_r \stackrel{\Delta}{=} \mathbb{E}[Y|r,x,a,U=1] - \mathbb{E}[Y|r,x,a,U=0]$ does not depend on $x$ or $a$.

Consider first the post-intervention mean, $\mathbb{E}[Y^*|R=r]$, which is identified as

$$\mathbb{E}[Y^*|R=r] = \int \Big[ \sum_{a=0,1} p^*(a|r,x)\big(\mathbb{E}[Y|x,r,a,U=1]\Pr[U=1|x,r] + \mathbb{E}[Y|x,r,a,U=0]\Pr[U=0|x,r]\big)\Big] dP(x|r).$$

By contrast, our estimator of $\mathbb{E}[Y^*|R=r]$ converges to

$$\begin{aligned}\tilde{\mathbb{E}}[Y^*|R=r] = \int \Big[ \sum_{a=0,1} p^*(a|r,x)\big(&\mathbb{E}[Y|x,r,a,U=1]\Pr[U=1|x,r,a] \\ &+ \mathbb{E}[Y|x,r,a,U=0]\Pr[U=0|x,r,a]\big)\Big] dP(x|r).\end{aligned}$$

Taking the difference of the above two quantities, we have

$$\begin{aligned}&\text{bias}(\mathbb{E}[Y^*|R=r]) \\ =&\int \Big[ \sum_{a=0,1} p^*(a|r,x)\big(\mathbb{E}[Y|r,x,a,U=1] - \mathbb{E}[Y|r,x,a,U=0]\big)\big(\Pr[U=1|x,r,a] - \Pr[U=1|x,r]\big)\Big] dP(x|r) \\ =&\int \Big[ \sum_{a=0,1} p^*(a|r,x)\big(\mathbb{E}[Y|r,x,a,U=1] - \mathbb{E}[Y|r,x,a,U=0]\big) \\ &\big(\Pr[U=1|x,r,a] - \Pr[U=1|x,r,1-a]\big)p(1-a|x,r)\Big] dP(x|r)\end{aligned}$$

Then, applying assumptions (i) and (ii), we have

$$\begin{aligned}&\text{bias}(\mathbb{E}[Y^*|R=r]) \\ =&\int \big[ \sum_{a=0,1} p^*(a|r,x)(-1)^{1-a}\alpha_r\beta_r p(1-a|x,r)\big] dP(x|r) \\ =&\alpha_r\beta_r \int \big(\Pr[A=0|x,r]\Pr[A^*=1|x,r] - \Pr[A=1|x,r]\Pr[A^*=0|x,r]\big) dP(x|r) \\ =&\alpha_r\beta_r \int \big(\Pr[A^*=1|x,r] - \Pr[A=1|x,r]\big) dP(x|r) \\ =&\alpha_r\beta_r\big(\Pr[A^*=1|r] - \Pr[A=1|r]\big).\end{aligned}$$

Since the interventional disparity ($\Delta^*$) is defined as $\mathbb{E}[Y^*|R=\text{White}] - \mathbb{E}[Y^*|R=\text{Black}]$, bias($\Delta^*$) can be written as:

$$\begin{aligned}\text{bias}(\Delta^*) =& \alpha_{\text{White}}\beta_{\text{White}}\big(\Pr[A^* = 1|\text{White}] - \Pr[A = 1|\text{White}]\big) \\ &- \alpha_{\text{Black}}\beta_{\text{Black}}\big(\Pr[A^* = 1|\text{Black}] - \Pr[A = 1|\text{Black}]\big).\end{aligned}$$

Similarly, the bias of the estimated gap reduced $(\Delta - \Delta^*)$ can be written as:

$$\begin{aligned}\text{bias}(\Delta - \Delta^*) =& -\alpha_{\text{White}}\beta_{\text{White}}\big(\Pr[A^* = 1|\text{White}] - \Pr[A = 1|\text{White}]\big) \\ &+ \alpha_{\text{Black}}\beta_{\text{Black}}\big(\Pr[A^* = 1|\text{Black}] - \Pr[A = 1|\text{Black}]\big).\end{aligned}$$

If we further assume that $\alpha_{\text{White}} = \alpha_{\text{Black}} = \alpha$ and $\beta_{\text{White}} = \beta_{\text{Black}} = \beta$, these bias formulas will reduce to

$$\text{bias}(\Delta^*) = \alpha\beta C,$$

and

$$\text{bias}(\Delta - \Delta^*) = -\alpha\beta C,$$

where $C = \big(\Pr[A^* = 1|\text{White}] - \Pr[A = 1|\text{White}]\big) - \big(\Pr[A^* = 1|\text{Black}] - \Pr[A = 1|\text{Black}]\big)$.

# D Parametric Estimates of the Generalized KOB Decomposition

In the main text, we estimated the generalized KOB decomposition using the doubly robust (DR) estimator for $\mu_{B,j}^{\text{pred}}$ combined with cross-fitting and machine learning estimates of the outcome and group membership models. The use of machine learning estimators in tandem with cross-fitting and the DR estimator is desirable because it leads to an instance of what Chernozhukov et al. (2018) call "double/debiased machine learning" (DML) — a procedure that is robust, efficient, and has theoretically justified analytical standard errors. As mentioned in the main text, it is also possible to obtain estimates of the generalized KOB decomposition with parametric models applied to the regression-imputation (RI), weighting, or DR estimator. In this case, the DR estimator enjoys the "double-robustness" property as described in the main text.

We have implemented these alternative estimators for our running example. The results are reported in Figures 6 and 7. In general, the results obtained under different estimation procedures are similar, although we see that the weighting estimator is generally less efficient than the parametric RI and DR estimators. As a combination of RI and weighting, the DR estimates tend to lie between estimates from the other two methods.

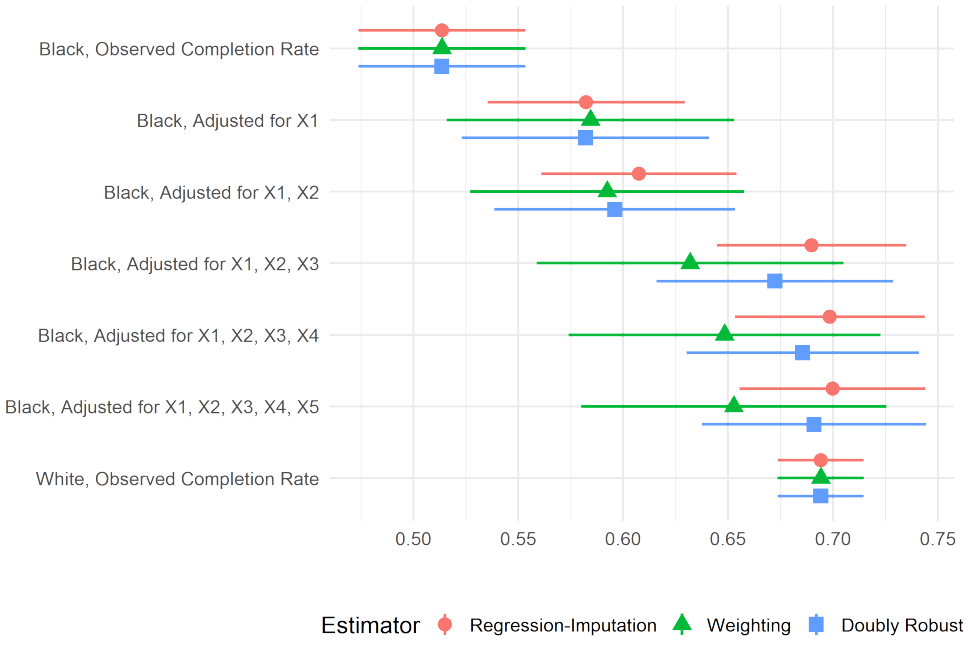


Figure 6: Parametric estimates of observed and adjusted completion rates. The outcome and group membership models are both fitted with logistic regression. Standard errors are obtained using the nonparametric bootstrap (with 250 replications).


Figure 7: Parametric estimates of the explained and unexplained components of the Black-White gap in college completion. The outcome and group membership models are both fitted with logistic regression. Standard errors are obtained using the nonparametric bootstrap (with 250 replications).

# E Parametric Estimates of the Gap Reduced ($\Delta - \Delta^*$)

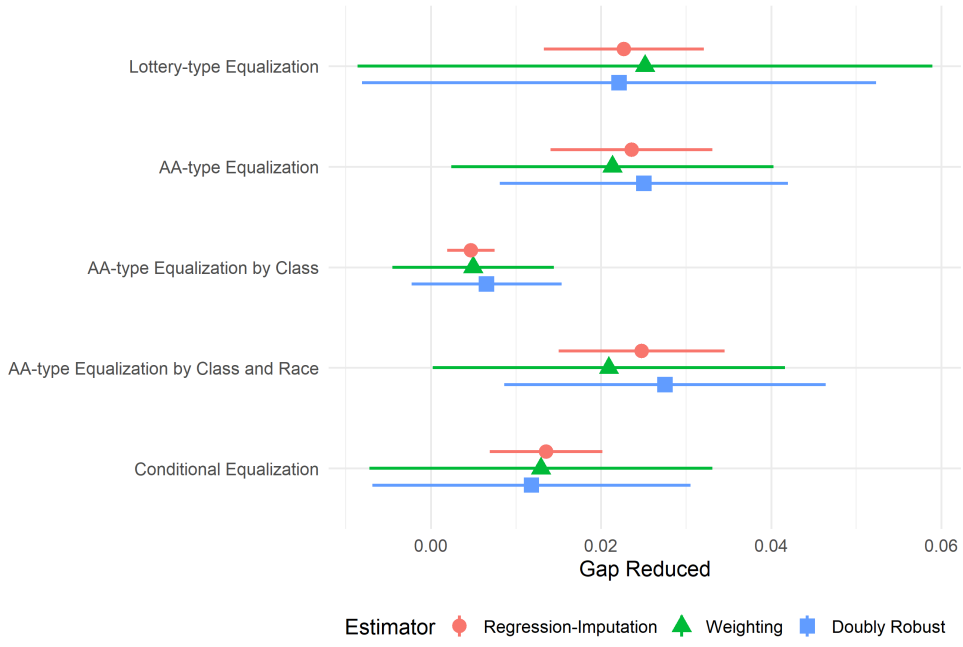


Figure 8: Parametric estimates of the gap reduced under different stylized interventions to selective college attendance.