The gap-closing estimand: A causal approach to study interventions that close disparities across social categories*

Ian Lundberg

Department of Sociology and California Center for Population Research

University of California, Los Angeles

Last updated: 31 January 2022

Abstract

Disparities across race, gender, and class are important targets of descriptive research. But rather than only describe disparities, research would ideally inform interventions to close those gaps. The gap-closing estimand quantifies how much a gap (e.g. incomes by race) would close if we intervened to equalize a treatment (e.g. access to college). Drawing on causal decomposition analyses, this type of research question yields several benefits. First, gap-closing estimands place categories like race in a causal framework without making them play the role of the treatment (which is philosophically fraught for non-manipulable variables). Second, gap-closing estimands empower researchers to study disparities using new statistical and machine learning estimators designed for causal effects. Third, gap-closing estimands can directly inform policy: if we sampled from the population and actually changed treatment assignments, how much could we close gaps in outcomes? I provide open-source software (the R package `gapclosing`) to support these methods.

Keywords: causal inference, disparities, stratification, race, class, gender

Word count: 11,465

*The final published version of this paper is available from *Sociological Methods and Research*: doi.org/10.1177%2F00491241211055769. The methods presented in this paper are implemented in the R package `gapclosing`, available from CRAN. Replication code is available on Dataverse: doi.org/10.7910/DVN/UWYAJD. A preprint is available on SocArxiv: doi.org/10.31235/osf.io/gx4y3. Address correspondence to Ian Lundberg, ianlundberg@ucla.edu, California Center for Population Research, 4284 Public Affairs Bldg, MC 723603, PO Box 957236, Los Angeles, CA 90095. For feedback relevant to this project, I thank Brandon Stewart, Matthew Salganik, Belén Unzueta, Christopher Felton, Daniela Urbina, Felix Elwert, Gillian Slee, Hannah Waight, Janet Xu, Rebecca Johnson, Simone Zhang, Xiang Zhou, members of the Stewart Lab, and reviewers. Research reported in this publication was supported by the National Science Foundation under Award Number 2104607 and by The Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Numbers P2CHD047879 and P2CHD041022.

1 Introduction

Gaps in socioeconomic outcomes are among the strongest indictments of inequality. For example, the net worth of the median black household in America in 2016 was 10 cents for every dollar held by the median white household (McIntosh et al., 2020). Among those employed full-time in 2014, the median woman earned 79 cents for every dollar earned by the median man (Blau and Kahn, 2017). People raised by high-income parents have higher incomes themselves as adults (Chetty et al., 2014). These disparities are compelling in part because they are simple: the difference in a summary statistic across populations. Yet having done the important work of describing a gap, one might turn to a further object of research: how we could intervene to close that gap.

Learning how to close a gap requires one to think about causal counterfactuals. What racial wealth gap would remain if access to homeownership were equalized? What sex gap in pay would remain if men and women were assigned equitably to occupations? What income mobility would be observed if access to college were equalized? Each of these questions involves a disparity across categories (race, sex, family origin) that would persist under counterfactual assignment of some treatment (homeownership, occupation, college). For some treatments (e.g. occupation), gap-closing estimands speak to scientific understanding. For instance, we can assess the degree to which the sex gap in pay is a causal consequence of occupational segregation. Other treatments (e.g. access to college) may be directly amenable to policy manipulation. Evidence about these treatments would directly inform interventions to promote equality.

A gap-closing estimand places categories of race, class, and gender within a causal framework. It begins with a set of units labeled by categories (e.g. race). Then it posits that we could take a sample of those units from the population and conduct a counterfactual intervention to a treatment variable (e.g. attaining a college degree). The gap-closing estimand is the expected disparity in an outcome (e.g. income) across categories of units in that sample who receive that counterfactual intervention to the treatment (Fig 1). This way

of thinking about the problem is useful for three reasons. First, we never have to posit a counterfactual manipulation of the category itself (e.g. race), thereby avoiding the pitfalls of estimating the causal effect of seemingly immutable characteristics (Holland, 1986; Kohler-Hausmann, 2018). Second, by focusing on what would happen if the treatment were applied to a sample we avoid the interference problems that would arise from giving treatment to the full population. To know what would happen in a sample, we do not need formal models of interference and can directly apply standard statistical and machine learning procedures. Third, this framing is useful for policy: we cannot change people’s race, but we can send them to college.

The gap-closing estimand complements other research approaches. For example, assessments of discrimination necessarily involve the causal effect of the category: when someone perceived as black is not hired for a job, would they have been hired if the prospective employer had perceived them as white? Audit studies assess discrimination by experimentally manipulating signals of a category like race while holding all else constant (Pager, 2003; Bertrand and Mullainathan, 2004). Doing so identifies the causal effect of one signal of that category, such as the name at the top of one’s resume (Kaufman, 2008; Greiner and Rubin, 2011; Sen and Wasow, 2016). But audit studies do not capture other aspects of race, such as how racism produces barriers to educational opportunity that create inequality in the content of resumes across racial categories. A similar argument would hold when the category is gender. In order to make a social category manipulable, audit studies reduce the operationalization to one small slice of the broader social and historical construct (Omi and Winant, 1994; Sen and Wasow, 2016; Kohler-Hausmann, 2018). Doing so is worthwhile: causal evidence of discrimination is extremely important. Likewise, descriptive research about disparities is extremely important. Gap-closing estimands complement each of those existing approaches: what disparity would persist under an intervention to a treatment variable?

Gap-closing estimands are rooted in causal decomposition analyses in epidemiology

The gap-closing estimand is the expected value of this disparity over hypothetical samples $\mathcal{S}$ from the population $\mathcal{P}$

Fig. 1. The gap-closing estimand uses observational data to emulate a hypothetical experiment. Social categories (e.g. class origin) are denoted by X and are conceptualized as collections of units: sets of observations over which we seek to estimate a disparity. We never consider the outcome that a unit of one category (e.g. a circle) would realize if they were counterfactually of another category (e.g. a diamond). Instead, we consider sampling from the population and intervening to set a manipulable treatment variable T to some value t . We would then observe the potential outcomes $y_i(t)$ and $y_j(t)$ for individuals in each category and the disparity across categories. The gap-closing estimand is the expected disparity over hypothetical samples $\mathcal{S}$. The target trial clarifies the intervention at the core of the claim and the scope of the intervention—applied to a sample $\mathcal{S}$ rather than the full population $\mathcal{P}$. The notion of formalizing causal claims with respect to a target trial comes from Hernán and Robins (2016).

(VanderWeele and Robinson, 2014; Jackson and VanderWeele, 2018) and are related to causal perspectives on fairness (Zhang and Bareinboim, 2018). But they apply in a much wider range of settings. The paper proceeds in several sections. Section 2 argues that gap-closing estimands formalize a widespread problem. Section 3 defines gap-closing estimands in the potential outcomes framework and discusses two important considerations: the scope of the claim and the credibility of the intervention. Section 4 presents causal assumptions. Section 5 presents estimation strategies. Section 6 illustrates the method by extending previous work that described gaps in pay by class origins (Laurison and Friedman, 2016). After detailing the method and how it can be applied, I present more detailed connections to previous research in Section 7. Section 8 concludes with implications for social science practice: gap-closing estimands would free researchers to make explicitly causal claims about interventions that could close gaps across social categories of race, gender, and class, thereby promoting transparency about research goals and assumptions as well as improving the relevance of social science to policy.

2 Gap-closing estimands address questions of widespread interest

Researchers frequently consider disparities across social categories. They also routinely consider how those disparities would change in some counterfactual scenario. Gap-closing estimands are therefore already implicit in the goals of existing research, especially in the study of race, class, and gender.

Those who study race frequently explore the degree to which racial inequality would be different under alternative institutional arrangements. Western (2006) studies the role of incarceration in the racial earnings gap among men. Although incarceration substantially harms earnings, it is rare enough that “the difference in earnings between blacks and whites would be reduced only by about 3 percent if the incarceration rate were zero,” (Western

2006:127). This is a gap-closing estimand because it involves descriptive components (the initial racial gap and the proportion incarcerated) but turns crucially on a causal effect (the effect of incarceration on the earnings of black and white men). Ciocca Eller and DiPrete (2018) examine the black-white gap in college degree completion. They conclude “if black students matched to colleges in the same way as white students with similar backgrounds, their dropout rate would decrease from 50.4 to 47.5 percent, and the BA attainment rate would increase from about 19 to 20 percent,” (Ciocca Eller and DiPrete 2018:1194). Some studies make explicit the fact that the estimand of interest involves the causal effect of some treatment. Killewald and Bryan (2016:123) identify the effect of home ownership on wealth and then conduct a simulation to show that “altering their [blacks’] homeownership experiences to be comparable to those of whites would substantially narrow race gaps in midlife wealth.” Each interpretation above appeals to the notion that gaps would change if an intervention occurred to some policy-amenable variable.

Studies of social class likewise routinely involve questions about the gap that would persist across class origins (some operationalization of one’s family background) if we intervened to help people attain some treatment that could potentially break them free from the constraints of birth. In the example to which this paper returns repeatedly, Laurison and Friedman (2016) study the earnings of British workers in higher managerial and professional occupations. Within this category, the authors examine the class gap in pay between the intergenerationally stable (those whose parents were in the professional class) and the upwardly mobile (those whose parents were in the working class). The authors control for a series of variables that might be consequences of class origins (parents’ occupational class) but causes of class destinations (own occupational class). One could interpret that research goal causally as a gap-closing estimand: the difference in pay between those of professional- and working-class backgrounds, if we intervened to lift them personally to the professional class. Similarly, studies about the role of education in social mobility often involve implicit claims about the gap-closing estimand between those of different family origins if we

intervened to send them to college (Hout, 1988; Torche, 2011; Zhou, 2019).

Gender inequality likewise begets claims about how gender gaps would be different under alternative conditions. For instance, the gender wage gap within jobs is very small. Petersen and Morgan (1995:338) motivate this estimand in language that closely resembles a gap-closing estimand: “Suppose sex segregation—by occupation, establishment, or occupation-establishment—were abolished; what then would the remaining gender relative wages be?” The gender wage gap is purely a descriptive quantity, but the hypothetical intervention of equalizing the distribution across occupations and establishments invokes a causal effect of occupations and establishments on wages. Given persistent gender segregation across occupations with unequal pay (Blau and Kahn, 2017), this gap-closing estimand remains central to the study of gender wage inequality today.

Many of the examples above explicitly state that the research goal is descriptive rather than causal. On one hand, that caveat is correct: absent causal assumptions, observational evidence is necessarily descriptive. Yet one reason these claims are compelling is because the descriptive evidence points toward a possible causal claim: if we intervened on the treatment, the gap might close. For researchers who want to make that causal claim, the gap-closing estimand provides a method to do so.

3 Define the goal: Gap-closing estimands provide a causal framework for social categories

A gap-closing estimand explicitly appeals to a counterfactual world in which a treatment variable was reassigned. Because it is a world that does not exist, great care is needed to define that world. This section begins by defining gap-closing estimands with fixed treatment assignments (assign treatment value t) before turning to stochastic treatment assignments (assign a treatment value t selected at random from a set $\mathcal{T}$). Then, I discuss two important considerations when defining the goal: choice of a credible scope and choice of a realistic

intervention. In the framework of Lundberg et al. (2021), this section defines the theoretical estimand and presents some considerations for choosing a theoretical estimand which is empirically tractable.

To make the gap-closing estimand precise, it is helpful to follow the advice of Hernán and Robins (2016) and specify a target trial: the hypothetical experiment which we hope to approximate by analyzing observational data. The motivation for a target trial comes from experimental settings, where the protocol for assigning the treatment makes the research goal unambiguous. Randomization is not possible in observational settings, yet we can still gain clarity about the research goal by specifying the procedure we would like to apply if it were possible. Fig 1 presents a target trial for the gap-closing estimand. Suppose you draw a sample $\mathcal{S}$ from a population $\mathcal{P}$. Then, you intervene to assign each sampled unit i to treatment value $T_i = t$ and observe the outcome $y_i(t)$ under that treatment. You then calculate a disparity $\bar{y}_{\mathcal{S},x'}(t) - \bar{y}_{\mathcal{S},x}(t)$ in the mean outcomes of units in sample $\mathcal{S}$ from the categories $X = x'$ and $X = x$, under exposure to treatment value t . That disparity is a random variable because it averages over a random sample $\mathcal{S}$ from the population $\mathcal{P}$. The gap-closing estimand is the expected disparity over hypothetical samples $\mathcal{S}$.

$$\begin{array}{ccc}
 & \text{Expectation over} \\
 & \text{hypothetical samples } \mathcal{S} \\
 & \text{from the population } \mathcal{P} \\
 & \downarrow \\
 \text{Gap-closing estimand: } & \tau_{x',x}(t) \equiv \mathbf{E}_{\mathcal{S}} \left(\bar{y}_{\mathcal{S},x'}(t) - \bar{y}_{\mathcal{S},x}(t) \right) & (1) \\
 & \uparrow \qquad \qquad \qquad \uparrow \\
 & \text{Mean outcome} \qquad \text{Mean outcome} \\
 & \text{in category } x' \qquad \text{in category } x \\
 & \text{if treatment were set to the value } t \\
 & \text{for units in sample } \mathcal{S}
 \end{array}$$

The right column of Fig 1 makes this concrete. Suppose we begin with those raised in professional and working-class families, defined by the occupation of one's father figure ($X = x'$ and $X = x$). Then we take a sample from the population and assign people in that sample to personally hold a professional occupation as an adult ($T = t$), thus cutting the

intergenerational transmission of class attainment. The gap-closing estimand is the expected pay disparity by class origin in the sample that would result from this hypothetical procedure.

Sometimes our theoretical question might not involve assigning everyone to a single treatment condition. Instead, we might want to know what would happen if we shuffled the treatment assignments (possibly as a function of covariates) while keeping their relative prevalence fixed. The gap-closing estimand extends to this type of stochastic treatment assignment rule. In the target trial, a unit i would be selected as part of the sample, the researcher would observe their covariate values $\vec{L}_i = \vec{\ell}$ (e.g. education is a college degree), and then the researcher would assign them to a treatment value t (e.g. an occupational class) probabilistically from a set of many possible treatment values $\mathcal{T}$ (e.g. professional and working-class) according to a set of probabilities $\{\pi(t, \vec{\ell})\}_{t \in \mathcal{T}}$. In this notation, there is one treatment probability $\pi(t, \vec{\ell})$ for each treatment value t among all candidate treatment values $\mathcal{T}$, and those probabilities could be functions of the observed covariate values $\vec{\ell}$. For instance, perhaps we would counterfactually assign professional occupations with higher probabilities to those with college degrees. Each unit i has an expected outcome of this experiment: the average of the set potential outcomes $\{y_i(t)\}_{t \in \mathcal{T}}$ where each $y_i(t)$ is the outcome unit i unit would realize under each treatment value t , with each outcome weighted by the counterfactual probability $\pi(t, \vec{L}_i)$ of assigning that treatment value. I use $\bar{y}_i(\pi)$ to denote this average for unit i under treatment assignment rule π .

$$\begin{array}{ccc}
 & \text{Sum over all} & \\
 & \text{treatment values } t & \\
 & \downarrow & \\
 \text{Expected outcome} & \bar{y}_i(\pi) \equiv \sum_t \pi(t, \vec{L}_i) y_i(t) & \\
 \text{for unit } i \text{ under a} & & \\
 \text{stochastic intervention:} & & \\
 & \nearrow \quad \nwarrow & \\
 \text{Counterfactual probability} & & \text{Potential outcome} \\
 \text{of assigning treatment} & & \text{for unit } i \text{ under} \\
 \text{value } t \text{ to unit } i & & \text{treatment value } t
 \end{array} \tag{2}$$

Given the unit-specific expected outcome $\bar{y}_i(\pi)$, let $\bar{y}_{S,x}(\pi)$ denote the mean of that unit-

specific expected outcome among units i in the sample $\mathcal{S}$ for whom $X_i = x$. The gap-closing estimand under a stochastic assignment rule $\tau_{x',x}(\pi)$ is then analogous to Eq. 1 but with the assignment rule π replacing the fixed treatment value t everywhere it appears. Instead of equalizing at a single treatment value, we are equalizing the distribution from which a treatment is drawn. This idea appears in VanderWeele and Robinson (2014) and is discussed in greater depth by Jackson (2018).

Regardless of whether treatments are hypothetically assigned by a fixed rule (assign one treatment value t) or a stochastic rule (assign a treatment value t probabilistically from the set $\mathcal{T}$), the target trial clarifies two important considerations. The first is the scope of the intervention: we are not giving treatment to the entire population $\mathcal{P}$, but only to a sample $\mathcal{S}$. The scope is most credible if $\mathcal{S}$ is a very small fraction of $\mathcal{P}$, because then the treatment assigned to one unit in $\mathcal{S}$ is unlikely to affect the outcome of another unit in $\mathcal{S}$ since they are dispersed over a much larger population $\mathcal{P}$ and thus less likely to interact. The second is the degree to which the intervention is realistic: we have to think carefully about whether each unit can realistically be assigned to each treatment $T = t$.

3.1 Consideration 1: Make a claim with credible scope

It can be tempting to interpret a gap-closing estimand in terms of a global claim: what would happen if every unit in the entire population received treatment value t (Fig 2). The examples highlighted previously dealt with global interventions to set the incarceration rate to zero (Western 2006:127), to abolish occupational sex segregation (Petersen and Morgan 1995:338), or to create “a utopian world of ‘college for all’,” (Zhou 2019:466). All of these claims involve interventions that simultaneously alter the treatment assignments of vast swaths of the population. It is difficult to provide empirical support for these global claims because the required assumptions are particularly stringent.

To take one example, only 35% of the U.S. population ages 25 and older in 2018 held a bachelor’s degree or higher (author’s tabulation from Table 3 in U.S. Census Bureau 2019).

A world in which *everyone* attended college would be radically different from the world in which we live. A college degree would no longer have the same meaning. Past studies of education expansion have shown that elites find ways to maintain their advantages (Raftery and Hout, 1993; Lucas, 2001) and that the benefits of being in the pool of college graduates may decline as the size of that pool grows (Horowitz, 2018). Thus, the outcome realized under a college degree depends on how many other people are also assigned to a college degree. There is a serious problem of interference: intervening to change the treatment value of unit i might change the effectiveness of the treatment for another unit j . We might credibly be able to infer what would happen if treatment were expanded to a few units sampled from a much large population, so that they would not interfere with each other. But to predict what would happen if everyone received the treatment would require formal theory about interference, which might border on speculation. Often, it would be better to reduce the scope of the claim.

The interference concerns that arise with global claims involve violations of an assumption which is often overlooked: the consistency assumption (Hernán and Robins, 2020). The consistency assumption defines the potential outcomes $y_i(t)$. One setting which violates consistency is interference, where the potential outcome $y_i(t)$ is defined as a function of the treatment value t assigned to unit i but in fact unit i responds to the treatments assigned to other units. Sometimes this is termed a violation of the Stable Unit Treatment Value Assumption (Imbens and Rubin, 2015), and it is often emphasized in other substantive contexts, as when an experiment is carried out on units embedded in social networks (e.g. Aronow et al. 2017). Yet interference is also in play when social interactions are less explicit. This is especially true in sociological settings where giving everyone the treatment would change the effectiveness of the treatment. Population-level interpretations of gap-closing estimands lean heavily on an individualistic definition of potential outcomes in a setting where that definition is particularly imperfect.

Researchers concerned with global claims have at least three options. One option

The **scope of the intervention** is a key consideration for gap-closing estimands.

The observational study seeks to approximate a hypothetical experiment (Fig 1).
In that experiment, for whom would we hypothetically reassign the treatment?

Fig. 2. Clarify the scope of the intervention. A gap-closing estimand invokes a counterfactual world where treatment is reassigned for a sample $\mathcal{S}$ from a population $\mathcal{P}$ (Fig 1). Inference about a world where treatment is reassigned for the entire population ($\mathcal{S} = \mathcal{P}$) is often only speculative. That global extreme is not empirically tractable unless one has strong theory about interference among units that can be encoded in a formal model. In practice, we often lack this theory. Thankfully, the policy-relevant claim is much more local: it is often infeasible for a policymaker to change treatment for the entire population at once. A claim relevant to policy is thus one about what would happen if treatment is reassigned for a very small sample drawn from the population. In the local extreme, that sample might be only one unit from each category. Empirical evidence speaks to the local extreme, and interpretations become increasingly speculative as one moves toward the global extreme.

is to theorize how the system would change under the intervention and encode that theory into a formal model like an agent-based simulation (Jackson and Arah, 2020). Yet, doing so requires strong social theory about how units interfere with each other, which may not be available in some substantive settings. A second option is to select an estimand for which the interactions among units may be less severe. For example, one of the most severe threats to validity arises when the effectiveness of the treatment is a function of the proportion of the population to receive the treatment. This is the case, for example, when college helps you secure access to an occupation for which the number of available positions is limited. This particularly severe threat can be averted by shifting to a stochastic estimand that keeps the marginal distribution of the treatment at its observed distribution. We might be more willing to speculate about a world where college degrees were randomly shuffled among the population (assignment rule $\pi(t, \ell) = P(T = t)$), thus changing *who* holds the degree but not changing the *prevalence* of the degree. There is still a leap to global inference—the data only speak to the local claim. But with stochastic interventions a researcher may be more willing to make that leap because the global counterfactual is more like the observed world, at least in the marginal distribution. Global claims are always speculative, but stochastic interventions may help to improve credibility.

In many settings, a third option is most promising: make a local (rather than global) version of the claim. Following this path, a gap-closing estimand is the expected result of a hypothetical experiment: if we sample a small fraction of the population and assign them to the treatment, what disparity would we expect to observe in that sample? The local extreme of Fig 2 is a conceptually helpful edge case: suppose we sample one unit from each category and carry out the intervention on only those two units. In that limiting case, interference problems are unlikely—if the treatment is a college degree, assigning two people to receive college degrees would not change the meaning of the treatment. Moving back from the limiting case, we may often be able to credibly infer what would happen if treatment were provided to a small sample from the population: if that small sample is randomly spread

throughout the much larger population, the risk of interference is reduced. That empirically tractable target is also relevant to certain types of policies. Policymakers generally cannot intervene on the entire population at once, so what would happen if treatment is provided to a sample may be what the policymaker wants to know. At least in principle, the researcher could update the evidence with new observational analyses conducted in real time as the policy expands to ever-greater shares of the population.

3.2 Consideration 2: Define a realistic intervention

One can speculate about the disparity under any treatment assignment $T = t$. But that speculation is only useful in the real world if the treatment assignment $T = t$ is plausible for each unit in question. The pre-treatment characteristics of an individual can make certain treatments implausible. For example, someone without a high school degree is factually unlikely to hold a professional occupation (e.g. as a doctor). It would be better to rule out this type of implausible intervention when defining a gap-closing estimand. Stochastic interventions $\pi(t, \vec{\ell})$ are extremely helpful in this regard: the researcher can define the assignment rule such that the probability of treatment $T = t$ is zero in subpopulations $\vec{L} = \vec{\ell}$ for whom treatment value t is implausible (see Nguyen et al. 2020 footnote 22 and the discussion of positivity in the identification section of Jackson 2021).

When defining a realistic intervention, researchers may face a tradeoff between an equitable intervention and a credible intervention. It might be most equitable to assign treatment irrespective of the covariates $\vec{L}$, thus giving everyone in the population equal access to each treatment. But doing so risks the implausible counterfactuals discussed above. The researcher could allow $\pi(t, \vec{\ell})$ to be a function of some of the covariates in $\vec{\ell}$, such as education. Jackson (2021) outlines how researchers could distinguish between one set of covariates that we would allow to determine treatment assignment (e.g. education) and non-allowable covariates over which unequal allocation of treatments might be considered especially discriminatory (e.g. race, sex). In some settings, one may be able to avoid im-

plausible treatment assignments by making the counterfactual assignments $\pi(t, \vec{\ell})$ a function only of allowable covariates. Third, the researcher could accept all covariates as allowable and assign treatment as a function of all covariates. For example, one could imagine shuffling treatments across the category X but maintaining the distribution of treatments within each covariate subgroup $\vec{L} = \vec{\ell}$. Doing so would maximize credibility (covariate subgroups are assigned the treatments as observed in that subgroup, thus ruling out implausible assignments) but may come at the cost of equity (it may be unjust to assign treatment as a direct function of some covariates). The possible tradeoff between equity and credibility is the kind of substantive debate one can only have after entering a causal framework with stochastic interventions.

4 Identify the estimand: Causal assumptions are agnostic about the effect of social categories

Like causal effects, gap-closing estimands involve potential outcomes that are not observed for some units (Holland, 1986). This section focuses on the link between the theoretical estimand and an empirical estimand defined in terms of observable data (Lundberg et al., 2021). The identification assumptions for gap-closing estimands are the same as those for estimating causal effects and are untestable—they must be defended on conceptual grounds. These assumptions include consistency, conditional mean independence, and positivity (following the terminology of Hernán and Robins 2020). Consistency defines the potential outcomes and equates the observed outcome with the potential outcome under the observed treatment condition ($Y_i = y_i(T_i)$). Conditional mean independence requires that the treatment T is unconfounded given a set of observed covariates (see Fig 3). Positivity requires that treatment value t occurs with non-zero probability in each stratum of those observed variables; without positivity, we might not observe any individuals from whom to learn the potential outcomes. Appendix A provides more details about these assumptions

and proves that they are sufficient for identification.

Fig. 3 illustrates key issues using Directed Acyclic Graphs (DAGs, Pearl 2000).¹ The central benefit of gap-closing estimands is that counterfactuals are defined over hypothetical interventions to the treatment T and not over the gap-closing category X , thereby allowing a range of assumptions about that category. For example, suppose the category X is race. Race may be understood as assigned prior to all other variables, so that covariates $\vec{L}$ are consequences of race (Panel A, see Sen and Wasow 2016). Race may be understood to have no causal effects (Panel B, see Rubin 1986; Holland 2008). The covariates may affect how one racially identifies (Panel C, see Saperstein and Penner 2012). In all of these cases, the gap-closing estimand is still identified because it involves the causal effect of the treatment, not the causal effect of race. The social category is understood agnostically as a marker of two collections of units across which we seek to summarize a gap.

The settings that threaten causal identification for gap-closing estimands are the same settings that threaten identification for the causal effect of the treatment T . If a backdoor path cannot be blocked due to an unobserved variable V that affects both the treatment and the outcome, then the gap-closing estimand is not identified (Panel D). Confoundedness can also arise through M -bias (Panel E, Greenland et al. 1999, Greenland 2003). If the pre-treatment covariates $\vec{L}$ are causally downstream from both an unobserved variable U that affects the outcome directly and an unobserved variable V that affects the treatment directly, then $\vec{L}$ is a collider variable and conditioning on it can open a backdoor path $T \leftarrow V \rightarrow \boxed{\vec{L}} \leftarrow U \rightarrow Y$, where the box around $\vec{L}$ indicates that we have conditioned on $\vec{L}$ (Elwert and Winship, 2014). This possibility illustrates the need for a clear translation between our beliefs about the causal structure of the world and the mathematical assumptions required

¹The DAGs in Fig 3 invoke assumptions that are stronger than needed for conditional mean independence. A DAG implies that treatment is independent of the potential outcomes given the conditioning set. The weaker condition that the mean of the potential outcomes is independent of treatment is sufficient. Second, because a gap-closing estimand focuses exclusively on one treatment level t , conditional mean independence only needs to hold for that treatment level. The other potential outcomes, which are not the object of study, may be related to treatment assignment without threatening identification. Despite making stronger assumptions than necessary, DAGs facilitate reasoned argument about these assumptions.

A gap-closing estimand is identified under a wide range of assumptions about the social category X (e.g. race, class, gender).

Above, there is no causal effect of X (B), or it is not identified (A, C) due to the backdoor path $X \leftarrow U \rightarrow Y$ through unobserved U . The gap-closing estimand is nonetheless identified.

A gap-closing estimand is **not** identified when $T \rightarrow Y$ is not identified.

Fig. 3. Identification of gap-closing estimands. Observed variables include the social category X (e.g. race, class, gender), the manipulable treatment variable T (e.g. college completion, occupational attainment), and other pre-treatment covariates $\vec{L}$. Nodes U and V are unobserved. The blue $T \rightarrow Y$ edge represents the causal effect that must be identified. The dashed red edges in (D–E) represent threats to identification. DAGs present assumptions that are stronger than needed; identification is possible under the slightly weaker assumption of conditional mean independence for the potential outcome $Y(t)$ under the treatment value t of interest (Appendix A).

for identification. Making the research goal explicitly causal creates new opportunities for clear reasoning about these assumptions.

5 Estimation: Learn from data

The central hurdle for estimation is that some potential outcomes $y_i(t)$ are unobserved. If the identification assumptions hold, we can learn the average of those potential outcomes for all units by examining the units we actually observe with treatment value $T_i = t$, within each subpopulation with a particular category value x and covariate set $\vec{\ell}$. Then, we can aggregate those subpopulation estimates across the population distribution of $\{X, \vec{L}\}$. In a limited sample size, however, each of those subpopulations may have very few observations. Statistical models and machine learning tools improve efficiency by sharing information across subpopulations. Choices about how to share information correspond to the estimation step of Lundberg et al. (2021).

To simplify the discussion of estimation, it is helpful to break the gap-closing estimand into two components, which I call post-intervention means.

$$\begin{array}{ccc}
 \tau_{x',x}(\pi) \equiv \mathbf{E}_{\mathcal{S}}(\bar{y}_{\mathcal{S},x'}(\pi)) - \mathbf{E}_{\mathcal{S}}(\bar{y}_{\mathcal{S},x}(\pi)) & & \\
 \nearrow & & \nwarrow \\
 \theta_{x'}(\pi) & & \theta_x(\pi) \\
 \text{Post-intervention mean} & & \text{Post-intervention mean} \\
 \text{in category } x' & & \text{in category } x \\
 \text{if treatment were assigned by the rule } \pi & &
 \end{array} \tag{3}$$

Once you have an estimator for each post-intervention mean θ , you have by extension an estimator for the gap-closing estimand τ . For simplicity, this section therefore focuses on estimators for post-intervention means: the expected outcome in a category $X_i = x$ under assignment to treatment by a rule π . I first discuss estimation by predicted outcomes (e.g. regression modeling) and by predicted treatment probabilities (e.g. propensity score

weighting). Then, I discuss how these two approaches can be brought together in a doubly robust estimator that is consistent if either estimator is consistent. The doubly robust estimator can be interpreted as either estimator paired with a bias correction. Double robustness becomes especially helpful when the prediction functions for treatment probabilities and outcomes are estimated by machine learning, which is biased by regularization. In that setting, it is also useful to carry out the bias correction of the doubly robust estimator with a sampling splitting step. The end of the section briefly discusses standard error estimation by resampling strategies.

5.1 Estimation by predicted outcomes

One estimation approach relies on a function $g()$ for predicting the outcome variable.

$$\text{Outcome prediction function:} \quad g(t, x, \vec{\ell}) = \mathbf{E}(Y \mid T = t, X = x, \vec{L} = \vec{\ell}) \quad (4)$$

The true conditional mean function $g()$ could be approximated by any statistical or machine learning algorithm (e.g. OLS) to predict an outcome Y as a function of T , X , and $\vec{L}$. If estimation is conducted on a complex survey sample, one should use survey weights in estimation of $g()$.

A function to predict outcomes can be converted to an estimate of the gap-closing estimand by an approach known as the parametric g -formula (Robins 1986, Hernán and Robins 2020:166). The researcher uses the estimated function to predict unobserved potential outcomes with estimates $\hat{g}()$, thereby imputing the potential outcome under treatment value t for every unit regardless of whether that particular unit is observed in this treatment

condition. It then becomes possible to aggregate over all observations by a weighted mean,

$$\hat{\theta}_x^{\text{Outcome}}(\pi) = \frac{1}{\eta_x} \sum_{i: X_i=x} \eta_i \sum_t \hat{\pi}(t, \vec{L}_i) \hat{g}(t, X_i, \vec{L}_i) \quad (5)$$

$\begin{array}{cccc} \text{Sample weight} & & \text{Counterfactual probability of treatment } t & & \text{Predicted outcome under treatment } t \\ & \searrow & \downarrow & & \swarrow \\ \uparrow & \nearrow & \nwarrow & & \swarrow \nearrow \\ \text{Estimate from predicted outcomes} & \text{Sum over all sampled units } i \text{ in category } x & \text{Sum over all treatment values } t & & \text{Category and covariates as observed} \end{array}$

where η_i are inverse probability of sampling weights and η_x is the sum of these weights among sampled units in category $X = x$. After estimating the post-intervention means $\{\hat{\theta}_{x'}^{\text{Outcome}}, \hat{\theta}_x^{\text{Outcome}}\}$ in each category $\{x', x\}$, the difference is an estimate of the gap-closing estimand.

5.2 Estimation by predicted treatment probabilities

Instead of predicting the unobserved outcomes, one can also reweight the observed outcomes to draw inference about the average outcome under treatment in the population. This approach begins with a function $m()$ to predict the probability of treatment, which could be estimated by any statistical or machine learning algorithm (e.g. logistic regression).

$$\text{Treatment prediction function:} \quad m(t, x, \vec{\ell}) = P(T = t \mid X = x, \vec{L} = \vec{\ell}) \quad (6)$$

The task then becomes analogous to sampling from a population. We only observe the outcome $y_i(t)$ for units who factually have $T_i = t$ (analogous to only observing outcomes for sampled units). We estimate the probability that $T_i = t$ from the prediction function above (analogous to the known probability of sample inclusion in a survey). Generalized propensity score methods (Imbens, 2000) draw on this analogy to estimate the population-average outcome under treatment $T = t$ by the sample mean with inverse probability of treatment weights, analogous to similar estimators for population means weighted by inverse

sampling probabilities (Horvitz and Thompson, 1952). For the more general setting of a stochastic assignment rule to set t to a value selected probabilistically from the set $\mathcal{T}$, the inverse probability weights involve the ratio of the assignment probability for the factual treatment T_i under the counterfactual rule $\pi(T_i, \vec{L}_i)$ and the factual rule $m(T_i, X_i, \vec{L}_i)$. The estimate is a weighted average of the observed outcomes,

$$\hat{\theta}_x^{\text{Treatment}}(\pi) = \frac{1}{\sum_{i: X_i=x} \hat{w}_i} \sum_{i: X_i=x} \hat{w}_i Y_i, \quad \text{where } \hat{w}_i = \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)} \quad (7)$$

where η_i accounts for the unequal sampling probabilities for units from the population. An estimate $\hat{m}()$ of the generalized propensity score thus translates to estimates $\{\hat{\theta}_{x'}^{\text{Treatment}}, \hat{\theta}_x^{\text{Treatment}}\}$ of the post-intervention means, which can be differenced to estimate the gap-closing estimand.

5.3 Doubly robust estimation

Doubly robust estimation combines predicted outcomes and treatment probabilities to produce an estimator that is consistent if either the estimator for the outcome prediction function is consistent for the true conditional mean function or the estimator for the treatment prediction function is consistent for the true conditional probability of treatment (Fig 4, see Robins et al. 1994; Bang and Robins 2005; Glynn and Quinn 2010).² Begin with an estimate $\hat{\theta}_x^{\text{Outcome}}(\pi)$ based on predicted outcomes from $\hat{g}()$. Suppose the outcome function $\hat{g}()$ has a functional form that is misspecified (e.g. missing an interaction term, so that $\hat{g}() \rightarrow \tilde{g}() \neq g()$). For each unit, we are able to observe the error $\hat{g}(T_i, X_i, \vec{L}_i) - Y_i$ for

²By stating “ $\hat{g}()$ is consistent,” I mean $\hat{g}(t, x, \vec{\ell}) \rightarrow g(t, x, \ell) \forall t, x, \vec{\ell}$. Likewise for consistency of $\hat{m}()$.

the potential outcome in the observed treatment condition T_i . By reweighting the *errors* by the inverse probability of treatment, we can estimate the average error in the target population (all units, regardless of their factual treatment). A doubly-robust augmented inverse probability weighting estimator subtracts the estimated bias off from the outcome estimator,

$$\hat{\theta}_x^{\text{Robust}}(\pi) = \hat{\theta}_x^{\text{Outcome}}(\pi) - \underbrace{\frac{1}{\sum_{i: X_i=x} \hat{w}_i} \sum_{i: X_i=x} \hat{w}_i \left(\hat{g}(T_i, X_i, \vec{L}_i) - Y_i \right)}_{\text{Estimated bias of the predicted outcomes reweighted to the target population}} \quad (8)$$

$\uparrow$ Doubly-robust estimate $\uparrow$ Estimate from predicted outcomes Inverse probability weight Predicted outcome Observed outcome

where $\hat{w}_i = \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}$ as in the treatment prediction estimator.³

Doubly robust estimates are consistent if either $\hat{g}()$ or $\hat{m}()$ is consistent. There are two important caveats. First, if both prediction functions are inconsistent, there is no guarantee that a doubly-robust estimator will outperform an estimator based solely on outcome predictions (see the exchange between Kang and Schafer 2007 and Robins et al. 2007). The next subsection addresses the setting where both the treatment and the outcome functions have unknown functional forms. Second, double robustness is about estimation of the factual outcome and treatment functions $g()$ and $m()$; there is no double robustness to misspecification of a counterfactual treatment assignment rule $\hat{\pi}()$ that is learned from the data.

³Doubly robust estimators are a general class with many specific types. Eq. 8 is an Augmented Inverse Probability Weighting estimator (AIPW, Robins et al. 1994). There are other possibilities. For example, the inverse probability of treatment can be included as a predictor in the function to predict outcomes instead of being used for reweighting (as in Scharfstein et al. 1999 and Bang and Robins 2005). Targeted Maximum Likelihood Estimators (TMLE, Van der Laan and Rose 2011) follow that strategy with a link function to compress predictions to the support of the outcome variable. The present paper is not the first to consider double robustness for gap-closing estimands. For the special case of simple random samples and fixed (non-stochastic) counterfactual assignment rules, Naimi et al. (2016) provide two doubly robust estimators. The first is a structural nested mean model (SNMM) which allows treatment-category interactions but assumes the absences of treatment-confounder interactions (their Eq. 9). The second is TMLE with a logit link for binary outcomes. Out of the many candidate doubly robust estimators, I derive AIPW estimators because this type of estimator applies to binary or continuous outcomes, can be carried out with or without parametric assumptions, and directly relates to double machine learning (Chernozhukov et al. 2018, see Section 5.4 of the present paper). Certain types of doubly robust estimators may be preferable in certain settings; this is an open topic for future work.

Fig. 4. Doubly-robust estimation helps when one of two prediction functions is correct. The data in this example are simulated, so the truth is known. The vertical line indicates the true gap-closing estimand $\tau_{1,0}(1)$ across binary categories $X = 1$ and $X = 0$ under an intervention set treatment to $T = 1$. Densities depict the empirical distribution of estimates $\hat{\tau}_{1,0}(1)$ by the three estimators on the vertical axis, over 1,000 simulations. Axis labels are omitted because the scale of the simulation is arbitrary; all three panels have the same scale. Estimation by predicted outcomes can be biased if the outcome prediction function is misspecified (in this case, missing an interaction term). Estimation by predicted treatment probabilities can be biased if the function to predict treatment probabilities is misspecified (in this case, missing a quadratic term). Doubly robust estimates are consistent and approximately unbiased as long as one of the two is correctly specified. In this simulation, there is a binary category X , a binary treatment T , and one continuous confounder L . The confounder L is a stochastic function of the category X , the treatment T is a stochastic function of the confounder, and the outcome Y is a stochastic function of the confounder and the treatment. Simulation details are provided in Appendix D.1.

For example, suppose you study the outcome when treatment is assigned proportional to its factual prevalence in each covariate subgroup, marginalized only over the category X . You might estimate those prevalences from the data with a logistic regression model. But if that logistic regression model is misspecified, then your estimated counterfactual assignment rule $\hat{\pi}()$ might be inconsistent for the true rule $\pi()$ about which you sought to learn. Thus, while double robustness protects against misspecification of the factual functions $m()$ and $g()$, it provides no protection against a researcher’s misspecification of the counterfactual rule $\pi()$. There are two common settings in which this is not a problem. First, this is not a problem when the counterfactual rule $\pi()$ is known a priori, as when a researcher is interested in giving the treatment to all units. Second, it is not a problem for relatively simple counterfactual rules $\pi()$ which can be learned nonparametrically (e.g. by means in subpopulations). While the world determines the complexity of the factual assignment rule $m()$, the complexity of the counterfactual rule $\pi()$ is entirely under the researcher’s control: you can choose the complexity of the $\pi()$ that you define. Thus, you can define it to involve a small set of categorical predictors linked to treatment by a function that is easy to learn. In contrast, the factual treatment function $m()$ must include all confounders with whatever functional form factually exists for them, and thus robustness to misspecification of $m()$ can be very useful.

5.4 Estimation with cross fitting

At the core of doubly robust estimation is an estimated bias that involves residuals $\hat{g}(T_i, X_i, \vec{L}_i) - Y_i$. But if $\hat{g}()$ is learned on a sample that includes case i , the residual could be misleading due to overfitting—the algorithm explicitly learned $\hat{g}()$ to minimize a loss function involving this residual. Sample splitting can produce a better bias correction: learn $\hat{g}()$ and $\hat{m}()$ in one sample and use another sample to convert them into an estimate of the causal estimand (including the bias correction). Cross fitting carries out this sample splitting procedure repeatedly: use subsample A for learning and B for estimation, then B for learning

and A for estimation, and average the results (Chernozhukov et al., 2018; Bickel, 1982).

When the treatment probabilities and expected outcomes follow unknown functional forms, cross fitting becomes especially important. In these settings, the researcher can learn the treatment and outcome functions with flexible machine learning estimators (as in McCaffrey et al. 2004; Lee et al. 2010; Hill 2011; Van der Laan and Rose 2011). For example, a random forest would automatically learn interactions and nonlinearities (Breiman, 2001). Machine learning estimators balance bias against variance to produce a result optimized for individual-level prediction. But the task when estimating a gap-closing estimand is not individual level prediction; a small but desirable bias for individual-level prediction could correspond to a large and undesirable bias once aggregated across individuals. For this reason, the bias correction of doubly robust estimation is important when using machine learning estimators, and is especially helpful when carried out with cross-fitting (Chernozhukov et al., 2018).

The most pronounced benefit of cross fitting with machine learning estimators is improved convergence toward the truth as the sample size grows. Fig 5 illustrates in a simulated setting. The true gap-closing estimand is known and we can observe the root mean squared error (RMSE) of the gap-closing estimator over many simulations at various sample sizes. As the sample size grows, the RMSE approaches zero more quickly when cross fitting is used in the estimation procedure (details in Appendix D.3).

5.5 Standard error estimation for gap-closing estimands

The estimators discussed above all involve multiple steps: fit a predictive algorithm for the treatment and/or outcome, make predictions, and report some function of those predictions. Analytical standard errors for this procedure are not straightforward. Instead, researchers should produce standard errors computationally by resampling-based methods that mimic the process by which the sample was drawn. If the sample is a simple random sample, then the variance can be estimated by the nonparametric bootstrap: sample `data*`

Fig. 5. Cross fitting can improve convergence rates. The data in this example are simulated, so the truth is known. Potential outcomes are a linear function of a binary treatment and 10 continuous confounders in this simulation. In the true data generating process, there is no disparity across categories $X = 1$ and $X = 0$ in the absence of treatment. The treatment effect is $+1$ if $X = 1$ and -1 if $X = 0$, so the gap-closing estimand under assignment to treatment is $\tau_{1,0}(1) = 2$. The simulation estimates $\hat{\tau}_{1,0}(1)$ on samples from this data generating process at each sample size (x -axis). Because a researcher would not know the true functional form, I learn the functional form from the data with random forests using the defaults of the **ranger** package in R (Wright and Ziegler, 2017). I learn (1) a forest for treatment probabilities and (2) a forest for the potential outcome under treatment, estimated on the subsample that was factually treated. Then I aggregate to a gap-closing estimate $\hat{\tau}_{1,0}(1)$ according to the three methods represented by curves. Performance is measured by root mean squared error (RMSE) across the 1,000 simulations, $\sqrt{\frac{1}{1000} \sum_{r=1}^{1000} (\hat{\tau}_{1,0}^r(1) - \tau_{1,0}(1))^2}$. At small sample sizes, doubly robust estimation with cross fitting is a suboptimal estimator because of its high variance. As the sample size grows, the doubly robust cross fitting estimator RMSE converges to zero most quickly. For simulation details, see Appendix D.3.

from `data` with replacement, calculate an estimate τ^* on the resampled data, and use the empirical variance over resampled estimates $\hat{V}(\hat{\tau}^*)$ as an estimator for the variance of the point estimate $V(\hat{\tau})$ (Efron and Tibshirani, 1994). In complex survey samples, the resampling procedure should mimic the procedure by which the original sample was drawn (Krewski and Rao, 1981; Rao and Wu, 1988; Rust and Rao, 1996; Lumley, 2011). Appendix C presents an example. Variance estimates can be converted into confidence intervals by a normal approximation to the sampling distribution.

6 Empirical example: Class ceiling in pay

Laurison and Friedman (2016) describe the log incomes of British workers who attain higher managerial and professional occupations (hereafter “professional class”). Among this high-attainment category, mean log income is still lower for those whose father held a working-class occupation. The authors coin the term “class ceiling” for this intriguing descriptive result. This section extends the idea to a related causal estimand: what pay gap by class origin would persist if we *intervened* to assign some individuals to professional class destinations? This gap-closing estimand makes no restrictions on the meaning or causal importance of class origin X (the social category), which simply denotes the collections of individuals with professional vs. working-class parents. The estimand directs attention to the manipulable treatment T : one’s own class destination (a professional vs. a working-class occupation). It becomes important to adjust for observed confounders $\vec{L}$ of the association between class destination and pay (e.g. one’s own education), but not of confounders between class origin and pay (e.g. one’s parents’ education). The end product is a gap-closing estimand that answers a straightforward question: whether an intervention to one’s own class attainment T can break one free from the constraints of class origin X .

I assess this new question in the U.S. context, analyzing data from the 1975–2018 General Social Survey (GSS, Smith et al. 2018), which is conducted each year on a national

probability sample ($N = 12,328$). I define professional class for the respondent and for the father as any occupation in Class I of the Erikson-Goldthorpe-Portocarero (EGP) schema (Erikson et al., 1979), which Laurison and Friedman (2016) highlight as the U.S. parallel to the U.K. National Statistics Socioeconomic Classification used in their study. Examples of professional occupations include manager, engineer, scientist, and lawyer. In the weighted sample, 10.2% of respondents report that their father held a professional occupation (the gap-defining category) and 9.6% report that they personally hold a professional occupation (the gap-closing treatment). The outcome variable is log annual income in 1986 dollars. Because the goal of this example is to illustrate the method rather than to provide the definitive answer to this substantive question, I keep the adjustment set as simple as possible and only include four covariates: race (white, black, other), sex, age, and highest degree (less than high school, high school, junior college, bachelor’s, or a graduate degree). Additional details about the data and sample restrictions are provided in Appendix E. In this example, the unit-specific quantity at the core of the claim is potential log income in a professional occupation. The target population is U.S. adults ages 30–45 in 1975–2018, with each year equally weighted. The theoretical estimand is the unit-specific quantity averaged over the target population.

6.1 Identification

Suppose we took a sample from the population and reassigned the class destinations of that sample. To what degree would the pay gap by class origin close for that sample? Answering this question requires us to identify the causal effect of one’s own occupational class on pay. As in Sec. 4, this requires the assumption that the population average potential log income that *would* be realized in a professional occupation is equal to the observable outcome among those who *factually* hold a professional occupation, within subgroups defined by race, sex, age, education, and father’s occupational class. Because education is such a strong determinant of both occupational attainment and pay, its inclusion in this adjustment set is

essential. Nonetheless, this conditioning set is likely to yield only imperfect identification. Because the aim of this analysis is only to illustrate the method, I proceed with this simple example and leave it to future work to conduct similar analyses in settings where the identification assumptions are more plausible but which would be more complex for illustrating the method.

6.2 Estimation

I estimate the outcome function g and the treatment propensity function m by ordinary least squares and logistic regression, respectively (Appendix Table 2). Using standard regression estimators is useful because it clarifies how the gap closing estimand is distinct from a coefficient, even when the conditional mean functions are estimated with coefficient-based regression models. I aggregate these functions to an estimate by the doubly robust estimator (Eq. 8). I estimate the variance by balanced repeated replication (BRR), a computational strategy for variance estimation in complex survey samples (Krewski and Rao 1981:1013). Appendix C presents the variance estimation procedure. Results with alternative estimators including are reported in Appendix Fig. 11, including machine learning estimators that are more data-adaptive.

6.3 Results

Fig. 6 presents results. Descriptively, log incomes are 0.32 points higher for those from professional class origins compared with working-class origins. One might argue that this pay gap is caused by the unequal rates at which people from these categories attain professional class destinations themselves: 24% among those from professional origins attain professional destinations compared with only 8% among those from working-class origins (Appendix Table 1). The gap-closing estimand allows us to estimate the gap that would persist if we took a sample and assigned them to professional class destinations, thus cutting off this potential explanation for the disparity in that sample. Under that intervention, the

disparity would remain at 0.27 (84% of its original size, Panel A). Assigning a professional destination would increase pay in both groups, but the causal effects are about the same size so that the gap is almost unchanged. Thus, the disparity in class destinations does not explain the pay disparity by class origin—the pay disparity would be almost the same size even if we equalized class destinations. This reinforces the general conclusion of Laurison and Friedman (2016): attaining a professional destination is insufficient to erase the disparity by class origin. It reinforces it to a much larger degree: intervening to assign a professional class destination would reduce the pay gap by class origin by a *tiny amount*. Likewise, a substantial disparity by class origin would persist if we intervened to assign a working-class occupation (Panel B), if we stochastically intervened to assign a random class proportional to their population prevalence (Panel C), or if we stochastically intervened to assign a random class within subgroups of covariates (Panel D). In all cases, the gap-closing estimand is nearly as large as the descriptive disparity.

These results speak to theories of social mobility. Theories often posit a status attainment process that begins with one’s family background as a constraint on life chances. Over the life course, attaining a high level of education or a professional occupation might gradually free one from those constraints, as one’s own status overpowers disparities determined by one’s family of origin (e.g. the discussion of college in Hout 1988). But the evidence here casts doubt on that set of theories. Even if we intervened to assign people to professional occupations, the pay disparity by class origin would almost entirely remain. Occupational attainment does not have the power to liberate individuals from the shadow of their family background.

The test of those theories in this empirical example is limited; all of these claims rely on identification assumptions involving no unobserved confounding, and these assumptions are unlikely to hold in the current example because of the very limited adjustment set. Yet unobserved selection into treatment may actually bias the estimate toward an overstatement of the degree to which an intervention to class destinations would close the pay gap by class

Fig. 6. An intervention to change one’s own social class would do little to close the pay gap across categories of one’s father’s social class. Among those whose father held a professional occupation, mean log income is 0.32 points higher than among those whose father held a working-class occupation. But what if we intervened on a sample to send people personally to professional occupations? Would the gap close? That intervention (Panel A) would causally increase pay in both categories but would leave the gap across categories almost unchanged. Similarly, the gap would be almost unchanged if we counterfactually assigned people to a working-class occupation (Panel B), to an occupational class selected randomly proportional to its prevalence in the population (Panel C), or to an occupational class selected randomly proportional to its prevalence among those who match the covariates of the person in question (Panel D). In no case does an intervention to class destination substantially close the pay gap by class origin. Appendix E presents details for this illustration, including sample selection, definitions of the interventions, and the regression specifications. Conceptually, this figure builds on Laurison and Friedman (2016). Data are pooled from the 1975–2018 General Social Survey ($N = 12,328$). The term “professional class” as used here refers to holding any occupation within Class 1 of the Erikson et al. (1979) class schema, with all other occupations termed “working class.” Results use doubly robust estimators (Eq. 8) with g and m estimated by weighted linear and logistic regression, respectively (Table 2). Confidence intervals are calculated by balanced repeated replication (Appendix C). Appendix Fig. 11 presents estimates under alternative estimation approaches.

origins. Those from working-class origins face greater barriers to occupational attainment, so those who attain professional destinations may be more positively selected among those from working-class origins than professional origins. This would upwardly bias the estimated post-intervention means, but more so among those of working-class origins, which would downwardly bias the gap-closing estimand (discussed in greater depth in Appendix E.5). Due to this bias, the disparity by class origin if we assigned people to professional class destinations might be even bigger than the estimate reported in this paper. Overall, there is good reason to believe that personally attaining a professional class occupation does very little to attenuate the pay disparity by class origin.

7 Contribution and related work

This paper introduces gap-closing estimands for social scientists, drawing on a growing literature on causal decomposition analysis in epidemiology and biostatistics (VanderWeele and Robinson, 2014; Jackson and VanderWeele, 2018; Jackson, 2018; Jackson and Arah, 2020; Jackson, 2021). Studies of fairness in machine learning are beginning to consider the causal process that produces an observed disparity (Zhang and Bareinboim, 2018). The present paper connects those research goals to a broader class of social science settings. In the service of that primary goal, there are three specific contributions: a conceptual contribution delimiting the scope of the intervention, a technical contribution deriving a doubly robust estimator for this setting, and a contribution to the accessibility of these methods by introducing them with examples intended to reach a broad social science audience.

The conceptual contribution delimits the scope of the intervention: the gap-closing estimand is the expected disparity in a sample $\mathcal{S}$ from a population $\mathcal{P}$ if treatment is reassigned for that sample (Fig 1 and 2). In experimental settings, this distinction is obvious: the protocol of any randomized controlled trial makes it clear that treatment is not randomized for the entire population but only for a sample. In observational settings, past work

has already emphasized how appeal to a hypothetical experiment clarifies issues like how the treatment is defined, the assignment process, and who would be eligible (Hernán and Robins, 2016; Hernán, 2016; Hernán et al., 2016). The target trial is important not only for those reasons but also because it allows one to distinguish an intervention applied to a sample $\mathcal{S}$ from one applied to a population $\mathcal{P}$. Even though the intervention sample $\mathcal{S}$ is purely hypothetical in observational settings, the distinction clarifies the degree to which interference (a violation of consistency) threatens the validity of interpretations. When we conceptualize $\mathcal{S}$ as a small fraction of $\mathcal{P}$, interference problems are greatly reduced because most of the people with whom any unit i interacts would not experience a change in treatment assignment under the intervention. When we conceptualize a treatment applied to the full population $\mathcal{S} = \mathcal{P}$, there is an enormous risk of interference that would need to be modeled formally (Jackson and Arah, 2020). A target trial that explicitly appeals to a sample $\mathcal{S}$ is not a methodological fix for this problem; it is a conceptual framework for explicitly navigating the tension between the credibility of the claim (most credible if the size of $\mathcal{S}$ is much less than $\mathcal{P}$) and the scope of the claim (the size of $\mathcal{S}$).

The technical contribution is a doubly robust estimator for this particular setting: stochastic treatment assignments in a complex survey sample. Doubly robust estimators have a long history in causal inference (Robins et al., 1994; Bang and Robins, 2005; Kang and Schafer, 2007), including for the setting with stochastic treatment assignments (see Dudík et al. 2014 Sec. 3.3 and Murphy et al. 2001 Sec. 5.2).⁴ There is also growing interest in applying doubly robust estimators in complex survey samples (Rudolph et al., 2014), albeit with a focus on non-stochastic treatment assignment rules. The present paper proves double robustness for stochastic treatment assignment rules when the sample contains units selected from the population with unequal probabilities (proofs in Appendix B). Given a doubly robust estimator, the extension to machine learning estimation of nuisance functions with unknown functional forms follows directly from the literature for causal effects (Van der Laan

⁴Continuous treatments are one setting in which doubly robust estimators for stochastic assignment rules have received considerable interest (Díaz Muñoz and van der Laan, 2012; Kennedy et al., 2017).

and Rose, 2011; Chernozhukov et al., 2018). This paper also argues that double robustness for the treatment and outcome prediction functions $m()$ and $g()$ is important because these can be complex, whereas the researcher can control the complexity of the counterfactual rule $\pi()$ for which there is no robustness to misspecification.

Each contribution above serves the third and main contribution: bringing a range of existing ideas together in an accessible framework designed to support social science inquiry to define interventions, estimate the resulting disparities, and inform policy to close gaps. By illustrating the relevance of gap-closing estimands to a range of substantive questions about inequality, this paper builds a bridge between the questions sociologists are already asking implicitly and a set of methods that can answer those questions more explicitly. In the service of this third contribution, the `gapclosing` R package provides publicly available and open source software to support the methods described in this paper.

7.1 Comparing coefficients across regressions does not estimate a gap-closing estimand

There is a common research practice that does not estimate a gap-closing estimand. Suppose a researcher estimates two regression models, one with and one without the treatment variable T .

$$\mathbf{E}(Y \mid X, \vec{L}) = \alpha_{\text{Model 1}} + \beta_{\text{Model 1}}X + \vec{\eta}'_{\text{Model 1}}\vec{L} \quad (9)$$

$$\mathbf{E}(Y \mid X, \vec{L}, T) = \alpha_{\text{Model 2}} + \beta_{\text{Model 2}}X + \vec{\eta}'_{\text{Model 2}}\vec{L} + \gamma_{\text{Model 2}}T \quad (10)$$

Suppose both functional forms are correct (matching the true conditional mean function). Suppose also that the adjustment set $\{X, \vec{L}\}$ identifies the causal effect of T (as in Fig 3). The researcher might interpret the difference $\beta_{\text{Model 1}} - \beta_{\text{Model 2}}$ as the amount of the disparity “attributable” to the treatment. In the class gap in pay example, adding class destination (T) to an OLS model of log income (Y) would reduce the coefficient on class origin (X)

from $\hat{\beta}_{\text{Model 1}} = 0.05$ to $\hat{\beta}_{\text{Model 2}} = 0.02$ (Appendix Table 3). One might infer (incorrectly) that over half of the pay gap by class origin is “attributable” to class destination. But that comparison speaks to a complicated quantity.

$$\beta_{\text{Model 1}} - \beta_{\text{Model 2}} = \overbrace{\left[\mathbf{E}(Y \mid X = 1, \vec{L} = \vec{\ell}) - \mathbf{E}(Y \mid X = 0, \vec{L} = \vec{\ell}) \right]}^{\text{Disparity within subgroup } \vec{L} = \vec{\ell}} - \underbrace{\left[\mathbf{E}(Y(t) \mid X = 1, \vec{L} = \vec{\ell}) - \mathbf{E}(Y(t) \mid X = 0, \vec{L} = \vec{\ell}) \right]}_{\text{Disparity within subgroup } \vec{L} = \vec{\ell} \text{ under intervention to set } T = t} \quad \forall t, \vec{\ell} \quad (11)$$

This quantity focuses on a subpopulation with a particular covariate set $\vec{L} = \vec{\ell}$. Then, it involves the disparity that would persist in that subpopulation if the treatment were set to some value $T = t$. Because there are no interactions in Eq. 9–10, the comparison between the two is (by assumption) the same at every value of $\vec{\ell}$ and t .

The comparison in Eq. 11 is difficult to interpret for two reasons. First, neither $\beta_{\text{Model 1}}$ nor $\beta_{\text{Model 2}}$ is the marginal descriptive disparity. Both adjust for covariates $\vec{L}$ that may lie along the causal path $X \rightarrow \vec{L} \rightarrow Y$ from the category to the outcome. Holding them constant blocks an important path. For the same reason, neither coefficient is a gap-closing estimand—an intervention to equalize the treatment would not equalize the covariates $\vec{L}$. You might still interpret Eq. 11 as the amount that the intervention would close the gap net of the covariates $\vec{L}$. But then there is a second complexity: endogenous selection bias opens a new path $X \rightarrow \vec{L} \leftarrow U \rightarrow Y$ which makes the disparity hard to interpret. For example, consider the pay disparity by class origin X among those who complete a college degree $\vec{L} = \vec{\ell}$. Because of barriers to college access, people from working-class origins $X = 0$ who nonetheless overcome the odds to complete college $\vec{L} = \vec{\ell}$ may have more work ethic (an unobserved U) than their counterparts who did not complete college. Adjusting for education $\vec{L}$ creates that association even if work ethic U is marginally unrelated to class origin X . The reason interpretations are difficult is because the covariates $\vec{L}$ are possibly consequences of the gap-defining category X . The gap-closing estimand avoids both of the

issues by not holding $\vec{L}$ constant. The covariates $\vec{L}$ are used to learn the algorithm to predict potential outcomes. But then those predictions are made for each unit i with the covariates at their observed values $\vec{L}_i$. Thus, the gap-closing estimand addresses an intervention to the treatment T while allowing the covariates $\vec{L}$ to remain as they factually exist.

7.2 Past work on neighboring descriptive and causal topics does not speak to gap-closing estimands

Descriptive work may seem related but is distinct from gap-closing estimands because it does not invoke a causal claim. Substantial scholarship within sociology has summarized disparities by Kitagawa-Blinder-Oaxaca decompositions (Kitagawa, 1955; Blinder, 1973; Oaxaca, 1973). This technique (or the extension to generalized linear models by Fairlie 2005) appears in examinations of inequality over numerous social categories in sociology: class origins (Laurison and Friedman, 2016), race (Ciocca Eller and DiPrete, 2018), disability (Shandra, 2018), sexual orientation (Mize, 2016), and gender (Weisshaar, 2017), to name a few. Absent causal assumptions, these decompositions are purely descriptive: they provide evidence about the disparity across categories among units who are identical along all covariates. Gap-closing estimands are different because they speak to whether an intervention would causally close a disparity. To use Kitagawa-Blinder-Oaxaca decompositions to estimate gap-closing estimands requires careful interpretation under a specific set of causal assumptions and a specific assumed functional form (Jackson and VanderWeele, 2018).⁵

Causal work on controlled direct effects may seem related but is distinct because it posits a different kind of intervention. A controlled direct effect (Pearl, 2001; Robins, 2003; Acharya et al., 2016; Zhou and Wodtke, 2019) is best understood in the context of a hypothetical experiment: the expected difference in an outcome Y if we intervene to assign someone to treatment $X = x'$ or control $X = x$ while intervening to hold a post-treatment

⁵An and Glynn (2019) extend Kitagawa-Blinder-Oaxaca decomposition to the related setting in which all covariates are pre-treatment and the decomposition is of a disparity over the treatment conditions, which is not a gap-closing estimand because the category is taken to be a causal treatment in that analysis.

intervention at some fixed value $T = t$. In potential outcomes notation, the CDE involves potential outcomes $y_i(x, t)$ defined over both the category and the treatment. In contrast, the gap-closing estimand involves potential outcomes $y_i(t)$ defined over the treatment only, which are then aggregated over collections of units i defined by X (which is never subject to a hypothetical intervention). This essential distinction makes the gap-closing estimand applicable in settings where mediation estimands are not: when the category X is a social category for which the causal effect is philosophically fraught. Nonetheless, methods developed to estimate controlled direct effects have analogs for gap-closing estimands. Structural nested mean models for controlled direct effects (Vansteelandt, 2009) are analogous to estimation by predicted outcomes; both depend on models of the conditional mean of the outcome. Inverse probability weighting (VanderWeele, 2009) is directly analogous to estimation by predicted treatments. Other methods developed for controlled direct effects, such as regression with residuals (Zhou and Wodtke, 2019), could point toward analogous procedures for gap-closing estimands.

8 Discussion

Gap-closing estimands provide opportunities to not only study disparities, but to learn about interventions to close them. By making an explicitly causal claim, researchers who estimate gap-closing estimands gain transparency about required causal assumptions needed for identification. They also gain opportunities to estimate by flexible predictive algorithms, which are implemented in open-source software (the `gapclosing` package). Gap-closing estimands have the potential to promote deeper understanding of gaps across social categories and how to close them.

The knowledge we can gain from gap-closing estimands complements existing bodies of research that focus on descriptive disparities and causal assessments of discrimination. Descriptive research could proceed in tandem with gap-closing research; a descriptive study

documenting the presence of a large disparity would set the stage for subsequent studies to explore interventions to close that disparity. Causal assessments of discrimination in audit studies provide one type of understanding about why a gap exists: a gap may exist because decision makers react differently when they perceive a person to be of one category versus another. Gap-closing estimands provide a complementary type of understanding about disparities: how would a gap change if some other treatment variable took a different value? To fully understand disparities, we need both types of understanding.

An embrace of gap-closing estimands would both sharpen theory and change the language with which social scientists discuss race, class, and gender in observational studies. Too often, theory about disparities involves vague claims about the role that some treatment T plays in generating the gap. Gap-closing estimands lend new theoretical precision to those claims: the treatment T plays a role in the sense that an intervention to reassign that treatment by a new rule in a sample would change the gap in a particular way. Thus, the treatment is a cause of the gap. Language for discussing the categories would also change. We would no longer speak of the “effect” or “influence” (in scare quotes) of categories like race, class, and gender in observational studies. After all, a hypothetical intervention to change someone’s race, class, or gender (not just a signal, but the entire construct) is neither straightforward to define nor especially relevant for policy. Instead, researchers would be clear that the only effect identified is the effect of a manipulable treatment. Doing so would create space for substantive arguments about causal assumptions and technical advances in estimation, both of which would be grounded in a research goal defined outside of the statistical algorithm (Lundberg et al., 2021).

Substantively, a shift away from conditional comparisons (e.g. coefficients that statistically adjust for many covariates) and toward gap-closing estimands (e.g. the outcome of an intervention on one variable) might reveal that post-intervention disparities are actually *larger* than researchers might have otherwise thought. This shift would improve rhetorical clarity and also contribute to the policy-relevance of research: policymakers can understand

that the research implies that an intervention to the variable studied might plausibly close a gap. Finally, formalizing the hypothetical experiment provides an opportunity to clarify the scope of the intervention (to a sample vs. to the population). Gap-closing estimands provide a framework for clarity about the target of statistical inference, thereby promoting transparent research and new estimators like the one developed in this paper to help us build evidence on interventions to close gaps.

References

- Acharya, A., M. Blackwell, and M. Sen 2016. Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review*, 110(3):512–529.
- An, W. and A. N. Glynn 2019. Treatment effect deviation as an alternative to Blinder-Oaxaca decomposition for studying social inequality. *Sociological Methods & Research*, 50:1006–1033.
- Aronow, P. M., C. Samii, et al. 2017. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947.
- Bang, H. and J. M. Robins 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.
- Bertrand, M. and S. Mullainathan 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013.
- Bickel, P. J. 1982. On adaptive estimation. *The Annals of Statistics*, Pp. 647–671.
- Blau, F. D. and L. M. Kahn 2017. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865.
- Blinder, A. S. 1973. Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, Pp. 436–455.
- Breiman, L. 2001. Random forests. *Machine Learning*, 45(1):5–32.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.
- Chetty, R., N. Hendren, P. Kline, E. Saez, and N. Turner 2014. Is the United States still a land of opportunity? Recent trends in intergenerational mobility. *American Economic Review*, 104(5):141–47.
- Ciocca Eller, C. and T. A. DiPrete 2018. The paradox of persistence: Explaining the black-white gap in bachelor’s degree completion. *American Sociological Review*, 83(6):1171–1214.
- Díaz Muñoz, I. and M. van der Laan 2012. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549.
- Dudík, M., D. Erhan, J. Langford, L. Li, et al. 2014. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511.
- Efron, B. and R. J. Tibshirani 1994. *An Introduction to the Bootstrap*. CRC Press.
- Elwert, F. and C. Winship 2014. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40:31–53.
- Erikson, R., J. H. Goldthorpe, and L. Portocarero 1979. Intergenerational class mobility in three Western European societies: England, France and Sweden. *The British Journal of Sociology*, 30(4):415–441.
- 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.
- Glynn, A. N. and K. M. Quinn 2010. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1):36–56.
- Greenland, S. 2003. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306.
- Greenland, S., J. Pearl, and J. M. Robins 1999. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

- Greiner, D. J. and D. B. Rubin 2011. Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93(3):775–785.
- Hernán, M. A. 2016. Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*, 26(10):674–680.
- Hernán, M. A. and J. M. Robins 2016. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764.
- Hernán, M. A. and J. M. Robins 2020. *Causal Inference: What if*. Chapman & Hall/CRC.
- Hernán, M. A., B. C. Sauer, S. Hernández-Díaz, R. Platt, and I. Shrier 2016. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology*, 79:70–75.
- Hill, J. L. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.
- Holland, P. W. 1986. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.
- Holland, P. W. 2008. Causation and race. *White Logic, White Methods: Racism and Methodology*, Pp. 93–109.
- Horowitz, J. 2018. Relative education and the advantage of a college degree. *American Sociological Review*, 83(4):771–801.
- Horvitz, D. G. and D. J. Thompson 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.
- Hout, M. 1988. More universalism, less structural mobility: The American occupational structure in the 1980s. *American Journal of Sociology*, 93(6):1358–1400.
- Imbens, G. W. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.
- Imbens, G. W. and D. B. Rubin 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- Jackson, J. W. 2018. On the interpretation of path-specific effects in health disparities research. *Epidemiology*, 29(4):517–520.
- Jackson, J. W. 2021. Meaningful causal decompositions in health equity research: definition, identification, and estimation through a weighting framework. *Epidemiology*, 32(2):282–290.
- Jackson, J. W. and O. A. Arah 2020. Invited commentary: making causal inference more social and (social) epidemiology more causal. *American Journal of Epidemiology*, 189(3):179–182.
- Jackson, J. W. and T. J. VanderWeele 2018. Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology*, 29(6):825–835.
- Kang, J. D. and J. L. Schafer 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.
- Kaufman, J. S. 2008. Epidemiologic analysis of racial/ethnic disparities: some fundamental issues and a cautionary example. *Social Science & Medicine*, 66(8):1659–1669.
- Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small 2017. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(4):1229.
- Killewald, A. and B. Bryan 2016. Does your home make you wealthy? *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 2(6):110–128.
- Kitagawa, E. M. 1955. Components of a difference between two rates. *Journal of the American Statistical Association*, 50(272):1168–1194.

- Kohler-Hausmann, I. 2018. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Northwestern University Law Review*, 113:1163.
- Krewski, D. and J. Rao 1981. Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, Pp. 1010–1019.
- Laurison, D. and S. Friedman 2016. The class pay gap in higher professional and managerial occupations. *American Sociological Review*, 81(4):668–695.
- Lee, B. K., J. Lessler, and E. A. Stuart 2010. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.
- Lucas, S. R. 2001. Effectively maintained inequality: Education transitions, track mobility, and social background effects. *American Journal of Sociology*, 106(6):1642–1690.
- Lumley, T. 2011. *Complex Surveys: A Guide to Analysis Using R*, volume 565. John Wiley & Sons.
- Lundberg, I., R. Johnson, and B. M. Stewart 2021. What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review*, 86(3):532–565.
- McCaffrey, D. F., G. Ridgeway, and A. R. Morral 2004. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403–425.
- McIntosh, K., E. Moss, R. Nunn, and J. Shambaugh 2020. Examining the black-white wealth gap. *Washington DC: Brookings Institutes*.
- Mize, T. D. 2016. Sexual orientation in the labor market. *American Sociological Review*, 81(6):1132–1160.
- Murphy, S. A., M. J. van der Laan, J. M. Robins, and C. P. P. R. Group 2001. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423.
- Naimi, A. I., M. E. Schnitzer, E. E. Moodie, and L. M. Bodnar 2016. Mediation analysis for health disparities research. *American Journal of Epidemiology*, 184(4):315–324.
- Nguyen, T. Q., I. Schmid, and E. A. Stuart 2020. Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods*.
- Oaxaca, R. 1973. Male-female wage differentials in urban labor markets. *International Economic Review*, Pp. 693–709.
- Omi, M. and H. Winant 1994. *Racial Formation in the United States*. Routledge.
- Pager, D. 2003. The mark of a criminal record. *American Journal of Sociology*, 108(5):937–975.
- Pearl, J. 2000. *Causality: Models, Reasoning and Inference*. Springer.
- Pearl, J. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, Pp. 411–420. Morgan Kaufmann Publishers Inc.
- Petersen, T. and L. A. Morgan 1995. Separate and unequal: Occupation-establishment sex segregation and the gender wage gap. *American Journal of Sociology*, 101(2):329–365.
- Raftery, A. E. and M. Hout 1993. Maximally maintained inequality: Expansion, reform, and opportunity in Irish education, 1921–75. *Sociology of Education*, Pp. 41–62.
- Rao, J. N. and C. Wu 1988. Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401):231–241.
- Robins, J. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512.
- Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky 2007. Comment: Performance of double-robust estimators when ‘inverse probability’ weights are highly variable. *Statistical Science*, 22(4):544–559.

- Robins, J. M. 2003. Semantics of causal DAG models and the identification of direct and indirect effects. *Oxford Statistical Science Series*, Pp. 70–82.
- Robins, J. M., A. Rotnitzky, and L. P. Zhao 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.
- Rubin, D. B. 1986. Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962.
- Rudolph, K. E., I. Díaz, M. Rosenblum, and E. A. Stuart 2014. Estimating population treatment effects from a survey subsample. *American Journal of Epidemiology*, 180(7):737–748.
- Rust, K. F. and J. Rao 1996. Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5(3):283–310.
- Saperstein, A. and A. M. Penner 2012. Racial fluidity and inequality in the United States. *American Journal of Sociology*, 118(3):676–727.
- Scharfstein, D. O., A. Rotnitzky, and J. M. Robins 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.
- Sen, M. and O. Wasow 2016. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19:499–522.
- Shandra, C. L. 2018. Disability as inequality: Social disparities, health disparities, and participation in daily activities. *Social Forces*, 97(1):157–192.
- Smith, T. W., M. Davern, J. Freese, and S. Morgan 2018. *General Social Surveys, 1972-2018 [machine-readable data file]*. Sponsored by National Science Foundation. –NORC ed.– Chicago: NORC, 2018: NORC at the University of Chicago [producer and distributor]. Data accessed from the GSS Data Explorer website at gssdataexplorer.norc.org.
- Torche, F. 2011. Is a college degree still the great equalizer? Intergenerational mobility across levels of schooling in the United States. *American Journal of Sociology*, 117(3):763–807.
- U.S. Census Bureau 2019. Educational attainment in the United States: 2018. <https://www.census.gov/data/tables/2018/demo/education-attainment/cps-detailed-tables.html>.
- Van der Laan, M. J. and S. Rose 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media.
- VanderWeele, T. and W. Robinson 2014. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology*, 25(4):473–484.
- VanderWeele, T. J. 2009. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, Pp. 18–26.
- Vansteelandt, S. 2009. Estimating direct effects in cohort and case-control studies. *Epidemiology*, Pp. 851–860.
- Weisshaar, K. 2017. Publish and perish? An assessment of gender gaps in promotion to tenure in academia. *Social Forces*, 96(2):529–560.
- Western, B. 2006. *Punishment and Inequality in America*. Russell Sage Foundation.
- Wright, M. N. and A. Ziegler 2017. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1).
- Zhang, J. and E. Bareinboim 2018. Fairness in decision-making—The causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- 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. and G. T. Wodtke 2019. A regression-with-residuals method for estimating controlled direct effects. *Political Analysis*, 27(3):360–369.

A Identification proof for gap-closing estimands

This section provides proofs for identification: the link between unobserved counterfactuals and statistical quantities involving observable data. This section does not address the estimation task—translating observable data into an estimate—which is the subject of proofs in Appendix B.

A gap-closing estimand $\tau_{x',x}(t)$ is identified under assumptions of consistency, positivity, and conditional mean independence. The first required assumption (consistency) links the observable data to the potential outcomes: the observed outcome for unit i is the potential outcome under the treatment assignment that unit i actually experiences.

Assumption 1. (Consistency) $Y_i = y_i(T_i) \quad \forall i$

The second assumption (conditional mean independence) addresses confounding: within each subpopulation of observed covariates $\{X = x, \vec{L} = \vec{\ell}\}$, the mean potential outcome is assumed to be the same for treated units ($T = t$) as for all units.

Assumption 2. (Conditional mean independence)

$$\mathbf{E}\left(Y(t) \mid X = x, \vec{L} = \vec{\ell}, T = t\right) = \mathbf{E}\left(Y(t) \mid X = x, \vec{L} = \vec{\ell}\right) \quad \forall x, \vec{\ell}$$

In Assumption 2, $Y(t)$ denotes a random variable that would take the specific value $y_i(t)$ when a unit i is selected from the population. The expectation is over units in the subpopulation defined by the conditional (e.g. the set of i such that $X_i = x$, $\vec{L}_i = \vec{\ell}$, and $T_i = t$).

The third assumption (positivity) is needed for nonparametric identification: it requires that treatment value t factually occurs with non-zero probability in every covariate stratum. As in Assumption 2, subscripts are omitted because the probability is taken over all units i in the subpopulation with $X_i = x$ and $\vec{L}_i = \vec{\ell}$.

Assumption 3. (Positivity) $P(T = t \mid X = x, \vec{L} = \vec{\ell}) > 0 \quad \forall x, \vec{\ell}$

A post-intervention mean $\theta_x(t)$ in category x under treatment value t is identified under consistency, positivity, and conditional mean independence by the following proof.

$$\theta_x(t) = \mathbf{E}(Y(t) \mid X = x) \quad (12)$$

$$= \mathbf{E} \left(\mathbf{E} \left(Y(t) \mid X = x, \vec{L} \right) \mid X = x \right) \quad \text{by the law of iterated expectation} \quad (13)$$

$$= \mathbf{E} \left(\mathbf{E} \left(Y(t) \mid X = x, \vec{L}, T = t \right) \mid X = x \right) \quad \text{by conditional mean independence} \quad (14)$$

$$= \mathbf{E} \left(\mathbf{E} \left(Y \mid X = x, \vec{L}, T = t \right) \mid X = x \right) \quad \text{by consistency} \quad (15)$$

The inner conditional mean is identified for each stratum of $\vec{L}$ by the positivity assumption, which guarantees that the population contains some units with $\{X = x, \vec{L} = \vec{\ell}, T = t\}$. Thus, $\theta_x(t)$ is nonparametrically identified.

A similar proof holds for identification with stochastic interventions.

$$\theta_x(\pi) = \mathbf{E} \left(\sum_t \pi(t, \vec{L}) Y(t) \mid X = x \right) \quad (16)$$

$$= \mathbf{E} \left(\sum_t \mathbf{E} \left(\pi(t, \vec{L}) Y(t) \mid X = x, \vec{L} \right) \mid X = x \right) \quad \text{by the law of iterated expectation} \quad (17)$$

$$= \mathbf{E} \left(\sum_t \pi(t, \vec{L}) \mathbf{E} \left(Y(t) \mid X = x, \vec{L} \right) \mid X = x \right) \quad \text{pull constant out of the inner } \mathbf{E} \quad (18)$$

$$= \mathbf{E} \left(\sum_t \pi(t, \vec{L}) \mathbf{E} \left(Y(t) \mid X = x, \vec{L}, T = t \right) \mid X = x \right) \quad \text{by conditional mean independence} \quad (19)$$

$$= \mathbf{E} \left(\sum_t \pi(t, \vec{L}) \mathbf{E} \left(Y \mid X = x, \vec{L}, T = t \right) \mid X = x \right) \quad \text{by consistency} \quad (20)$$

These proofs also imply identification in settings beyond simple random samples as long as the conditional means $\mathbf{E}(Y \mid X = x, \vec{L}, T = t)$ are identified. This includes samples selected with unequal but known probabilities, for which a weighted mean in the sample can estimate the conditional mean in the population.

B Proofs that estimators are consistent

This section provides proofs about the proposed estimators, showing that under certain conditions they are consistent for the true unknown parameters. Section B.1 defines the notation to be used throughout the proofs. Section B.2 proves that estimation with predicted outcomes is consistent if the outcome prediction function is consistent. Section B.3 proves that estimation with predicted treatment probabilities is consistent if the treatment prediction function is consistent. Sections B.4 and B.5 show that doubly-robust estimation is consistent if the prediction function is consistent for outcomes or treatment probabilities, respectively.

B.1 Preliminaries

These proofs assume for simplicity that the population is infinite. The goal is to show that sample estimates converge to population parameters as the sample size grows. To denote the population parameters without repeatedly invoking limits, I use conditional expectation and probability operators.

$$P(\text{Event} \mid \text{Condition}) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i:\text{Condition}_i=\text{True}} (\text{Event}_i = \text{True}) \quad (21)$$

$$E(\text{Outcome} \mid \text{Condition}) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i:\text{Condition}_i=\text{True}} \text{Outcome}_i \quad (22)$$

I use the convergence operator $\rightarrow$ to denote convergence in probability, as in

$$\frac{1}{n} \sum_{i:\text{Condition}_i=\text{True}} \text{Outcome}_i \rightarrow E(\text{Outcome} \mid \text{Condition}) \quad (23)$$

where the quantity at the left is stochastic (because of sampling n units from the population) while the quantity on the right is a fixed but unknown population parameter.

A sampling indicator S_i indicates whether unit i is included in the observed sample,

and S_i is taken to be a random variable for each i . All other variables subscripted by i are taken to be fixed quantities (e.g. X_i). When not subscripted by i (e.g. X), these are taken to be random variables with stochasticity deriving from the random sampling of a unit i from the population. I assume throughout that sample inclusion is inversely proportional to a known weight η_i so that all variables are independent of sample inclusion given η .

Assumption 4. (Ignorable sample inclusion) $S \perp\!\!\!\perp \{Y, X, T, \vec{L}\} \mid \eta$

To simplify notation, I use shorthand for strata-specific sums of the sample weight η .

$$\eta_x = \sum_{i: X_i=x} \eta_i \quad (24)$$

$$\eta_{x\vec{\ell}} = \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}} \eta_i \quad (25)$$

$$\eta_{x\vec{\ell}t} = \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \eta_i \quad (26)$$

All proofs assumes that if the assignment rule $\pi(t, \vec{\ell})$ has to be learned from the data, that a consistent estimator $\hat{\pi}(t, \vec{\ell}) \xrightarrow{P} \pi(t, \vec{\ell})$ is available. I also assume consistency of either the outcome prediction function

$$\hat{g}(t, x, \vec{\ell}) \xrightarrow{P} g(t, x, \vec{\ell}) = \mathbf{E} \left(Y \mid T = t, X = x, \vec{L} = \vec{\ell} \right) \quad \forall t, x, \vec{\ell} \quad (27)$$

or the treatment prediction function,

$$\hat{m}(t, x, \vec{\ell}) \xrightarrow{P} m(t, x, \vec{\ell}) = \mathbf{P} \left(T = t \mid X = x, \vec{L} = \vec{\ell} \right) \quad \forall t, x, \vec{\ell} \quad (28)$$

B.2 Estimation by predicted outcomes

Suppose we have consistent estimators for the counterfactual assignment rule $\hat{\pi}(t, \vec{\ell}) \xrightarrow{P} \pi(t, \vec{\ell})$ and for the conditional mean of the outcome $\hat{g}(t, x, \vec{\ell}) \xrightarrow{P} g(t, x, \vec{\ell}) = \mathbf{E}(Y \mid T = t, X = x, \vec{L} = \vec{\ell})$. This would be the case, for instance, if $\pi(t, \vec{\ell})$ assigns each treatment t according to its population prevalence, for which a consistent nonparametric estimator is the mean of a simple random sample, and if $g(\cdot)$ is a correctly-specified outcome OLS regression. Then the estimator by predicted outcomes is consistent for the post-intervention mean.

$$\hat{\theta}_x^{\text{Outcome}}(\pi) = \frac{1}{\eta_x} \sum_{i: X_i = x} \eta_i \sum_t \hat{\pi}(t, \vec{L}_i) \hat{g}(t, X_i, \vec{L}_i) \quad (29)$$

Make the sum over covariate strata $\vec{\ell}$ explicit

$$= \frac{1}{\eta_x} \sum_{\vec{\ell}} \sum_{i: X_i = x, \vec{L}_i = \vec{\ell}} \eta_i \sum_t \hat{\pi}(t, \vec{L}_i) \hat{g}(t, X_i, \vec{L}_i) \quad (30)$$

Multiply by $\frac{\eta_{x\vec{\ell}}}{\eta_{x\vec{\ell}}}$ and rearrange

$$= \sum_{\vec{\ell}} \frac{\eta_{x\vec{\ell}}}{\eta_x} \sum_t \underbrace{\frac{1}{\eta_{x\vec{\ell}}} \sum_{i: X_i = x, \vec{L}_i = \vec{\ell}} \eta_i \overbrace{\hat{\pi}(t, \vec{L}_i) \hat{g}(t, X_i, \vec{L}_i)}^{\text{Constant within inner-most sum}}}_{\text{Weighted sample mean of a constant}} \quad (31)$$

$$= \sum_{\vec{\ell}} \underbrace{\frac{\eta_{x\vec{\ell}}}{\eta_x}}_{\rightarrow \mathbf{P}(\vec{L} = \vec{\ell} \mid X = x)} \sum_t \underbrace{\hat{\pi}(t, \vec{\ell})}_{\rightarrow \pi(t, \vec{\ell})} \underbrace{\hat{g}(t, x, \vec{\ell})}_{\rightarrow g(t, x, \vec{\ell})} \quad (32)$$

$$\rightarrow \sum_{\vec{\ell}} \mathbf{P}(\vec{L} = \vec{\ell} \mid X = x) \sum_t \pi(t, \vec{\ell}) g(t, x, \vec{\ell}) \quad (33)$$

Replace $g(\cdot)$ with its definition

$$\rightarrow \sum_{\vec{\ell}} \mathbf{P}(\vec{L} = \vec{\ell} \mid X = x) \sum_t \pi(t, \vec{\ell}) \mathbf{E}(Y \mid T = t, X = x, \vec{L} = \vec{\ell}) \quad (34)$$

Invoke causal identification assumptions

$$\rightarrow \sum_{\vec{\ell}} \mathbf{P}(\vec{L} = \vec{\ell} \mid X = x) \sum_t \pi(t, \vec{\ell}) \mathbf{E}(Y(t) \mid X = x, \vec{L} = \vec{\ell}) \quad (35)$$

Replace the sum with an expectation over the random $\vec{L}$

$$\rightarrow \mathbf{E} \left(\sum_t \pi(t, \vec{L}) Y(t) \mid X = x \right) \quad (36)$$

$$\rightarrow \theta_x(\pi) \quad (37)$$

B.3 Estimation by predicted treatment probabilities

Suppose we have consistent estimators for the counterfactual assignment rule $\hat{\pi}(t, \vec{\ell}) \xrightarrow{P} \pi(t, \vec{\ell})$ and for the conditional treatment probabilities $\hat{m}(t, x, \vec{\ell}) \xrightarrow{P} m(t, x, \vec{\ell}) = P(T = t \mid X = x, \vec{L} = \vec{\ell})$. This would be the case, for instance, if $m(\cdot)$ is a correctly-specified logistic regression and if $\pi(t, \vec{\ell})$ is $m(\cdot)$ marginalized over the gap-defining category X and possibly some of $\vec{L}$. Then the estimator with predicted treatment probabilities is consistent for the post-intervention mean.

$$\hat{\theta}_x^{\text{Treatment}} = \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)} Y_i \quad (38)$$

Make the sum over covariate strata $\vec{\ell}$ and treatments t explicit

$$= \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{\vec{\ell}} \sum_t \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \eta_i Y_i \quad (39)$$

Multiply by one: $\frac{\eta_x \vec{\ell} t}{\eta_x \vec{\ell} t}$

$$= \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{\vec{\ell}} \sum_t \frac{\eta_x \vec{\ell} t}{\eta_x \vec{\ell} t} \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \underbrace{\frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})}}_{\text{Constant within inner sum}} \eta_i Y_i \quad (40)$$

Pull the constant out of the inner sum and rearrange

$$= \sum_{\vec{\ell}} \sum_t \underbrace{\frac{\eta_x \vec{\ell} t}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}}}_{\substack{\rightarrow P(\vec{L}=\vec{\ell}, T=t \mid X=x) \\ \text{Proof in Eq. 46-52}}} \underbrace{\frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})}}_{\substack{\rightarrow \frac{\pi(t, \vec{\ell})}{P(T=t, \mid X=x, \vec{L}=\vec{\ell})}}} \underbrace{\frac{1}{\eta_x \vec{\ell} t} \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \eta_i Y_i}_{\rightarrow E(Y \mid X=x, \vec{L}=\vec{\ell}, T=t)} \quad (41)$$

$$\rightarrow \sum_{\vec{\ell}} \sum_t \underbrace{\frac{P(\vec{L}=\vec{\ell}, T=t \mid X=x)}{P(T=t, \mid X=x, \vec{L}=\vec{\ell})}}_{\substack{= P(\vec{L}=\vec{\ell} \mid X=x) \\ \text{Proof in Eq. 53-54}}} \pi(t, \vec{\ell}) E(Y \mid X=x, \vec{L}=\vec{\ell}, T=t) \quad (42)$$

Invoke causal identification assumptions for the expectation

$$\rightarrow \sum_{\vec{\ell}} \sum_t P(\vec{L}=\vec{\ell} \mid X=x) \pi(t, \vec{\ell}) E(Y(t) \mid X=x, \vec{L}=\vec{\ell}) \quad (43)$$

Replace the sum with an expectation over the random $\vec{L}$

$$\rightarrow E \left(\sum_t \pi(t, \vec{L}) Y(t) \mid X=x \right) \quad (44)$$

$$\rightarrow \theta_x(\pi) \quad (45)$$

Eq 41 above posits that a ratio of sums of weights converges to $P(\vec{L} = \vec{\ell}, T = t \mid X = x)$. That is proven below.

$$\frac{\eta_x \vec{\ell} t}{\sum_{i: X_i = x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} = \frac{\eta_x \vec{\ell} t}{\sum_{\vec{\ell}} \sum_t \sum_{i: X_i = x, \vec{L}_i = \vec{\ell}, T_i = t} \eta_i \frac{\hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \quad (46)$$

Pull out constants from the sum

$$= \frac{\eta_x \vec{\ell} t}{\sum_{\vec{\ell}} \sum_t \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \sum_{i: X_i = x, \vec{L}_i = \vec{\ell}, T_i = t} \eta_i} \quad (47)$$

Simplify the inner-most sum

$$= \frac{\eta_x \vec{\ell} t}{\sum_{\vec{\ell}} \sum_t \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \eta_x \vec{\ell} t} \quad (48)$$

Multiply by $\frac{1}{\eta_x}$ on top and bottom and by $\frac{\eta_x \vec{\ell}}{\eta_x \vec{\ell}}$ on the bottom

$$= \frac{\frac{1}{\eta_x} \eta_x \vec{\ell} t}{\sum_{\vec{\ell}} \frac{\eta_x \vec{\ell}}{\eta_x} \sum_t \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \frac{1}{\eta_x \vec{\ell}} \eta_x \vec{\ell} t} \quad (49)$$

Replace estimates with their probability limit parameters

$$\rightarrow \frac{P(\vec{L} = \vec{\ell}, T = t \mid X = x)}{\sum_{\vec{\ell}} P(\vec{L} = \vec{\ell} \mid X = x) \sum_t \frac{\pi(t, \vec{\ell})}{P(T=t \mid X=x, \vec{L}=\vec{\ell})} P(T = t \mid X = x, \vec{L} = \vec{\ell})} \quad (50)$$

Cancel terms in the internal sum in the denominator

$$\rightarrow \frac{P(\vec{L} = \vec{\ell}, T = t \mid X = x)}{\underbrace{\sum_{\vec{\ell}} P(\vec{L} = \vec{\ell} \mid X = x)}_{=1} \underbrace{\sum_t \pi(t, \vec{\ell})}_{=1}} \quad (51)$$

$$= P(\vec{L} = \vec{\ell}, T = t \mid X = x) \quad (52)$$

Eq. 42 above posits that a ratio involving conditional probabilities has a simpler form. That is proven below by expanding the numerator and then canceling terms.

$$\frac{P(\vec{L} = \vec{\ell}, T = t \mid X = x)}{P(T = t, \mid X = x, \vec{L} = \vec{\ell})} = \frac{P(T = t \mid X = x, \vec{L} = \vec{\ell}) P(\vec{L} = \vec{\ell} \mid X = x)}{P(T = t, \mid X = x, \vec{L} = \vec{\ell})} \quad (53)$$

$$= P(\vec{L} = \vec{\ell} \mid X = x) \quad (54)$$

B.4 Doubly robust with a consistent prediction function for outcomes

Suppose we have consistent estimators $\hat{\pi}(t, \vec{\ell}) \xrightarrow{P} \pi(t, \vec{\ell})$ and $\hat{g}(t, x, \vec{\ell}) \xrightarrow{P} g(t, x, \vec{\ell})$. Suppose the function to predict treatment probabilities converges to the wrong value: $\hat{m}(t, x, \vec{\ell}) \rightarrow \tilde{m}(t, x, \vec{\ell}) \neq m(t, x, \vec{\ell})$. The robust estimator is equivalent to the estimator by predicted outcomes with a correction term,

$$\hat{\theta}_x^{\text{Robust}}(\pi) = \hat{\theta}_x^{\text{Outcome}}(\pi) - \underbrace{\frac{1}{\sum_{i: X_i=x} \hat{w}_i} \sum_{i: X_i=x} \hat{w}_i \left(\hat{g}(T_i, X_i, \vec{L}_i) - Y_i \right)}_{\text{Correction } \hat{\delta}_x^{\text{Outcome}}(\pi)} \quad (55)$$

where $\hat{w}_i = \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}$. By the assumed consistency of $\hat{g}()$, the proof in B.2 implies $\hat{\theta}_x^{\text{Outcome}}(\pi) \rightarrow \theta_x(\pi)$. To show that $\hat{\theta}_x^{\text{Robust}}(\pi) \rightarrow \theta_x(\pi)$, it suffices to show that $\hat{\delta}_x^{\text{Outcome}}(\pi) \rightarrow 0$.

$$\hat{\delta}_x^{\text{Outcome}}(\pi) = \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)} \left(\hat{g}(T_i, X_i, \vec{L}_i) - Y_i \right) \quad (56)$$

Make the sum over covariate strata $\vec{\ell}$ and treatments t explicit

$$= \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{\vec{\ell}} \sum_t \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \eta_i \left(\hat{g}(T_i, X_i, \vec{L}_i) - Y_i \right) \quad (57)$$

Pull the constant $\frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})}$ out of the inner-most sum

$$\frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{\vec{\ell}} \sum_t \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \eta_i \left(\hat{g}(T_i, X_i, \vec{L}_i) - Y_i \right) \quad (58)$$

Multiply by $\frac{\eta_{x\vec{\ell}t}}{\eta_{x\vec{\ell}t}}$

$$= \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{\vec{\ell}} \sum_t \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \eta_{x\vec{\ell}t} \underbrace{\frac{1}{\eta_{x\vec{\ell}t}} \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \eta_i \left(\hat{g}(T_i, X_i, \vec{L}_i) - Y_i \right)}_{\xrightarrow{0} \text{Weighted mean of outcome prediction errors within sample stratum } x, \vec{\ell}, t} \quad (59)$$

$$\rightarrow 0 \quad (60)$$

Therefore $\hat{\theta}_x^{\text{Robust}}(\pi) \rightarrow \theta_x(\pi)$ when $\hat{g}()$ is consistent even if $\hat{m}()$ is misspecified, because the correction term involving the misspecified $\hat{g}()$ converges to zero.

B.5 Doubly robust with a consistent prediction function for treatment probabilities

Suppose we have consistent estimators $\hat{\pi}(t, \vec{\ell}) \xrightarrow{P} \pi(t, \vec{\ell})$ and $\hat{m}(t, x, \vec{\ell}) \xrightarrow{P} m(t, x, \vec{\ell})$. Suppose the prediction function for outcomes converges to the wrong value: $\hat{g}(t, x, \vec{\ell}) \rightarrow \tilde{g}(t, x, \vec{\ell}) \neq g(t, x, \vec{\ell})$. The robust estimator is equivalent to an estimator by predicted treatment probabilities with a correction term,

$$\hat{\theta}_x^{\text{Robust}}(\pi) = \hat{\theta}_x^{\text{Outcome}}(\pi) - \frac{1}{\sum_{i: X_i=x} \hat{w}_i} \sum_{i: X_i=x} \hat{w}_i \left(\hat{g}(T_i, X_i, \vec{L}_i) - Y_i \right) \quad (61)$$

$$= \underbrace{\frac{1}{\sum_{i: X_i=x} \hat{w}_i} \sum_{i: X_i=x} \hat{w}_i Y_i}_{=\hat{\theta}_x^{\text{Treatment}}(\pi)} + \underbrace{\hat{\theta}_x^{\text{Outcome}}(\pi) - \frac{1}{\sum_{i: X_i=x} \hat{w}_i} \sum_{i: X_i=x} \hat{w}_i \hat{g}(T_i, X_i, \vec{L}_i)}_{\text{Correction } \hat{\delta}_x^{\text{Treatment}}(\pi)} \quad (62)$$

where $\hat{w}_i = \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}$. By the assumed consistency of $\hat{m}()$, the proof in B.3 implies $\hat{\theta}_x^{\text{Treatment}}(\pi) \rightarrow \theta_x(\pi)$. To show that $\hat{\theta}_x^{\text{Robust}}(\pi) \rightarrow \theta_x(\pi)$, it suffices to show that $\hat{\delta}_x^{\text{Treatment}}(\pi) \rightarrow 0$. This is proven on the next page.

$$\hat{\delta}_x^{\text{Treatment}}(\pi) = \hat{\theta}_x^{\text{Outcome}}(\pi) - \frac{1}{\sum_{i: X_i=x} \hat{w}_i} \sum_{i: X_i=x} \hat{w}_i \hat{g}(T_i, X_i, \vec{L}_i) \quad (63)$$

Replace $\hat{\theta}_x^{\text{Outcome}}(\pi)$ and $\hat{w}_i$ with their definitions

$$\begin{aligned} &= \frac{1}{\eta_x} \sum_{i: X_i=x} \eta_i \sum_t \hat{\pi}(t, \vec{L}_i) \hat{g}(t, X_i, \vec{L}_i) \\ &\quad - \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, X_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, X_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)} \hat{g}(T_i, X_i, \vec{L}_i) \end{aligned} \quad (64)$$

Make sums over $\vec{\ell}$ and t explicit. Note that sums over i simplify.

$$\begin{aligned} &= \sum_{\vec{\ell}} \left(\frac{1}{\eta_x} \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}} \eta_i \overbrace{\sum_t \hat{\pi}(t, \vec{\ell}) \hat{g}(t, x, \vec{\ell})}^{\text{Constant with respect to } i} \right. \\ &\quad \left. - \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, X_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_t \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \eta_i \overbrace{\frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \hat{g}(t, x, \vec{\ell})}^{\text{Constant with respect to } i} \right) \quad (65) \\ &\quad \quad \quad = \eta_x \vec{\ell} t \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \hat{g}(t, x, \vec{\ell}) \end{aligned}$$

Simplify and note limits of each component, with $\tilde{g}() \neq g()$ denoting the incorrect limit of $\hat{g}()$

$$\begin{aligned} &= \sum_{\vec{\ell}} \left(\underbrace{\frac{\eta_x \vec{\ell}}{\eta_x}}_{\rightarrow P(\vec{L}=\vec{\ell} | X=x)} \sum_t \underbrace{\hat{\pi}(t, \vec{\ell})}_{\rightarrow \pi(t, \vec{\ell})} \underbrace{\hat{g}(t, x, \vec{\ell})}_{\rightarrow \tilde{g}(t, x, \vec{\ell})} - \sum_t \underbrace{\frac{\eta_x \vec{\ell} t}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, X_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}}_{\substack{\rightarrow P(\vec{L}=\vec{\ell}, T=t | X=x) \\ \text{by Eq. 46-52}}} \underbrace{\frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})}}_{\rightarrow \frac{\pi(t, \vec{\ell})}{m(t, x, \vec{\ell})}} \underbrace{\hat{g}(t, x, \vec{\ell})}_{\rightarrow \tilde{g}(t, x, \vec{\ell})} \right) \quad (66) \\ &\rightarrow \sum_{\vec{\ell}} \left(P(\vec{L}=\vec{\ell} | X=x) \sum_t \pi(t, \vec{\ell}) \tilde{g}(t, x, \vec{\ell}) - \sum_t \underbrace{P(\vec{L}=\vec{\ell}, T=t | X=x)}_{=m(t, x, \vec{\ell}) P(\vec{L}=\vec{\ell} | X=x)} \frac{\pi(t, \vec{\ell})}{m(t, x, \vec{\ell})} \tilde{g}(t, x, \vec{\ell}) \right) \quad (67) \end{aligned}$$

Cancel $m()$ at the right and rearrange terms

$$\rightarrow \sum_{\vec{\ell}} P(\vec{L}=\vec{\ell} | X=x) \underbrace{\left(\sum_t \pi(t, \vec{\ell}) \tilde{g}(t, x, \vec{\ell}) - \sum_t \pi(t, \vec{\ell}) \tilde{g}(t, x, \vec{\ell}) \right)}_{=0} \quad (68)$$

$$\rightarrow 0 \quad (69)$$

Therefore $\hat{\theta}_x^{\text{Robust}}(\pi) \rightarrow \theta_x(\pi)$ when $\hat{m}()$ is consistent even if $\hat{g}()$ is misspecified, because the component that involves the misspecified $\hat{g}()$ differences out to zero in the limit.

C Variance estimation with complex survey samples

The GSS survey design samples households within geographic strata, with two Primary Sampling Units (PSUs) selected at random from each stratum. The distribution of the sample across strata is fixed by the survey design and would not change under new draws of the sample. The source of sampling uncertainty is randomness in which PSUs are selected within each stratum. Balanced Repeated Replication (BRR, Krewski and Rao 1981:1013) is a strategy to simulate this uncertainty. BRR creates many simulated replicate samples that include only one of the two PSUs within each stratum. Similar to the bootstrap for simple random samples, in complex surveys BRR results in many estimates $\hat{\tau}^*$ with an empirical variance $\hat{V}(\hat{\tau}^*)$ across the balanced repeated replicates. This empirical variance is an estimate of the variance of the original estimator over repeated samples. I use the **survey** package in R (Lumley, 2019) to generate the balanced repeated replicates and construct 95% confidence interval by a normal approximation centered at $\hat{\tau}$ with the estimated variance $\hat{V}(\hat{\tau}^*)$. A key advantage of this strategy is that it is entirely computational. Within each replicate I repeat the entire data analysis pipeline: sample restrictions, imputation of missing values, estimation of g and m , and aggregation to the target parameter. The variance estimator thus incorporates the consequences of sampling uncertainty throughout the entire process without requiring variance formulas designed specifically for the estimation algorithm applied.

Resampling-based strategies for complex survey samples encompass a broad class of methods for a range of survey designs (Rao and Wu, 1988; Rust and Rao, 1996), many of which are implemented in open-source software (Lumley, 2011, 2019). BRR is one special case which works well for samples that contain two sampled PSUs in each stratum. There is no single method that will work in all settings; researchers should carefully consider the design of their sample when selecting a method for inference.

D Simulations

Three simulations illustrate the properties of the estimators. The first simulation (Fig 4 in the main text) illustrates how the doubly-robust estimator is approximately unbiased if the prediction function for either outcomes or for treatment probabilities is correctly specified. The second repeats the first simulation but in the case of a sample selected from the population with unequal probabilities of selection, so that sampling weights are needed. The third simulation (Fig 5 in the main text) shows that cross fitting improves on double robustness in a setting with random forests used for estimation.

D.1 Simulation 1: Doubly-robust estimates are approximately unbiased when at least one of the two prediction functions is correct

Suppose we have a population following the simple causal structure of Fig 7 in which there is one confounding variable $\vec{L}$. Maintaining a focus on this simple setting for identification is helpful because it directs attention to errors that relate to functional forms, rather than to the selection of covariates. That is the correct focus of attention because the selection of covariates is an identification question best addressed by assumptions encoded in DAGs; the same set of covariates are needed for identification regardless of whether one conducts estimation by predicted outcomes or by predicted probabilities of treatment. Double robustness is instead solely directed to uncertainty about the statistical procedure by which to conduct that adjustment. In this causal structure, conditioning on L is sufficient to identify the post-intervention mean under any treatment value t (e.g. 0 or 1).

Fig. 7. Causal structure for double robustness simulation.

For each simulation $r = 1, \dots, 1000$, I draw a sample of size $n = 500$ from the following data generating process, which is summarized visually in Fig 8. The gap-defining category X is a binary indicator.

$$X \sim \text{Bernoulli}(.5) \tag{70}$$

A continuous confounder L is tightly linked with X . When $X = 0$, L is uniformly distributed on $[-1,0]$. When $X = 1$, L is uniformly distributed on $[0,1]$. Note that this does not create a positivity violation because X is not the treatment.

$$L \sim \begin{cases} \text{Uniform}(-1, 0) & \text{if } X = 0 \\ \text{Uniform}(0, 1) & \text{if } X = 1 \end{cases} \tag{71}$$

The true probability $m(1, x, \vec{\ell})$ of treatment $T = 1$ is quadratic in log odds (Fig 8).

$$m(1, X, L) \equiv \text{logit}^{-1}(-2 + 4L^2) \quad (72)$$

$$T \sim \text{Bernoulli}(m(1, x, \ell)) \quad (73)$$

The outcome conditional mean $g(t, x, \ell)$ involves an interaction between L and T (Fig 8).

$$g(T, X, L) \equiv 4 + 4L + 4T - 2TL \quad (74)$$

$$Y \sim \text{Normal}(\text{Mean} = g(T, X, L), \text{SD} = 1) \quad (75)$$

Given this data generating process, I then fit prediction functions for which the functional form is either correctly or incorrectly specified. For the outcome, these specifications are OLS regressions that either do or do not include the interaction between T and L .

$$g(t, x, \ell) = \begin{cases} \beta_0 + \beta_1 t + \beta_2 \ell + \beta_3 t\ell & \text{correct specification} \\ \eta_0 + \eta_1 t + \eta_2 \ell & \text{incorrect specification} \end{cases} \quad (76)$$

For the treatment probabilities, the specifications are logistic regressions that either do or do not include a quadratic term for L .

$$m(t, x, \ell) = \begin{cases} \text{logit}^{-1}(\lambda_0 + \lambda_1 \ell + \lambda_2 \ell^2) & \text{correct specification} \\ \text{logit}^{-1}(\kappa_0 + \kappa_1 \ell) & \text{incorrect specification} \end{cases} \quad (77)$$

Fig 8 illustrates the true data generating process and the approximations to which the incorrect specifications converge at large sample sizes. Fig 4 (main text) shows simulation results: the doubly-robust estimator for the gap closing estimand over categories $X = \{1, 0\}$ under treatment $T = 1$ produces estimates $\hat{\tau}_{1,0}(1)$ distributed around the truth $\tau_{1,0}(1)$ even when one of the two prediction functions is misspecified.

Fig. 8. Data generating process and functional form misspecifications to illustrate double robustness. Results of this simulation are shown in Fig 4 in the main text. The two categories X (solid and dashed lines) have disjoint support on the continuous confounding variable L . The log odds of treatment are a quadratic function of the confounder L ; the incorrect specification omits this quadratic and fits with the best linear approximation. The true outcomes are centered on a linear function of L with different slopes in the treatment and control groups. The incorrect specification omits this interaction. Over 1,000 simulations fit to samples of size $n = 500$, the doubly robust estimator produces estimates around the true gap-closing estimand even when one of the two prediction functions is misspecified (main text Fig 4).

D.2 Simulation 2: Doubly robust estimation in weighted samples

This simulation is identical to Simulation 1 except that samples are drawn from the population as a function of the pre-treatment covariate L . Specifically, I generate a population of $N = 100,000$ and repeatedly draw samples of size $n = 500$ where the probability of inclusion in the sample is proportional to the probability of treatment $m()$ from Simulation 1. I then estimate (A) as though this were a simple random sample and (B) with all analyses weighted by the known inverse sampling probability. Fig 9 shows results. Simulated estimates of the gap-closing estimand under assignment to treatment 1 are severely biased when sampling weights are ignored (Panel A), but the analysis with sampling weights as derived in this paper produces estimates that are approximately unbiased (Panel B).

- A) If the sample is selected from the population with unequal but known probabilities, analyzing it as a simple random sample produces bias for all estimators.

Error Distribution Estimates Across Simulations

- B) If the sample is selected from the population with unequal but known probabilities, weighted estimators are approximately unbiased.

Error Distribution Estimates Across Simulations

Fig. 9. Simulation showing estimation with sampling weights. In this simulation, units are sampled from the population with known but unequal probabilities of inclusion. Simulation details are provided in Sec. D.2. Other than the sample selection, the simulation is identical to that reported in Fig 4 in the main text.

D.3 Simulation 3: Cross-fitting can improve convergence when estimation is via random forest

Double robustness is beneficial when one or both prediction functions is correctly specified. In practice, however, we may not know either functional form. This simulation illustrates the benefit of doubly-robust estimation with cross fitting when the prediction functions have unknown functional forms and are estimated by a flexible machine learning tool (random forest).

The causal structure for this simulation is given in Fig 10. There are 10 confounding variable $\vec{L}$ that are uniform on $[-1,1]$, all of which are independent of each other and of the binary gap-defining category X . The confounders $\vec{L}$ cause the binary treatment T and the continuous outcome Y . The gap-defining category X also causes the outcome Y .

Fig. 10. Causal structure for cross fitting simulation.

Mathematical details of the simulation are given in Eq. 78–83. The true treatment assignment is a linear function of $\vec{L}$ in the space of log odds. The true expected outcomes are a linear function of $\vec{L}$ and T , with a treatment effect of $+1$ when $X = 1$ and -1 when $X = 0$.

$$X \sim \text{Bernoulli}(.5) \quad (78)$$

$$L_1, \dots, L_{10} \sim \text{Uniform}(-1, 1) \quad (79)$$

$$m(1, X, \vec{L}) = \begin{cases} \text{logit}^{-1}(.1L_1 + \dots + .1L_{10}) & \text{if } X = 1 \\ \text{logit}^{-1}(.3L_1 + \dots + .3L_{10}) & \text{if } X = 0 \end{cases} \quad (80)$$

$$T \sim \text{Bernoulli}\left(m(1, X, \vec{L})\right) \quad (81)$$

$$g(T, X, \vec{L}) = \begin{cases} L_1 + \dots + L_{10} + T & \text{if } X = 1 \\ L_1 + \dots + L_{10} - T & \text{if } X = 0 \end{cases} \quad (82)$$

$$Y \sim \text{Normal}\left(\text{Mean} = g(T, X, \vec{L}), \text{SD} = 10\right) \quad (83)$$

These functional forms could be estimated by OLS and logistic regression, but in practice the researcher would not know the correct functional forms. These simulations therefore use a random forest instead, fit with the **ranger** package with default options (Wright and Ziegler, 2017). I conduct 1,000 simulations at each sample size $n = \{100, 250, 500, 750, 1000, 1250, 1500, 1750, 2000\}$, including the category X and the confounders $\vec{L}$ as predictors in both cases. The outcome prediction function is fit on the subsample with observed treatment $T = 1$, thereby enforcing the possibility of interactions with the treatment.

In the data generating process if Eq. 78–83, the true gap-closing estimand under assignment to $T = 1$ is $\tau = 2$. Fig 5 in the main text reports root mean squared error (RMSE) of each estimator at each sample size.

$$\text{RMSE} = \sqrt{\frac{1}{1000} \sum_{r=1}^{1000} (\hat{\tau}_{1,0}^r(1) - \tau_{1,0}(1))^2} \quad (84)$$

As is visible in Fig 5, RMSE converges to zero more quickly for doubly-robust estimation than for an approach with only predicted outcomes, and more quickly for doubly-robust

estimation with cross fitting than for doubly robust estimation in a single sample.

E Details of the empirical example

E.1 Data

I analyze data from the pooled 1975–2018 General Social Survey (GSS, Smith et al. 2018). This excludes the 1972 and 1973 waves, in which respondent incomes were not recorded, and the 1974 wave, in which the sample design variables used to calculate standard errors are unavailable. The advantage of the GSS is that it has consistently asked questions about the respondent’s occupation and the occupation of the father figure with whom the respondent lived at age 16, making it a standard dataset in the study of social mobility (Jonsson et al., 2009; Beller, 2009; Pfeffer and Hertel, 2015). Because the analyses focus on respondent-level variables and only one respondent per household was interviewed, those residing in larger households had lower probabilities of selection. All regressions are therefore weighted. Because I pool across years, I adjust the weight variable to give equal total weight to each survey year in the analytic sample. Pooling across years yields a sufficiently large sample to illustrate the gap-closing estimand; future research might explore how the gap-closing estimand has changed over time. I restrict the sample to those in the prime working ages of 30–45 ($N = 19,617$) for whom the occupation of the father figure is not a valid skip (i.e. a father figure exists, $N = 15,929$), own occupation is not a valid skip (i.e. respondent has been employed, $N = 15,502$), and income in the past year is non-zero ($N = 12,328$). Missing values on other variables (less than 3% per variable, weighted) are singly imputed for constructing each point estimate, with imputation conducted separately on each balanced repeated replicate to capture uncertainty (Appendix C).

As discussed in the main text, class is operationalized with “professional” defined as any occupation in Class I of the Erikson-Goldthorpe-Portocarero (EGP) schema (Erikson et al., 1979). Class I occupations are a broad set of higher managerial and professional jobs including managers, engineers, scientists, and lawyers. The GSS harmonizes occupations according to 2010 Census codes, and I categorize these into EGP classes according to a

crosswalk provided in a GSS methodological memo (Morgan, 2017). For simplicity of discussion, I use the term *professional class* in reference to all EGP Class I occupations and *working class* in reference to all other occupations.

E.2 Example: Specification of $g()$ and $m()$

In the main text, I estimate the outcome function $g()$ and the treatment function $m()$ by linear and logistic regression, respectively.

$$\text{logit}\left(m\left(t, x, \vec{\ell}\right)\right) = \text{logit}\left(\text{P}\left(T = t \mid X = x, \vec{L} = \vec{\ell}\right)\right) \quad (85)$$

$$\underbrace{=}_{\text{By parametric approximation}} \alpha_m + x\beta_m + \vec{\ell}^T \vec{\gamma}_m \quad (86)$$

$$g\left(t, x, \vec{\ell}\right) = \text{E}\left(Y \mid T = t, X = x, \vec{L} = \vec{\ell}\right) \quad (87)$$

$$\underbrace{=}_{\text{By parametric approximation}} \alpha_g + \underbrace{x\beta_g + t\lambda_g + (xt)\nu_g}_{\text{Category, treatment, and interaction}} + \underbrace{\vec{\ell}^T \vec{\gamma}_g}_{\text{Pre-treatment covariates}} \quad (88)$$

I estimate these regressions with survey weights in order to arrive at a parametric approximation that prioritizes closeness to the true function in the regions of the covariate space which are more heavily weighted. If the parametric specification of $g()$ and $m()$ was known to be correct and the survey weights were a function of the predictors only, then maximal efficiency for coefficient estimates would be realized by estimating $\hat{g}()$ and $\hat{m}()$ without sampling weights (Winship and Radbill, 1994). But if the survey weights are a function of Y conditional on the predictors, then weighting is necessary because even a correctly specified functional form could yield the wrong predicted values at the population level.

E.3 Example: Specification of $\pi()$

Fig 6 reports estimates for two fixed and two stochastic assignment rules.

$$\text{Assign a professional occupation: } \pi(\text{Professional}, \vec{\ell}) = 1 \quad (89)$$

$$\text{Assign a working-class occupation: } \pi(\text{Professional}, \vec{\ell}) = 0 \quad (90)$$

$$\text{Assign a random occupational class: } \pi(\text{Professional}, \vec{\ell}) = P(T = \text{Professional}) \quad (91)$$

$$\text{Assign randomly within covariates: } \pi(\text{Professional}, \vec{\ell}) = P(T = \text{Professional} \mid \vec{L} = \vec{\ell}) \quad (92)$$

For the stochastic assignment rule (Eq. 91 and 92), $\pi()$ must be learned from the data. Because there is no double robustness to misspecifying this function, I estimate nonparametrically. For the rule in Eq. 91, I estimate by the weighted sample mean in a professional occupation. For the rule in Eq. 92, I estimate by the weighted sample mean within each stratum of the covariates.

E.4 Example: Nonparametric estimation of $g()$ and $m()$

For nonparametric estimation of $g()$ and $m()$, I apply a random forest as implemented in the **ranger** package in R with the default options (Wright and Ziegler, 2017). Because of the complex survey design, I incorporate weights with the **case.weights** option, which modifies the algorithm to select a weighted bootstrap sample rather than a simple random sample for learning each tree. Because $g()$ and $m()$ are both conditional mean functions for which the true function would minimize squared prediction error, I estimate all forests with the outcome coded numerically so that the forest seeks to minimize squared prediction error.

To conduct cross-fitting, I first block the sample by the GSS sampling strata (**VSTRAT**) and randomly assign half of all observations to fold 1 within each stratum. I learn $\hat{g}()$ and $\hat{m}()$ on the first fold and convert them to an estimate using data from the second fold. Then, I learn these same functions on the second fold and convert them to an estimate using

data from the first fold. I average the two resulting estimates to arrive at the cross-fitting estimate. For variance estimation, I carry out the entire procedure within each balanced repeated replicate and calculate the variance across the replicates as with the parametric approaches.

E.5 Example: Reasoning about unobserved confounding

Any causal claim should be read cautiously in light of the possibility of unobserved confounding. For gap-closing estimands, selection bias arising from unobserved confounding depends on the differential bias of the post-intervention means for the two categories of interest. For example, consider the bias of an estimator for the post-intervention gap between those from social categories denoted 1 and 0 (subscripts) under intervention to assign treatment value 1 (parentheses).

$$\text{Bias}\left(\hat{\tau}_{1,0}(1)\right) = \overbrace{\mathbf{E}\left(\hat{\tau}_{1,0}(1) - \tau_{1,0}(1)\right)}^{\text{Bias of gap-closing estimator under treatment value 1}} \quad (93)$$

$$= \mathbf{E}\left(\overbrace{\left(\hat{\theta}_1(1) - \hat{\theta}_0(1)\right)}^{\text{Difference in post-intervention mean estimates}} - \overbrace{\left(\theta_1(1) - \theta_0(1)\right)}^{\text{Difference in post-intervention mean true values}} \right) \quad (94)$$

$$= \underbrace{\mathbf{E}\left(\hat{\theta}_1(1) - \theta_1(1)\right)}_{\substack{\text{Bias}(\hat{\theta}_1(1)) \\ \text{Bias of post-intervention mean} \\ \text{estimator for units in category 1}}} - \underbrace{\mathbf{E}\left(\hat{\theta}_0(1) - \theta_0(1)\right)}_{\substack{\text{Bias}(\hat{\theta}_0(1)) \\ \text{Bias of post-intervention mean} \\ \text{estimator for units in category 0}}} \quad (95)$$

Using the concrete example of the paper, the categories (subscripts) denote those from professional origins (1) and working-class origins (0). For both categories, we imagine assignment to a professional destination (parenthetical 1 notation). Attainment of a professional occupation is difficult, so among both categories those who actually attain professional occupations are likely to be positively selected with respect to potential pay in these occupations, even net of covariates: $\text{Bias}(\hat{\theta}_1(1)) > 0$ and $\text{Bias}(\hat{\theta}_0(1)) > 0$. However,

professional class attainment is especially difficult for those of working-class origins, who face many barriers. Those from this category who attain professional occupations may be especially positively selected along potential earnings: $\text{Bias}(\hat{\theta}_0(1)) > \text{Bias}(\hat{\theta}_1(1))$. If so, then the bias of the gap-closing estimate would be negative: $\text{Bias}(\hat{\tau}_{1,0}(1)) < 0$. Substantively, the disparity that would persist by class origins if we intervened to assign professional occupational destinations is likely to be larger than what this paper estimates—if anything, I have understated the gap and overstated the ability of professional occupational attainment to close the gap. This possible selection bias only reinforces the main takeaway from the analysis: professional occupational attainment does very little to close the pay disparity by class origin. This example illustrates how one must reason carefully to understand selection bias for gap-closing estimands—the direction of the overall bias depends on the relative selection bias in the two categories.

F Supplemental tables and figures

		Professional origin $X = 1$	Working-class origin $X = 0$
Y	Log income (1986 dollars)	9.97	9.65
T	Professional class destination (own occupation)	0.24	0.08
L_1	Race: White	0.87	0.83
L_2	Race: Black	0.04	0.11
L_3	Race: Other	0.08	0.06
L_4	Degree: Less than high school	0.02	0.12
L_5	Degree: High school	0.27	0.54
L_6	Degree: Junior college	0.07	0.08
L_7	Degree: Bachelor's	0.38	0.18
L_8	Degree: Graduate	0.26	0.09
L_9	Female	0.45	0.48
L_{10}	Age	37.22	37.31
N		1,374	10,954

Table 1. Weighted means of variables in the analytic GSS sample. Columns are subsamples defined by the respondent's father's occupational class (the gap-defining category).

	Outcome prediction function $\hat{g}$ (OLS predicting log income)	Treatment prediction function $\hat{m}$ (logit predicting professional class)
Professional class origin (father figure held a professional occupation)	-0.01 (0.04)	0.50 (0.09)
Treated (respondent holds a professional occupation)	0.38 (0.03)	
Treated $\times$ professional origin	0.09 (0.07)	
Age	0.02 (0.00)	0.02 (0.01)
Highest degree (less than high school omitted)		
High school	0.46 (0.03)	2.22 (1.08)
Junior college	0.62 (0.04)	3.05 (1.08)
Bachelor's	0.87 (0.04)	4.04 (1.07)
Graduate	1.06 (0.05)	5.19 (1.07)
Race (white omitted)		
Black	-0.02 (0.03)	-0.44 (0.16)
Other	-0.15 (0.04)	0.40 (0.13)
Female	-0.71 (0.02)	-0.80 (0.08)
Intercept	9.40 (0.08)	-5.36 (1.12)

Table 2. Coefficients of parametric outcome and treatment prediction functions. All regressions are estimated with survey weights. These functions are used to generate the doubly robust estimates reported in Fig 6.

	Not conditional on treatment	Conditional on treatment
Professional class origin (father figure held a professional occupation)	0.04 (0.03)	0.01 (0.03)
Treated (respondent holds a professional occupation)		0.40 (0.03)
Age	0.02 (0.00)	0.02 (0.00)
Highest degree (less than high school omitted)		
High school	0.47 (0.03)	0.46 (0.03)
Junior college	0.65 (0.04)	0.62 (0.04)
Bachelor's	0.93 (0.04)	0.86 (0.04)
Graduate	1.21 (0.04)	1.06 (0.05)
Race (white omitted)		
Black	-0.03 (0.03)	-0.02 (0.03)
Other	-0.13 (0.05)	-0.15 (0.04)
Female	-0.74 (0.02)	-0.71 (0.02)
Intercept	9.41 (0.08)	9.40 (0.08)

Table 3. Coefficients of comparison additive linear regressions. All regressions are estimated with survey weights. These regressions are only included to illustrate a common research practice: fit regressions with and without the treatment and examine the change in the coefficient on the gap-defining category ($0.04 - 0.01 = 0.03$). This procedure does not estimate a gap-closing estimand (see Sec. 7.1).

Fig. 11. Gap-closing estimands by alternative estimators. Estimation by predicted outcomes follows Eq. 5. Estimation by predicted treatment probabilities reweights observed outcomes for treated units to estimate the population average potential outcome as in Eq. 7. Doubly robust estimation combines the two and is consistent if either is consistent (Eq. 8). Cross fitting is doubly robust estimation conducted within a cross-fitting procedure with one subsample used to learn $\hat{g}()$ and $\hat{m}()$ and another subsample used to convert them into gap-closing estimates (see Sec. 5.4 and Chernozhukov et al. 2018). For each approach, estimation with generalized linear models refers to OLS for the outcome and logit for the treatment (see Table 2). Estimation with Generalized Additive Models is analogous except that the conditional association with age is estimated by a thin-plate spline rather than a linear term (Wood, 2017). Estimation with random forests is conducted with the **ranger** package in R (Wright and Ziegler, 2017), with the outcome forest fit separately on treated and untreated units to enforce interactions with the treatment.

Appendix

References

- Beller, E. 2009. Bringing intergenerational social mobility research into the twenty-first century: Why mothers matter. *American Sociological Review*, 74(4):507–528.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.
- Erikson, R., J. H. Goldthorpe, and L. Portocarero 1979. Intergenerational class mobility in three Western European societies: England, France and Sweden. *The British Journal of Sociology*, 30(4):415–441.
- Jonsson, J. O., D. B. Grusky, M. Di Carlo, R. Pollak, and M. C. Brinton 2009. Micro-class mobility: Social reproduction in four countries. *American Journal of Sociology*, 114(4):977–1036.
- Krewski, D. and J. Rao 1981. Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, Pp. 1010–1019.
- Lumley, T. 2011. *Complex Surveys: A Guide to Analysis Using R*, volume 565. John Wiley & Sons.
- Lumley, T. 2019. *survey: Analysis of Complex Survey Samples*. R package version 3.36.
- Morgan, S. L. 2017. A coding of social class for the general social survey. *Chicago, Illinois: GSS Methodological Report No. 125*.
- Pfeffer, F. T. and F. R. Hertel 2015. How has educational expansion shaped social mobility trends in the United States? *Social Forces*, 94(1):143–180.
- Rao, J. N. and C. Wu 1988. Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401):231–241.
- Rust, K. F. and J. Rao 1996. Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5(3):283–310.
- Smith, T. W., M. Davern, J. Freese, and S. Morgan 2018. *General Social Surveys, 1972-2018 [machine-readable data file]*. Sponsored by National Science Foundation. –NORC ed.– Chicago: NORC, 2018: NORC at the University of Chicago [producer and distributor]. Data accessed from the GSS Data Explorer website at gssdataexplorer.norc.org.
- Winship, C. and L. Radbill 1994. Sampling weights and regression analysis. *Sociological Methods & Research*, 23(2):230–257.
- Wood, S. N. 2017. *Generalized Additive Models: An Introduction with R*. Chapman and Hall / CRC.
- Wright, M. N. and A. Ziegler 2017. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1).