

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Sociology

2021

Research Begins with a Goal: A Framework Illustrated by Gap-Closing Estimands

Lundberg, Ian Daniel

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp011j92gb558>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

RESEARCH BEGINS WITH A GOAL:
A FRAMEWORK ILLUSTRATED BY
GAP-CLOSING ESTIMANDS

IAN DANIEL LUNDBERG

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
SOCIOLOGY
ADVISER: BRANDON M. STEWART

MAY 2021



This work is licensed under a Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International License.

To view a copy of this license, visit
<http://creativecommons.org/licenses/by-nc-nd/4.0/>

Abstract

Quantitative social science is at a moment of tremendous opportunity. Computational power has been multiplying for decades, accelerating the development of new statistical and machine learning tools. But tools are only as good as the questions they are used to answer: research always begins with a goal. This dissertation argues that renewed attention to social science goals can point the way toward transparent assumptions and the use of predictive algorithms for estimation. Chapter 2 (with Rebecca Johnson and Brandon Stewart) presents a framework for quantitative research in which a precise goal—the theoretical estimand—guides all methodological choices. Our framework creates opportunities for advances that are both methodological and substantive. Chapter 3 presents a methodological advance—the gap-closing estimand—which quantifies how an intervention to a treatment (e.g. occupational class attainment) would close a gap (e.g. income by class origin). Chapter 4 applies that methodological advance to quantify the contribution of occupational segregation to a racial disparity in health. Chapter 5 concludes. Together, these chapters illustrate the fruitfulness of a framework for quantitative social science that unlocks computational tools to answer research questions that are conceptually precise. The key to this framework is clear reasoning about the research goal.

Acknowledgements

This dissertation owes a debt of gratitude to the community in which it was produced. As my primary adviser, Brandon Stewart walked this journey alongside me. Brandon welcomed my sometimes contrarian perspectives on methodology, always open to the possibility that one or both of us might be wrong about something we believed. We challenged each other, pushing our perspectives to their limits to learn from the settings in which we could break them. My approach to quantitative research has evolved dramatically over the past six years, due in no small part to my conversations with and extensive feedback from Brandon.

Matthew Salganik has also enormously influenced my own perspective for the better. From Matt, I have learned the value of reaching across fields to learn from others, especially when I have a problem that someone else has thought about before. Matt always challenges me to ask important questions and to welcome ideas from others. In each of these ways, I aspire to be like Matt.

Sara McLanahan and Dalton Conley have also taught me so much. As I engage with new and complex methods, Sara has often helped me to translate methods to interpretations that speak to a broad audience. I have also learned so much about research through our collaborative projects using the Fragile Families and Child Well-being Study, which do not appear in the dissertation. Dalton has always helped me search for the social science puzzle that lies behind the methods I encounter and develop. I am grateful to Filiz Garip and Yu Xie for their guidance at other moments of my academic journey, and for serving as outside readers on this dissertation.

Many others contributed to this work directly or indirectly. Chapter 2 is equally co-authored with Rebecca Johnson and Brandon Stewart, and would not have been the same without our conversations with Simone Zhang. Many ideas in this dissertation are influenced by conversations with the students in the Ph.D. statistics sequence for sociologists at Princeton, whose questions in section and office hours taught me

as much as I taught them. The Stewart Lab provided extensive feedback on all aspects of this dissertation. The work could not have been carried out apart from the supportive community of students, faculty, and staff in the Department of Sociology and the Office of Population Research.

Academic achievements require non-academic support, for which I am thankful to my community in the Episcopal Church at Princeton. This work also owes gratitude to my parents and my sister, whose love and support has been strong throughout.

Research reported in this publication was supported by The Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD047879. Replication code is available on Dataverse for Chapter 2 (doi.org/10.7910/DVN/ASGOVU) and on GitHub for Chapters 3 and 4 (github.com/ilundberg/replication).

Contents

1	Introduction	1
2	What is Your Estimand?	
	Defining the Target Quantity Connects Statistical Evidence to Theory	3
3	The gap-closing estimand: A causal approach to study interventions that close disparities across social categories	60
4	Quantifying the contribution of occupational segregation to racial disparities in health: A gap-closing perspective	100
5	Conclusion	130
A	Supplementary materials for Chapter 2	131
B	Supplementary materials for Chapter 3	159
C	Supplementary materials for Chapter 4	189
	Bibliography	208

Chapter 1

Introduction

The set of statistical and machine learning tools now available to social scientists would have been unthinkable decades ago. We have opportunities that did not exist before, and those opportunities call for a revolution in social science methodology. Much has been learned from the mode of quantitative social science that dominated the latter half of the 20th century: research questions equated with coefficients in multivariate regression models. When data or computational power are scarce, there may still be good reason to direct attention to hypotheses about coefficients. But often, data and computational power are not scarce. A move beyond research questions about $\hat{\beta}$ opens new doors.

Computational social scientists have led the charge away from $\hat{\beta}$ questions and toward $\hat{y}$ questions that use statistical and machine learning tools for the task they are most directly designed to do: predict outcomes (Hofman et al., 2017; Salganik et al., 2020). But many social science questions are not about prediction. Often we care about describing population means, estimating average causal effects, or estimating average outcomes under some counterfactual condition. None of those questions is inherently a $\hat{y}$ question about prediction, nor is it inherently a $\hat{\beta}$ question about interpreting a particular model. They do not fit neatly into these two cultures of modeling (Breiman, 2001; Shmueli, 2010). Instead, those questions are about population quantities defined outside of any statistical model. They are perhaps best answered by a combined approach that uses a statistical or machine learning method ($\hat{\beta}$) to predict individual outcomes ($\hat{y}$) as a stepping stone toward the real goal: summarizing something factual or counterfactual about a population.

The chapters of this dissertation collectively present a framework for quantitative social science and illustrate its implications for methodological and substantive advances. Chapter 2 (co-authored with Rebecca Johnson and Brandon Stewart) introduces a framework in which a precise statement of the research goal—the theoretical estimand—motivates all methodological choices. We argue that many social science questions can be usefully defined in terms of a unit-specific quantity aggregated over a target population. In this framework, the entire purpose of a statistical or machine learning model is to predict the unobserved unit-specific quantities that we need in order to answer the research question. Doing so facilitates transparency about identification assumptions and unlocks the possibility of estimation by new machine learning tools. Chapter 3 illustrates a methodological advance within that framework. Disparities in socioeconomic outcomes across categories of race, class, and gender are central in the study of inequality. Yet we could also ask questions about disparities in the counterfactual outcomes that would be realized if a treatment variable were reassigned. I introduce the term gap-closing estimands for those types of questions about counterfactual disparities. Gap-closing estimands become straightforward in a research framework centered on unit-specific quantities aggregated over target populations, and framing the problem in this way allows me to develop new doubly-robust estimators that open the door to estimation by machine learning. Chapter 4 applies gap-closing estimands to a health disparity to yield a new substantive result: if we counterfactually eliminated occupational segregation for a sample from the U.S. adult population, we could reduce the Black-white disparity in a health condition related to work by approximately one-third for that sample. Occupational segregation is an important cause of that disparity. Together, these chapters show how new methodological advances and substantive insights become possible if we begin the research process in a particular way: with a well-stated goal.

Chapter 2

What is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory

Equally co-authored with Rebecca Johnson and Brandon Stewart¹

Chapter abstract. We make only one point in this article. Every quantitative study must be able to answer the question: what is your estimand? The estimand is the target quantity—the purpose of the statistical analysis. Much attention is already placed on how to do estimation; a similar degree of care should be given to defining the thing we are estimating. We advocate that authors state the central quantity of each analysis—the theoretical estimand—in precise terms that exist outside of any statistical model. In our framework, researchers do three things: (1) set a theoretical estimand, clearly connecting this quantity to theory, (2) link to an empirical estimand, which is informative about the theoretical estimand under some identification assumptions, and (3) learn from data. Adding precise estimands to research practice expands the space of theoretical questions, clarifies how evidence can speak to those questions, and unlocks new tools for estimation. By grounding all three steps in a precise statement of the target quantity, our framework connects statistical evidence to theory.

¹Forthcoming in *American Sociological Review*. Previously titled “Setting the Target: Precise Estimands and the Gap Between Theory and Empirics.” A full draft including appendices is available on SocArxiv: <https://doi.org/10.31235/osf.io/ba67n>. Replication code is available on Dataverse: <https://doi.org/10.7910/DVN/ASGOVU>. This manuscript was presented at the 2020 Annual Meeting of the American Sociological Association. For helpful discussions and feedback relevant to this project, we thank Dalton Conley, Matt Desmond, Felix Elwert, Adam Goldstein, Justin Grimmer, Tod Hamilton, Erin Hartman, Daniel Karell, Gary King, Sarah Mustillo, Matt Salganik, Gillian Slee, Chris Winship, and members of the Stewart Lab. This manuscript was greatly improved by advice from editors Rory McVeigh and Omar Lizardo as well as three anonymous reviewers. Special thanks to Simone Zhang for both many useful comments and excellent collaboration on related projects. Research reported in this publication was supported by The Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD047879.

2.1 Introduction

In every quantitative paper we read, every quantitative talk we attend, and every quantitative article we write, we should all ask one question: what is the estimand? The estimand is the object of inquiry—it is the precise quantity about which we marshal data to draw an inference. Yet, too often social scientists skip the step of defining the estimand. Instead, they leap straight to describing the data they analyze and the statistical procedures they apply. Without a statement of the estimand, it becomes impossible for the reader to know whether those procedures were appropriate. The methodological approach becomes tautological: if the thing to be estimated is defined within a statistical model, it cuts off productive consideration of a broader class of models that could accomplish the same goal. Further, a goal defined entirely within a model bears a connection to theory that is questionable at best. This paper presents a methodological framework for quantitative social science in which a precise statement of the research goal motivates all steps of the empirical analysis. The estimand unlocks new research tools and can resolve statistical disputes about methodological choices.

Our framework stands in contrast to the currently dominant mode of quantitative inquiry: hypotheses about regression coefficients. That mode of inquiry defines the research goal *inside* a particular statistical model. If your research goal is a coefficient of a particular model, then you are committed to that model: it becomes impossible to reason about other approaches to achieve the goal. By contrast, we advocate a statement of the goal *outside* the statistical model—like an average causal effect or a population mean—which opens the door to alternative estimation procedures that could answer the research question under more credible assumptions. More importantly, stating the research goal outside the model frees us to ask more interesting theoretical questions; the scope of theory is no longer bound to the space of questions

involving the best linear approximation to the conditional association between two variables with all else held constant (i.e. a regression coefficient).

We introduce a term for the goal stated outside the model—the *theoretical estimand*—which has two components. The first is a unit-specific quantity, which could be a realized outcome (whether person i is employed), a potential outcome (whether person i would be employed if they received job training), or a difference in potential outcomes (the effect of job training on the employment of person i). It could also be a potential outcome that would be realized under intervention to more than one variable (whether person i would be employed if they received job training and child care), thus unlocking numerous new causal questions. The unit-specific quantity clarifies whether the research goal is causal, and if so what counterfactual intervention is being considered. The second component of the theoretical estimand is the target population: over whom do we aggregate that unit-specific quantity? The unit-specific quantity and target population combine to define the theoretical estimand: the thing we would like to know if we had data for the full population in all factual or counterfactual worlds of interest. A paper may have multiple theoretical estimands.

Each theoretical estimand is linked to an *empirical estimand* involving only observable quantities (e.g. a difference in means in a population) by assumptions about the relationship between the data we observe and the data we do not. These identification assumptions can be conveyed in a Directed Acyclic Graph (DAG). Finally, one chooses an *estimation strategy* to learn the empirical estimand (e.g. a regression model). This paper uses the general term “estimands” to refer to both the theoretical and the empirical estimands.

Stating both the theoretical estimand and the empirical estimand separately from the estimation strategy partitions the link between theory and evidence into three steps involving different modes of argument (Fig. 2.1). The distinction between the theoretical and empirical estimands is subtle but important: the former may involve

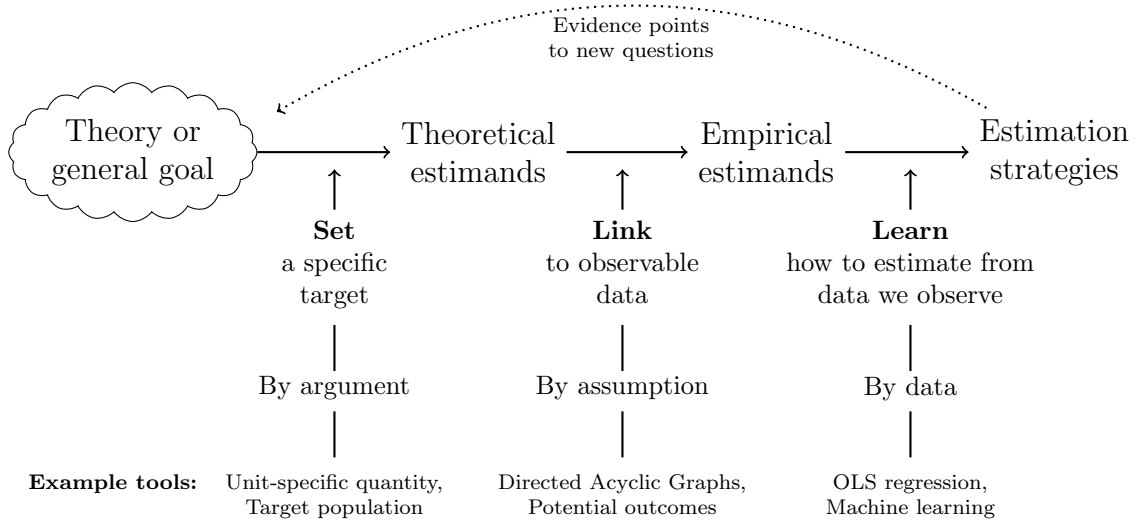


Figure 2.1: Three critical choices in quantitative social science arguments.

The first choice is the theoretical estimands, which set the targets of inference. Argument is required to link the theoretical estimands to the broader theory. The second choice is the empirical estimands, which link the targets to observable data. The connection requires substantive assumptions that can be formalized in Directed Acyclic Graphs. The third choice is the estimation strategies, which captures what we will actually do with data. We select estimation strategies based on the data.

unobservable quantities such as counterfactuals while the latter involves only observable data. Our full argument for the separate statement of the theoretical and empirical estimands appears in the section that introduces the empirical estimand. The choice of theoretical estimands requires substantive argument about the theory and goals, while the choice of empirical estimands requires conceptual argument about unobserved data. The choice of estimation strategies is distinct because it can be at least partially data-driven. Separating these steps helps researchers make principled choices, allows readers to evaluate claims, and enables the community to build on research findings. Too often, research papers involve pages of rich theory followed by pages of procedures applied to data, with a vague link between the two. The theoretical and empirical estimand fill the void by making precise both the theoretical quantity we would like to know and the empirical quantity that our procedures are most directly designed to approximate. Our most emphatic argument is that the

field has much to gain from a precise statement of the true target of inquiry even if the assumptions required to estimate it hold only imperfectly and the empirical tools available are limited. Stating the goal allows the reader and the community to engage meaningfully with and build on the work.

The paper proceeds in several sections. We first introduce our framework, highlighting each step of our proposed research process: setting the theoretical estimand (2.2.1), linking to an empirical estimand (2.2.2), and learning an estimate from data (2.2.3). Section 2.3 demonstrates the prevalence of the problems we address through a review of the 2018 volume of *ASR* and illustrates how our framework would transform quantitative research through two in-depth examples. Section 2.4 concludes by describing how estimands clarify methodological issues for analysts, readers, and the broader community.

2.2 Estimands Link Theory and Evidence

This section details each step in our proposed link between theory and evidence. Figure 2.2 presents the examples that we use to introduce this general framework.

2.2.1 The Theoretical Estimand: Set the Target

The most important step in quantitative empirical research is a clear statement of the research question. The clarity of the question is paramount because no single analysis can prove or undermine an entire sociological theory (Lieberson and Horwich, 2008). When estimating causal effects, for instance, one must state the population over which heterogeneous effects are averaged (Brand and Xie, 2010; Xie, 2013). When estimating associations, one must be clear about whether the target of inference is causal (Hernán, 2018), and if so to be clear about the hypothetical intervention at the core of the claim (Greiner and Rubin, 2011; Hernán et al., 2016a; Morgan

Set the target: The theoretical estimand		Link to observables	Learn from data
Unit-specific quantity	Target population of units	Identification	Estimation
Difference in whether application i would be called back if it signaled white with a felony vs. black without	Applications to jobs in Milwaukee	<pre>graph TD; R1[Random] --> AR[Applicant race]; R2[Random] --> SF[Signals felony]; AR --> CBI[Called back for interview]; SF --> CBI; CBI --> Y[Outcome]</pre>	Logistic regression
Difference in whether mother i would be employed if she had three vs. two children	Those who would have a third birth only if first two of the same sex	<pre>graph TD; FTS[First two same sex] --> TB[Third birth]; TB --> E[Employed]; FTS -.-> U E</pre>	Two-stage least squares
Difference in whether person i would be employed if convicted vs. if not	Those who would be convicted only under certain judges	<pre>graph TD; SJ[Strict judge] --> C[Convicted]; C --> E[Employed]; SJ -.-> U E</pre>	Two-stage least squares
Difference in whether person i would be stopped if perceived as black vs. white	Those stopped by police	<pre>graph TD; PR[Perceived race] --> S[Stopped]; S --> SH[Shot]; PR -.-> U SH; S --> SH</pre>	Logistic regression
Difference in whether applicant i would be admitted if perceived as male vs. female	Applicants to Berkeley	<pre>graph TD; S[Sex] --> A[Applied]; A --> Ad[Admitted]; PS[Perceived sex] --> Ad; A -.-> U Ad; Ad --> Y[Outcome]</pre>	Difference in proportions
Adult income that person i would be realized if childhood income took a particular value	U.S. population	<pre>graph TD; R[Race] --> CI[Childhood income]; CI --> AI[Adult income]; CI -.-> U AI; CI --> AI</pre>	OLS
Wage that mother i would realize if she were an employed mother vs. an employed non-mother	U.S. civilian women ages 25–44 in March 2019	<pre>graph TD; M[Motherhood] --> E[Employed]; E --> W[Wage]; M --> W; E -.-> U2 W; M -.-> U1 W; W --> Y[Outcome]</pre>	OLS Parametric g -formula

Figure 2.2: Estimands are relevant to a broad range of social science studies. White boxes on the diagonal are the focus of the main text, but every study implicitly involves all four steps. Some steps (i.e. DAGs for identification, see Section 2.2.2) are simplified to fit in the table. In the identification step, blue edges represent the causal effect at the center of the paper and dashed red edges represent threats to identification.

and Winship, 2015; Sen and Wasow, 2016). A lack of clarity can lead to a table of regression coefficients that are, at best, weakly informative about theory (Keele et al., 2020; Westreich and Greenland, 2013). Before you apply a statistical procedure, you have to define the thing you are trying to estimate or measure (Katz et al., 2020). Without the language to make a more precise statement, researchers find themselves constrained to questions stated in terms of regression coefficients. We join a long line of increasingly urgent calls to think beyond the constraints of regression as it is commonly practiced (Duncan 1984; Lieberman 1987; Abbott 1988; Freedman 1991; Berk 2004).

Our framework provides researchers with the language they need for the thing they already want: a precise statement of the research goal. The first step of quantitative empirical research—whether descriptive, predictive, or causal—is to state a theoretical estimand that exists outside of the statistical model. We propose that the goal is often a quantity involving two components: a *unit-specific quantity* subscripted by i aggregated over a *target population* of units. For instance, we might study the employment rate among U.S. adults.

$$\frac{1}{\text{Size of U.S. adult population}} \sum_{\substack{i \text{ in U.S.} \\ \text{adult population}}} \text{Employed}_i \quad (2.2.1)$$

Causal goals follow similar notation. How would the probability of employment differ if we enrolled a randomly chosen individual in job training or not? We can define this causal goal using potential outcomes notation (Imbens and Rubin, 2015) as the difference in the potential employment each person would realize if enrolled in job training—denoted $\text{Employed}_i(\text{Job training})$ —versus if they did not—denoted $\text{Employed}_i(\text{No job training})$.

$$\frac{1}{\text{Size of U.S. adult population}} \sum_{i \text{ in U.S. adult population}} \left(\underbrace{\text{Employed}_i(\text{Job training})}_{\text{Employment if received job training}} - \underbrace{\text{Employed}_i(\text{No job training})}_{\text{Employment if did not receive job training}} \right) \quad (2.2.2)$$

Just like a descriptive estimand (the employment rate), the causal estimand (the effect of job training) sums over unit-specific quantities (subscripted by i). Stating the theoretical estimand in this form clarifies two critical components: (1) the unit-specific quantity and (2) the target population over which the unit-specific quantity is aggregated.

2.2.1.1 Specify the Unit-Specific Quantity

The first building block of a theoretical estimand is a quantity defined for each unit in the population. That quantity might be descriptive: the factual value (e.g. Y_i) that some variable actually would take for unit i in the absence of intervention. It might be causal: the value some variable would take if a treatment variable D was set to a particular value d , producing the potential outcome $Y_i(d)$. It might involve interventions to multiple variables, as in the case of a mediation claim about the outcome $Y_i(d, m)$ that unit i would realize if the treatment were set to $D = d$ and some mediator were set to $M = m$. A unit-specific quantity can be any function of realized variables or potential outcomes particular to unit i . It sits outside of any statistical model and involves a substantive question: what factual or counterfactual thing would we like to know for each unit in the population?

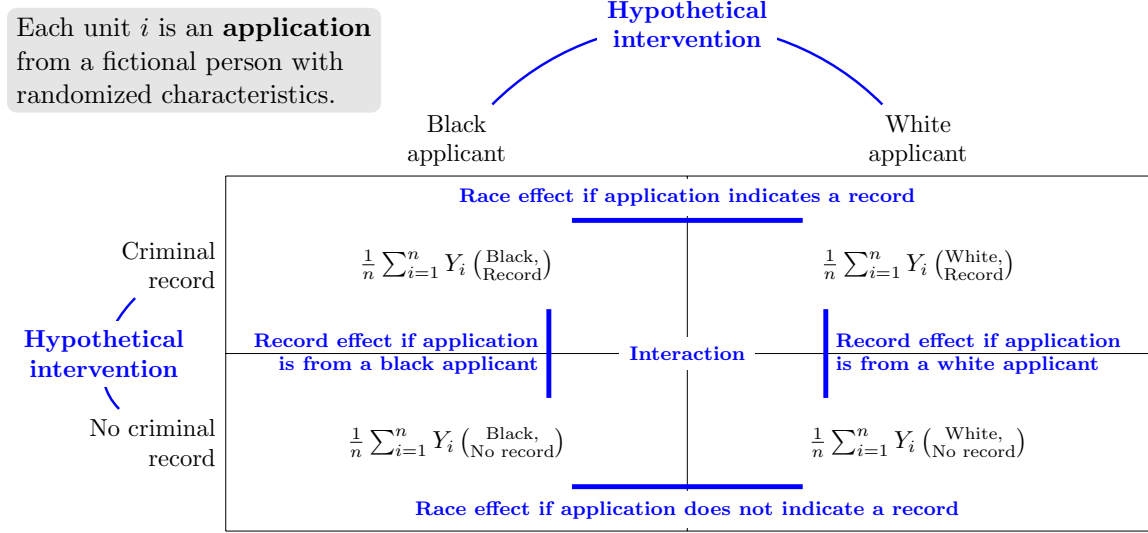
The unit-specific quantities in sociological research can be complex. For instance, Pager (2003) explores “the ways in which the effects of race and criminal record interact to produce new forms of labor market inequalities,” (p. 938). The study navigates this difficult topic through a randomized design. Even before randomization, however, the real novelty of the study is the definition of the unit of analysis as an *application*

rather than as a *person*. In the experiment, job postings were randomly assigned to receive applications from a white or black pair of applicants. Each member of the pair approached the employer at a different time to apply for the job posting, a combination which we call an application. For each posting, one application was randomly assigned to signal a felony conviction for possession of cocaine. For each application i , an outcome Y_i was observed: whether that application received a callback. Each application was thus randomized to one of the four treatment conditions captured by the 2×2 table in Figure 2.3 Panel A, each of which has a potential outcome $Y_i(\text{Treatment Condition})$. Taking the unit of analysis as the application rather than the person sidesteps problems that plague the study of race within a causal framework. It may be very difficult to disentangle race from individual identities in order to consider a counterfactual world in which a person signaled a different racial category (Kohler-Hausmann, 2018). It is reasonably straightforward, however, to imagine an application signaling a different racial category (Sen and Wasow, 2016). Pager (2003) therefore makes progress by studying the application rather than the person as the unit of analysis. Randomization is made possible by the pivot in how the unit of analysis is defined.

With the unit defined as an application, it becomes easier to define the unit-specific quantity: any of four potential outcomes under the two interventions (race and criminal record). The striking result of the study—a lower callback rate for a black applicant without a criminal record than for a white applicant with a criminal record—is meaningful because the unit-specific quantity involves potential outcomes over both of these inputs. The result can only be attributed to the bias of the person evaluating the applications.

The scientific insight of Pager (2003) contrasts with what could be learned in an observational study focused on a different unit-specific quantity (Fig. 2.3 Panel B). Suppose we defined the unit of analysis as a real flesh-and-blood applicant in

A) Causal interaction: Intervention to two variables averaged over *one* population



B) Effect heterogeneity: Intervention to one variable averaged over *two* populations

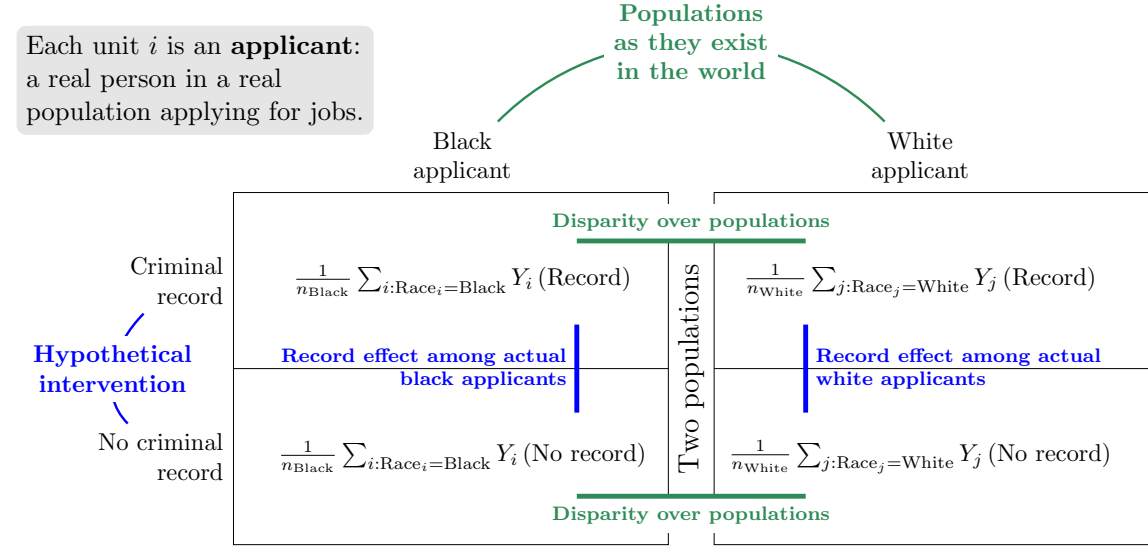


Figure 2.3: Two estimands with different unit-specific quantities and different target populations. Both estimands could be termed the effect of a criminal record on the probability of a callback among black and white applicants, yet the two are quite different. A design targeting causal interaction (Pager, 2003) would randomly assign units (applicant-application pairs) to a cell of the 2×2 table that combines all values of both treatments. A design targeting effect heterogeneity would take applications in the real-world distribution for each subgroup and estimate the outcome they would realize if they signaled or did not signal a criminal record. Both estimands are of substantive interest.

an actual population of those applying for jobs. For real people (as opposed to applications), it is difficult to conceptualize all the things that would have to change in a world where a real person was counterfactually of another racial category—access to schooling, earlier experiences of discrimination, and innumerable opportunities that strengthen a resume. For these reasons, viewing race as a causal treatment may not be straightforward in a study where the unit of analysis is a person. Racial categories could instead denote two populations that differ in myriad ways due to systemic racism. Potential outcomes could be defined as a function of a criminal record *only*. The unit-specific quantity would involve two potential outcomes: the outcome each person would realize if they had a criminal record or if they did not. One could compare the causal effect of a criminal record across subpopulations of black and white applicants.

The colloquial term “moderation” could describe the research goal in both the observational design and the Pager (2003) design, but the meanings of the two estimands are distinct. The policy implications of the former would focus on preventing hiring decision makers from directly considering race and criminal histories when evaluating identical applications. The policy implications of the latter would focus on how equalizing criminal histories (or signals of those histories) could reduce disparities across populations of actual job applicants of different racial categories. Both are worthwhile goals. Distinguishing them requires clarity about whether the unit-specific quantity is a potential outcome as a function of one or two treatments.

The unit-specific quantity provides an opportunity to clarify the causal component (if any) of our research claims. It invites us to be precise about the causes we are studying (e.g. a signal of race) and those that we are not (e.g. how racial disparities in access to education create differences in resumes). It allows us to be precise about the levels of the treatment being contrasted (e.g. a particular resume line about a felony conviction for possession of cocaine). Finally, precision about the unit-specific

quantity facilitates the study of questions involving interventions to multiple variables; the causal interaction between race and a criminal record is only one example of many such questions (Table 2.1). An experimental protocol provides a perhaps unparalleled opportunity for clarity in these regards, but nothing prevents observational studies from aspiring to similar clarity. Even if the assumptions necessary to estimate the goal are doubtful, there is never a reason for the author’s intention for the statistical evidence to be in doubt.

2.2.1.2 Define the Target Population

The second building block of a theoretical estimand is the target population: the set of units over which the unit-specific quantity is aggregated. Statistical evidence often speaks directly to only a limited population, producing a tension: authors must either argue that the population that is empirically tractable is of theoretical interest in itself, or they must argue that the population that is empirically tractable is informative about a broader population. We consider this tension in three contexts: randomized experiments, instrumental variable designs, and strategies that adjust for measured confounding.

The target population is a widely-recognized issue in experiments. For example, Pager (2003:965) explicitly notes that “one key limitation of the audit study design is its concentration on a single metropolitan area.” The true target population may involve a larger set, such as all entry-level job openings in the U.S. If so, the researcher must argue that a particular piece of empirical evidence (an experiment in Milwaukee with entry-level job openings advertised in one newspaper and one website) is informative about that broader target population. Alternatively, one could define the target population more narrowly as entry-level job openings in Milwaukee. Then, the researcher must motivate not what Milwaukee tells us about other places but why we

Estimand name	Mathematical statement	DAG (Sec. 2.2.2)	Reference	Colloquial terms
Average treatment effect	$\frac{1}{n} \sum_i (Y_i(d') - Y_i(d))$	$D \rightarrow Y$	Morgan and Winship (2015)	Effect
Conditional average treatment effect	$\frac{1}{n_x} \sum_{i: X_i=x} (Y_i(d') - Y_i(d))$	$X \rightarrow D \rightarrow Y$	Athey and Imbens (2016)	Effect heterogeneity or moderation
Causal interaction	$\frac{1}{n} \sum_i \left(\left(Y_i(a', d') - Y_i(a', d) \right) - \left(Y_i(a, d') - Y_i(a, d) \right) \right)$	$A \nearrow Y$ $D \nearrow$	VanderWeele (2015)	Joint treatment effect
Controlled direct effect	$\frac{1}{n} \sum_i \left(Y_i(d', m) - Y_i(d, m) \right)$	$M \nearrow$ $D \longrightarrow Y$	Acharya et al. (2016)	Mediation (Sec. 2.3.3)
Natural direct effect	$\frac{1}{n} \sum_i \left(Y_i(d', M_i(d)) - Y_i(d, M_i(d)) \right)$	$M \nearrow$ $D \longrightarrow Y$	Imai et al. (2011)	Mediation (Appendix A.2)
Effect of time-varying treatment	$\frac{1}{n} \sum_i (Y_i(d'_1, d'_2) - Y_i(d_1, d_2))$	$D_1 \rightarrow D_2 \rightarrow Y$	Wodtke et al. (2011)	Cumulative effect

Table 2.1: Unit-specific quantities defined in potential outcomes unlock many causal estimands for inquiry. Social scientists who define the research goal before moving to regression uncover more possible questions than those who confine themselves to regression parameters. This provides a non-exhaustive list of common causal estimands.

should care about Milwaukee specifically. A clear statement of the target population allows a researcher to clarify which approach they are taking.

The target population is also an issue with instrumental variables designs. For example, Angrist and Evans (1998) examine the effect of having three vs. two children on women's employment under an instrumental variables (IV) design: having the first two children of the same sex (the instrument) causes some families to have a third birth (the treatment) without directly affecting employment (the outcome). The IV design offers strong causal identification at a cost: the estimated causal effect is an average not over the full population, but only over the subpopulation of compliers whose treatment status is causally affected by the instrument (Imbens and Angrist, 1994). In this case, the complier population contains those with at least two births who would have a third birth if and only if their first two children are of the same sex. The authors provide enough information to imply that this is only 4% of all mothers (Appendix A.3). If that is the target population, then the biggest leap between theory and evidence lies in the first step: motivating the theoretical importance of that complier population. If instead the target population is all mothers, then the biggest leap lies in the second step of the process: motivating why an estimate for 4% of mothers is informative about all mothers. Setting the target population would clarify which tack the authors are taking.

Sometimes, one could defend the complier population as being of genuine theoretical interest in itself. Harding et al. (2018) estimate the effect of prison on labor market outcomes by leveraging random variation in judges' propensities to sentence those convicted of felonies to probation versus prison. The target population is offenders who would have been sentenced to probation rather than prison if they had they faced a more lenient judge (Harding et al. 2018:67). This is a subpopulation that is conceptually interesting: individuals whose sentences might plausibly change if judges were encouraged to be more lenient in sentencing.

Observational studies that adjust for observed confounders also face challenges with the target population arising from common support problems. For example, any method would struggle to assess the causal effect of probation on offenders who committed a very serious crime (e.g. terrorism) because no one sentenced for that crime would receive probation. A lack of common support arises whenever some subpopulation defined by a confounder (e.g. terrorists) contains no treated units or contains no untreated units (e.g. those on probation or not). Common support problems leave the researcher three options. They can argue that the feasible subpopulation—those with covariates at which both treated and control units are observed—is theoretically interesting (a leap at the link between theory and the theoretical estimand); they can argue that the feasible subpopulation is informative about the broader population (a leap between the theoretical and empirical estimand); or, they can lean heavily on a parametric model and extrapolate what is observed in the feasible subpopulation to what they think would happen in the space beyond common support (a leap in estimation). As in experiments and IV, there is no free lunch. A statement of the target population is an opportunity for the author to put the difficulty in the pages of the article and clarify how they address it.

The target population clarifies debates about which methods are most credible. Some feel that econometric approaches like IV and regression discontinuity designs provide evidence that is too limited because the leap from the identified population to the full population of interest is too severe (Deaton, 2010; Deaton and Cartwright, 2018; Heckman and Urzua, 2010). Others argue that the causal identification problems in the full population are so difficult that we are better off focusing on the subpopulation for whom we can identify an effect (Imbens, 2018, 2010; Samii, 2016). Both sides have fair points. A target population allows authors to navigate this tension directly, either by arguing that the subpopulation they identify is informative about the general population or is theoretically important in its own right.

The target population appears in past work, but renewed attention is needed. Xie (2013:6262) calls for “recognition of inherent individual-level heterogeneity” and Morgan and Winship (2015:47) write that the target population is “crucial” to the definition of average causal effects. We should not expect “all-powerful theories operating with such force that they will make their presence felt regardless of countervailing conditions,” (Lieberson and Horwich 2008:11). Yet few studies state the target population. We therefore make a renewed call: the link between theory and evidence would be greatly improved if authors stated the population of units over which they seek to draw inference about the unit-specific quantity.

To summarize, the theoretical estimand states the study aim in precise terms involving a unit-specific quantity aggregated over a target population. The theoretical estimand exists outside of any statistical model and liberates us to make complex research questions precise. Descriptive estimands can be stated even if some of the population would refuse all survey attempts or is structurally missing from administrative records. Causal estimands can be stated in terms of counterfactuals we could never observe. In contrast to the constraints of regression coefficients, a theoretical estimand allows us to formalize the quantity most relevant to theory.

2.2.2 Identification: Link to an Empirical Estimand

The same quality that makes a theoretical estimand liberating—it can involve unobservable data—also means that strong assumptions will be required to learn about that estimand from statistical procedures, which can only be applied to observable data. The second step of our framework links the theoretical estimand to an empirical estimand: a target of inference that only involves observable data. That link can be formalized with tools like Directed Acyclic Graphs (DAGs) (Morgan and Winship, 2015; Pearl, 2009). Yet, despite decades of methodological advice to focus not only on technical fixes in regression but also on scientific issues like selection into treatment

(Freedman, 1991), DAGs or other equivalent statements of conditional independence appear in only a minority of sociological studies. One reason causal assumptions are missing from research practice may be that authors believe their research goals are not causal and therefore lie outside the scope of problems for which those assumptions are needed. We argue that a clear statement of both the research goal (the theoretical estimand) and the concrete target of the statistical analysis (the empirical estimand) would clarify the identification assumptions are needed in a much wider range of questions. We introduce the idea of two estimands (theoretical and empirical) with a causal example before turning to more complex examples from demography and from the study of disparities.

2.2.2.1 Two estimands: An introduction with a causal effect

A causal example from Pager (2003) illustrates how the theoretical and empirical estimands are distinct. One theoretical estimand is the average difference in whether an application would receive a callback if it came from a white applicant with a criminal record as compared with from a black applicant without a criminal record.

$$\tau = \underbrace{\frac{1}{n} \sum_{i=1}^n}_{\text{Average over all applications}} \left(\underbrace{Y_i(\text{White, Record})}_{\text{Potential outcome under one condition}} - \underbrace{Y_i(\text{Black, No record})}_{\text{Potential outcome under another condition}} \right) \quad (2.2.3)$$

For each unit, it is not possible to observe both potential outcomes, so the theoretical estimand τ is not an empirical quantity. The empirical estimand is the difference in the observed outcomes between job applications actually assigned to each of these experimental conditions. This involve only observable quantities (no potential outcomes).

$$\begin{aligned}
\theta = & \overbrace{\frac{1}{n_{\text{White, Record}}} \sum_{i \in S_{\text{WR}}} Y_i}^{\text{Average among those assigned to one condition}} - \overbrace{\frac{1}{n_{\text{Black, No record}}} \sum_{j \in S_{\text{BN}}} Y_j}^{\text{Average among those assigned to another condition}} \\
& \quad \uparrow \qquad \qquad \qquad \uparrow \\
& \text{Sum over all applications } i \text{ in} \qquad \text{Sum over all applications } i \text{ in} \\
& \text{the set } S_{\text{WR}} \text{ that signal a} \qquad \text{the set } S_{\text{BN}} \text{ that signal a} \\
& \text{White applicant with a record} \qquad \text{Black applicant with no record}
\end{aligned} \tag{2.2.4}$$

The empirical estimand θ is informative about the theoretical estimand τ under a key assumption that the signals of race and of a felony conviction are assigned independently of the callback that would be realized if they were different. Like many identification assumptions, this assumption involves counterfactual outcomes and thus must be defended on conceptual grounds rather than checked empirically. In Pager (2003), the design—randomization of treatment assignment—makes the identification assumption highly plausible. Observational studies often seek to condition on variables to address confounding—the failure of this key assumption.

These two types of estimands (theoretical and empirical) appear in different spaces of the methodological literature. If you opened a textbook on causal inference or missing data (e.g. Imbens and Rubin 2015), the authors would use the word “estimand” to mean things like the average treatment effect. Because these involve unobservable data, we would term them theoretical estimands. In a standard probability or statistics textbook (e.g. Blitzstein and Hwang 2019), the authors would talk about estimators as tools to estimate unknown parameters of random variables for which it is possible to observe realizations. These are empirical estimands in our framework. In social science, we need both. The theoretical estimand clarifies the social science goal and the empirical estimand clarifies the quantity that our statistical procedures are designed to recover.

Stating both estimands is important because there is no 1-1 mapping between a theoretical and empirical estimand. One could examine a particular empirical

estimand—for example, the difference in the mean callbacks of black and white applicants in administrative records with no adjustment—which could correspond to theoretical estimands as diverse as a descriptive disparity or a causal effect. The argument of a research study has to clarify to which of many possible theoretical estimands they intend the empirical estimand to speak so that the reader can adequately evaluate the available evidence for the claim. This is especially true in more complex settings.

2.2.2.2 Setting 1: Demographic standardization

Consider standardized mortality rates. We might compare the age-specific mortality rate (e.g. deaths per thousand among those ages 50–54) in Mexico and the U.S. A demographer might then aggregate age-specific estimates to a summary statement: the mortality rate in the U.S. compared with the mortality rate in the Mexican population aggregated over the age distribution of the U.S. (Preston et al., 2000). At this point, there are at least two possible theoretical estimands. One is the descriptive disparity between U.S. mortality and Mexican mortality aggregated over the U.S. age distribution. For that estimand, the link to theory is weak: why exactly do we care about that reweighting of the Mexican population, given that Mexico does not have the age distribution of the U.S.? A second theoretical estimand is the causal difference between U.S. mortality and the counterfactual mortality that U.S. individuals would experience under an intervention to move them to Mexico. That would clarify why we aggregate over the U.S. age distribution; we are making an estimate for which the target population is the U.S. That estimand might have a strong link to theory: it assesses how societal context affects mortality; however, the link to evidence is weak. It is hard to believe the causal claim when only age has been adjusted and not other contributors to mortality like differences in educational attainment between the populations. Although a demographer would rarely state the goal in explicitly causal

terms, they might discuss what “would” happen in a “counterfactual” population. Without such an explicit statement, the goal is unclear.

Sociologists fall prey to the same problem when, for example, they deploy Kitagawa-Blinder-Oaxaca decompositions (Kitagawa, 1955) and related methods to discuss what would happen in counterfactual populations in which covariates took different values (e.g. Ciocca Eller and DiPrete 2018; Mize 2016; Storer et al. 2020). Like a standardized mortality rate, the methods used in those articles allow us to back out the empirical estimand from the procedures applied to the data. But we are left wondering what the theoretical estimand was, and how the authors navigate the link between the two. Rather than only discussing the procedures applied to the data, authors who state both the theoretical and empirical estimands get to clarify exactly what they are after and how their evidence speaks to that quantity.

2.2.2.3 Setting 2: Disparities in the presence of selection processes

Few sociology papers reason explicitly about identification assumptions. One reason may be that the objects of sociological inquiry appear on the surface to be descriptive sample quantities, which may be valid under weaker assumptions. Yet results which seem to be descriptive empirical regularities, or stylized facts (Hirschman, 2016), often take on a theoretical meaning only under identification assumptions. We review three examples (Table 2.2) with a common style. The authors cite a descriptive disparity—police shootings by race, graduate admissions by sex, and adult incomes by race—but control for a third variable that is a consequence of the demographic characteristic of interest. This produces problems from conditioning on a collider variable (Elwert and Winship, 2014). These examples highlight the need to state the theoretical estimand, the empirical estimand, and the identification assumptions under which the two are equal even when the target quantity may not appear to be causal at first glance. A

precise statement of the theoretical estimand can inform the assumptions to identify that estimand.

Fryer (2019) examines police interactions by race in several administrative data sources. In records from New York City, the use of sub-lethal force was higher for blacks than non-blacks. Yet data from Houston on the most extreme form of force, police-involved shootings, showed no differences across racial groups. In both of these settings, the theoretical estimand (racial bias) is the difference in force if we intervene to change an officer’s perception of an individual’s race, averaged over those stopped by police. The empirical estimand is the difference in force used against black and white individuals who are involved in police interactions. Knox et al. (2020) highlight a key issue: the sample only includes those who interacted with police, either due to a stop or a 911 call, yet race affects whether these events occur (Table 2.2). If being black increases the risk of being stopped, then black individuals with a range of behaviors are stopped while only the most dangerous white individuals are stopped. Because the white individuals who are stopped are more dangerous than the black individuals who are stopped, an unbiased officer might actually use lethal force against whites at a *higher* rate among those who have been stopped. That is, equivalent rates are actually consistent with racial discrimination.²

The core empirical fact has not changed; one would calculate the same probability of a police-involved shooting given race of the stopped suspect in the sample. The *theoretical implication* of that empirical fact, however, has changed quite dramatically if we accept the assumption that being stopped by police is a consequence of both race and behavior. Black individuals are shot at equal rates despite good reason to suspect that their behavior (among those stopped) is less dangerous. What seemed to be a descriptive empirical regularity is best interpreted in light of causal assumptions

²Fryer (2019) discusses sample selection in the section “A note on potential selection into police data sets.” He controls for available measures including precinct and officer characteristics, but this cannot adjudicate selection on unmeasured factors.

Study	Empirical Regularity	Misleading Conclusion	Directed Acyclic Graph
Fryer (2019)	Among those they stop, police shoot the same proportion of black individuals as white individuals.	Police do not discriminate against black individuals when using lethal force.	<pre> graph LR PB[Perceived as black] --> SBP[Stopped by police] SBP --> LF[Lethal force] CA[Criminal activity] --> SBP CA --> LF PB -.-> LF </pre>
Bickel et al. (1975)	Among those who apply, Berkeley departments admit a higher proportion of women than of men.	Admissions committees do not discriminate against women.	<pre> graph LR F[Female] --> AB[Applied to Berkeley] AB --> A[Accepted] SC[Strong candidate] --> AB SC --> A F -.-> A </pre>
Chetty et al. (2020)	Among those with equal childhood incomes, black and white women earn similar amounts as adults.	Equalizing childhood incomes would eliminate the racial gap in women's adult incomes.	<pre> graph LR B[Black] --> CI[Childhood income] CI --> AI[Adult income] OFA[Other family advantages] --> CI OFA --> AI B -.-> AI </pre>

Table 2.2: Empirical regularities can be misleading without estimands.

Each example reports an empirical regularity with a vague connection to a theoretical claim. The empirical regularity supports the misleading conclusion only under identification assumptions that the node at the bottom of each Directed Acyclic Graph (DAG, Pearl 2009) does not affect both the variable that the researchers hold constant (boxed) and the outcome (at right). We draw the Fryer (2019) example from a critique by Knox et al. (2020) which highlights this and other issues with the original paper. In the first row, equal use of lethal force against black individuals stopped by police may stem from the fact that being stopped is a collider: among those stopped, the behavior of blacks is likely to be less dangerous. In the second row, equal or higher acceptance rates among female candidates who apply to Berkeley could result because applying to Berkeley is a collider: among women, only the strong candidates apply. In the third row, childhood income is a collider: black families who overcome discrimination to attain incomes comparable to those of white families likely have other advantages that may contribute to their children's incomes in adulthood. When we state the theoretical and empirical estimands, the DAG makes clear that they are not equal and thus the descriptive quantity does not support the conclusions drawn.

that clarify the jump from the observed association to a theoretical conclusion about racial bias.

In response to a comment by Durlauf and Heckman (2020), Fryer (2020) claims that he never sought to study racial bias, but only “racial differences” by repeatedly caveating the results with the phrase “conditional on interaction,” (Fryer 2020:1). He writes, “I am not sure how many more ways we would have needed to caveat our results to satisfy [Durlauf and Heckman],” (Fryer 2020:1). But caveats are exactly the problem. No one is well-served when methods make empirical evidence transparent (disparities in shooting conditional on a stop) but the theoretical quantity that motivates that evidence remains vague. Retreating from theoretical claims does not make the link between theory and evidence stronger. Rather, directly confronting the gap between a precise goal and the available evidence opens the door to transparent discussions and new tools to address that gap, as demonstrated by Knox et al. (2020). This is just one reason why it is essential to transparently state the study’s true goals.

This problem is more general than the use of administrative data to study police bias. Bickel et al. (1975) study graduate admissions at Berkeley and discover that, although men are admitted at rates nine percentage points higher than women school-wide, women are admitted at higher rates than men within departments. The theoretical estimand is the difference in admission if we intervene to change a committee’s perception of an applicant’s sex. However, the empirical estimand—the disparity among students who actually apply to Berkeley—is not well situated to speak to that counterfactual. If, due to discrimination at the undergraduate level, many men apply to Berkeley but only the most qualified women apply to Berkeley, equal rates of admission among the men and women we observe could actually be consistent with sex-based discrimination against women.³

³Bickel et al. (1975) explicitly assume away this problem: “in any given discipline male and female applicants do not differ in respect of their intelligence, skill, qualifications, promise, or other attribute deemed legitimately pertinent to their acceptance as students. It is precisely this assumption that makes the study of ‘sex bias’ meaningful, for if we did not

As a third example, Chetty et al. (2020) show that black and white women who are raised in families with similar incomes have similar earnings as adults. At face value, one might interpret this in terms of a theoretical estimand: if we intervened to equalize the childhood incomes of black and white women, the racial income gap in adulthood would disappear. Yet this would be misleading because family income is a consequence of both race and other family advantages; the black families who overcome discrimination to achieve incomes comparable to those of whites are likely to be advantaged in many other ways. In other words, childhood income is a collider variable (Table 2.2). The racial income gap in adulthood that would persist if we equalized childhood family incomes (a theoretical estimand, see Lundberg 2021) is likely to be different from empirical evidence about the racial gap in adult incomes among those observed with equal childhood incomes (an empirical estimand).

In all three cases, what appears to be a descriptive empirical regularity may not tell us what we want to know about racial bias, sex bias, and the transmission of racial inequality across generations. These issues are more complex because of selection into the sample along a variable—application to Berkeley, being stopped by police, and childhood income—which is a consequence of the category of interest. The key to using description to update our theories is to translate the theory into an implication about the world. But when selection limits us to observing only a slice of the world, we can get counterintuitive results. Issues of sample selection may grow in importance as sociology explores new data sources. We expect causal reasoning about sample selection will play a pivotal role in the transparent presentation of descriptive claims.

Table 2.2 formalizes these selection problems in causal DAGs (Pearl, 2009). Our framework aligns well with DAGs because they are nonparametric: they allow us to focus on one set of considerations (causal relationships) while delaying questions

hold it any differences in acceptance of applicants by sex could be attributed to differences in their qualifications, promise as scholars, and so on”(p. 398). While we applaud the explicitness of the assumption, it is questionable when discrimination affects decisions to apply for graduate school.

about the shape of statistical associations for the subsequent choice of an estimation strategy. We have argued here that identification assumptions that can be stated in DAGs apply to a wider set of sociological problems than applied researchers may think (including missing data problems for purely descriptive inferences). We refer readers to other pedagogical sources for an introduction to identification using DAGs (Morgan and Winship, 2015; Pearl and Mackenzie, 2018).

2.2.3 Estimation: Learn the Empirical Estimand from Data

Generalized linear models are far and away the primary estimation tool deployed in quantitative sociology, yet many sociologists will admit that the functional form assumptions of these models are far from perfect. The field’s awareness of this problem is evident in proposals to assess robustness across model specifications (Young and Holsteen, 2017) as well as perspectives that view any regression model as an approximation (Aronow and Miller, 2019; Berk et al., 2021; Buja et al., 2019). The appeal of new machine learning tools (Molina and Garip, 2019) and predictive exercises (Watts, 2014) derives from how these tools present an opportunity to break out of the parametric models that we all know are imperfect. But for sociologists trained to ask research questions in terms of regression coefficients, it can be difficult to see how new computational tools (which may not involve coefficients) can answer our social science questions. The path forward requires us to change the way we think about estimation. Instead of thinking about estimating the parameters of a model, we must think of the estimation algorithm as a tool to estimate the unknown components (e.g. conditional means) that appear in the empirical estimand. Doing so allows empirical evidence to inform the choice of an estimation strategy. While conceptual argument is central to the statement of the theoretical and empirical estimands, selection of an estimation strategy can be largely data-driven.

For example, identification might lead to an empirical estimand θ that is the sample-average difference between the expected outcome among those treated $D = 1$ and among those untreated $D = 0$ conditional on pre-treatment covariates $\vec{X}$.

$$\theta = \underbrace{\frac{1}{n} \sum_{i=1}^n}_{\text{Sample average}} \left(\underbrace{\mathbf{E}(Y \mid \vec{X} = \vec{x}_i, D = 1)}_{\text{Expected outcome among those treated}} - \underbrace{\mathbf{E}(Y \mid \vec{X} = \vec{x}_i, D = 0)}_{\text{Expected outcome among those untreated}} \right) \quad (2.2.5)$$

Estimation is the step of learning an estimate $\hat{\theta}$ from data. One straightforward approach is to estimate the conditional mean function and plug in estimates (e.g. $\hat{\mathbf{E}}(Y \mid \vec{X} = \vec{x}_i, D = 1)$) wherever it appears.

$$\hat{\theta} = \underbrace{\frac{1}{n} \sum_{i=1}^n}_{\text{Sample average}} \left(\underbrace{\hat{\mathbf{E}}(Y \mid \vec{X} = \vec{x}_i, D = 1)}_{\text{Regression prediction if treated}} - \underbrace{\hat{\mathbf{E}}(Y \mid \vec{X} = \vec{x}_i, D = 0)}_{\text{Regression prediction if untreated}} \right) \quad (2.2.6)$$

Estimation is tightly linked with prediction. Suppose we wanted to predict the outcome value Y for random units sampled from the subpopulation with covariates $\vec{X} = \vec{x}_i$ and treatment $D = 1$, and we fit our models on a different set of units sampled from the same population (e.g. we are not forecasting the future). For example, we might observe employment for a sample and want to predict whether a new person sampled from the same population who is 43 years old is employed. The prediction that would be as close as possible to the true values (in terms of expected squared error) is the conditional mean in the population, $\mathbf{E}(Y \mid \vec{X} = \vec{x}_i, D = 1)$. One could therefore select the estimator $\hat{\mathbf{E}}(Y \mid \vec{X} = \vec{x}_i, D = 1)$ out of many candidate estimators by selecting the one that empirically minimizes out-of-sample mean squared prediction error. A researcher could consider the empirical mean among those observed with those covariates, the prediction of a regression model with or without interactions and squared terms, or some machine learning tool. Rather than arguing among these model specifications conceptually, we could decide among them empirically: the one

that best estimates $\mathbf{E}(Y \mid \vec{X} = \vec{x}_i, D = 1)$ is the one that minimizes expected squared prediction error in out-of-sample cases. That procedure provides an empirical basis to adjudicate choices about functional forms.

Stating the empirical estimand creates further opportunities to improve the model selection metric. Empirical mean squared error optimizes the fit of predictions where we have data, but that may not be where we want to make predictions. Perhaps only 10% of the observed cases are treated, but we want to predict the outcome under treatment and control for all cases. In that setting, an estimator that predicts poorly for the treated cases but well for the untreated cases might perform well on average in the data we observe (which are almost all untreated). But in fact the estimand suggests that we should care equally about predictive performance in the treatment and control conditions, regardless of how often these appear in the observed data. The estimand could therefore guide us to a modified performance metric that adapts predictive performance to focus on the predictions we actually need to make. This is what the rapidly developing literature in machine learning and causal inference accomplishes (Van der Laan and Rose, 2011); modifying machine learning tools for social science goals requires us to specify those goals precisely.

Before assessing a set of candidate estimators, we have to develop or select those estimators. The key choice here involves how information will be shared across nearby units. Social science theory often provides only a limited guide for this task. For instance, suppose we want to estimate the proportion of 43-year-olds who are employed, and we have a simple random sample of people of various ages. We could estimate the proportion employed by the empirical mean among those who are actually 43, but our sample size in that exact age cell might be small. Social science theory could suggest some amount of smoothness: we could expect employment among 43-year-olds to be similar to the employment of those who are 42 and 44. We might share information

across these covariate values by averaging over all those ages 42–44, thus producing a slightly more precise estimator by drawing on our assumption of smoothness.

Social scientists often leap to very strong assumptions for information sharing. By assuming that the association between age and employment follows a linear or quadratic functional form, one could pool information across all ages to estimate the employment of 43-year-olds. That would produce a low-variance estimator, but only under doubtful assumptions: it is difficult to defend a linear or quadratic functional form from theory alone. This is why empirical evidence is so useful for selecting an estimator. In a very small sample, a linear regression that pools a lot of information might be the best predictor. In the Census, the empirical mean within each subgroup might be the best estimator because it makes minimal assumptions. Out-of-sample predictive performance provides an empirical tool to assess the best option.

2.2.3.1 Concrete Estimation Example: The Family Gap in Pay

To illustrate the estimation step, we conduct an exercise inspired by Pal and Waldfogel’s (2016) examination of the effect of motherhood on women’s hourly wages. Following the authors, we analyze data from the Annual Social and Economic Supplement of the March Current Population Survey, with details deferred to Appendix A.4. We focus on the most recent data collected in 2019, thereby updating the original results with the most current evidence. Our conclusion bolsters the claims of the original authors, showing that their conclusions hold under milder estimation assumptions than those maintained in the original paper.

Our focus in this example is estimation. However, as argued throughout this paper, clear reasoning about estimation requires that we first define the theoretical and empirical estimand. The original paper is not entirely clear; while the authors deploy “causal estimation techniques” (Pal and Waldfogel 2016:108), they also define the target quantity in a way that seems to appeal to a descriptive disparity between

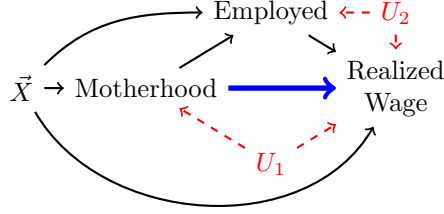


Figure 2.4: Identification assumptions for the motherhood wage penalty. To identify the controlled direct effect of motherhood on the wages that would be realized under employment, one must assume that the covariates $\vec{X}$ are sufficient to block all confounding of motherhood and of employment. The red nodes U_1 and U_2 represent threats to identification.

two populations, as “the differential in hourly wages between women with children and women without children,” (Pal and Waldfogel 2016:104). We take the goal to be causal. However, we cannot simply define the goal as the average causal effect of motherhood on the wages of mothers, because wages are not defined for those who are not employed (Pal and Waldfogel 2016:109–110 acknowledge this complexity in a footnote). In order to target a well-defined unit-specific quantity for every mother in the population, we take the theoretical estimand to be the controlled direct effect of motherhood on the wages women would realize if they were employed, averaged over the population of mothers (Line 1 in Figure 2.5). We take the empirical estimand to be the descriptive gap between employed mothers and non-mothers conditional on covariates (Line 2 in Figure 2.5). The theoretical and empirical estimands are equal under the assumptions presented in Fig 2.4, which we emphasize includes a very strong assumption of no mediator-outcome confounding. Appendix A.4 discusses alternative ways to frame the problem. Our focus here is on estimating the empirical estimand by using regression to predict the unknown conditional expectations (Line 3 in Figure 2.5). This estimation strategy is known as the parametric g -formula in biostatistics (Hernán and Robins, 2020, Ch. 13) and the imputation estimator in econometrics (e.g. Hahn 1998:321).

The imputation estimator illustrates how an empirical estimand guides the choice of an estimation strategy. Mechanically, it first involves fitting a model for log wages (the outcome variable) as a function of covariates and motherhood among those who are employed. Then, we predict log wages for all mothers with their observed covariates. Third, we predict wages in the same dataset but with the motherhood variable changed from the value `mother` to the value `non-mother`. Finally, we difference these predictions for each mother and average over the sample of mothers with survey weights to draw inferences about the target population. The imputation estimator is a general strategy that can be used to estimate any average causal effect by imputing the potential outcomes under each treatment condition for each unit. If we had measured mediator-outcome confounding, the imputation estimator could still be used with some additional modifications (Acharya et al., 2016).

The imputation estimator unlocks new tools: the same procedure holds regardless of whether the algorithm used to predict the outcome variable is OLS regression, logistic regression for a binary outcome, or a machine learning strategy. In the OLS case, it simplifies back to a familiar result: the estimated treatment effect is the coefficient $\hat{\beta}$ on motherhood. However, that simplification is only possible under the (doubtful) no-interactions assumption that the treatment effect is the same value $\hat{\beta}$ at the covariate value $\vec{x}_i$ of every unit i . In other words, a coefficient $\hat{\beta}$ is only consistent for the average effect under special cases, such as when the effect is the same in every subgroup.⁴ By motivating the estimation procedure as a tool for predicting conditional means rather than estimating coefficients, researchers open the door to alternative strategies that do not rely on the implausible assumption of linearity with no interactions.

⁴When the effect is not constant, $\hat{\beta}$ can be reinterpreted as a weighted average of strata-specific estimates (Elwert and Winship, 2010). The weighted average can equal the unweighted average, but it is not true in general.

1) **Set** the target. Define a theoretical estimand. Requires substantive **argument**.

Average difference in the **potential outcome** each woman i would realize

$$\tau = \frac{1}{n} \sum_{i=1}^n \left(\begin{array}{c} \text{if she were an employed mother} \\ Y_i(\text{Mother, Employed}) \end{array} \quad \text{versus} \quad \begin{array}{c} \text{if she were an employed non-mother} \\ Y_i(\text{Non-mother, Employed}) \end{array} \right)$$

2) **Link** to observables. Define an empirical estimand. Requires conceptual **assumptions**.

Average difference in the **realized outcomes** of women with the covariates $\vec{x}_i$ of women i who

$$\theta = \frac{1}{n} \sum_{i=1}^n \left(\begin{array}{c} \text{actually are mothers} \\ \mathbf{E} \left(Y \mid \begin{array}{l} \text{Motherhood} = \text{Mother,} \\ \text{Employment} = \text{Employed,} \\ \text{Covariates } \vec{X} = \text{Observed } \vec{x}_i \end{array} \right) \end{array} \right) - \begin{array}{c} \text{actually are not mothers} \\ \mathbf{E} \left(Y \mid \begin{array}{l} \text{Motherhood} = \text{Non-mother,} \\ \text{Employment} = \text{Employed,} \\ \text{Covariates } \vec{X} = \text{Observed } \vec{x}_i \end{array} \right) \end{array}$$

3) **Learn** from data. Select an estimation strategy. Requires statistical **evidence**.

Average difference in the **regression prediction** at the covariates $\vec{x}_i$ of women i if we

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n \left(\begin{array}{c} \text{recode as a mother} \\ \hat{\mathbf{E}} \left(Y \mid \begin{array}{l} \text{Motherhood} = \text{Mother,} \\ \text{Employment} = \text{Employed,} \\ \text{Covariates } \vec{X} = \text{Observed } \vec{x}_i \end{array} \right) \end{array} \right) - \begin{array}{c} \text{recode as not a mother} \\ \hat{\mathbf{E}} \left(Y \mid \begin{array}{l} \text{Motherhood} = \text{Non-mother,} \\ \text{Employment} = \text{Employed,} \\ \text{Covariates } \vec{X} = \text{Observed } \vec{x}_i \end{array} \right) \end{array}$$

$\uparrow$ $\uparrow$ $\uparrow$
 estimate of estimated $\hat{Y}_i(\text{Mother})$ estimated $\hat{Y}_i(\text{Non-mother})$
 the estimand

Does the estimate $\hat{\theta}$ equal a coefficient?

If that regression is
ordinary least squares *If that regression is*
anything else

Simplifies to a coefficient
 (by an overly simplistic parametric model)

Does not simplify
 (because the model is realistically complex)

$$\begin{aligned} & \hat{\mathbf{E}} \left(Y \mid \begin{array}{l} \text{Motherhood,} \\ \text{Employment} = \text{Employed,} \\ \text{Covariates } \vec{X} = \text{Observed } \vec{x}_i \end{array} \right) \\ &= \begin{cases} \hat{\alpha} + \vec{x}_i' \hat{\gamma} & \text{if Motherhood = Non-Mother} \\ \hat{\alpha} + \hat{\beta} + \vec{x}_i' \hat{\gamma} & \text{if Motherhood = Mother} \end{cases} \\ & \quad \begin{array}{l} \text{Intercept} \quad \text{Coefficient on} \quad \text{Coefficients on} \\ \text{By parametric} \quad \text{motherhood} \quad \text{other covariates} \\ \text{approximation} \end{array} \\ & \hat{\theta} = \frac{1}{n} \sum_{i=1}^n \left(\overbrace{\left(\hat{\alpha} + \hat{\beta} + \vec{x}_i' \hat{\gamma} \right)}^{\hat{Y}_i(\text{Mother})} - \overbrace{\left(\hat{\alpha} + \vec{x}_i' \hat{\gamma} \right)}^{\hat{Y}_i(\text{Non-mother})} \right) \\ &= \frac{1}{n} \sum_{i=1}^n \left(\underbrace{\hat{\alpha} + \vec{x}_i' \hat{\gamma} - \hat{\alpha} - \vec{x}_i' \hat{\gamma}}_{\text{Cancels because model assumes no interactions}} \right) + \frac{1}{n} \sum_{i=1}^n \hat{\beta} \\ &= \hat{\beta} \quad \leftarrow \text{coefficient on motherhood} \end{aligned}$$

$$\begin{aligned} & \hat{\mathbf{E}} \left(Y \mid \begin{array}{l} \text{Motherhood,} \\ \text{Employment} = \text{Employed,} \\ \text{Covariates } \vec{X} = \text{Observed } \vec{x}_i \end{array} \right) \\ &= \hat{f}(\text{Motherhood, Employed, Observed } \vec{x}_i) \\ & \text{where } \hat{f}() \text{ is a prediction function.} \end{aligned}$$

$\hat{f}()$ may involve interactions. The gap between mothers and non-mothers may depend on $\vec{x}_i$. This is a feature; a flexible $\hat{f}()$ can capture complexities that actually exist in the world.

There may be no parameter of $\hat{f}()$ that estimates the estimand. This is equally true for logistic regression as for complex machine learning methods.

Nonetheless, a summary statistic is possible. The empirical estimand guides the extraction of an estimate from a complex model: predict each observation's outcome under each treatment condition, take the difference, and average over the sample.

Figure 2.5: Estimands unlock new estimation tools. Once we state the estimand, we can use a predictive algorithm to impute unknown conditional means. It could be a linear model (e.g. OLS) or a nonlinear model (e.g. logit, random forest). The result equals a coefficient for linear models only. The empirical estimand guides the use of a realistically complex functional form.

In this example, we can actually conduct estimation without *any* assumptions about the functional form. The result holds even when we share no information across observations, imputing the expected wage within a covariate cell by the mean wage of those observed with exactly that set of covariates (far left of Fig. 2.6). So why would one assume a functional form, like a parametric model where the estimand is estimated by a coefficient (far right of Fig. 2.6)? In this case, the OLS model is clearly misspecified: it assumes that the family gap in pay does not vary by age (see lower right panel of Fig. 2.6), despite evidence in the other panels that the gap is larger in magnitude at younger ages. Yet one might prefer the OLS coefficient if the sample size were very small so that the stratification estimator at the left was infeasible or produced extremely uncertain estimates. Stating the estimand therefore does not preclude the use of a regression model as an approximation; rather, it provides a precise statement of the research goal so that we can begin to reason about the best empirical approximation. In a large sample, we may often be able to estimate by more credible assumptions than parametric models.

Using the stratification estimator weakens the *estimation* assumptions but cannot weaken the *identification* assumptions. For example, no estimator will get us out of making assumptions about unobserved confounding that affects both employment and wages (U_2 in Figure 2.4); those assumptions are out of scope for an algorithm because no algorithm can see the non-existent wages of the non-employed. While the choice of the best estimation method could be made using the data, we need subject matter knowledge to assess whether the identification assumptions are plausible.

2.2.3.2 Estimands reveal two estimation issues that are often overlooked: Statistical inference and common support

The purpose of an estimation strategy is to estimate the empirical estimand. We would like valid procedures for statistical inference that produce, for example, a 95%

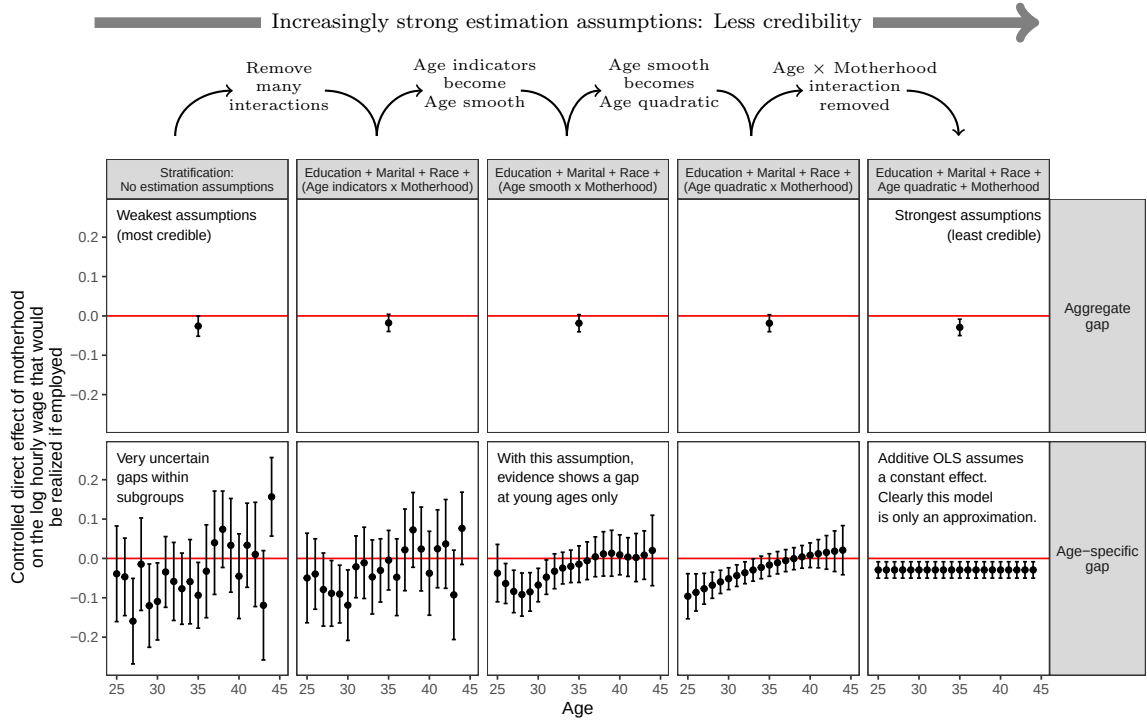


Figure 2.6: A series of estimation strategies (columns) for two estimands (rows). Each estimand is the gap in log hourly wages between mothers and childless women, conditional on age, education, marital status, and race and aggregated over the covariate distribution of mothers. Estimands differ by aggregating over ages (top row) or not (bottom row). Estimation strategies range from weakest assumptions (left) to strongest assumptions (right). In the notation of the top titles, terms such as (Age indicators $\times$ Motherhood) represent an interaction and its lower-order terms. Provided that samples sizes are large enough to yield estimates that are sufficiently precise, one would prefer the estimation strategies to the left because they are more credible. Machine learning approaches such as the Generalized Additive Model (center column, Wood 2017) represent a middle ground between parametric models (OLS, far right) and nonparametric approaches (stratification, far left). Some findings, such as the population average gap (top row), are relatively invariant to the estimation strategy and can be defended under minimal estimation assumptions (far left). Other findings, such as the age-specific gap (bottom row), require modeling assumptions to achieve adequate precision. We suggest that the tendency to define estimands by a regression coefficient has prevented social scientists from recognizing setting when inference can proceed from more minimal estimation assumptions (at left). Instead of beginning from the right and moving left, we propose that researchers default to the left side and move right, motivating each choice to add an assumption.

confidence interval that actually would contain the empirical estimand in 95% of hypothetical samples. A valid interval is elusive in both parametric models and machine learning approaches. Parametric models (e.g. OLS) come with readily-available confidence intervals for the coefficient of interest. Those intervals would provide the expected coverage for the coefficient that would be estimated if the regression were estimated on the full population. But that coefficient is not the empirical estimand: if the functional form assumed by the model is a poor approximation to the truth, then a confidence interval for the coefficient may have very poor coverage for the empirical estimand. It is therefore easy to produce a confidence interval for a coefficient, but the properties of that interval with respect to the estimand rely on the assumed functional form, which may be questionable. Flexible machine learning approaches avoid this problem by learning the functional form from the data. Yet they face a different problem: the statistical theory to place standard errors around machine learning estimates can be lacking. This is an area of active research (e.g. Wager and Athey 2018) and can sometimes be overcome by computational approaches such as bootstrapping. Appendix A.4 details the procedure that produces standard errors in our example about the family gap in pay. Parametric models and machine learning estimators thus lead to distinct issues for statistical inference.

Second, all estimation approaches can be hindered by problems of common support; if there are covariate strata $\vec{X}$ for which one or more treatment levels is not observed, the estimator must somehow extrapolate from other observations to impute a potential outcome. The flexibility of some machine learning approaches means that it can be difficult to summarize how the model extrapolates to accomplish this task, a problem that can be particularly acute in high-dimensional data (D’Amour et al., 2020). The extrapolation may be more transparent in parametric models (extrapolate a line), albeit still doubtful because the assumption of a linear relationship may be

difficult to defend. Both settings therefore call for careful consideration of common support.

A precise estimand is the first step toward productive dialogues on both of these fronts. Confidence intervals may provide imperfect coverage of the estimand and estimates may rely on questionable extrapolations. Yet we do ourselves no favors by hiding these problems behind an assumed parametric model that we know is actually only an approximation. Stating the estimand brings issues of statistical inference and common support out of the shadows and onto the page of the research paper, thereby facilitating arguments about these difficult issues.

Interactive parametric models and flexible machine learning approaches have a lot to offer the social sciences. The cost of these approaches is that the treatment effect is no longer equated with a coefficient. This cost falls on the researcher, who must conduct post-processing steps to convert an estimated model into predicted values and then to an estimate of the estimand. Because these tasks are carried out by the researcher, more flexible models impose almost no burden on the reader. In exchange, both the researcher and the reader have the benefit of substantive conclusions estimated under weaker (more credible) functional form assumptions.

2.2.4 Summary of Research Framework

To summarize, our proposed research framework involves three key choices. (1) Choose a theoretical estimand and defend its relationship to a general theory. This is likely to require specificity about the hypothetical intervention (if causal) and the target population (in all cases). (2) Choose an empirical estimand that can be linked to the theoretical estimand by a set of identification assumptions. (3) Choose an estimation strategy to learn the empirical estimand from data. Together, these three steps make a clear linkage between theory and empirical evidence in which each step can involve a principled choice.

2.3 Illustrations: How Existing Work Conflicts with Our Framework and Estimands Improve Practice

Our contention is that greater attention to estimands could revolutionize substantive claims and reorient methodological guidance. A necessary condition for this argument is that current quantitative practice in sociology does not already explicitly or implicitly specify the theoretical and empirical estimands. To investigate this, we review the 2018 volume of the *American Sociological Review* and show that we cannot consistently determine the theoretical estimand. We then turn to what it would mean for the field to reorient methodological choices around our framework. Unlike new statistical adjustments, changing the theoretical estimand can set the research on a completely different path, making it difficult to produce general statements about what would happen. Instead, we demonstrate in two specific examples: a descriptive example about the gender gap in college completion and a causal example about the effect of paternal incarceration on maternal depression.

2.3.1 Statistical Practice in a Top Journal Does Not Follow Our Framework

Fig. 2.7 summarizes our review of all 32 articles using quantitative data in the 2018 volume of the *American Sociological Review*. The goal of this review was to assess whether our proposed framework merely introduces new terminology for existing practices: can we already translate standard summaries of quantitative analyses into unambiguous theoretical estimands involving unit-specific quantities aggregated over target populations even if they are not stated explicitly? Because the theoretical estimand links statistical analyses to theory, we considered not only the procedures

applied to the data but also how the authors interpreted the procedures and results. Two of us read each paper and iterated to come to a joint assessment. Our determinations on each paper are summarized in Appendix A.8. For zero papers were we completely certain of both the unit-specific quantity and the target population. The fact that past research does not fit into our framework is unsurprising: we had not yet proposed this framework. Yet the conflicts between standard practice and our framework are nonetheless troubling: when there is disagreement about what the research goal is, it is difficult to adjudicate downstream debates about identification and estimation.

2.3.1.1 Unit-specific quantity

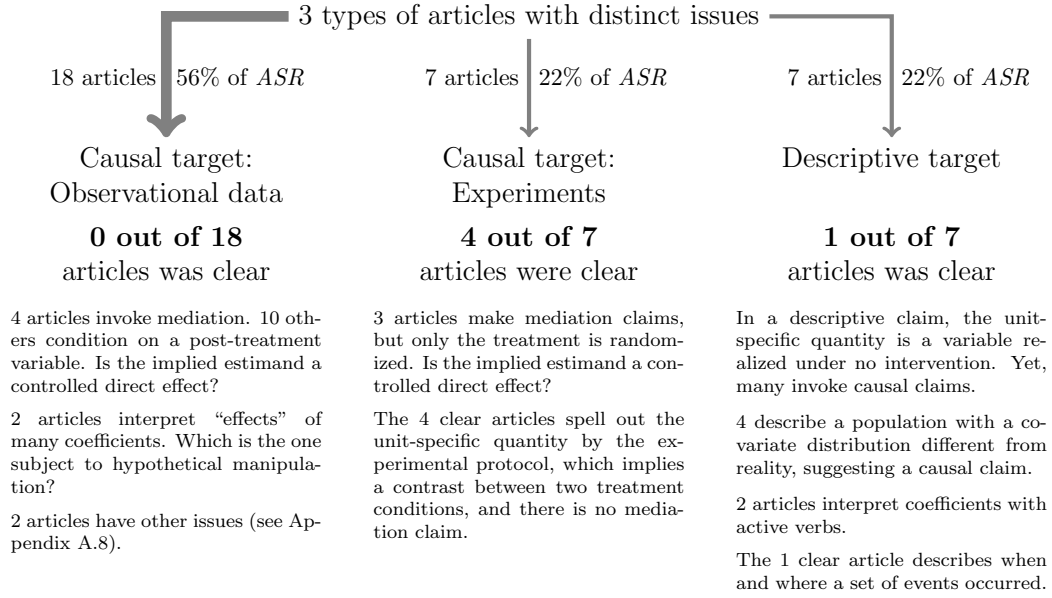
Our framework advocates the statement of unit-specific quantities as either realized random variables (for descriptive goals) or random variables that would be realized if one or more treatments were fixed to values they would not have otherwise taken (for causal goals). In our framework, every unit-specific quantity involves components like those in Eq. 2.3.1.

$$\text{Unit-specific quantity: } \underbrace{\text{Outcome}_i}_{\text{Descriptive}} \quad \text{or} \quad \underbrace{\text{Outcome}_i(\text{intervention})}_{\text{Causal}} \quad (2.3.1)$$

A descriptive unit-specific quantity supports interpretations about outcomes among sets of units. A causal unit-specific quantity supports interpretations about what would happen to a given unit if predictors took a different set of values; that requires clarity about the exact values to which predictors are hypothetically fixed.

We can confidently state the unit-specific quantity in the form of Eq. 2.3.1 in only 5 out of 32 papers (16%). Ambiguities in this quantity differ across three categories of papers. In studies drawing causal claims from observational data (56% of *ASR*), zero articles provide enough detail for us to be entirely confident of the intended

A) Is there a unit-specific quantity (e.g. Eq 2.3.1)? Details in Appendix Table A.1



B) Is there a target population (e.g. Eq 2.3.2)? Details in Appendix Table A.2

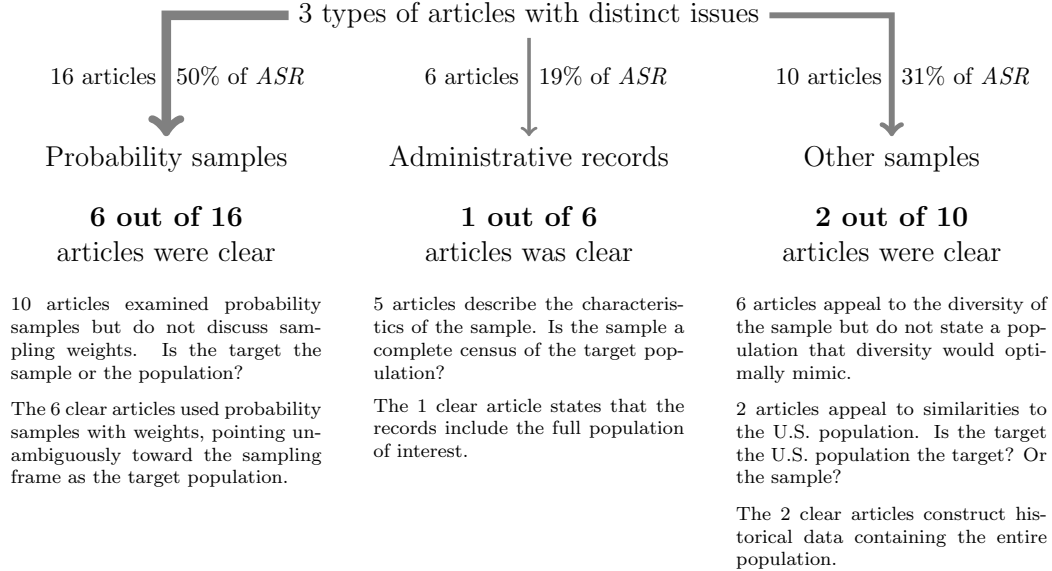


Figure 2.7: Our methodological framework differs from standard practice.

In our framework, all estimands are functions of actual outcomes Y_i (for descriptive estimands) or potential outcomes $Y_i(d)$ (for causal estimands), aggregated over well-defined populations of units indexed by i . Zero articles in the 2018 volume of *ASR* wrote the estimand this way. Some articles used sufficiently precise language that either the unit-specific quantity or the target population could be inferred unambiguously, but for no article could we infer both without some ambiguity. Details are in Appendix A.8.

intervention. Most of this category (78%) conducts an analysis that conditions on a post-treatment variable, often targeting a regression coefficient net of this variable. Four articles explicitly mention mediation; ten do not explicitly discuss mediation but nonetheless condition on a post-treatment variable. For all 14, we are unsure which of the many possible mediation estimands (Sec. 2.3.3) is the object of inquiry. Experiments (22% of *ASR*) involve a well-defined causal contrast for the treatment effect as specified by the experimental protocol. However, mediation claims with non-randomized mediators (appearing in 43% of such studies) are subject to the same issues common among observational studies (discussed in Sec. 2.3.3).

Descriptive studies constitute a minority of articles (22% of *ASR*). However, 6 of these 7 studies conduct at least one analysis reaching beyond pure description to claims that we consider to be at least implicitly causal. For example, Ciocca Eller and DiPrete (2018:1187) do not discuss causal effects or identification assumptions and yet examine disparities in college completion after “counterfactually shifting the dropout risk distribution of entering black students so that it more closely resembles the distribution for white students.” In our framework, there are two types of claims: descriptive claims about unit-specific quantities in an observed population and causal claims about unit-specific quantities that would be realized if each individual were exposed to a hypothetical intervention. Both types of claims are important. Descriptive claims might include a comparison of rates between two groups—e.g., college completion rates for Black and White students. Causal claims involve a hypothetical causal intervention and identification assumptions. What does not fit in our framework is the middle ground: claims about what would happen under some **condition** (e.g., if Black students’ dropout risk was similar to that of White students) which present a regression prediction but do not make causal assumptions explicit. The middle-ground claims only tell us how our model-specific predictions would change if we alter the input for **condition** in the model. Crucially, without identification

assumptions and a hypothetical causal intervention, these are counterfactuals *of the model*, but they need not correspond to the effect of `condition` being realized in the world. Because the estimand does not describe the world as it is, or the world as it might be under a clearly stated intervention, we do not consider non-causal counterfactual estimands to provide a compelling link between the model and the theory.⁵

2.3.1.2 Target Population

In our framework, a theoretical estimand involves a precise statement of the target population about whom claims are made. A target population is a set of existing units (indexed by i) such that the theoretical estimand can be defined as some aggregation over that population, such as the average in Eq. 2.3.2.

$$\text{Aggregate over target population: } \frac{1}{\text{Number of units in the target population}} \sum_{\substack{i \text{ in the} \\ \text{target population}}} (\text{Unit-Specific Quantity})_i \quad (2.3.2)$$

The target population in our framework is rarely the set of units in the data; it is the set of units in the population about which the theoretical claims are made. A detailed statement of how the data were collected does not constitute a statement of the target population.

We can confidently state the target population in only 9 out of 32 papers (28%). Half of *ASR* articles draw on probability samples, but 62% of those articles do not discuss how (if at all) survey weights are incorporated into the main analyses. It is then ambiguous whether the target population is the sample, the sampling frame,

⁵Often these analyses have clear empirical estimands, but they are about parameters of a specific model. Our objection then, is primarily about the lack of assumptions to translate between that model and a claim about the world outside of the model. Without these assumptions, readers have no way to adjudicate between competing estimates of the same non-causal counterfactual. In practice, we also think readers interpret counterfactual claims in a causal way even when explicitly cautioned not to.

or some broader population. For instance, Liu (2018) uses data from Framingham, Massachusetts. Is this particular town of theoretical interest, or is the hope that it is informative about a broader population? With administrative records (19% of *ASR*), authors are often remarkably clear about who is in the records, but only 1 out of 6 articles (Font et al., 2018) explicitly states whether the set of units in the records is the entire population about which they seek to draw inference. Finally, other samples such as Amazon Mechanical Turk and datasets constructed by the author appear in 31% of *ASR*. Most of these studies defend the chosen sample on the grounds of its diversity. Yet diversity is only helpful if that diversity matches the diversity of some target population of interest, enabling weighted inferences to be valid for the target. What target population should our diverse sample mimic? It is difficult to assess things like confidence intervals and significance tests (typically based on the idea of sampling from a target population) when authors have described the sample but left the target population unstated.

2.3.1.3 Summary of the review of *ASR*

As readers, we often ask “what is the estimand?” and cannot reverse-engineer an answer from published articles. In its purest form, our framework proposes mathematical precision to resolve these ambiguities. The causal contrast becomes clear when estimands are written as functions of unit-specific quantities: either actual outcomes Y_i for descriptive estimands or potential outcomes $Y_i(d)$ for causal estimands. Explicit aggregation over a well-defined set of units indexed by i leaves no ambiguity about the target population. But mathematical formalization is not absolutely necessary; we would be happy if authors stated the estimand in words with sufficient precision that we could translate the description to a particular estimand. This might be possible through description of an experimental protocol or a clear hypothetical intervention in an observational study, paired with a precise statement of the target

population. What is troubling is our inability to unambiguously translate the description provided by authors into a precise research goal. As a result, it is difficult to know what was learned or to reason about methodological procedures by which it could be learned better. Productive scientific exchange is difficult when articles do not make clear what question was answered.

The present state of the field means that sociology stands at a remarkable opportunity. We can answer more precise research questions and unlock new tools for estimation through a clear statement of the estimand. The next sections use specific examples to show how the proposed framework could transform quantitative sociology.

2.3.2 Specific Example 1: Descriptive Estimands Can Be More Compelling Without Multiple Regression

Buchmann and DiPrete (2006) summarize a gender reversal: while men historically completed college degrees at higher rates than women, the disparity reversed over the second half of the 20th century. In one analysis, the authors fit a logistic regression for college completion as a function of gender, birth cohort, and father’s education, with interactions (original Table 2 Model 1). They conclude that “the emergence of a female advantage in education is attributable to a reversal in the gender-specific effects of father status,” (Buchmann and DiPrete 2006:525). While the statement evokes something more meaningful, the quantity in question is relatively opaque—the coefficient on an interaction capturing change over time in a difference in log odds between two subgroups.

Suppose the researchers instead summarized a series of descriptive estimands: the probability of college completion as a function of gender, birth cohort, and parental

characteristics, stated without any appeal to regression coefficients.

$$\tau(g, p, c) = P(\text{College} = 1 \mid \text{Gender} = g, \text{Parent characteristics} = p, \text{Cohort} = c) \quad (2.3.3)$$

Each person has a unit-specific realized outcome indicating whether that person completed college or not. The conditional probability above averages that unit-specific outcome over the subpopulation with a given set of predictor variables, among the larger population of white U.S. adults born in 1947–1984 and observed at ages 25–34.

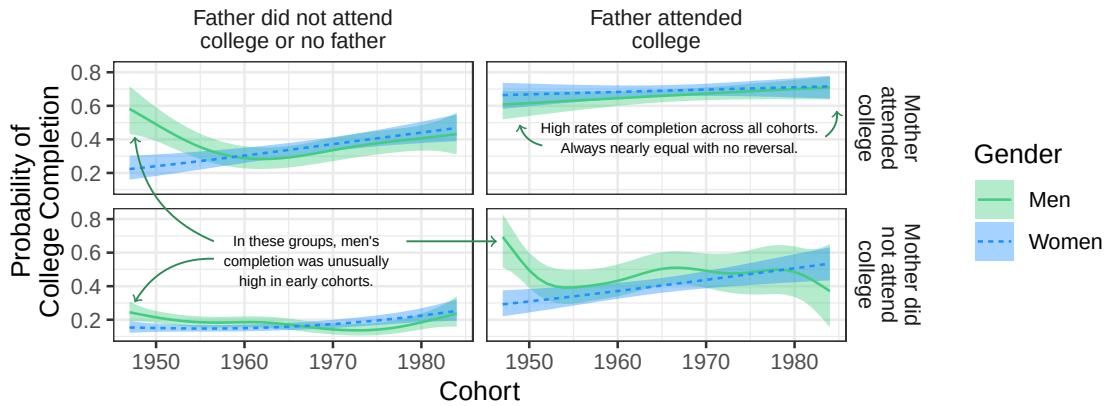
Our empirical estimand is the same quantity, conditional on the fact that one is alive and willing to respond to the survey. We will return to discuss how selective death may call into question the equality of τ and θ .

$$\theta(g, p, c) = P(\text{College} = 1 \mid \text{Gender} = g, \text{Parent characteristics} = p, \text{Cohort} = c, \text{Alive and willing to respond}) \quad (2.3.4)$$

Fig. 2.8 summarizes these descriptive estimands with gender g represented by colors, parent characteristics p represented by the grid of plots, and cohorts c represented by the x -axis. We extend the series to include all data now available. Instead of dichotomizing birth cohorts at 1966 (the original specification), we use generalized additive models (GAMs, Wood 2017) to estimate smooth but flexible curves. In one respect, our result reproduces the original authors’ claims: the disparity reversed the most among those for whom at least one parent did not attend college (panels other than the top right). This descriptive statement involves no appeal to effects defined as regression parameters: it is simply a description within subgroups.

Flexible descriptions can spark theoretical questions that would otherwise remain buried by parametric specifications. The original authors focus on why the gap closed, but our results raise a different question: why was college completion so common

Descriptive claim: Men historically completed college degrees at higher rates than women. The reversal over cohorts born 1947–1984 differed across subgroups.



Vague estimand reaching beyond description: “The emergence of a female advantage in education is attributable to a reversal in the gender-specific effects of father status,” (Buchmann and DiPrete 2006:525).

Figure 2.8: Descriptive estimands are worthwhile goals. The language of “effects” common in multiple regression models can produce confusion. The figure is purely descriptive, presenting estimates of the mean within subgroups with no control variables. The only model serves to smooth over cohorts (Wood, 2017). The evidence base in the figure is analogous to the logistic regression model from Buchmann and DiPrete (2006, Table 2 Model 1): the predictors of that model define the subgroups in this plot (gender, cohort, and parent characteristics). It is clear that these subgroups produce an interesting description. It is not clear that this description, or the analogous logit model, allows one to attribute the reversal to any particular “effect.” We propose that descriptive estimands—means within subgroups—can specify the research goal while avoiding the tendency to state results in vaguely-defined effects and attributions.

among men born in the early 1950s whose father did not attend college? One candidate explanation is that these are the same cohorts in which men were conscripted to serve in the Vietnam War (especially birth years 1950–1952, Angrist and Chen 2011). The Vietnam War could have produced especially high college completion rates for this cohort because veterans received scholarships upon their return under the GI Bill. The high completion rate could also arise in part from an identification problem: the theoretical estimand involves everyone born in 1950, but those killed in the war were not around decades later to complete the GSS survey. Perhaps the men who would not have completed college were disproportionately drafted and killed in the war, contributing to a gap between the theoretical and empirical estimand.

In this example, a clear statement of the theoretical estimand took us down a very different interpretive road than that taken by the original authors. When puzzling over the gender reversal in college completion, much has been learned by decades of scholarship about the role that fathers play in sons’ educational attainment. Yet, much could also be learned by closer examination of the gendered effects of the Vietnam War on college completion. A clear statement of a descriptive estimand—free from colloquial “effects” terminology—has the power to remove blinders from our collective eyes and promote the development of new theory and new research questions.

2.3.3 Specific Example 2: Causal Estimands Facilitate Interpretable Effect Sizes and Clarify Claims to Mediation

The subtle nature of counterfactual statements means that causal work would particularly benefit from clear estimands. Colloquial terms like “mediation” can obscure the claim being made.

2.3.3.1 Example: Effects of Paternal Incarceration

Wildeman et al. (2012) estimates how incarceration of a child’s father has collateral consequences across the family by increasing the probability that the child’s mother will be depressed. Using a design that adjusts for observed sources of confounding, the authors report that paternal incarceration increases the log odds of maternal depression by 0.32. Our replication recovers a similar estimate of 0.28 (details in Appendix A.6).⁶ We convert the model into an estimate of a nonparametric estimand: paternal incarceration increases the probability of maternal depression by 4 percentage points (95% CI: -0.02, 0.10) on average. This illustrates a first advantage of an approach centered on estimands: it becomes clear that the effect size is small, although important nonetheless. The original authors proceed to a series of mediation claims, which we use to make broader points about causal mediation.

2.3.3.2 Direct Effects: Defining a Mediation Estimand

We focus on one mediator from Wildeman et al. (2012): paternal incarceration may cause the mother to reside with a new partner, which could in turn affect her depression. Claims about this mechanism would invoke counterfactuals for both the treatment (what if the father had not been incarcerated?) and the mediator (what if the mother had not repartnered?). The potential outcomes, $Y_i(d, m)$, are thus functions of both the treatment value d and the mediator value m . This definition of potential outcomes allows one to target many mediation estimands defined as contrasts over the outcomes that would realized at different values of d and m (Imai et al., 2011; Pearl, 2001). We focus on one particular set of mediation estimands—controlled direct effects—which can be estimated under more credible assumptions

⁶The version of the underlying data currently available is different from that used by the original authors, and complete replication code was unavailable. The original estimate was statistically significant at the .05 level, whereas our 95% confidence interval (-0.14 to 0.74) contains zero.

than other mediation estimands because they can be identified even in the presence of treatment-induced mediator-outcome confounding (Acharya et al., 2016). A controlled direct effect $\tau(m)$ compares the outcome under two different treatment values that would persist if we intervened to fix the treatment to a particular value m .

$$\underbrace{\tau(m)}_{\text{Controlled direct effect}} = \underbrace{\frac{1}{n} \sum_{i=1}^n}_{\text{Average over sample}} \left(\underbrace{Y_i(1, m)}_{\substack{\text{Mother } i\text{'s depression} \\ \text{if father were incarcerated} \\ \text{and maternal repartnering} \\ \text{was set to value } m}} - \underbrace{Y_i(0, m)}_{\substack{\text{Mother } i\text{'s depression} \\ \text{if father were } \mathbf{not} \text{ incarcerated} \\ \text{and maternal repartnering} \\ \text{was set to value } m}} \right) \quad (2.3.5)$$

For instance, what would be the effect of paternal incarceration on maternal depression if the mother did not repartner ($m = 0$)? What if she did repartner ($m = 1$)? The controlled direct effects $\tau(0)$ and $\tau(1)$ are different estimands with different true values: the “direct effect” is undefined until the value m is stated. This section shows that estimates of these two estimands can be remarkably different.

2.3.3.3 Identifying Direct Effects: The Threat of Mediator-Outcome Confounding

Because mediation invokes counterfactual assignments of both the treatment and the mediator, it is necessary to adjust for variables that confound the assignment of both of these variables. Sociologists frequently discuss confounding of the treatment but almost never discuss confounding of the mediator. Wildeman et al. (2012, p. 222) “adjust for preexisting differences between mothers who have and have not experienced the incarceration of their child’s father,” but they do not say anything to address concerns that an unobserved variable U might affect the mediator (maternal repartnering) and the outcome (maternal depression, Fig. 2.9). This threat persists even in randomized experiments where the treatment (but not the mediator) is randomized (see Appendix Table A.6 for examples from *ASR*). The assumption of no treatment-outcome or mediator-outcome confounding is often doubtful. We will

nonetheless proceed under this assumption in order to further illustrate complexities of mediation that arise (at least implicitly) in many studies.

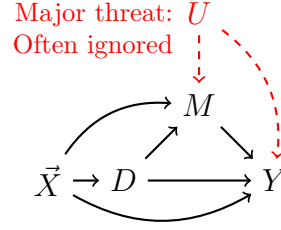


Figure 2.9: Causal structure for mediation estimands. These require identifying the effect of the treatment D and the effect of the mediator M on the outcome Y . The unobserved mediator-outcome confounder, U , is a threat to inference.

2.3.3.4 Estimating Direct Effects

Due to our identification assumptions, direct effects can be estimated by the imputation estimator (Figure 2.5): fit a statistical model for Y as a function of $\{\vec{X}, D, M\}$, plug in the new values d and m for the variables D and M , predict $\hat{Y}_i(d, m)$ for each unit i , and average over units as specified by the estimand.

$$\underbrace{\hat{\tau}(m)}_{\text{Estimated controlled direct effect}} = \underbrace{\frac{1}{n} \sum_i}_{\text{Average over sample}} \left(\underbrace{\hat{Y}_i(1, m)}_{\substack{\text{Predicted probability of} \\ \text{depression for mother } i \\ \text{if father were incarcerated and} \\ \text{repartnering was set to value } m}} - \underbrace{\hat{Y}_i(0, m)}_{\substack{\text{Predicted probability of} \\ \text{depression for mother } i \\ \text{if father were not incarcerated and} \\ \text{repartnering was set to value } m}} \right) \quad (2.3.6)$$

We predict the potential outcomes using logistic regression specified similarly to the original authors, but we add an interaction between the treatment and the mediator.

2.3.3.5 Results

The four plots in the upper left of Fig. 2.10 correspond to average potential outcomes: the proportion of mothers who would be depressed under each possible value of the treatment (father incarceration) and the mediator (mother having a new partner). The direct effects (right column) are the difference in the average potential outcomes

across treatment conditions within a mediator condition. The direct effect to which this estimand corresponds is subtle. Paternal incarceration would *reduce* maternal depression by 1 percentage point under a subsequent intervention to repartner any mother who would not otherwise repartner. This is a direct effect because the intervention would remove the causal pathway through repartnering. Meanwhile, paternal incarceration would *increase* maternal depression by 5 percentage points under a subsequent intervention to prevent any mother from repartnering (middle right). The “direct effect” is an ambiguous estimand until we state the value to which the mediator is fixed.

2.3.3.6 Mediation: Concluding Remarks

Mediation claims invoke a chain of causal effects from the treatment to the mediator to the outcome. The required assumptions are more stringent than those required for causal effects because the effect of the mediator must be identified. A precise statement of the unit-specific quantity—a causal contrast between the outcomes realized under two treatment conditions in a world where the mediator was set to some value—both clarifies the goal and the required assumptions.

If there are many mediators, then the question is even more complex. In Table 4 Model 5, Wildeman et al. (2012) “consider all the mechanisms simultaneously, and the relationship is reduced by approximately half,” (Wildeman et al. 2012:233). A precise version of this claim would require defining the potential outcomes as functions of the treatment and all 11 mediators, stating the values to which these mediators are fixed, and arguing that all 11 mediators are unconfounded. Such an argument would be extremely difficult. Mediation is one setting where we worry that reasoning about the research goal in terms of coefficients has led scholars to a false sense of simplicity about the target of inference.

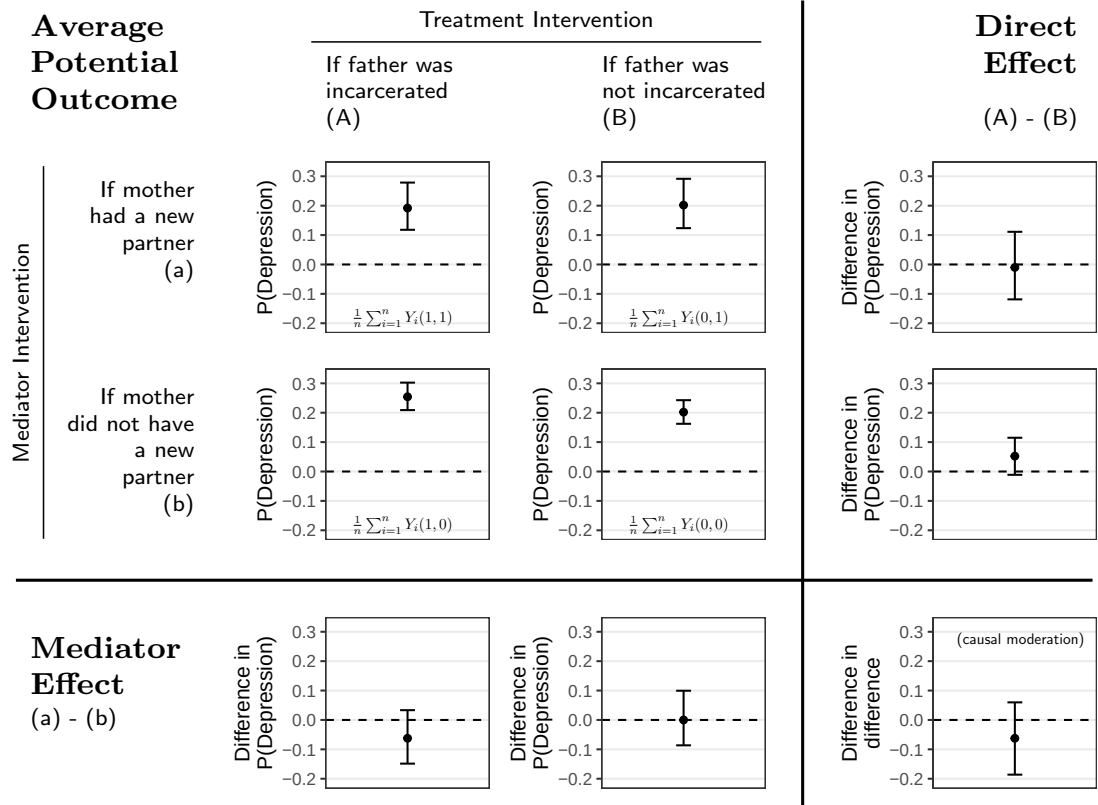


Figure 2.10: Controlled direct effects involve an intervention to the treatment and the mediator. Figure explores the degree to which the effect of paternal incarceration on maternal depression operates through the mother residing with a new romantic partner. In a possibly counterfactual world in which a mother had a new partner (top row), paternal incarceration would reduce her probability of depression. In a possibly counterfactual world in which a mother did not have a new partner (middle row), then paternal incarceration would increase her probability of depression. Implicit in these claims is that we can identify the effect of the mediator (bottom row) that would exist under each intervention to paternal incarceration. To estimate, we fit a logistic regression model, predicted the potential outcomes for each mother, and averaged over the sample. Estimates are the mean and the 0.025 and 0.975 quantiles of 10,000 simulated draws calculated by 100 likelihood-based samples from parameters estimated in each of 100 multiply-imputed datasets.

2.4 Discussion: Estimands are Useful to Analysts, Readers, and the Community

Estimands are stepping stones between theory and evidence: they clarify the component of a theory at stake and provide a clear purpose for each statistical analysis. We advocate a three step research process which involves (1) choosing one or more theoretical estimands, (2) choosing empirical estimands which are informative about the theoretical estimands under a set of identification assumptions, and (3) choosing estimation strategies to learn the empirical estimands. We argue that following this three-step process will clarify the goals of sociological research and clearly delineate which parts of the argument are conceptual and which are empirical.

So what is your estimand? This should be a default question for those who produce, consume, and evaluate quantitative research. If you do not answer this question, you have missed an opportunity to clarify your contribution to knowledge. Much of existing quantitative sociology provides an inadequate answer. Instead, authors define the research goal as the result of a statistical procedure, as when hypotheses are made about regression coefficients. Stating the research goal within the model leaves substantial ambiguity: what goal outside the model was the target of inquiry? Our review of the 2018 volume of *ASR* reveals that we often cannot reverse-engineer the estimand from published papers. Our examples show that underspecified estimands can lead to deeply misleading conclusions. Clear statements of the estimands can address these issues, improve how authors make methodological choices, allow readers to engage meaningfully with the author's claims, and provide a basis for the research community to accumulate knowledge. A precise research goal can also bring us back to what we all wanted to do: begin from a theoretically-motivated question and let all methodological choices follow from the aim of producing a credible answer. Bring-

ing methodological choices under the umbrella of estimands yields benefits for the analyst, the reader, and the broader community.

2.4.1 For Analysts: Estimands Ground Methodological Choices

For the analyst, the estimand (step 1) guides all subsequent methodological choices about identification (step 2) and estimation (step 3). Without an estimand clarifying the objective of the research, it is impossible to answer methodological questions such as ‘what variables should I include?’ (Raftery, 1995), ‘should I report a predicted probability?’ (Breen et al., 2018), or ‘should I use fixed or random effects?’ (Firebaugh et al., 2013). Answers to all of these first require that we answer the essential question: what is the estimand?

For example, researchers estimating binary outcome models are often confused about whether interaction terms are needed and how to interpret them if so. The extensive methodological debate on this topic frames these issues within the terms of logit and probit models (Berry et al., 2010; Nagler, 1991; Rainey, 2016). Yet the problem is more fundamental: the researcher must begin by stating what they mean by the term “interaction.” Fig. 2.11 illustrates that a treatment may multiply the probability of an outcome by a fixed amount (no interaction) while increasing that probability by an additive amount that depends on a pre-treatment covariate (interaction). Whether or not interaction is present depends on the estimand. No argument about a model can resolve this question; it is fundamentally a question about the research goal itself.

Estimands likewise offer guidance about which variables to include in a model. Mood (2010:67) warns that problems arise in logistic regression because “we can seldom include in a model all variables that affect an outcome.” Breen et al. (2018:47) write that logit coefficient estimates “are lower bounds to the true or underlying co-

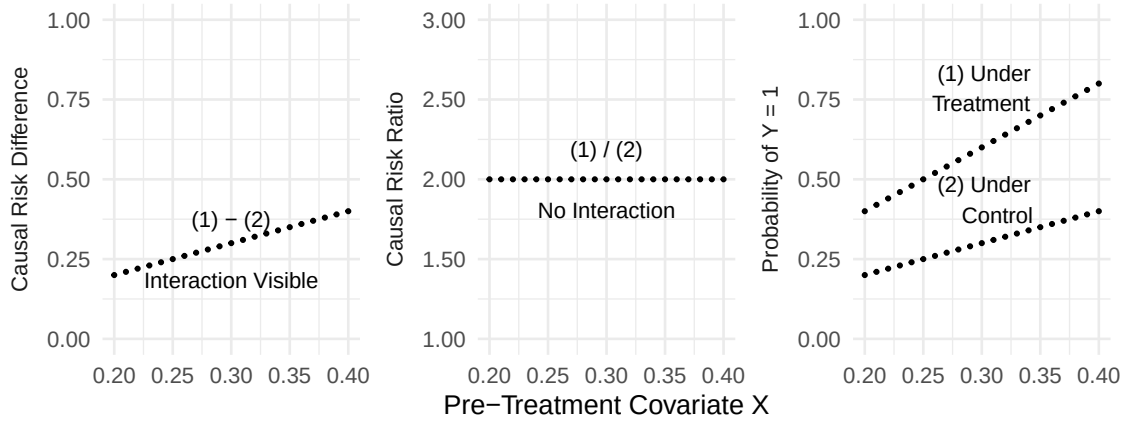


Figure 2.11: Illustration: The presence or absence of interaction may depend on whether the estimand is a multiplicative or an additive effect. In this example, the treatment multiplies the probability of $Y = 1$ by a constant value (2.00) at all values of X . If the baseline outcome is a function of X , however, then a constant multiplicative effect implies an additive effect that is a function of X : the probability of $Y = 1$ in this example increases by a greater amount when X is greater. Whether interaction is present depends on the estimand. No amount of literature about the proper interpretation of coefficients in binary outcome models can help a researcher as long as that researcher’s goal of assessing the presence or absence of interaction remains underspecified. Appendix A.7 provides simulation details.

efficients unless all relevant covariates are included.” But are these “true” coefficients even a well-defined estimand? What would it mean to include all relevant covariates? Our framework instead offers a straightforward answer: you must include the covariates needed to identify the estimand. For descriptive estimands, that might only be the predictors of interest.

Presentation of results can also be guided by estimands. If the estimand is a difference in probabilities, then the researcher would naturally present predicted probabilities rather than reporting coefficient estimates (Breen et al., 2018; Mize, 2019; Mood, 2010). Questions about which model to use, what assumptions are required, and how to present results all become clearer once the research goal has been stated.

2.4.2 For Readers: Estimands Clarify the Author’s Claim

Readers can only evaluate a paper when they clearly understand the claim it is making. No quantitative paper should leave the reader wondering ‘is the claim causal or descriptive?’, ‘to whom does it apply?’, or ‘what assumptions are needed to believe the results?’ Readers would gain clear answers to these questions if every paper provided a straightforward answer to our guiding question: what is the estimand?

Readers often like to see that results are ‘robust’. In the 2018 volume of the *American Sociological Review*, at least 20 papers reported a robustness check (Appendix A.1). Our framework asks: to what do we want results to be robust? Some forms of robustness focus on the theoretical estimand (e.g. a different outcome), others focus on the identification strategy (e.g. a different set of variables to control for), and still others focus on the estimation strategy (e.g. a different functional form like logit versus linear regression). These forms of robustness are very different. Robustness across unit-specific quantities and target populations may provide useful context for our theoretical understanding. Robustness across conditioning sets only matters for those sets which credibly identify the causal effect. Robustness across estimation strategies is only important among those methods which are comparably accurate.

In general, robustness checks provide useful information only in the context of a well-defined target and clarity about the alternatives to which we are evaluating robustness. In contrast, robustness checks as currently applied treat all specifications as equally valid. Young and Holsteen (2017) provides tools to automate this procedure. Yet when an unguided search for robustness is taken to its logical extreme with thousands of specifications, it is impossible to defend each individual specification. The resulting benchmark for methodological rigor devolves into a requirement that sociologists report only the results that survive a test of methodological invariance: they are the same even if we target several different estimands through several

different estimation strategies. This is not a requirement of a credible claim. Conversations about robustness would be more productive if they centered on how each check resolves a specific point of uncertainty in the link between theory and evidence.

2.4.3 For the Community: Estimands Partition the Role of Evidence in Social Science

For the community, clarity about estimands illuminates how studies relate to each other. If the field aspires build cumulative knowledge, a critical first step is to achieve clarity about our key question: what is the estimand in study A, and how does it relate to the estimands in studies B and C?

Distinguishing between differences in estimation strategies versus differences in estimands is essential in the case of replication. Questions of replication often focus on questions about statistical power and hypothesis testing. Yet a replication can also fail because it targets a different estimand from the original study, as when the replication uses a different pool of experimental subjects.⁷ A replication focused on the same estimand as the original study provides evidence about statistical issues like false positives. A replication focused on a different estimand provides evidence about theoretical issues regarding the generality of the phenomenon across settings. Specificity about the estimand of each paper is key to the advancement of general theories of social life that produce results that replicate across many studies in many distinct settings. For the field as a whole to grow, it would help if each paper's contribution to knowledge is stated precisely.

⁷Freese and Peterson (2017) also discuss replications that vary in their degree of similarity to the original study. In our terms, a replication may investigate the same estimand, or it may investigate whether two related estimands (such as the same quantity in slightly different populations) yield similar results. Apparent failure to replicate can stem from statistical anomalies or from differences between the original estimand and the replication estimand.

Certain communities may lack the theoretical closure necessary to agree that a given set of estimands can inform a multifaceted theory. Readers may fear that following our framework will lead them to get stuck in debates with colleagues and reviewers about the most appropriate estimand. In our view, this is exactly how the community makes progress—focusing on what quantities are most important to theory rather than talking past each other about methodological choices most appropriate for studying different things.

2.4.4 Concluding remarks: Estimands prepare us for the future of quantitative sociology

A renewed focus on estimands will be important as sociology navigates a methodological landscape that is changing rapidly. A pivot to new sources of ‘big data’ creates an ever-greater need for clarity about the gap between the theoretical goal and selection issues that constrain the data available (as in Section 2.2.2). As sociologists increasingly engage with predictive tasks (Salganik et al., 2020), estimands will clarify key distinctions among different types of prediction: for cases from the same data generating process as the training data (standard prediction), for the outcome that would be realized under an intervention (causal prediction), or for future events that have not yet occurred (forecasting). Each setting corresponds to a different theoretical estimand and requires a different set of identification assumptions. Estimands can also improve the use of inductive measurement strategies such as latent class analysis for surveys, methods for text as data, and unsupervised machine learning. Explicit statements of the estimands in these settings would both clarify what these procedures are learning from data and what evidence would be necessary to contradict the finding. They also highlight the importance of choosing an estimand of interest in one sample and then evaluating the estimand on a second sample (Egami et al., 2018).

Broadly, estimands will be key to whatever methods quantitative sociology develops in the coming years.

Important questions often require a leap from the empirical evidence to the theoretical claim. Sociology stands out from other social sciences for its willingness to tackle hard questions even when they require such a leap; however, burying the estimand obscures those decisions and confuses the link between theory and evidence. At best, this creates an uncomfortable ambiguity about the author's intentions. At worst, it can mislead. Rather than a call for sociologists to narrow their ambitions, our framework is a call for sociologists to be explicit about the goals that motivate their projects and transparent about the assumptions needed to believe them. A paper that develops a compelling theoretical estimand but relies on less-than-perfect identification assumptions should be recognized for making an important contribution: it sets the stage for future work to explore that theoretical estimand under different identification assumptions. While it may be simpler, obfuscation of the true goal does not make an argument more compelling. If we want to make progress on big theoretical questions, we should begin every quantitative analysis with a question that makes its purpose precise: what is the estimand?

Chapter 3

The gap-closing estimand: A causal approach to study interventions that close disparities across social categories¹

Chapter 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 produce results that may directly inform policy: they tell us the degree to which an intervention applied to a sample would close a gap. I provide open-source software (the R package `gapclosing`) to support these methods.

¹A draft is available on SocArxiv: doi.org/10.31235/osf.io/gx4y3. The methods presented in this paper are implemented in the R package `gapclosing`; a development version is available at ilundberg.github.io/gapclosing/. Replication code is available at github.com/ilundberg/replication and will be posted on Dataverse if the paper is accepted for publication. This manuscript was presented at the 2020 Annual Meeting of the Population Association of America. Address correspondence to Ian Lundberg, ilundberg@princeton.edu, 227 Wallace Hall, Princeton University, Princeton, NJ 08544. I thank Brandon Stewart, Matthew Salganik, Xiang Zhou, Felix Elwert, Janet Xu, Daniela Urbina, Belén Unzueta, and Gillian Slee for comments on drafts and Hannah Waight, Rebecca Johnson, Simone Zhang, and the Stewart Lab for helpful discussions. Research reported in this publication was supported by The Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD047879.

3.1 Introduction

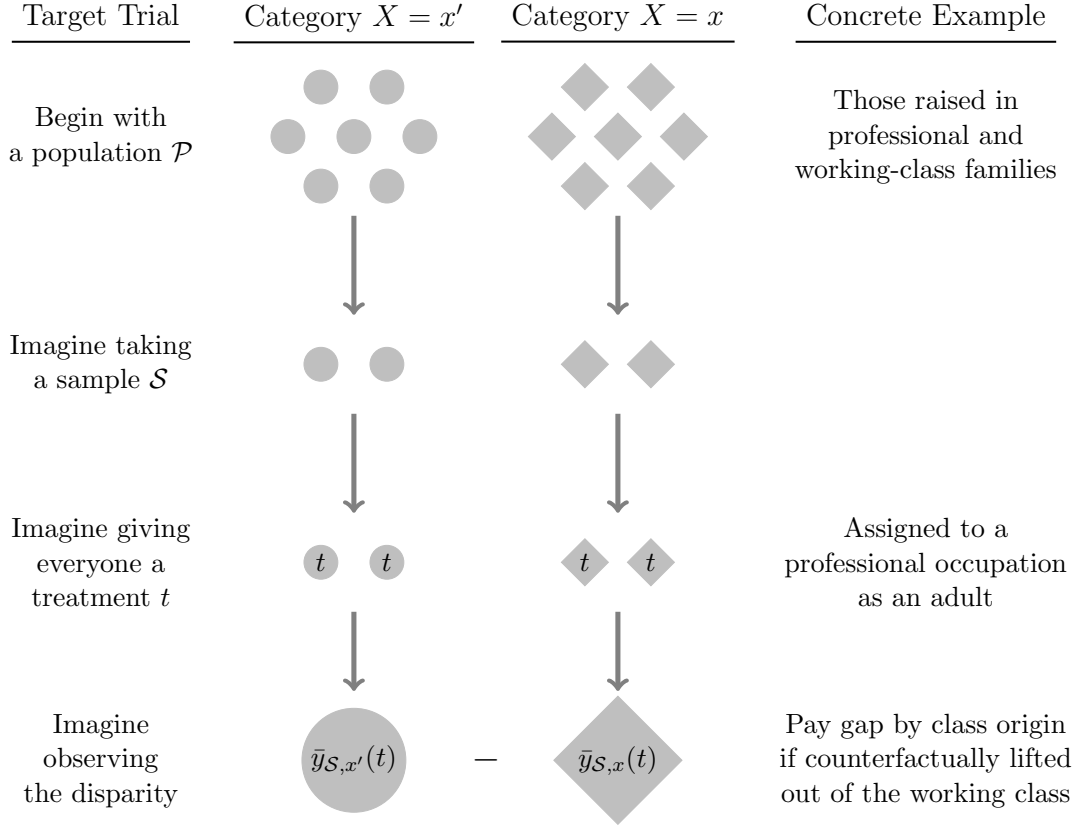
Gaps in socioeconomic outcomes are among the strongest indictments of inequality. For example, the net worth of the median black household in America in 2016 was 10 cents for every dollar held by the median white household (McIntosh et al., 2020). Among those employed full-time in 2014, the median woman 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 question: 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. For other treatments (e.g. access to college), a gap-closing estimand also speaks to policy. We can assess whether an expansion of college would change income mobility.

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. college). The gap-closing es-

timand 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 3.1). 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). 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. Statistical and machine learning estimators can yield credible evidence without formal models of interference; these estimators are implemented in the publicly-available and open-source R package `gapclosing` (available at [redacted for peer review]). 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: would this person have been hired if they had been white? Audit studies assess discrimination by experimentally manipulating signals of a category like race while holding all else constant (Bertrand and Mullainathan, 2004; Pager, 2003). Doing so identifies the causal effect of one signal of that category, such as the name at the top of one’s resume (Greiner and Rubin, 2011; Kaufman, 2008; 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 (Kohler-Hausmann, 2018; Omi and Winant, 1994; Sen and Wasow, 2016). Doing so is worthwhile: causal evidence of discrimination is extremely important. Likewise, descriptive research about disparities is extremely important. Gap-closing estimands



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

Figure 3.1: 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 et al. (2016a).

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 (Jackson and VanderWeele, 2018; VanderWeele and Robinson, 2014) 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 3.2 argues that gap-closing estimands formalize a widespread problem. Section 3.3 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 3.4 presents causal assumptions. Section 3.5 presents estimation strategies. Section 3.6 illustrates the method by extending previous work that described gaps in pay by class origins (Laurison and Friedman, 2016). To make the paper accessible to a broad audience, I defer a precise statement of the relationship between the current contribution and existing methods to Section 3.7. Section 3.8 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.

3.2 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 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 method provides a method to do so.

3.3 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.

For either fixed or stochastic treatment assignments, causal questions can often be made more precise through an appeal to a target trial: a hypothetical experiment that may be infeasible but which clarifies the goal of a study (Hernán et al., 2016a). This is also the case for a gap-closing estimand (Fig 3.1). 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 $\mathcal{S}$.

$$\begin{array}{c}
 \text{Expectation over} \\
 \text{hypothetical samples } \mathcal{S} \\
 \text{from the population } \mathcal{P} \\
 \downarrow \\
 \text{Gap-closing estimand: } \tau_{x',x}(t) \equiv \mathbf{E}_{\mathcal{S}} \left(\bar{y}_{\mathcal{S},x'}(t) - \bar{y}_{\mathcal{S},x}(t) \right) \\
 \begin{array}{cc}
 \nearrow & \nwarrow \\
 \text{Mean outcome} & \text{Mean outcome} \\
 \text{in category } x' & \text{in category } x \\
 \text{if treatment were set to the value } t & \\
 \text{for units in sample } \mathcal{S} &
 \end{array}
 \end{array} \tag{3.3.1}$$

Fig 3.1 presents an example. Suppose we begin with those raised in professional and working-class families, defined by the occupation of one's father figure ($X = x'$ and $X = x$). Suppose 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.

The gap-closing estimand extends to stochastic treatment assignment rules. Perhaps instead of assigning people to a professional class destination, we imagine assigning people to a random draw from the population distribution of class destinations, so that some are assigned to professional and some to working-class destinations. 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 $\{\pi(t, \vec{\ell})\}_{t \in \mathcal{T}}$. 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 potential outcomes $\{y_i(t)\}_{t \in \mathcal{T}}$ that unit would realize under each treatment value t , with each outcome weighted by the probability $\pi(t, \vec{L}_i)$ of assigning that treatment value.

$$\begin{array}{ccc}
& \text{Sum over all} \\
& \text{treatment values } t \\
& \downarrow \\
\text{Expected outcome} & \bar{y}_i(\pi) \equiv \sum_t \pi(t, \vec{L}_i) y_i(t) & \\
\text{for unit } i \text{ under a} & & \\
\text{stochastic intervention:} & \begin{array}{cc} \nearrow & \nwarrow \\ \text{Counterfactual probability} & \text{Potential outcome} \\ \text{of assigning treatment} & \text{for unit } i \text{ under} \\ \text{value } t \text{ to unit } i & \text{treatment value } t \end{array} & \\
& & (3.3.2)
\end{array}$$

Given the unit-specific expected outcome $\bar{y}_i(\pi)$, let $\bar{y}_{\mathcal{S},x}(\pi)$ 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. 3.3.1 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) as a randomized interventional analog.

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 $\mathcal{P}$. 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 the treatment $T = t$.

3.3.1 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 3.2). 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. Yet that world is very far from the available data. 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.

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 (Lucas, 2001; Raftery and Hout, 1993) 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. While speculation about a

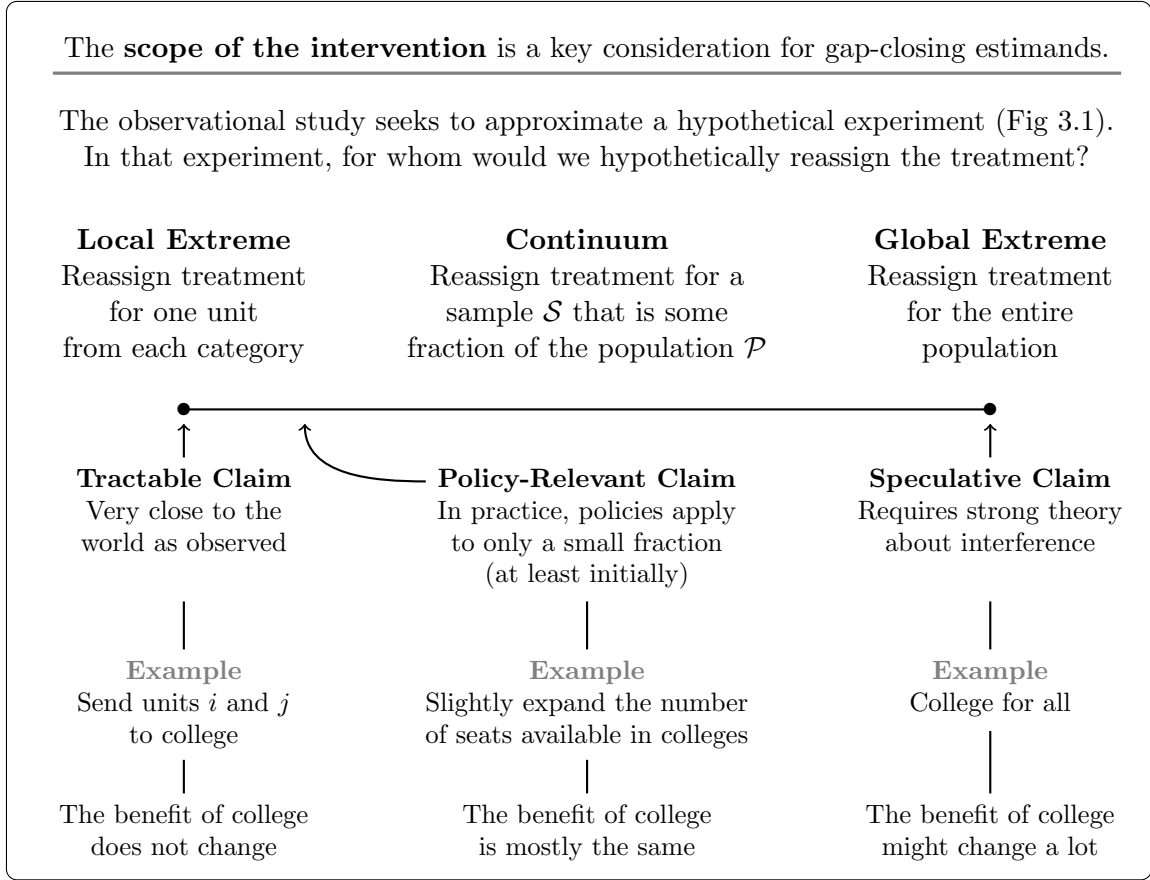


Figure 3.2: Clarify the scope of the intervention. A gap-closing estimand invokes a counterfactual world where treatment is reassigned for a sample $\mathcal{S}$ from a population $\mathcal{P}$ (Fig 3.1). Inference about a world where treatment is reassigned for the entire population ($\mathcal{S} = \mathcal{P}$) 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.

population-level intervention can be suggestive, it must be paired with transparency about the limited empirical base for these sweeping claims.

The difficulty of global claims hinges on an assumption which is often overlooked: the Stable Unit Treatment Value Assumption (SUTVA). That assumption requires that each unit’s outcome depends only on the treatment to which that unit is assigned, and not on the treatment assignments of other units (no interference). SUTVA 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 SUTVA in a setting where it is particularly imperfect.

Researchers concerned with global claims have at least three options. One option 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 value. 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 3.2 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.

3.3.2 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 allowable covariates that might justly determine treatment assignment (e.g. education) and non-allowable covariates over which unequal allocation of treatments might be considered discriminatory (e.g. race, sex). In some settings, one may be able to avoid implausible 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 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.

3.4 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). For this reason, strategies to learn gap-closing estimands are the same as those for estimating causal effects and rely on untestable identification assumptions that must be defended on conceptual grounds: 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 3.3). 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 B.1 provides more details about these assumptions and proves that they are sufficient for identification.

Fig. 3.3 illustrates key issues using Directed Acyclic Graphs (DAGs, Pearl 2000).² 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

²The DAGs in Fig 3.3 invokes 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.

variables, so that covariates $\vec{L}$ are consequences of race (Panel A, see Sen and Wassow 2016). Race may be understood to have no causal effects (Panel B, see Holland 2008; Rubin 1986). 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.

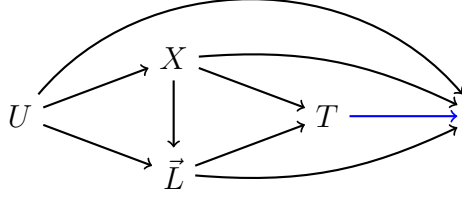
The settings that threaten causal identification for gap-closing estimands are the same settings that threaten identification for the causal effect of the treatment T . If a backdoor path cannot be blocked due to an unobserved variable V that affects both the treatment and the outcome, then the gap-closing 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 $T \leftarrow V \rightarrow \boxed{\vec{L}} \leftarrow U \rightarrow Y$ (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 for identification. Making the research goal explicitly causal creates new opportunities for clear reasoning about these assumptions.

3.5 Estimation: Learn from data

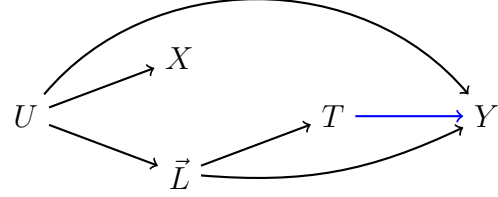
The central hurdle for estimation is that some potential outcomes $y_i(t)$ are unobserved. If the identification assumptions hold, we can learn the average of those potential outcomes for all units by examining the units we actually observe with treatment value $T_i = t$, within each subpopulation with a particular category value

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

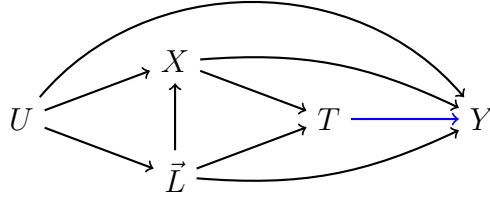
A) X can have causal effects



B) X can have no causal effects



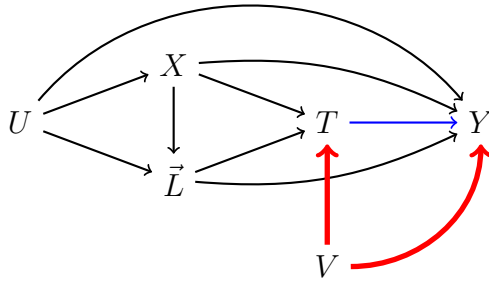
C) X can be a consequence of $\vec{L}$



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.

D) Classic confounding



E) M -bias (Greenland et al., 1999)

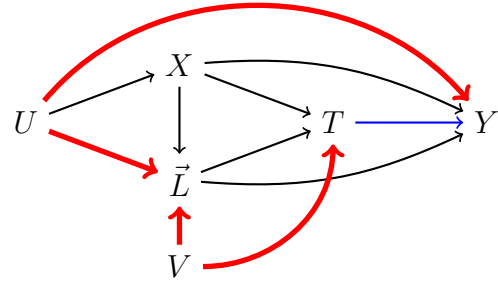


Figure 3.3: Identification of gap-closing estimands. Observed variables include the social category X (e.g. race, class, gender), the manipulable treatment variable T (e.g. college completion, occupational attainment), and other pre-treatment covariates $\vec{L}$. Nodes U and V are unobserved. The blue $T \rightarrow Y$ edge represents the causal effect that must be identified. The 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 B.1).

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.

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

$$\begin{array}{ccc} \tau_{x',x}(\pi) \equiv \mathbf{E}_{\mathcal{S}}(\bar{y}_{\mathcal{S},x'}(\pi)) - \mathbf{E}_{\mathcal{S}}(\bar{y}_{\mathcal{S},x}(\pi)) & & \\ \nearrow & & \nwarrow \\ \theta_{x'}(\pi) & & \theta_x(\pi) \end{array} \quad (3.5.1)$$

Post-intervention mean Post-intervention mean
in category x' in category x
if treatment were assigned by the rule π

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. This section ultimately describes a sample splitting procedure which can be paired with doubly robust estimation to make machine learning tools useful for estimating gap-closing estimands. The end of the section briefly discusses standard error estimation by resampling strategies.

3.5.1 Estimation by predicted outcomes

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

$$\text{Outcome prediction function:} \quad g(t, x, \vec{\ell}) = \mathbf{E}(Y \mid T = t, X = x, \vec{L} = \vec{\ell}) \quad (3.5.2)$$

The function $g()$ to predict outcomes can be 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 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,

$$\hat{\theta}_x^{\text{Outcome}}(\pi) = \frac{1}{\eta_x} \sum_{i: X_i=x} \eta_i \sum_t \hat{\pi}(t, \vec{L}_i) \hat{g}(t, X_i, \vec{L}_i) \quad (3.5.3)$$

↑

Estimate from
predicted outcomes

↗

Sum over all
sampled units i
in category x

↘

Sum over all
treatment
values t

↖ ↗

Category and
covariates as
observed

↘

Sample
weight

↓

Counterfactual
probability of
treatment t

↙

Predicted
outcome under
treatment t

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 estimator of the gap-closing estimand.

3.5.2 Estimation by a 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 $m(\cdot)$ to predict the probability of treatment, which could be estimated by any statistical or machine learning algorithm (e.g. logistic regression).

$$\text{Treatment prediction function:} \quad m(t, x, \vec{\ell}) = P(T = t \mid X = x, \vec{L} = \vec{\ell}) \quad (3.5.4)$$

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 a 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 involves 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,

$$\begin{array}{c}
\hat{\theta}_x^{\text{Treatment}}(\pi) = \frac{1}{\sum_{i: X_i=x} \hat{w}_i} \sum_{i: X_i=x} \hat{w}_i Y_i, \quad \text{where } \hat{w}_i = \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)} \\
\begin{array}{ccccccc}
& & & \text{Sampling} & \text{Counterfactual probability} & & \\
& & & \text{weight} & \text{of the observed treatment} & & \\
& & & \searrow & \swarrow & & \\
& \uparrow & & \nearrow & \nwarrow & \uparrow & \\
\text{Estimate from} & \text{Reweighted to the} & \text{Outcome in the} & \text{Factual probability} & & & \\
\text{predicted probabilities} & \text{target population} & \text{observed treatment} & \text{of the observed treatment} & & & \\
\text{of treatment} & & & & & &
\end{array}
\end{array} \tag{3.5.5}$$

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.

3.5.3 Doubly robust estimation

Doubly robust estimation combines predicted outcomes and treatment probabilities to produce an estimator that is consistent if either the outcome prediction function $\hat{g}()$ is consistent for the true conditional mean function $g()$ or if the estimated treatment prediction function $\hat{m}()$ is consistent for the true conditional probability of treatment $m()$ (Fig 3.4, see Bang and Robins 2005; Glynn and Quinn 2010; Robins et al. 1994). 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 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,

$$\begin{array}{c}
 \hat{\theta}_x^{\text{Robust}}(\pi) = \hat{\theta}_x^{\text{Outcome}}(\pi) - \underbrace{\frac{1}{\sum_{i: X_i=x} \hat{w}_i} \sum_{i: X_i=x} \hat{w}_i \left(\hat{g}(T_i, X_i, \vec{L}_i) - Y_i \right)}_{\text{Estimated bias of the predicted outcomes reweighted to the target population}} \\
 \begin{array}{ccc}
 \uparrow & \uparrow & \begin{array}{l} \text{Inverse probability} \\ \text{weight} \end{array} \quad \begin{array}{l} \text{Predicted} \\ \text{outcome} \end{array} \quad \begin{array}{l} \text{Observed} \\ \text{outcome} \end{array} \\
 \text{Doubly-robust} & \text{Estimate from} & \\
 \text{estimate} & \text{predicted outcomes} &
 \end{array}
 \end{array}
 \tag{3.5.6}$$

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.⁴ There are two important caveats. First, if both prediction functions are inconsistent, there is no guarantee that a doubly-robust estimator will outperform 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

³Doubly robust estimators are a general class with many specific types. Eq. 3.5.6 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 3.5.4 of the present paper). Certain types of doubly robust estimators may be preferable in certain settings; this is an open topic for future work.

⁴By 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 $\hat{m}()$.

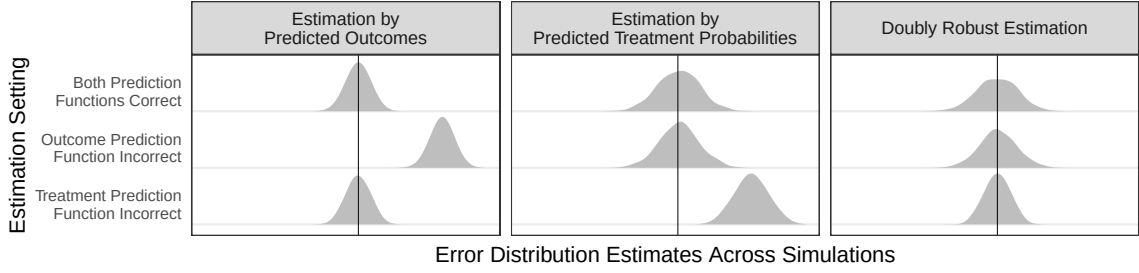


Figure 3.4: Doubly-robust estimation helps when one of two prediction functions is correct. The data in this example are simulated, so the truth is known. The vertical line indicates the true gap-closing estimand $\tau_{1,0}(1)$ across binary categories $X = 1$ and $X = 0$ under an intervention set treatment 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. 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 B.4.1.

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 (e.g. assign proportional to the unknown population prevalence of each treatment). The gap-closing estimator will only be consistent if $\hat{\pi}(t, \vec{\ell})$ is consistent for $\pi(t, \vec{\ell})$, the counterfactual rule about which the researcher wants to learn. There are two common settings in which this is not a problem. First, it is not a problem when the counterfactual rule $\pi()$ is known a priori, as when a researcher is interested in giving all units the treatment. Second, it is not a problem for relatively simple counterfactual rules $\pi()$ which can be learned nonparametrically (i.e. 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.

3.5.4 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 (Bickel, 1982; Chernozhukov et al., 2018).

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 Hill 2011; Lee et al. 2010; McCaffrey et al. 2004; 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. Because of the flexibility of machine learning estimators, overfitting is also a potential concern. For

this reason, cross fitting is important when implementing doubly robust estimation. Double robustness with cross fitting is a key that improves the usefulness of machine learning for estimation in settings that aggregate across individuals (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 3.5 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 B.4.3).

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*` 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; Lumley, 2011; Rao and Wu, 1988; Rust and Rao, 1996). Appendix B.3 presents an example. Variance estimates can be converted into confidence intervals by a normal approximation to the sampling distribution.

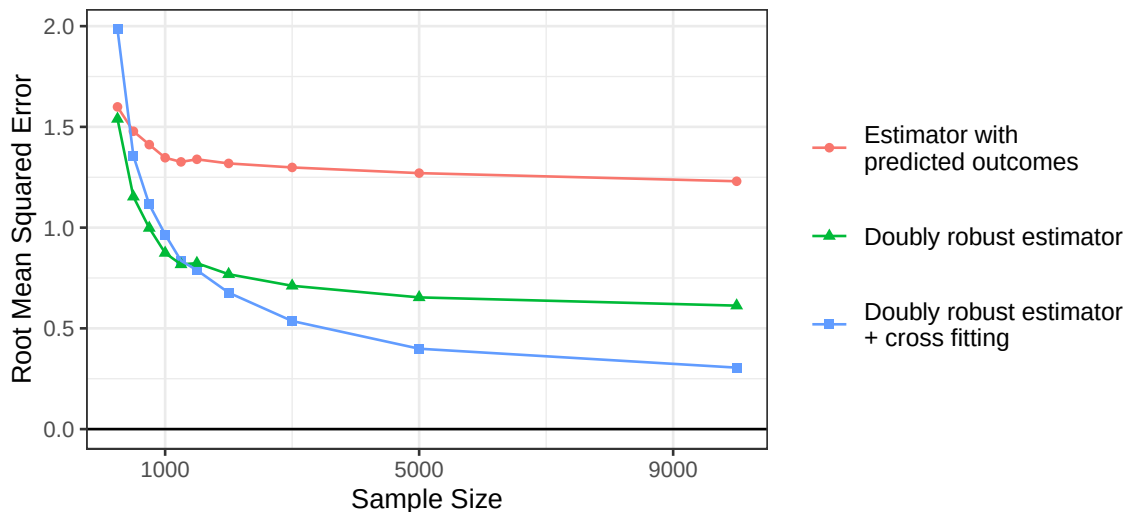


Figure 3.5: Cross fitting can improve convergence rates. The data in this example are simulated, so the truth is known. Potential outcomes are a linear function of a binary treatment and 10 continuous confounders in this simulation. In the true data generating process, there is no disparity across categories $X = 1$ 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 B.4.3.

3.6 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, 10.1% 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 B.5.

3.6.1 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. 3.4, 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.

3.6.2 Estimation

I estimate the outcome function g and the treatment propensity function m by ordinary least squares and logistic regression, respectively (Appendix Table B.2). I ag-

gregate these functions to an estimate by the doubly robust estimator (Eq. 3.5.6). 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 B.3 presents the variance estimation procedure. Results with alternative estimators are reported in Appendix Fig. B.5.

3.6.3 Results

Fig. 3.6 presents results. Descriptively, log incomes are 0.33 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 B.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 0.29 (86% 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 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 moreso among those of working-class origins, which would downwardly bias the gap-closing estimand (discussed in greater depth in Appendix B.5.5). 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.

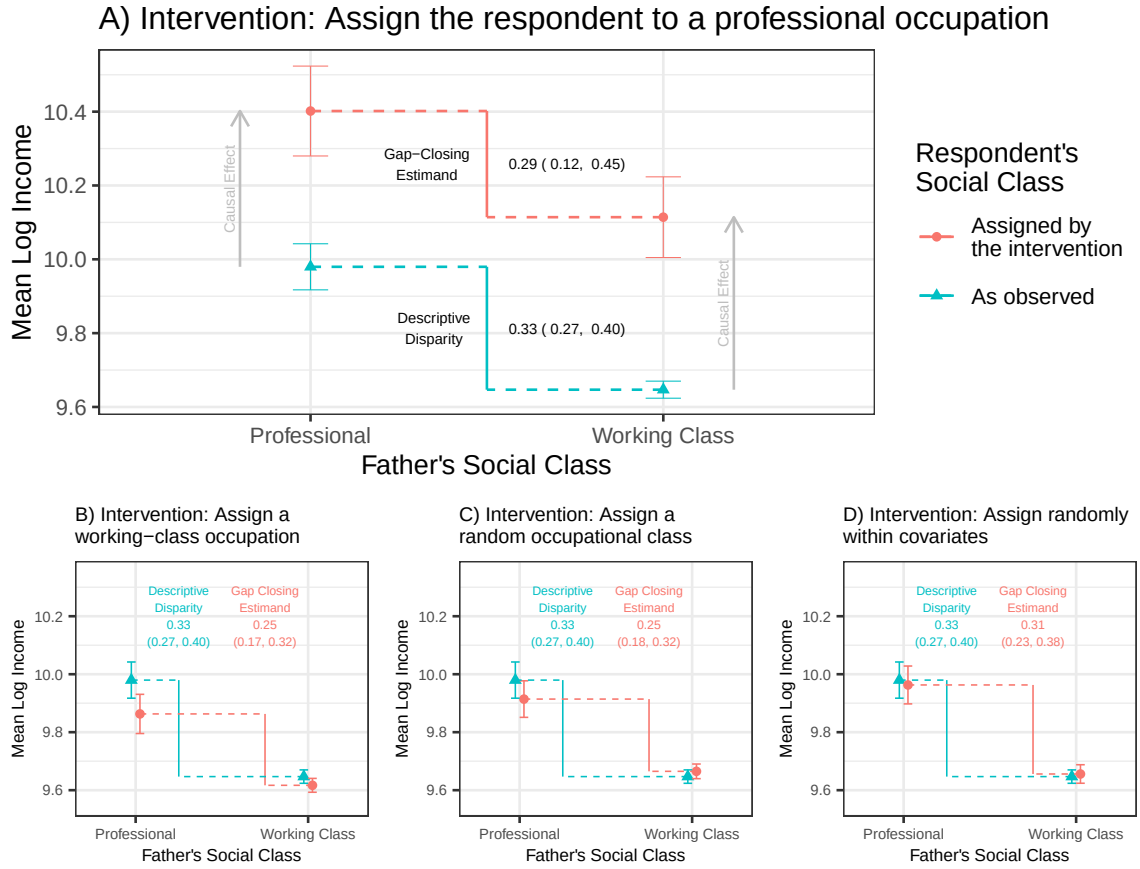


Figure 3.6: An intervention to change one’s own social class would do little to close the pay gap across categories of one’s father’s social class. Among those whose father held a professional occupation, mean log income is 0.33 points higher than among those whose father held a working-class occupation. 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 B.5 presents details for this illustration, including sample selection, definitions of the interventions, and the regression specifications. Results use doubly robust estimators (Eq. 3.5.6) with g and m estimated by weighted linear and logistic regression, respectively (Table B.2). Conceptually, this figure builds on Laurison and Friedman (2016).

3.7 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 (Jackson, 2018, 2021; Jackson and Arah, 2020; Jackson and VanderWeele, 2018; VanderWeele and Robinson, 2014). 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 3.1 and 3.2). 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, 2016; Hernán et al., 2016a,b). 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 SUTVA) 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 (Bang and Robins, 2005; Kang and Schafer, 2007; Robins et al., 1994), including for the setting with stochastic treatment assignments (see Dudík et al. 2014 Sec. 3.3 and Murphy et al. 2001 Sec. 5.2).⁵ 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.2). 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 (Chernozhukov et al., 2018; Van der Laan and Rose, 2011). 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

⁵Continuous treatments are one setting in which doubly robust estimators for stochastic assignment rules have received considerable interest (Kennedy et al., 2017; Muñoz Díaz and van der Laan, 2012).

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 implicitly asking 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.

3.7.1 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 .

$$\mathbf{E}(Y \mid X, \vec{L}) = \alpha_{\text{Model 1}} + \beta_{\text{Model 1}}X + \vec{\eta}'_{\text{Model 1}}\vec{L} \quad (3.7.1)$$

$$\mathbf{E}(Y \mid X, \vec{L}, T) = \alpha_{\text{Model 2}} + \beta_{\text{Model 2}}X + \vec{\eta}'_{\text{Model 2}}\vec{L} + \gamma_{\text{Model 2}}T \quad (3.7.2)$$

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 3.3). 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 B.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.

$$\begin{aligned}
\beta_{\text{Model 1}} - \beta_{\text{Model 2}} = & \overbrace{\left[\mathbf{E}(Y \mid X = 1, \vec{L} = \vec{\ell}) - \mathbf{E}(Y \mid X = 0, \vec{L} = \vec{\ell}) \right]}^{\text{Disparity within subgroup } \vec{L} = \vec{\ell}} \\
& - \underbrace{\left[\mathbf{E}(Y(t) \mid X = 1, \vec{L} = \vec{\ell}) - \mathbf{E}(Y(t) \mid X = 0, \vec{L} = \vec{\ell}) \right]}_{\text{Disparity within subgroup } \vec{L} = \vec{\ell} \text{ under intervention to set } T = t} \quad \forall t, \vec{\ell}
\end{aligned} \tag{3.7.3}$$

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. 3.7.1–3.7.2, the comparison between the two is (by assumption) the same at every value of $\vec{\ell}$ and t .

The comparison in Eq. 3.7.3 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. 3.7.3 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.

3.7.2 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 (Blinder, 1973; Kitagawa, 1955; 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).⁶

⁶An 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.

Causal work on controlled direct effects may seem related but is distinct because it posits a different kind of intervention. A controlled direct effect (Acharya et al., 2016; Pearl, 2001; Robins, 2003; 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 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.

3.8 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., 2021a).

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.

Chapter 4

Quantifying the contribution of occupational segregation to racial disparities in health: A gap-closing perspective¹

Chapter abstract. Racism causes racial disparities in health, and structural racism has many components. Focusing on one of those components, this paper addresses occupational segregation. I document high onset of work-limiting disabilities in occupations where many workers identify as non-Hispanic Black or as Hispanic. I then pivot to a causal question. Suppose we took a sample from the population and reassigned their occupations to be a function of education alone. To what degree would health disparities narrow for that sample? Using observational data, I estimate that the disparity between non-Hispanic Black and white workers would narrow by one-third. This estimate is credible because of adjustment for lagged measures of demographics, human capital, and health carried out under transparent causal assumptions. The result contributes to understanding about inequality and health by quantifying the contribution of occupational segregation to a disparity: if we took a sample and reassigned occupations, the disparity would narrow but would not disappear. The paper contributes to methodology by illustrating an approach to macro-level claims (how segregation affects a population disparity) that draws on explicitly causal micro-level analyses (potential outcomes for individuals) for which data are abundant.

¹A draft of this paper is available on SocArxiv: doi.org/10.31235/osf.io/x9evk. Replication code is available at github.com/ilundberg/replication and will be posted on Dataverse when the paper is forthcoming. For helpful discussions and feedback relevant to this project, I thank Brandon Stewart, Matthew Salganik, Rebecca Johnson, and Gillian Slee. Research reported in this publication was supported by The Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD047879.

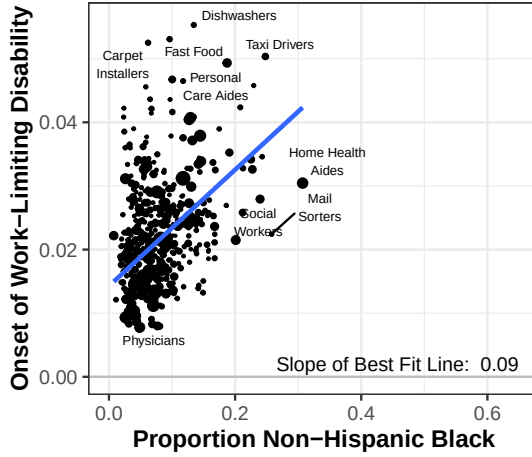
4.1 Introduction

From 1930–1932, roughly 5,000 men drilled a tunnel through Gauley Mountain in West Virginia. The drills dispersed clouds of silica dust through the air and into the lungs of the workers, who were given no protective equipment. At least 764 workers died of silicosis, a serious lung condition. The Hawk’s Nest Tunnel has been called the worst industrial disaster in American history (Cherniack, 1986). Further, the roughly 5,000 men employed to dig the tunnel were not a random cross-section of society: two-thirds were Black. The tragedy of the Hawk’s Nest Tunnel was not only the failure to use protective equipment; it was also the disproportionate allocation of Black workers into a job with high risk.

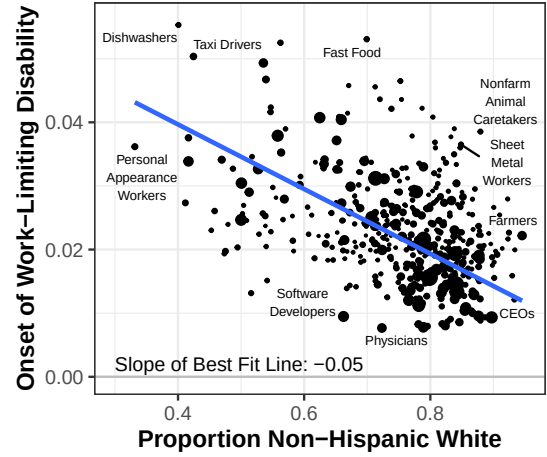
Disparate placement into dangerous occupations occurs in smaller form throughout the U.S. economy (Fig 4.1). Home health aides endure risk to physically support others; 31% identify as non-Hispanic Black. Agricultural workers face hazards to produce the food we consume; 46% identify as Hispanic. From industrial truck operators (19% Black, 21% Hispanic) to personal care aides (19% Black, 17% Hispanic) to roofers (5% Black, 41% Hispanic), people of color hold occupations with health hazards. In fact, the placement of people of color in harm’s way has been part of the American economy since the time of slavery.

This paper contributes a new perspective on inequality and health by examining the degree to which occupational segregation contributes to racial health disparities in the onset of work-limiting disabilities. It may seem uncontroversial to say that occupational segregation plays an important role, but it is not the only force. Health disparities are also the product of experiences of racism, residential segregation, access to healthy food, and many other factors. The degree to which one specific input—occupational segregation—is responsible for health disparities is an open topic for research. The magnitude matters: if occupational segregation accounts for 20% versus

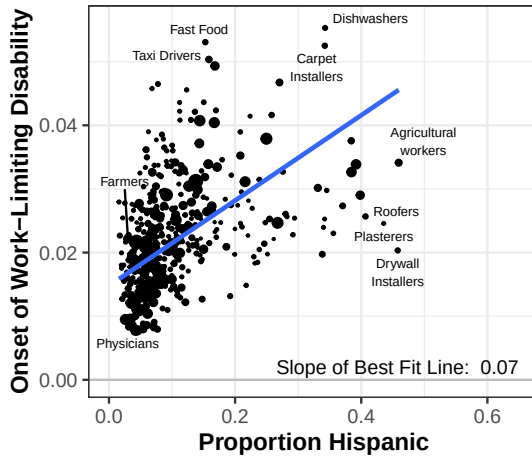
A) Occupations arranged by proportion non-Hispanic Black



B) Occupations arranged by proportion non-Hispanic white



C) Occupations arranged by proportion Hispanic



D) Occupations arranged by proportion of other categories

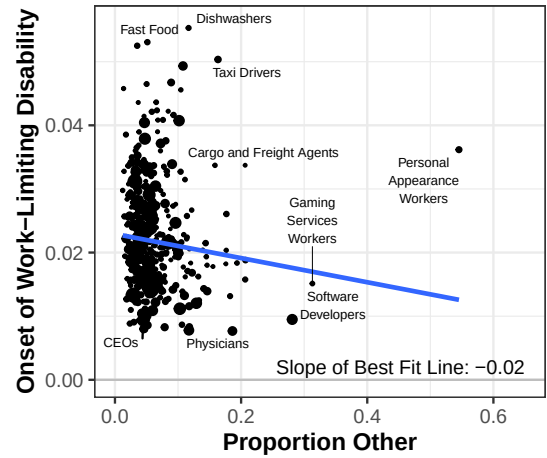


Figure 4.1: Dangerous occupations are disproportionately filled by people of color. Each point is an occupation scaled proportional to its size in the full population. The horizontal dimension is an estimate of the proportion identifying with a given racial and ethnic category in the occupation in year t . The vertical dimension is the proportion of all people in that occupation for whom a disability or health condition limiting the kind or amount of work they can do is reported in year $t + 1$. Both the horizontal and vertical dimensions are estimated by multilevel models with random intercepts using data from the 2005–2020 Annual Social and Economic Supplement of the Current Population Survey (details in Appendix C.2).

80% of the disparity, those would point to different areas for future research and different implications for what types of policies could help to close health disparities.

The paper also contributes to quantitative methodology by illustrating one approach for linking micro-level evidence to macro-level claims. Micro-level evidence is abundant: we can observe many people in many occupations, with disparate health outcomes. But sociological interest may lie in the macro-level implications of those micro-level outcomes: how occupational segregation (at the population level) contributes to a health disparity (at the population level). This paper presents one approach for making that connection. Each individual i has a potential health outcome $y_i(d)$ that they would realize if exposed to each possible occupation d . They also have an expected outcome $\bar{y}_i(\pi)$ realized if they were counterfactually assigned to an occupation from an equitable assignment rule π . Aggregating over individuals, we can return to the macro level to estimate a counterfactual population health disparity. The cost of a micro-level approach is that the result should be interpreted as the outcome of a hypothetical experiment: if we drew a sample from the population and desegregated occupations for that sample, what is the expected health disparity we would observe in the sample? This procedure draws on recent methodological advances to place racial disparities in a causal perspective (Jackson and VanderWeele, 2018; Lundberg, 2021; VanderWeele and Robinson, 2014).

The paper proceeds in several sections. Section 4.2 argues that occupational segregation contributes to health disparities but is not the whole story: the contribution is likely to be much more than 0% but much less than 100%. Section 4.3 introduces the analytical framework for linking micro-level evidence to a macro-level claim. Section 4.4 present the data, causal assumptions, and estimation methods. Section 4.5 shows that occupational segregation is partially responsible for the Black-white disparity in reported onset of work-limiting disabilities, accounting for 35% of the disparity.

Section 4.6 concludes by discussing implications for inequality and health as well as for quantitative methodology.

4.2 Occupational segregation likely contributes to health disparities but is not the whole story

Occupational segregation has been deeply intertwined with racial inequality throughout American history. In the colonial and antebellum periods, white traders reaped profit by forcing African people into bondage and transporting them to America to fill an exploitative labor market role in Southern agriculture. The Civil War brought legal emancipation, yet short-lived Reconstruction-era policies did little to produce economic opportunity. Sharecropping became pervasive in the South, and even in Northern cities economic barriers maintained racial inequality (Du Bois, 1899). Nor did the rapid industrialization of the early 20th century eliminate occupational segregation, as illustrated by the Hawk’s Nest Tunnel (1930–1932) from the opening of this paper.

Against this historical backdrop, the Civil Rights Act of 1964 proclaimed an end to hiring discrimination on the basis of race. Much of the sociological literature on occupational segregation begins the historical journey at that point. Examining racial segregation across nine occupational categories (e.g. service workers) recorded by the Equal Employment Opportunity Commission (EEOC), past studies document only a partial integration of occupations since 1964 (Ferguson and Koning, 2018; Stainback and Tomaskovic-Devey, 2012; Tomaskovic-Devey et al., 2006). The period after the Civil Rights Act thus continued the same pattern that had already characterized the long run of American history: movements toward occupational integration that were incomplete and faltering. But to what degree does incomplete integration contribute to health disparities?

4.2.1 Much more than 0%: Occupational segregation shapes health

In 2019, 2.8 million Americans experienced workplace injuries (Bureau of Labor Statistics, 2020). Compared with white workers, Black workers disproportionately hold occupations with high injury rates (Seabury et al., 2017; Stanbury and Rosenman, 2014). Case studies reveal specific links between particular occupational hazards and health outcomes, from heat-related illness among farmworkers (Fleischer et al., 2013) to physical trauma among day laborers (Lowry et al., 2010) and hotel workers (Buchanan et al., 2010). Occupations also expose individuals to unequal toxins with effects that may accumulate over time; for example, Juon et al. (2021) find that Black workers hold occupations that place them at higher exposure to dangers such as silica and asbestos, which increase the risk of lung cancer.

The total contribution of the work environment to health is likely to be even greater than those studies suggest because direct injuries and exposures are only a small slice of all the links between work and health. For example, random assignment to a work environment that allowed flexible hours increased sleep and exercise among workers in one corporate case study (Moen et al., 2011). In the opposite direction, workers in precarious employment situations with constantly shifting work schedules report lower quality sleep (Harknett et al., 2020). Work characteristics are related to health behaviors like alcohol consumption (Prins et al., 2019). An exclusive focus on injuries and illnesses tied to specific workplace experiences would miss these indirect consequences. Scholars have therefore called for a more comprehensive research agenda including both direct and indirect links between work and health (Ahonen et al., 2018; Burgard and Lin, 2013; Lipscomb et al., 2006; Schulte et al., 2017). In response to that call, it is important to study the effect of occupations as a whole rather than focusing solely on specific occupational hazards.

4.2.2 Much less than 100%: Structural racism is multifaceted

But occupational segregation is not the only input that produces racial health disparities; structural racism has many components (Williams et al., 2019). The inputs to racial health disparities range from exposures to environmental contaminants at home (Banzhaf et al., 2019) to experiences of racism in everyday life (Lewis et al., 2015; Williams and Mohammed, 2009). An extensive literature has documented how these stressors get under the skin in a cumulative process of weathering (Geronimus, 1992) that creates a physiological stress response measurable in indicators such as elevated blood pressure and cholesterol (Geronimus et al., 2006; McEwen, 1998; Merkin et al., 2009). Even if we were able to eliminate occupational segregation, that would not eliminate the plethora of other sources that would still produce racial health disparities. For this reason, we should not expect that occupational segregation would fully account for the magnitude of health disparities.

4.2.3 Quantifying the magnitude: An open question

Occupational segregation is likely responsible for substantially more than 0% but substantially less than 100% of the magnitude of racial health disparities. But it is difficult a priori to predict precisely how important it is. Even conceptualizing that question is somewhat difficult: what does it mean to say that segregation contributes to a disparity? The next section advances one possible definition: occupational segregation contributes in the sense that the disparity would be different under some hypothetical intervention to desegregate occupations.

4.3 Analytical framework: Bringing micro-level evidence to bear on macro-level claims

Occupational segregation can be conceptualized as a macro-level phenomenon: the unit of analysis would be the U.S. population, for which occupational segregation takes some particular value. A common approach to macro-level questions is to examine variation across social contexts such as neighborhoods, states, or countries (Heymann, 2003; Krieger, 2001; Marmot and Wilkinson, 2005; Ólafsdóttir and Beckfield, 2020). For example, Torrats-Espinosa (2021) follows that strategy to estimate the causal effect of residential segregation (defined at the county level) on racial disparities in COVID-19 mortality (defined at the county level), net of control variables (defined at the county level). That works because one can observe many counties—the entire analysis is at the macro level. But for some theoretical questions there is only one macro-level context. In the present case, there is only one U.S. society. Empirical considerations in this setting therefore require that the macro-level question be linked to a micro-level analysis for which the unit of analysis is a person.

A long theoretical tradition has focused on the linkage between micro and macro. Much of this tradition has called for (1) theories about how individual actors select actions and (2) theories about how those actions interact to produce collective phenomena (Coleman, 1990). For example, Bearman et al. (2004) theorize rules by which people select romantic ties and show by simulation that those rules aggregate up to network patterns similar to those observed in reality. The subfield of analytical sociology has most explicitly carried forth a theoretical aspiration toward “grounding the macro in the purposeful actions of individuals,” (Hedström and Bearman 2009:37).

But sometimes our social science knowledge is inadequate to theorize the behavior of actors or how they interact with each other. In this setting, we can still do something useful. This paper defines causal effects at the individual level (without

theorizing how they derive from specific actor behaviors) and aggregates those to a population-level claim (without addressing interactions among units). The cost of this more limited claim is that it does not address situations where the whole may be more complex than the sum of the parts. But, there are settings in which the result is still useful for science and for policy. In the absence of strong theory about the interactions among units that would make the whole different from the sum of the parts, an estimate that focuses on the sum of the parts is an important first step. The analytical framework proceeds in three steps: (1) define potential outcomes for individuals, (2) theorize an equitable treatment assignment rule, and (3) interpret results in terms of the initial response to a policy change.

4.3.1 Define potential outcomes for individuals

Suppose each person i has a potential outcome $y_i(d)$: the health (healthy or with a work-limiting disability) that they would report if assigned to occupation d (Imbens and Rubin, 2015). Because there are 451 occupations in the Census coding system, that corresponds to 451 potential outcomes for each person. Yet we observe only one: the outcome in the occupation they factually held. Nonetheless, defining all of the potential outcomes allows us to move beyond descriptive disparities and toward disparities in a counterfactual world with occupations allocated differently, under a hypothetical intervention to change the occupation to which each individual is assigned.

To illustrate, suppose we have a hypothetical population of eight people assigned to one of three occupations (Fig 4.2). We are interested in the health disparity across two categories of people, here labeled abstractly as circles and diamonds but which could correspond to categories of race, class, or gender. In this hypothetical setting, Panel A shows a health disparity: only $\frac{1}{4}$ of the circles will develop a disability next year compared with $\frac{3}{4}$ of the diamonds. That disparity could be a consequence of

segregation: the diamonds are disproportionately home health aids, an occupation which (in this hypothetical example) is very hazardous. What if each person was instead assigned to a random occupation, so that the disparity in that counterfactual world could not be attributed to occupational segregation? If we knew the full table of potential outcomes (Panel B), we could answer that question: randomly assigning occupations would reduce the disparity from $\frac{1}{2}$ to $\frac{1}{4}$. That provides a metric of the degree to which occupational segregation contributes to a disparity. One can extend this idea to more complex counterfactual rules for assigning occupations (next subsection) and must make causal assumptions in order to learn about unobserved potential outcomes (next section).²

Importantly, this approach makes no appeal to change a circle into a diamond or vice versa; the causal effect of the category (circles and diamonds) remains undefined. Rather, circles and diamonds are simply viewed as labels that distinguish two subpopulations of people. That perspective is enormously helpful when the disparity is defined over a category like race. Because it is difficult to conceptualize an intervention to change someone’s race, statisticians refer to race as an “immutable” characteristic for which causal effects are undefined (Greiner and Rubin, 2011; Holland, 1986). You might be able to manipulate one component like the name at the top of a resume (Bertrand and Mullainathan, 2004; Sen and Wasow, 2016), but some have proposed that doing so is too limiting (Kohler-Hausmann, 2018) because it reduces a

²The treatment here is occupation: we are imagining that each person could counterfactually have held a different occupation. This operationalization stands in contrast with a common approach in existing work: scaling occupations to a continuous score. For example, Fujishiro et al. (2017) estimate the degree to which racial disparities in mortality are attributable to the causal effect of the substantive complexity of one’s occupation (a scale involving the types of reasoning tasks done on the job). The scaling approach may seem more general, since one might imagine an intervention to change the tasks carried out in an occupation without changing the assignment of people to occupations. But in practice when a continuous scale is a deterministic function of occupational categories, there is no way to change someone’s value on the continuous scale without changing their occupation. This is why it is helpful to think about the treatment as the occupation itself, rather than a score to summarize the occupation.

The contribution of occupational segregation can be theorized as the difference between a factual and a counterfactual disparity

Legend: + Disability onset next year   Observed (factual occupation)
✓ Healthy next year   Unobserved (counterfactual occupation)
Category 1 Category 2
(e.g. Black) (e.g. white)

A) Factual disparity with occupational segregation as observed

	Administrative Assistant	Home Health Aid	Sales Supervisor	Probability of Onset of Disability	Disparity
Person 1				0	The average probability among circles is 1 / 4
Person 2				0	
Person 3				1	
Person 4				0	
Person 5				1	The average probability among diamonds is 3 / 4
Person 6				0	
Person 7				1	
Person 8				1	

The factual disparity is partly **because of segregation**:
Diamonds are disproportionately home health aids.

B) Counterfactual disparity if occupations were assigned at random

	Administrative Assistant	Home Health Aid	Sales Supervisor	Probability of Onset of Disability	Disparity
Person 1				1 / 3	The average probability among circles is 1 / 4
Person 2				0 / 3	
Person 3				2 / 3	
Person 4				0 / 3	
Person 5				1 / 3	The average probability among diamonds is 1 / 2
Person 6				1 / 3	
Person 7				2 / 3	
Person 8				2 / 3	

The counterfactual disparity is **not due to segregation**:
Diamonds and circles are assigned equitably to occupations.

Figure 4.2: Theoretical framework for consequences of occupational segregation. In this conceptual figure, each person has a potential health outcome in each possible occupation. The difference between the factual health disparity (A) and the counterfactual health disparity (B) corresponds to the contribution of occupational segregation to the health disparity.

socially and historically constructed category (Omi and Winant, 1994) to one signal. In short, any statement about the outcome a given person would realize if they were white versus if they were Black leads into heavily contested questions about what it means to be white or Black. The approach taken in this paper avoids that problem by defining counterfactuals over occupations, thus only requiring an answer to questions like what it would mean for a given person to be a home health aid or a sales supervisor. We get to maintain race in all its complexity. This pivot appears in epidemiology (Jackson, 2021; Jackson and VanderWeele, 2018; VanderWeele and Robinson, 2014) and sociological methodology (Lundberg, 2021), and the present paper applies those ideas in the specific setting of occupational segregation.

4.3.2 Theorize an equitable assignment rule

The illustration above proposed assigning people to occupations at random. To define a more complex assignment rule, a helpful abstraction is to consider occupations as though they were assigned by lotteries (Fig 4.3). In the world as it exists, each person has their own lottery: the probability that they hold any given occupation is a function of all of their characteristics, some of which might be deemed fair (e.g. education) and some of which might be deemed forms of unfair discrimination (e.g. race). When designing an equitable counterfactual lottery, we get to specify the counterfactual probability of assigning each occupation d to each person i such that the probability is only a function of the former, which Jackson (2021) terms allowable covariates. In this case, we can imagine counterfactually shuffling the assignment of people to occupations within subpopulations defined by education.

This counterfactual lottery is particularly helpful because educational credentials constrain access to occupations (Blau and Duncan, 1967; Hout, 2012). Those without a college degree are not able to be counselors or social workers, for example. A counterfactual that shuffles occupations within educational categories rather than

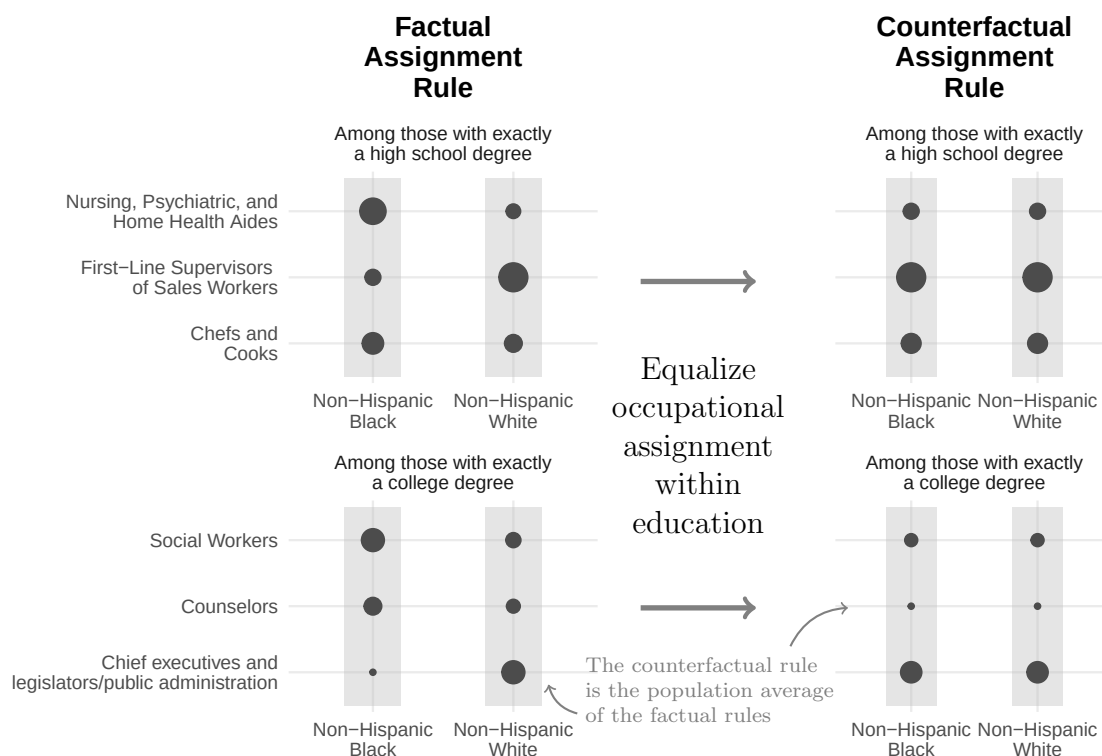


Figure 4.3: A counterfactual assignment rule illustrated with selected occupations. Each gray box corresponds to a lottery from which an individual is assigned an occupation. Circle sizes represent the probability of that particular occupation. The factual assignment rule at the left is segregated and inequitable: the probability of receiving an occupation is a function of race even within educational categories. The counterfactual rule at the right is equitable: the probability of assignment to an occupation is only a function of one’s education. The empirical analysis to follow carries out this exercise with 451 occupational categories counterfactually assigned by probabilities defined within four levels of education (less than high school, high school, some college, college).

in the population as a whole avoids unrealistic counterfactuals such as assigning a person with less than a high school degree to be a counselor. Eq 4.3.1 formalizes this equitable assignment rule mathematically as a function $\pi()$ that takes as inputs an occupation d and a value of education (less than high school, high school, some college, college) and returns a counterfactual probability of assignment to occupation d given that education category.³

$$\pi(d, e) \equiv P(\text{Occupation} = d \mid \text{Education} = e) \quad (4.3.1)$$

The counterfactual probability for assigning occupation d to a person with education e
 is defined to equal
 the population proportion in occupation d
 among people with education level e

Not a function of race.
 The counterfactual assignment rule varies by race only because educational levels are unequal by race.

Importantly, the equitable assignment rule above does *not* fully equalize occupations: the probability of assignment to an occupation remains a function of race indirectly because education is a function of race. But this rule allows us to focus on one step of the stratification process—occupational assignment given education—and leaves to future work questions about how educational inequality contributes to health disparities.

4.3.3 Study the initial response to a policy change

There is a cost to taking a micro-level approach to a macro-level question. A focus on individuals lends itself to claims about individuals—what would happen to a given person if they held a different occupation? It becomes much harder to make claims

³The choice to define the lotteries by these four educational categories is a conceptual rather than a data-driven choice. One could use a more granular definition of education or include other covariates in the assignment rule. These four educational categories are chosen as a parsimonious definition of categories within which all observed occupations are conceptually realistic for any given individual.

about populations—what would happen if everyone held a different occupation? The latter question faces problems because changing the occupation of one person might affect another person’s outcome. Purely macro-level analyses (comparisons across countries with different levels of occupational segregation) would avoid this problem by taking the entire country as the unit of analysis, but they also come with drawbacks: they cannot be used to make claims specific to the U.S., they are difficult when race means something different across countries, and they render it difficult to ask questions about counterfactuals with integrated occupations because no country has integrated occupations. It is therefore worthwhile to find an interpretation of the micro-level analysis that avoids some of the interference problems that inhibit its ability to speak to the full population.

Lundberg (2021) proposes a solution that appeals to hypothetical samples: suppose we took a small sample from the population and reassigned occupations for that sample. What disparity would remain in the sample? That procedure avoids many interference problems because we are only intervening on a fraction of the population, who may not interact with each other. The target of inference is then the expected disparity across all hypothetical samples from the population. The sampling-based target is relevant to policy: when creating a new policy, we can at best apply that policy to only a fraction of the population at first. For instance, it is easier to think about a policy to move one Black worker out of the Hawk’s Nest Tunnel by helping that worker find a different job. It is harder to think about a policy to change the entire composition of the Hawk’s Nest workforce. A local target also lends itself to continued updating of the evidence: after applying the policy to some fraction of the population, we could re-estimate the gap-closing estimand before applying it to another sample. A micro-level analysis of a macro-level claim cannot tell us what would happen if everyone’s occupation changed at once, but it can tell us the initial response to a policy change.

Carefully scoped interpretations are especially important for racial health disparities, for which population-level interventions could easily change the structure of the problem. Suppose thousands of white workers were counterfactually assigned to dig the Hawk’s Nest Tunnel instead of thousands of Black workers. In that case, racist supervisors might have felt compelled to distribute protective equipment. In the language of Jackson (2021), that intervention is not system preserving: it not only changes the distribution of the treatment (occupation), but also changes the mapping between treatments and outcomes (how dangerous the occupation is). In health disparities, many population-level interventions are not system preserving because those with resources find new ways to maintain their advantages (Link and Phelan, 1995). This makes it all the more important to address interference problems head-on, either by encoding interactions in a formal model (Jackson and Arah, 2020) or by delimiting the scope of the claim to focus on the outcome of an intervention to a hypothetical sample. In this case, the latter is more tractable because it is difficult to theorize how integrating occupations would change the mapping between occupations and health outcomes.

4.4 Methods: Disentangle causation from selection to estimate counterfactual health disparities

The previous section introduced an analytical framework for studying segregation. In that framework, each person has a potential outcome in every occupation and thus an expected potential outcome if occupations were assigned equitably. However, it is only possible to observe people in the occupations they actually hold. To learn about the potential outcomes, we need to describe something in the factual world which

implies (under assumptions) an answer about the counterfactual world. This section details that process.

Data come from the 2005–2020 Current Population Survey (hereafter CPS) on the civilian U.S. population ages 25–60, harmonized by the Integrated Public Use Microdata Series (IPUMS, Flood et al. 2020). The outcome variable is whether each individual reports “a health problem or a disability which prevents him/her from working or which limits the kind or amount of work.” I refer to this as a report of a work-limiting disability.⁴ Occupation (the treatment) is measured in 3-digit codes harmonized by IPUMS (Flood et al., 2020) to a standard based on the 2010 Census coding system, which contains 451 unique non-military occupations (e.g. chefs and cooks, elementary and middle school teachers, cashiers, roofers). Racial category is self-reported (non-Hispanic Black, non-Hispanic white, Hispanic, other). The data also contain covariates including sex (men, women), education (less than high school, high school, some college, college), foreign born, age, year, and a lagged measure of the standard five-point scale of self-rated health.⁵

All predictor variables (racial category, occupation, and covariates) are measured at year t . The outcome is measured 12 months later, taking advantage of the linked panel structure of the CPS (Appendix Fig. C.6, see Rivera Drew et al. 2014). The main analyses restrict the sample to those who report at year t that they have never

⁴This measurement of work-limiting disability is widely used in studies using the CPS (e.g. Brucker et al. 2015; Jajtner et al. 2020) as well as studies relying on similar questions in the Health and Retirement Study (e.g. Burkhauser et al. 1999), the Panel Study of Income Dynamics (e.g. Meyer and Mok 2019), and the National Longitudinal Survey of Youth (e.g. Pais 2014). Appendix C.5 discusses limitations of this operationalization of work-limiting disability: it only captures respondent perceptions of their own health, and the CPS survey instrument changed slightly over time.

⁵Supplementary analyses contain a more thorough adjustment set at the cost of narrowing the scope of years involved. Beginning in 2009, the CPS added a series of six questions in which respondents reported whether they had any difficulty with hearing, vision, remembering, walking or climbing stairs, performing basic activities outside the home alone, or taking care of their personal needs. Analyses on the subset of people who answer no to all of those questions yield substantively similar results (Appendix Fig C.9 Panel C).

quit a job for health reasons, are currently employed, and do not have a work-limiting disability (Fig 4.4). This restriction on the lagged outcome and employment reduces the risk that differences in the outcome observed one year later are consequences of latent health conditions. The analytical sample contains 246,658 persons. For each person in the sample, the empirical task can be broken into two components: making an estimate for each person and then aggregating over people.

We first want to know the outcome that person would realize if assigned an occupation from the equitable rule. Consider a subpopulation in which every person has exactly the covariate set of person i . Within that subpopulation, there is variation in occupations and in health outcomes. To make an estimate for person i , we could take a weighted average of the health outcomes of all the people with covariates that match person i , where the weights correspond to the counterfactual occupational distribution of interest. Doing so answers a descriptive question: what health outcome is observed for people who look like person i , but who hold different occupations? We might be willing to make the additional assumption that the outcome is a good estimate for what would happen to person i if counterfactually assigned to an occupation by the equitable rule. That assumption is depicted in the Directed Acyclic Graph in Fig. 4.5A and may be credible because this population is identical along lagged health, demographic, and human capital variables so that the most obvious sources of selection bias are blocked. For instance, the adjustment set accounts for unequal selection into occupations on the basis of lagged work-limiting disability or on the basis of education. One of the main threats to identification is that lagged health (an important confounder) is a latent construct for which lagged reports of health are a coarse measurement; Appendix C.3 discusses two forms of this threat. The threat is mitigated in part by the fact that lagged health is reported with several constructs, one of which involves the exact same question as the one by which the outcome is measured.

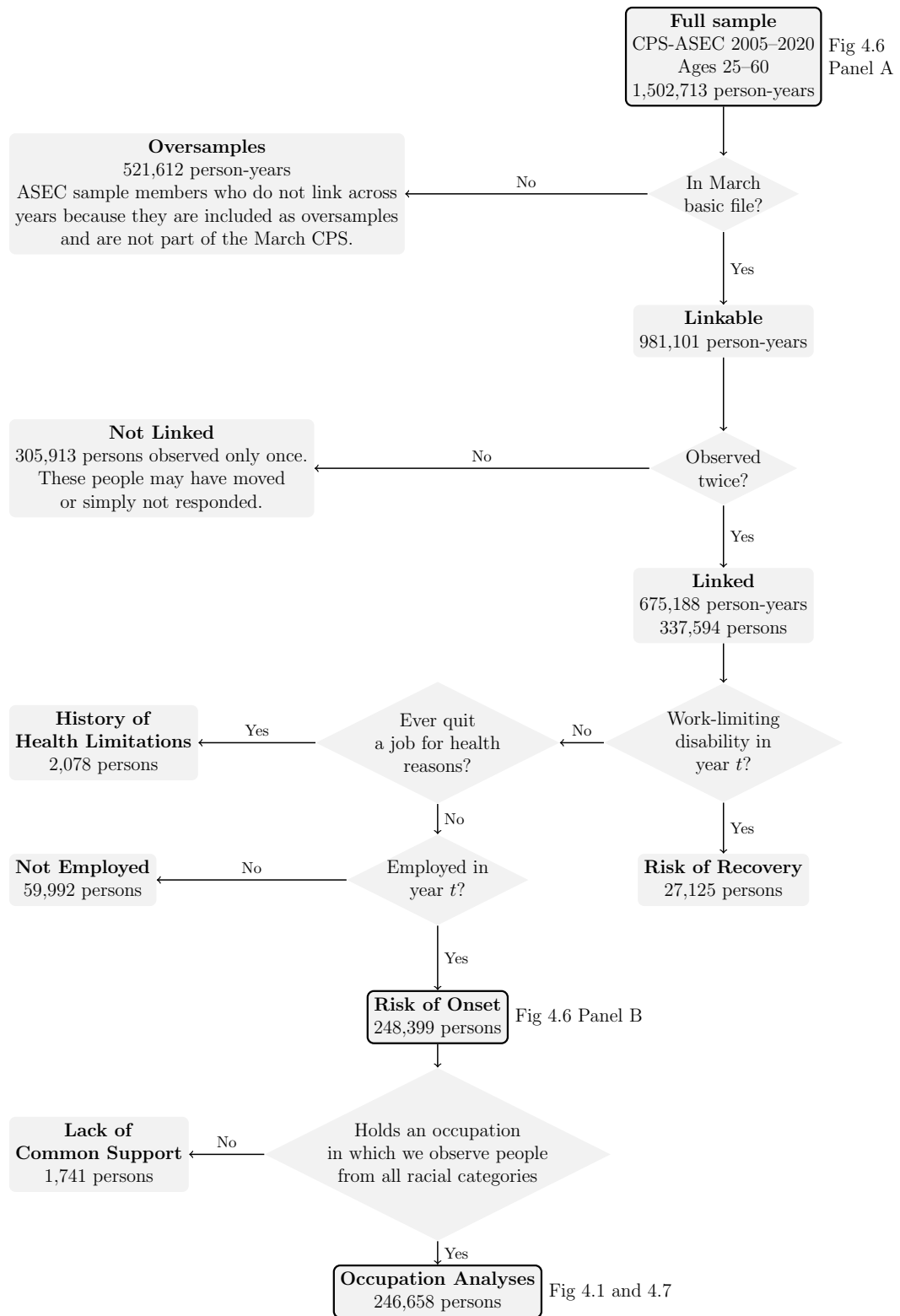


Figure 4.4: Sample restrictions. The three rectangles with black outlines correspond to samples analyzed in the main text.

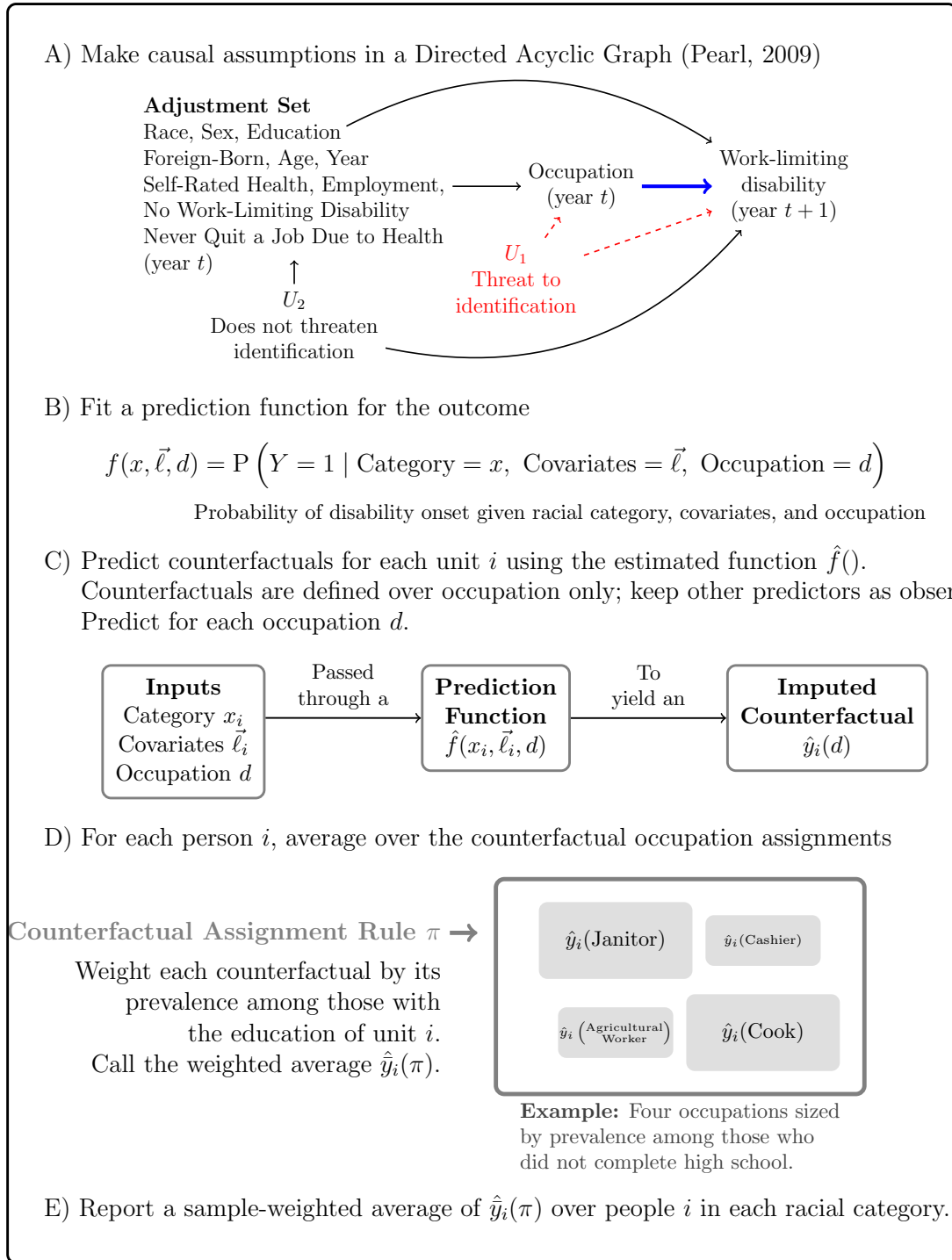
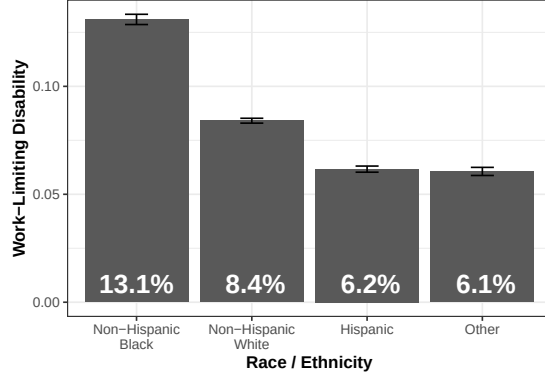


Figure 4.5: Procedure to estimate counterfactual disparities. Under causal assumptions (A), a predictive algorithm (B) can impute the counterfactual probability of work-limiting disability onset that each person i would realize if exposed to any occupation d (C). Weighting those counterfactuals by a new rule for assigning occupations (D) and averaging over subpopulations (E) yields counterfactual disparity estimates.

The argument above has conceptually focused on a subpopulation that is identical to person i along all of the covariates. In practice, that subpopulation may be extremely small in the data. Instead of actually creating subpopulations, I estimate a prediction function for the mean outcome that we would observe if we found people identical to person i . Specifically, I estimate a Generalized Additive Model (GAM, Wood 2017), which is analogous to a generalized linear model but allows non-linear terms for continuous predictors and supports hierarchical methods for more efficient estimation of the occupation-specific intercepts. I estimate four separate models on subpopulations defined by the four racial categories, thereby allowing the function mapping predictors to outcomes to differ across race. Appendix C.4 presents modeling details. Given the predictive model (Fig. 4.5B), I make predictions for each person at their observed covariates but at each possible occupation (Fig. 4.5C). Then, I take a weighted average for each person i over occupations (Fig. 4.5D). The strategy of using a predictive algorithm rather than strata of the population is advantageous because it improves the precision of estimates by sharing information from people with similar covariate values.

Once we have an estimate for each person, we can proceed to the second component of estimation: averaging over people (Fig. 4.5E). At this point, we are not comparing (for example) Black and white people who are identical along all covariates, as one would in a regression model. Instead, we are comparing Black and white people who differ along all the covariates as those populations differ in the real world. The covariates are used to make an estimate for each person, but the average of those person-specific estimates across populations does not hold the covariates constant. This makes the goal of the empirical analysis different from the goal of a single regression model. The final estimates do not hold all covariates constant; they emulate a hypothetical intervention to occupation alone.

A) In the full population, work-limiting disabilities exhibit large disparities by race and ethnicity



B) Among those employed with no work-limiting disability this year, there exist severe disparities in **onset** of disability next year

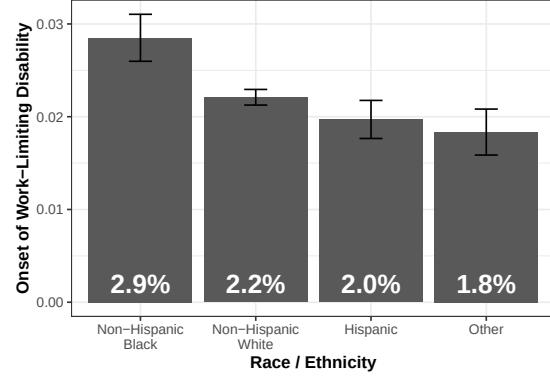


Figure 4.6: Disparities in work-limiting disability across population subgroups. Each panel pools data from the 2005–2020 March Supplement of the Current Population Survey. Error bars depict 95% confidence intervals based on variance estimates calculated using replicate weights (Appendix C.6).

;

4.5 Results

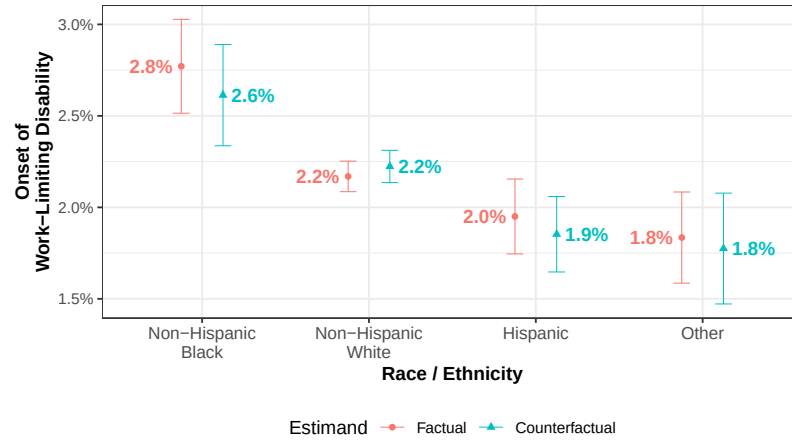
Among the U.S. civilian population ages 25–60 in 2005–2020, work-limiting disabilities are common and disparate across categories of race/ethnicity: a work-limiting disability is reported for 13.1% of those who identify as non-Hispanic Black but only 8.4% of non-Hispanic whites, 6.2% of Hispanics, and 6.1% of those who identify with another racial category (Fig. 4.6 Panel A). Focusing on the subpopulation who are employed and do not report a disability in year t , there are also disparities in the onset of disability by year $t + 1$, which is reported by 2.9% of non-Hispanic Black workers compared with only 2.2% of non-Hispanic white workers, 2.0% of Hispanic workers, and 1.8% of those who identify with another racial category (Fig. 4.6 Panel B). Although the prevalence of disability onset (Panel B) is lower than the overall prevalence of disability (Panel A), focusing on onset is useful because it directs attention to the process by which those overall disparities arise.

To what degree are the disparities in Fig 4.6 Panel B a consequence of occupational segregation? Fig 4.7 summarizes the onset of work-limiting disability in the factual population and under the counterfactual assignment rule. Equalizing occupational assignments within education categories would reduce the onset of work-limiting disability among non-Hispanic Black individuals from 2.8% to 2.6% and increase it slightly among non-Hispanic white individuals (who are now exposed to riskier occupations) (Panel A). The counterfactual Black-white disparity is 35% smaller than the factual disparity: a reduction from a disparity of 0.6 percentage points to 0.4 percentage points (Panel B). A reduction of 0.2 percentage points corresponds to a large number of people: it implies that roughly 39,200 Black workers who experienced onset of a work-limiting disability in a given year would not have done so if occupations were allocated equitably.⁶ Unequal allocation to occupations is therefore responsible for some of the Black-white disparity in the onset of work-limiting disability. But the amount accounted for by occupation is far from all of the racial disparity. Structural racism shapes health in many forms, and other contributors to the racial disparity (e.g. neighborhood environments) expose people to differential hazards outside of work. Occupational segregation is a substantial part of the story, but it is not the whole story.

The two other counterfactual disparities (Hispanic vs. non-Hispanic white and other vs. non-Hispanic white) follow a different pattern. Each of these factual disparities runs in the opposite direction of the Black-white disparity: these groups have *lower* onset of work-limiting disability compared to non-Hispanic whites. Yet they also work in more dangerous occupations. Equalizing occupational assignments would *increase* the Hispanic-white gap by 69% and the other-white gap by 34%. The

⁶In the fourth quarter of 2019 there were 19.6 million employed Black Americans (Edwards and Smith, 2020). The above estimate (based on 2005–2020) implies that 0.2% of these individuals experience disability onset and would not have experienced that onset if occupations were allocated equitably. That corresponds to $0.2\% \times 19.6 \text{ million} = 39,200$ people.

A) Equalizing occupational assignments for a sample would reduce the onset of work-limiting disability for all groups except non-Hispanic whites



B) Equalizing occupational assignments for a sample would reduce the Black-white disparity by 35%. It would increase the health advantage of the other categories compared to non-Hispanic whites.

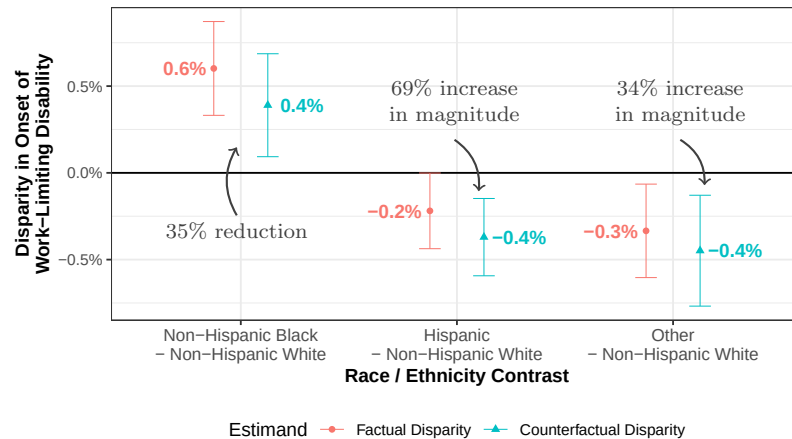


Figure 4.7: Factual and counterfactual probabilities of work-limiting disability onset. The counterfactual quantities refer to the subgroup average of the individual average potential outcome realized if occupations were assigned proportional to their prevalence within one's educational category. In Panel B, the 35% reduction and 69% increase are statistically significantly different from zero at the 0.05 level.

ability to make those claims is a key benefit of a causal perspective: even though these groups have descriptively better health than non-Hispanic whites, we would expect their health to be *even better* in a counterfactual world where occupations were allocated equitably. Hispanic individuals in the U.S., for example, possess other covariates that are associated with good health, such as a younger age distribution and a higher proportion immigrants.⁷ They do not fully realize the better health we might expect given those covariates in part because they are employed in more hazardous occupations. Just because a category has good health does not imply that the labor market treats that category well.

4.6 Discussion

Racial health disparities are partially rooted in economic structures. That was true when the American economy rested on the backs of slaves. It was true in the industrial tragedy of the Hawk’s Nest Tunnel, when a company profited by drilling through a mountain at breakneck speed by neglecting even the most basic protections for the workers involved. And it is true today, as people of color continue to work disproportionately in the occupations with the greatest hazards.

This paper presented three claims. First, work-limiting disabilities are highly disparate by race. Among Americans ages 25–60, only 8.4% of those identifying as non-Hispanic white report a health condition that limits the kind or amount of work they can do. The same statistic among those identifying as non-Hispanic Black is 13.1% (Fig 4.6). Drawing on panel data linked over periods of 12 months, this paper focused attention on inflows into those statistics: the onset of work-limiting disability in year $t + 1$ among those who were employed without a work-limiting disability

⁷In an alternative specification restricted to the native born population, the onset of work-limiting disabilities is factually slightly higher for Hispanics compared with non-Hispanic whites (Appendix Fig C.9 Panel D).

in year t . The onset of work-limiting disability is highest among those employed in occupations with high proportions non-Hispanic Black or Hispanic (Fig 4.1). Finally, Fig 4.7 presents this paper’s central claim: if we took a sample from the population and reallocated occupations equitably (as a function of education alone), the Black-white gap in the onset of work-limiting disability would close by 35% for that sample. The health advantages of the Hispanic population and of other categories relative to non-Hispanic whites would become more pronounced. In short, the unequal assignment of people to occupations plays an important role in the production of racial disparities in work-limiting disabilities.

While important, occupational segregation is far from the whole story. In fact, it accounts for less than half of the total disparity in the onset of work-limiting disabilities. Racism in America involves many structural components, ranging from occupational segregation to residential environments to personal experiences of racism. That is why quantifying the magnitude of one specific contributor is so important: we now know that occupational segregation alone is responsible for roughly one-third of the disparity in onset of work-limiting disabilities. Eliminating occupational segregation would likely move us in the direction of health equity, but even a perfect policy to desegregate occupations would be incomplete in the absence of interventions to other causes.

These results are informative despite the unavoidable limitations of drawing causal claims from observational data. While the adjustment set is surely insufficient to block all forms of selection, incorporating lagged indicators of health and employment into the estimation procedure moves us toward a causal claim. It is infeasible to actually randomize occupations, and natural experiments with exogenous shocks to occupation would speak only to one specific context. For the goal of producing a claim representative of the population, observational data is the only option. In this setting, adjustment for confounding variables yields a useful claim even if it is imperfect.

Results speak most directly to a local sampling-based interpretation: what would happen if we reallocated occupations for a sample. To make a global claim about what would happen under an economy-wide restructuring to eliminate occupational segregation, one would have to make many additional assumptions about how actors interact with each other. Absent those assumptions, the local claim provides an important point of empirical evidence from which those who desire to do so can theorize about what might happen under a more global intervention.

4.6.1 Implications for inequality and population health

Scholars of population health have long emphasized race as a fundamental causes of disease (Link and Phelan, 1995; Phelan and Link, 2015; Williams et al., 2019). Racism is too complex to be appropriately operationalized by the coefficient on race with a set of predictors held equal. Further, analyses that seek to jointly assess the contributions of many components of racism (e.g. residential segregation and occupations and discriminatory experiences) are bound to yield incomplete understanding: each of these inputs is sufficiently complicated to require an analysis all its own. Drawing on causal approaches (Jackson and VanderWeele, 2018; Lundberg, 2021; VanderWeele and Robinson, 2014), this paper illustrates what one can learn by focusing on one input: occupational segregation. For scientific understanding of the Black-white disparity in work-limiting disabilities, it is helpful to know that occupational segregation is responsible for about one-third of that disparity. Future work could undertake similar analyses for other structural causes.

By emphasizing occupational segregation, this paper directs attention to an overlooked aspect of health: the allocation of people to occupations. The safety of each occupation has long been the focus of government policy. For example, the mission of the Occupational Safety and Health Administration is “to assure so far as possible every working man and woman in the Nation safe and healthful working conditions,”

(OSH Act of 1970, Section 2). The agency does that by enforcing rules about the safety of occupations. But there is an additional set of policies one could also consider which might focus on providing all people with equal access to occupations, such as training programs and enforcement of laws against hiring discrimination. Those types of policies would not eliminate racial health disparities, but they might address some of the one-third of the Black-white disparity attributed to occupational segregation in this paper. In short, future research and policy should not only explore how to reduce risks (make occupations safer) but also how to ensure that risks are distributed equitably across population subgroups (e.g. reduce barriers to occupational entry).

4.6.2 Implications for quantitative methodology

A recurring question in sociological methodology is the linkage between micro- and macro-level processes. One can study macro-level processes by remaining entirely at the macro level: how population-level outcomes vary across social contexts (the unit of analysis) with different structural conditions. But sometimes we do not observe any macro-level context with the structural conditions of interest (e.g. racially integrated occupations). Comparisons across contexts are also limited when variables of interest (e.g. race) have definitions that may vary across those contexts. By focusing on the outcomes of individuals exposed to different structural conditions (e.g. assigned to different occupations), one gains opportunities to explore new theoretical questions, like the outcome of an equitable rule for assigning occupations which the researcher defines. One also gains new empirical opportunities: causal claims require us to find units who are similar but with different treatment conditions, and that task is easier when the unit is an individual rather than a population because there may be hundreds of thousands of individuals but only a handful of populations. There is much to be gained by defining the consequences of a macro-level phenomenon (e.g. segregation) in terms of individual-level outcomes. The only cost of doing so comes

in interpretation: in the absence of strong theoretical knowledge about interference, we must scale back the scope of claims to focus on what would happen for a sample (rather than the whole population) exposed to alternative structural conditions. Macro-level analyses have an important place in scholarship, but for many questions new insights may be possible through complementary approaches that consider the explicitly causal micro-level outcomes that individuals would experience if exposed to different macro-level conditions.

Because a pivot to micro-level data potentially unlocks sample sizes that are orders of magnitude larger (individuals instead of societies), this approach creates new opportunities for estimation via predictive algorithms as part of the growing field of computational social science (Edelmann et al., 2020; Lazer et al., 2020). In this paper, the predictive algorithm was a Generalized Additive Model (Wood, 2017) applied separately in subpopulations defined by race, and used to predict unobserved potential outcomes under transparent causal assumptions. Formalizing macro-level questions in terms of a population average of unit-specific potential outcomes (as in the general framework of Lundberg et al. 2021b) thus connects predictive algorithms to a macro-level sociological question.

4.6.3 Conclusion

The path to health equity requires that we understand the root causes of racial disparities. One of those root causes is occupational segregation. The Hawk’s Nest Tunnel was one egregious example, but the systematic placement of people of color into harm’s way by virtue of their occupational assignments is ubiquitous in the American economy. Occupational segregation accounts for roughly one-third of the racial disparity in the onset of work-limiting disabilities. Steps to eliminate segregation would move us part of the way toward health equity. Additional scholarship targeting other counterfactual health disparities under modified structural conditions

could help scholarship to contribute to policy changes that might advance a more equitable future.

Chapter 5

Conclusion

The future of quantitative social science is bright. There exist tremendous opportunities to bring ideas together that transcend disciplinary boundaries to yield new insights. The illustrations in this dissertation brought together theories like the social construction of race (Kohler-Hausmann, 2018; Omi and Winant, 1994) in synergy with formal language and assumptions of causal inference (Hernán and Robins, 2020; Imbens and Rubin, 2015; Pearl, 2009) and estimation ideas from statistics and machine learning (Bang and Robins, 2005; Chernozhukov et al., 2018; Van der Laan and Rose, 2011). The toolkit available to us is likely to continue multiplying in the years to come. Making the most of that toolkit is possible when we work together across domains of expertise to both define new research questions (often drawing on social science theory) and develop new estimators to answer those questions (often drawing on the principles of data science). With the tools now available, we should not ask research questions the way social scientists asked them in decades past—doing so would miss opportunities. As tools continue to develop in the decades to come, there will be new possibilities to ask and answer research questions in new ways. But in all cases, the tools alone are not what yields strong social science research—that research is best if guided toward well-stated questions that motivate all subsequent methodological choices. The world of quantitative social science is always changing, but one thing stays the same: research begins with a goal.

Appendix A

Supplementary materials for Chapter 2

A.1 Papers using robustness checks in ASR 2018

We arrived at this list by searching for the word “robust” and manually validating each case. Every entry in this list conducts some form of robustness check, but the list is not intended to be exhaustive.

- | | |
|------------------------------------|-------------------------------------|
| 1. Bloome et al. (2018) | 11. Kadivar (2018) |
| 2. Ciocca Eller and DiPrete (2018) | 12. Ludwig and Brüderl (2018) |
| 3. Desmond and Travis (2018) | 13. McDonnell and King (2018) |
| 4. Font et al. (2018) | 14. Mize and Manago (2018) |
| 5. Gauchat and Andrews (2018) | 15. Quadlin (2018) |
| 6. Goldstein (2018) | 16. Schilke and Rossman (2018) |
| 7. Gutierrez (2018) | 17. Schneider et al. (2018) |
| 8. Hahl et al. (2018) | 18. Villarreal and Tamborini (2018) |
| 9. Horowitz (2018) | 19. Weisshaar (2018) |
| 10. Inanc (2018) | 20. Wilmers (2018) |

A.2 Mediation involves interventions to two variables

This section briefly expands on the mediation estimands listed in Table 2.1. While Section 2.3.3 discussed mediation, it touched on only one type of mediation estimand (a controlled direct effect). This section compares that estimand to a different mediation estimand: the natural direct effect. This example also provides an additional example of a widely-read sociological study which implicitly involves a mediation claim.

We consider the discussion in Western (2006:30) of the “post-release effect of incarceration” on employment. Two variables in this claim are the subject of an intervention: a treatment D (incarceration at time 1) and a mediator M (incarceration at time 2). Each unit i has several potential outcomes (employment at time 2) denoted $Y_i(d, m)$, one for each combination of a treatment value d (incarcerated or not at time 1) and a mediator value m (incarcerated or not at time 2). The post-release effect of incarceration can be formalized as a controlled direct effect that compares the outcome under incarceration versus no incarceration at time 1, under an intervention to no incarceration at time 2.

$$\text{CDE}(0) = \frac{1}{N} \sum_{i=1}^N \left(Y_i(1, 0) - Y_i(0, 0) \right) \quad (\text{A.2.1})$$

A different controlled direct effect would be the effect that would persist if we intervened to assign you to prison at the time of observation regardless of your history. If employment is uncommon in prison, whether one has previously been incarcerated may have a negligible controlled direct on employment in this scenario: you would

not be employed regardless whether or not you were previously incarcerated.

$$\text{CDE}(1) = \frac{1}{N} \sum_{i=1}^N \left(Y_i(1, 1) - Y_i(0, 1) \right) \quad (\text{A.2.2})$$

A third alternative, the natural direct effect, would be the effect of time 1 incarceration if we intervened to hold time 2 incarceration at the value it would naturally have taken in the absence of prior incarceration, denoted as the potential mediator $M_i(0)$.

$$\text{NDE} = \frac{1}{N} \sum_{i=1}^N \left(Y_i(1, M_i(0)) - Y_i(0, M_i(0)) \right) \quad (\text{A.2.3})$$

Once we step away from models that are linear and additive, it becomes clear that the definition of mediation-based estimands depends on the value at which the mediator is held. For accessible introductions, see Acharya et al. (2016) on controlled direct effects and Imai et al. (2011) on natural direct effects. Once a researcher defines the estimand—a particular quantity that captures the specific aspect of mediation most relevant to the study—then it will be almost second-nature to conduct significance tests as advised in other commentaries (Mustillo et al., 2018). This fact reinforces the importance of stating the causal contrast at the heart of the unit-specific quantity in precise terms at the start of the analysis.

A.3 Target population details in Angrist and Evans (1998)

To calculate the target population size in Angrist and Evans (1998), we rely on the descriptive statements in the text. Between 37 and 53 % of mothers ages 21–35 have two or more children (Table 1 in Angrist and Evans 1998:453), and the proportion of women having a third child was between 6 and 7 percentage points greater among those whose first two children were of the same sex compared with those whose first

two children were of different sexes (bottom row of Table 3 in Angrist and Evans 1998:457). The size of the target population is therefore at most only 4 % (0.53×0.07) of all mothers ages 21–35.

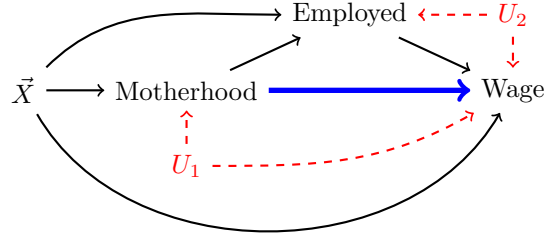
A.4 Family gap in pay: Example details

This appendix provides details about the family gap in pay (Pal and Waldfogel, 2016), which we used to illustrate the estimation step.

Fig A.1 Panel A presents the causal assumptions underlying the research question as it is presented in the main text. Because one cannot earn a wage unless one is employed, we assume that employment is a cause of wages. We focus on the direct effect of motherhood on the wages women would realize if they were employed (blue edge). Identifying this direct effect relies on two key assumptions. The first is that the covariates $\vec{X}$ (age, education, marital status, and race) are sufficient to block all backdoor paths between motherhood and wages. This assumption entails that there are no unobserved variables like U_1 that affect both motherhood and wages. The second assumption is that the same covariates $\vec{X}$ are sufficient to block all backdoor paths between employment and wages. This assumption entails that there are no unobserved variables like U_2 that affect both employment and wages. This latter assumption is often ignored in practice, but it is important because it is doubtful. The quality of one’s resume, for instance, might both increase the probability of employment and increase the wage that one would realize if employed. The main text proceeds to estimation to illustrate that step of the process, but selection along variables like U_2 is a serious concern in any study for which the outcome variable (e.g. wage) is only defined for those taking a particular value of a mediator (e.g. employment).

Transparency about the assumed causal structure opens dialogues about alternative causal assumptions. For example, Elwert and Winship (2014:41–42) discuss

A) Basic DAG illustrating key assumptions in the motherhood example



B) Detailed DAG with a more complex causal process

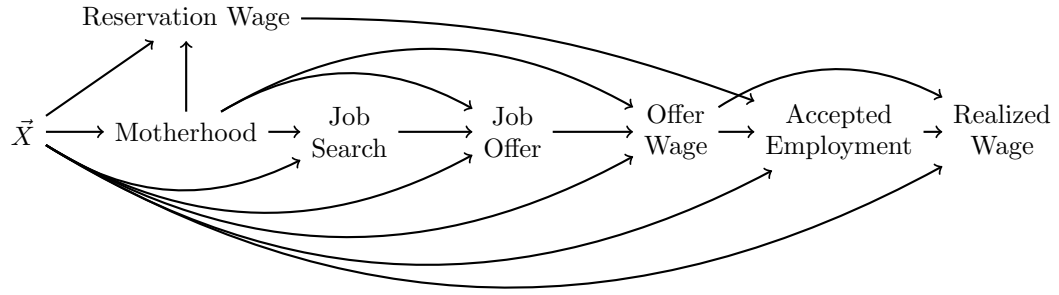


Figure A.1: DAGs showing causal assumptions for the motherhood wage penalty. Panel (A) shows the basic assumptions of our main analysis. We focus on the direct effect (blue edge) of motherhood on the wages that women would realize in a possibly counterfactual world in which they were employed. In this structure, we assume that employment causes wages: one first has to be employed in order to earn a wage. Identification requires that the covariates $\vec{X}$ (age, education, marital status, and race) block all backdoor paths between motherhood and wage (e.g. U_1 does not exist) and between employment and wages (e.g. U_2 does not exist). These assumptions are very strong; in particular, it seems likely that some unobserved variable U_2 affects both employment and wages. Panel (B) expands the DAG to a more detailed causal process: motherhood affects a woman's reservation wage (the lowest wage at which she would accept employment), her job search, whether she is offered a job, the wage she is offered, whether she accepts the offer, and the wage she realizes. In the assumptions of (B), the offer wage (an unobserved variable) confounds the relationship between accepting employment and the wage realized after negotiation, thereby contradicting the assumption of no U_2 in (A). The two sets of assumptions are therefore distinct. Broadly, a research framework centered on the estimand could lead the motherhood penalty literature to consider a more detailed set of questions about how motherhood affects all of these downstream labor market outcomes. We are grateful to Felix Elwert and Christopher Winship for suggesting the more detailed DAG.

the effect of motherhood on wages but assume that wages cause employment—the opposite of our assumption that employment causes wages. In that view, the outcome of interest is *offer wage*: the pay employers offer, after which individuals decide whether to accept the offer. In that framework, wages cause employment rather than the reverse. In the process of circulating drafts of this paper, we corresponded with Elwert and Winship and arrived at the possibility that the true causal process may be a more complex structure as presented in Fig A.1 Panel B. Motherhood affects whether women search for a job, whether they are offered a job, whether they accept that offer, and the wage they ultimately realize (possibly after negotiations). One way that motherhood shapes accepted employment may be by affecting women’s reservation wages: the lowest wage level such that one would accept an offer of employment rather than remaining unemployed. This more complex set of causal assumptions illustrates how the literature on the motherhood wage penalty could ask many distinct research questions about effects on reservation wages, offer wages, and realized wages. Further, the process in Panel B illustrates another problem with the assumptions in Panel A: offer wages can play the role of U_2 , by affecting employment and the wage one would actually realize in the job.

Our dialogue about the assumptions of Fig A.1 Panels A and B illustrates the kind of scholarly communication that can happen in the framework we propose. If we had instead followed the standard approach and defined the goal as the coefficient on motherhood, it would have been much more difficult to have discussions about the causal process as the reader would be unable to discern which causal process we were working with. This is one of the main benefits that the field stands to gain from our approach. Some readers may agree with your causal assumptions and others may not, and making those assumptions explicit is a good step toward productive conversation about those issues. The main text maintains the simpler causal assumptions for

the purpose of using this example to illustrate issues with estimation rather than identification.

Taking the goal to be the direct effect of motherhood on wages that would be realized under employment (Fig A.1 Panel A), we next proceed to estimate that quantity. We draw data from the 2019 Annual Social and Economic Supplement (ASEC) of the March Current Population Survey, in the form supplied by the Integrated Public Use Microdata Series (IPUMS, Ruggles et al. 2019). We restrict the sample to women ages 25–44 (16,977 mothers, 8,183 non-mothers) who fall in the region of common support (16,659 mothers, 8,025 non-mothers, see next paragraph). To estimate the potential wage that would be realized if one were an employed mother or an employed non-mother, we fit models on the subsample who worked for wages or salary in the last year (11,555 mothers, 6,277 non-mothers). To draw inference about the target population of mothers regardless of actual employment, we make predictions for all 16,659 mothers on the region of common support (discussed two paragraphs below). To construct the outcome (log hourly wage), we divide each individual’s annual wage and salary income by the number of hours worked in the last year. To determine the number of hours worked, we multiply the report of the usual hours worked per week by the number of weeks worked in the last year. Because division can produce extreme values, we truncate hourly wages at the 1st and 99th percentile (unweighted). The outcome variable is the log of truncated hourly wages. We operationalize motherhood by whether the respondent reports any of their own children residing in the household. Following Pal and Waldfogel (2016), we include four other predictors: age, education (less than high school, high school, some college, college or more), marital status (married with spouse present, versus not), and race (white, black, other).

To calculate the proportion of observations in the region of common support, we separate the data into subgroups of age, education, marital status, and race. We note whether each subgroup is on the region of common support (at least one employed

mother and one employed non-mother is observed) or not (no one is observed in at least one of these two conditions). We then aggregate over mothers and over non-mothers. The region of common support includes 98.1% of mothers and 98.1% of non-mothers. All analyses restrict to this region of common support. Common support is often a problem when the covariates have a very large number of values (D’Amour et al., 2020). For example, if even one covariate is continuous (with a unique value for each observation), then common support will always be a problem and some additional assumptions will be needed.

To produce point estimates of the family gap in pay, we apply the parametric g -formula (Hernán and Robins, 2020, Ch. 13), an approach involving several steps. We first apply each algorithm (discussed in next paragraph) to learn the conditional mean of log wages given all the predictors among employed women. Then, we predict the outcome for each mother (including those that are not employed) with her observed covariates, and with motherhood changed to zero and all other covariates left unchanged. For each mother, we difference the predictions to estimate the family gap in pay at her specific covariate set. Finally, we aggregate over all mothers by a weighted mean using the ASEC survey weights. To estimate the controlled direct effect of motherhood on the wages women would realize if they were employed, this procedure requires the identification assumptions that potential wages are independent of both motherhood and employment within subgroups of the observed covariates. Although this assumption is likely to be imperfect, we focus this example on issues of estimation and assume for the sake of illustration that the identification assumption holds.

The only difference across the estimators is the algorithm used to learn the conditional mean of log wages given predictors. These algorithms correspond to the columns of Figure 2.6. Column 1 shows results from a nonparametric stratification estimator which estimates the mean by the observed mean of individuals observed with each covariate set, weighted by the survey weights. Column 2 shows results

from a linear regression of log wages on all predictors, with age included as a series of indicator variables as well as their interaction with motherhood. Column 3 shows results from a generalized additive model (GAM, Wood 2017) that adds an assumption that the association between age and log wages follows a smooth functional form, which is allowed to differ for mothers and non-mothers. The thin-plate spline applied follows the defaults of the `mgcv` package in R, with automatic smoothing penalty. Column 4 shows results from an OLS regression model that requires that the association between age and log wage follows a quadratic form interacted with motherhood. Column 5 shows results from an OLS regression model that maintains the quadratic form for age but removes the interaction between age and motherhood. In all of the algorithms, we apply the ASEC survey weights in order to prioritize fitting of the conditional expectation function in regions of the covariate space that are heavily weighted.

To place confidence intervals on the point estimates, we repeat the procedure above 160 times, once for each set of replicate weights provided by the study. The reason for using the replicate weights is because the ASEC follows a complex survey design rather than a simple random sample. As a result, techniques like the nonparametric bootstrap that are designed for the simple random sample setting may not yield valid inference for samples like the ASEC. The replicate weights are analogous to bootstrap samples in that they are intended to approximate repeated draws from the population according to the ASEC sampling scheme. They are commonly used by survey samples because they allow those who collect the sample to design resamples without revealing some key elements of the sampling process, such as geographic identifiers. After re-applying our entire estimation procedure on each of the 160 replicate weights, we have 160 draws of each estimated estimand. We pool these estimates into a variance estimate according to the formula provided by the data

administrators at <https://cps.ipums.org/cps/repwt.shtml>,

$$\hat{V}(\hat{\theta}) = \frac{4}{160} \sum_{r=1}^{160} \left(\hat{\theta}_r - \hat{\theta} \right)^2 \quad (\text{A.4.1})$$

where $\hat{\theta}_r$ is each replicate-specific estimate for $r = 1, \dots, 160$ and $\hat{\theta}$ is the overall point estimate from the main ASEC survey weights. We produce confidence intervals by a normal approximation, taking the middle 95% of a normal distribution centered on the point estimate $\hat{\theta}$ with the estimated variance $\hat{V}(\hat{\theta})$.

The specification with a quadratic for age interacted with motherhood achieves the best predictive performance. To arrive at this result, we sorted the sample by the ASEC survey weights and assigned observations to five folds systematically: the first set of five observations were assigned to folds 1–5, respectively, the second set of five observations were also assigned to folds 1–5, and so on. This split was systematic rather than random, with the advantage that all folds have similar distributions of sample weights. Removing each fold in turn, we fit all the prediction algorithms on the remaining four folds and stored the predictions for the held-out fold. After running through all five folds, we pooled all predictions and calculated the weighted mean squared error over the full sample. This procedure provides a valid estimate of out-of-sample predictive performance because each observation’s prediction was estimated on a sample that excluded that observation. Because this entire procedure can be wrapped in a single function, we can also pass it through the same variance estimation procedure as the point estimates. The result suggests that the variance of the estimated mean squared errors is high, so that we are hesitant to conclude that the data strongly point toward one estimation approach over another, with the exception that nonparametric stratification has substantially worse predictive performance. Nonetheless, if one sought a purely data-driven approach to select

only one estimator, the cross-validation approach would point toward the smooth age specification interacted with motherhood, with all other predictors entered additively.

A.5 Buchmann and DiPrete (2006):

Example details

Our replication of Buchmann and DiPrete (2006) focuses on one descriptive component of the paper: the trend across birth cohorts in college completion for men and women within subgroups of parents' education, which appears in Fig. 2.8 of this paper and corresponds to results presented in Table 2 of the original paper. Following the original authors, we analyze data from the General Social Survey. While they used data collected from 1972–2002, we extend through all data currently available (1972–2018). Following the original authors, we focus on white respondents ages 25–34 ($N = 10,601$). We restrict to the 1947–1984 birth cohorts so that every cohort can be observed at every age from 25–34 ($N = 8,679$). Because the original authors do not conduct multiple imputation, we restrict to observations without missing values on any of the variables ($N = 7,605$). Following the original authors, we define mother's education as a binary indicator of whether the mother attended college. Father's education is more complicated in the original authors' specification: their model includes three categories (father attended college, father did not attend college, or there was no father) but for some interaction terms they combine fathers who did not attend college and those with no father into a single category. To avoid complexity, we use the binary variable definition throughout, categorizing fathers as either (1) attended college or (2) either did not attend college or are unknown.

We estimate the probability of college completion within subgroups of the covariates using a generalized additive logistic regression model (Wood, 2017), weighted by the GSS sampling weights. Our specification allows all interactions among gender,

mother’s education, father’s education, and a smooth function of birth cohort. The appeal of this highly flexible specification is that in a very large sample it would converge toward estimating the proportion completing college within each subgroup (gender $\times$ father’s education $\times$ mother’s education $\times$ cohort) by the simple average of college completion within that subgroup. The smooth term for cohorts pools information from observations in nearby birth cohorts, but is much more flexible than a line or a binary indicator for the cohort exceeding a particular value (the original authors dichotomize at 1966). The values reported in Fig. 2.8 are point estimates and confidence intervals based on standard errors calculated by the `predict` function in the `mgcv` package in R (Wood, 2017).

A.6 Wildeman et al. (2012): Example details

Our replication of Wildeman et al. (2012) is imperfect because replication code and the version of the data used by the original authors are unavailable. We are grateful to the authors for their correspondence with us, which made it possible for us to replicate the main results of the paper. The analyses use the Fragile Families and Child Wellbeing Study, a birth cohort study of 4,898 children born in large U.S. cities in 1998–2000, with an oversample of non-marital births (Reichman et al., 2001). The analyses restrict to the subsample for whom the mother completed interviews when the child was approximately 3 and 5 years old and have non-zero weights from the mother survey at age 5 ($N = 3,770$) and for whom the outcome variables (maternal depression and life satisfaction, where the latter is not examined in this replication) were non-missing ($N = 3,760$). Following the main specification of the original authors, we further restrict to those for whom the father had a history of incarceration at some point up through the age 3 interview ($N = 1,595$). Analyses use the city weights provided by the survey. The target population is therefore mothers of children

born in the 20 sampled U.S. cities in 1998–2000, for whom the father was incarcerated at some point up to child age 3. Following the original authors, we can think of this as a population at risk of paternal incarceration between child ages 3 and 5.

We construct variables to match the way their construction was described by the original authors as closely as possible. The outcome variable is a binary indicator of maternal depression at the age 5 interview, constructed by survey administrators based on maternal self-reports to questions that comprise a depression scale. Mothers are depressed in 21% of cases. The treatment variable is a binary indicator of recent paternal incarceration, defined as any incarceration between the age 3 and age 5 interviews, including those still incarcerated at age 5. Recent incarceration is reported in 37% of cases, with 7% of reports missing. Because our goal is simply to replicate the original paper, we direct the reader to Wildeman et al. (2012) for a description of these variables and to our replication code for details of our implementation of this description.

We multiply impute missing value of all predictors using the *Amelia* package in R (Honaker et al., 2011). While the original authors conduct multiple imputation by chained equations, we use *Amelia* because it is faster, thereby allowing us to conduct analyses on 100 imputed datasets, as compared with the 20 imputations used by the original authors. More imputations may improve the performance of the procedures, for which statistical guarantees are asymptotic in the number of imputations. Most variables are missing for less than 15% of observations. One variable missing in 25% of cases is whether the mother reports that either of her biological parents ever experienced a two-week period of feeling depressed, down in the dumps, or blue; this is sometimes missing because the mother has no knowledge of one of her own parents. Variables involving father reports are also often missing because survey nonresponse and attrition were higher for fathers, some of whom never participate in a survey. Father’s foreign-born status and age are missing in 22% of observations, and paternal

impulsivity is missing in 40% of observations because this variable is based on a scale administered at child age 1 to fathers in only 18 of the 20 sample cities. An argument can be made that the model would be more transparent by excluding these variables with high rates of missingness, but we include them to maintain maximal comparability to the specification of the original authors.

We fit three logistic regression models: a marginal association model in which recent paternal incarceration is the only predictor, a total causal effect model that adjusts for pre-treatment variables, and a direct effect model that additionally includes whether the mother resides with a new partner by the age 5 interview (the mediator), interacted with recent paternal incarceration (the treatment). We fit these models without survey weights for maximal comparability because survey weights are not discussed by the original authors. To aggregate to estimates of the estimands of interest in our framework, however, we make predictions for each observation under various interventions to the treatment and the mediator and aggregate over observations by a weighted mean. This procedure would be valid if the logistic regression model is the correct parametric form for the probability of maternal depression. The procedure is also a valid approximation when this assumption holds only imperfectly; this is a key benefit of stating the estimand separately from the imperfect methods used for estimation.

A.7 Risk ratios and risk difference: An example of interaction estimands that are distinct

Researchers often explore how a treatment effect varies as a function of pre-treatment covariates. Fig. 2.11 illustrates one setting in which the presence or absence of an interaction depends on the estimand. This section provides details about that simulation.

Suppose there exists a pre-treatment covariate X taking the values $\{.20, .21, \dots, .39, .40\}$ with equal probabilities in the population. Suppose a binary treatment variable D is randomly assigned with probability 0.50 for all units. Suppose the potential outcomes under each treatment are binary with probability proportional to X_i , but with a slope that is twice as steep under the treatment condition.

$$X_i \sim \text{Discrete Uniform}(\{.20, .21, \dots, .40\}) \quad (\text{A.7.1})$$

$$D_i \sim \text{Bernoulli}(0.50) \quad (\text{A.7.2})$$

$$Y_i(0) \sim \text{Bernoulli}(X_i) \quad (\text{A.7.3})$$

$$Y_i(1) \sim \text{Bernoulli}(2X_i) \quad (\text{A.7.4})$$

One might wonder whether interaction is present: does the effect of the treatment depend on the value of X ? Two definitions of the “effect” are the risk difference and the risk ratio. The risk difference examines the mean difference in the outcome under treatment and under control, among those in the subgroup $X = x$. The risk ratio examines the ratio of the mean outcomes under each treatment.

$$\text{Causal Risk Difference}(x) = \underbrace{\frac{1}{n_x} \sum_{i: X_i=x} Y_i(1)}_{\text{Mean potential outcome under treatment among those with } X_i = x} - \underbrace{\frac{1}{n_x} \sum_{i: X_i=x} Y_i(0)}_{\text{Mean potential outcome under control among those with } X_i = x} \quad (\text{A.7.5})$$

$$\text{Causal Risk Ratio}(x) = \frac{\frac{1}{n_x} \sum_{i: X_i=x} Y_i(1)}{\frac{1}{n_x} \sum_{i: X_i=x} Y_i(0)} \leftarrow \text{ratio of the means above} \quad (\text{A.7.6})$$

We might say that interaction is present if the causal effect of the treatment is a function of the pre-treatment covariate x . However, the conclusion could depend on what estimand is meant by “effect”. Interaction may be present for the risk difference but not for the risk ratio (Fig. 2.11). This example illustrates the broader point that

the language commonly used to state sociological goals (e.g. effect heterogeneity, interaction) is insufficiently precise.

A.8 Review of *ASR* 2018

We reviewed the 2018 volume of *ASR* to determine if our framework was redundant: can we already unambiguously translate research goals into theoretical estimands defined as unit-specific quantities aggregated over target populations? Table A.1 and A.2 state our coding of the unit-specific quantity and target population, respectively, for all 32 articles published in the 2018 volume of the *American Sociological Review*. Because the theoretical estimand is directly related to how statistical results are interpreted with respect to theory, we considered not only the statistical procedures applied to the data but also the ways the authors interpreted those procedures. Justifications for the codings are provided for the unit-specific quantity in Tables A.3–A.8 and for the target population in Tables A.9–A.12. We arrived at these codings after two of the authors read every paper and iterated to come to agreement about how it should be coded.

Author	Type of article	Difficulty translating the unit-specific quantity to our framework	Details in
Liu	Causal observational	Underspecified mediation	Table A.3
Wedow et al.	Causal observational	Underspecified mediation	Table A.3
Barber et al.	Causal observational	Underspecified mediation	Table A.3
Desmond and Travis	Causal observational	Underspecified mediation	Table A.3
McDonnell and King	Causal observational	Post-treatment controls	Table A.4
Mun and Jung	Causal observational	Post-treatment controls	Table A.4
Font et al.	Causal observational	Post-treatment controls	Table A.4
Kadivar	Causal observational	Post-treatment controls	Table A.4
Schneider et al.	Causal observational	Post-treatment controls	Table A.4
Inanc	Causal observational	Post-treatment controls	Table A.4
Gauchat and Andrews	Causal observational	Post-treatment controls	Table A.4
Gutierrez	Causal observational	Post-treatment controls	Table A.4
Villarreal and Tamborini	Causal observational	Post-treatment controls	Table A.4
Ludwig and Brüderl	Causal observational	Post-treatment controls	Table A.4
Barr et al.	Causal observational	Bundled treatment	Table A.5
Wilmers	Causal observational	Linear continuous treatment	Table A.5
Goldstein	Causal observational	Many treatments	Table A.5
Horowitz	Causal observational	Many treatments	Table A.5
Schilke and Rossman	Causal experiment	Underspecified mediation	Table A.6
Simpson et al.	Causal experiment	Underspecified mediation	Table A.6
Quadlin	Causal experiment	Underspecified mediation	Table A.6
Hahl et al.	Causal experiment	Ok	
Mize and Manago	Causal experiment	Ok	
Weisshaar	Causal experiment	Ok	
Flores and Schachter	Causal experiment	Ok	
Ferguson and Koning	Descriptive	Describes another population	Table A.7
Ciocca Eller and DiPrete	Descriptive	Describes another population	Table A.7
Guinea-Martin et al.	Descriptive	Describes another population	Table A.7
Bloome et al.	Descriptive	Describes another population	Table A.7
Offer and Fischer	Descriptive	Causal verbs	Table A.8
Freeland and Hoey	Descriptive	Causal verbs	Table A.8
Mullins et al.	Descriptive	Ok	

Table A.1: Review of unit-specific quantities. Systematic coding of individual papers from *ASR* 2018. Table A.3–A.8 justify the coding of each paper.

Author	Type of article	Difficulty translating the target population to our framework	Details in
Barr et al.	Probability sample	Does not discuss weights	Table A.9
Flores and Schachter	Probability sample	Does not discuss weights	Table A.9
Gauchat and Andrews	Probability sample	Does not discuss weights	Table A.9
Guinea-Martin et al.	Probability sample	Does not discuss weights	Table A.9
Gutierrez	Probability sample	Does not discuss weights	Table A.9
Horowitz	Probability sample	Does not discuss weights	Table A.9
Inanc	Probability sample	Does not discuss weights	Table A.9
Liu	Probability sample	Does not discuss weights	Table A.9
Schneider et al.	Probability sample	Does not discuss weights	Table A.9
Wedow et al.	Probability sample	Does not discuss weights	Table A.9
Barber et al.	Probability sample	Ok	
Bloome et al.	Probability sample	Ok	
Ciocca Eller and DiPrete	Probability sample	Ok	
Desmond and Travis	Probability sample	Ok	
Ludwig and Brüderl	Probability sample	Ok	
Villarreal and Tamborini	Probability sample	Ok	
Ferguson and Koning	Administrative records	Are records the full population?	Table A.10
Goldstein	Administrative records	Are records the full population?	Table A.10
McDonnell and King	Administrative records	Are records the full population?	Table A.10
Mun and Jung	Administrative records	Are records the full population?	Table A.10
Wilmsers	Administrative records	Are records the full population?	Table A.10
Font et al.	Administrative records	Ok	
Freeland and Hoey	Other sample	Diversity of sample	Table A.11
Mize and Manago	Other sample	Diversity of sample	Table A.11
Offer and Fischer	Other sample	Diversity of sample	Table A.11
Quadlin	Other sample	Diversity of sample	Table A.11
Simpson et al.	Other sample	Diversity of sample	Table A.11
Weisshaar	Other sample	Diversity of sample	Table A.11
Hahl et al.	Other sample	Similarity to the U.S. population	Table A.12
Schilke and Rossman	Other sample	Similarity to the U.S. population	Table A.12
Kadivar	Other sample	Ok	
Mullins et al.	Other sample	Ok	

Table A.2: Review of target populations. Systematic coding of individual papers from *ASR* 2018. Table A.9–A.12 justify the coding of each paper.

Authors	Title	Underspecified mediation claim
Liu	Social and genetic pathways in multigenerational transmission of educational attainment	Targets mediation estimands such as “the effect of parents’ genotypes on children’s education net of children’s genotype” (p. 283) or how the relationship between parent and child education is mediated by child genotype. Focusing on the latter, this could be a mediation claim about the effect of parents’ education on children’s education that would persist in a counterfactual world where children’s genotype took a particular value. While the authors use Sobel’s test to make claims about the “strength” of this mediation—for instance, “The parent-child association in education is reduced by...7.2 percent when child’s PGS is controlled” (p. 288), the direct effect might be different at different mediator values. To what value is the mediator counterfactually set?
Wedow et al.	Education, smoking, and cohort change: Forwarding a multidimensional theory of the environmental moderation of genetic effects	“Mediated pleiotropy exists when a genetic variant has a direct effect on only one of the phenotypes under consideration; its association with the second phenotype is mediated by the first. For example, mediated pleiotropic effects will arise if genes affect educational attainment, and education affects smoking,” (p. 803). The direct effect of genes on smoking may be different in counterfactual worlds in which education is set to different values. To which value is education counterfactually set?
Barber et al.	The dynamics of intimate partner violence and the risk of pregnancy during the transition to adulthood	“Panel B in Table 6 presents a summary of the results of formal tests of mediation (using the ‘ldecompose’ command in Stata), which indicates the percent of the total effect of recent IPV on pregnancy that is mediated by (i.e., an indirect effect through) pregnancy desire, sex, and contraceptive use,” (p. 1033). The authors cite a software package to be clear about what was done to data, and note, for instance, that the three mediators “mediate 95 percent of the total effect of recent physical assault on pregnancy rates” (p. 1039). But indirect effects require conceptual definitions of counterfactual outcomes that would have been realized if the mediators took different values. And as we discuss in the subsection immediately preceding this paper’s discussion, in cases of multiple mediators, the authors need to state these levels for each combination—for instance, women who desire pregnancy but who have sex with a consistent contraceptive; women who do not desire pregnancy and who have sex with a consistent contraceptive; and so on. With these unstated, the unit-specific quantity at the core of the causal claim is not clear.
Desmond and Travis	Political consequences of survival strategies among the urban poor	“To better understand the mediation process proposed in Figure 1, we decompose the direct, indirect, and total effects of local assistance on perceived political capacity (Table 6)...31 percent of the total effect of local assistance on political capacity is suppressed by perceptions of suffering in one’s neighborhood,” (p. 886). To be precise in our framework, these claims would need to formalize the unit-specific quantities at the core of the claims, which would involve counterfactuals defined over the treatment and mediator. The authors cite a paper on natural direct effects (Imai et al., 2011), but they frame it as a robustness check and (1) do not discuss that it entails holding the mediator to a value it would realize under a particular treatment or (2) that they need to assume away any confounding of the $M \rightarrow Y$ relationship in addition to the $D \rightarrow Y$ relationship.

Table A.3: Difficulty translating the unit-specific quantity to our framework: Underspecified mediation in observational studies (4 articles). As discussed in Sec. 2.3.3, we interpret any claim about a “direct effect” or an effect “net of” or “mediated by” something as a causal claim about the effect of a treatment on an outcome in a possibly counterfactual world where the mediator was set to a specified value. In our framework, direct effects therefore require a statement of the value at which the mediator is fixed by a hypothetical intervention. The magnitude of the direct effect may be different at different values of the mediator.

Authors	Title	Post-treatment conditioning
McDonnell and King	Order in the court: How firm status and reputation shape the outcomes of employment discrimination suits	The treatment is corporate prestige or status (measured by surveys administered in year $t - 1$ and released in year t), but the authors condition on current-year (t) measures such as a firm's cash, logged assets, and number of employees that could be influenced by $t - 1$ perceptions of that firm.
Mun and Jung	Policy generosity, employer heterogeneity, and Women's employment opportunities: The welfare state paradox reexamined	The treatment is a change in paid childcare leave policies in a particular year (1992). The authors control for time-varying covariates, such as the percent female among white collar employees, in order to look at factors that "affect a firm's response to family policy reforms" (p. 518). However, since these variables are measured in years after 1992, they may be consequences of changes in childcare leave policy rather than causes. The authors thus end up conditioning on a post-treatment variable. The authors could have examined how the policy's effects varied across firms with different levels of these characteristics pre-reform or targeted a mediation claim.
Font et al.	Permanency and the educational and economic attainment of former foster children in early adulthood.	Examines the effect of foster care exit type on high school and college enrollment. In one set of models, the authors condition on high school completion: "The conditional (on high school graduation) models suggest that these associations are not entirely due to differences in high school graduation," (p. 728). This is a mediation claim about how the effect of exit type on college enrollment may be mediated by high school graduation but is not defended as such.
Kadivar	Mass mobilization and the durability of New democracies	The treatment is the duration of mobilization for democracy, the outcome is failure of the democracy, and the controls include GDP per capita. Controls like GDP, which are potentially impacted by mobilization, are either measured in the same year as the treatment or after. Because the controls are not lagged temporally relative to the treatment, we interpret them as possibly post-treatment.
Schneider et al.	Income inequality and class divides in parental investments	The treatment is inequality in state t at year $t - 1$, but the authors control for post-treatment variables like work hours of each parent in year t .
Inanc	Unemployment, temporary work, and subjective well-being: the gendered effect of spousal labor market insecurity	The treatment is partner's labor market insecurity, but the author controls for partner's life satisfaction in the same survey wave, which seems likely to be a consequence of the treatment.
Gauchat and Andrews	The cultural-cognitive mapping of scientific professions	One of the treatments, in addition to a scientific literacy scale, is the number of college science courses the GSS cross-sectional respondents took. For most respondents, these courses were taken in the past, but the authors control for contemporaneous variables that the respondent's past college courses might impact, including political identity and family income.
Gutierrez	The institutional determinants of health insurance: Moving away from labor market, marriage, and family attachments under the ACA	The treatment is the passage of the Affordable Care Act, comparing the pre-ACA period (2009-2012) to the post-ACA period (2014-2016). But the author conditions on variables that could be consequences of the treatment such as income and health in the post-ACA period.
Villarreal and Tamborini	Immigrants' economic assimilation: evidence from longitudinal earnings records	The treatment is years since an immigrant's arrival in the U.S., but the authors control for variables that could be consequences of immigration such as years of work experience
Ludwig and Brüderl	Is there a male marital wage premium? New evidence from the United States	Examines the effect of marriage on men's wages but controls for consequences of marriage including number of children, tenure in the current job, and work experience.

Table A.4: Difficulty translating the unit-specific quantity to our framework: Post-treatment conditioning (10 articles). These studies potentially condition on a post-treatment variable; they either adjust for (1) a covariate that temporally follows a treatment, (2) a covariate measured in the same time period as the treatment and is potentially caused by it, or (3) are not clear enough about the temporal and causal ordering of covariates and treatments to determine. Is the target a mediation claim about a direct effect? If so, then making the target explicit would facilitate clear reasoning about the goal and the identification assumptions required to estimate the causal effect of the mediator. If not, then the authors have made an identification error by conditioning on a post-treatment variable. With these papers, we are unsure which has occurred.

Authors	Title	Issue
Barr et al.	Sharing the burden of the transition to adulthood: African American young adults' transition challenges and their mothers' health risk	The primary independent variable is a continuous measure reflecting different domains of challenges that a mother's young adult child (YAC) may face: "These challenges include unemployment, high levels of racial discrimination, educational disengagement, troubled romantic relationships, unmarried parenthood, and arrest" (p. 152). It is difficult to believe that the effect on health outcomes is the same from having an increase of 1 in this count of challenges regardless of which challenge is added—for instance, a YAC being arrested versus not being enrolled in higher education. The authors' secondary specifications—any challenge versus no challenge; 2+ challenges versus no challenge—still groups challenges together. In each case, the treatment is a bundle of things that is not clear; a precise statement of the unit-specific quantity would formalize each treatment being assigned to a particular value.
Wilmers	Wage stagnation and buyer power: How buyer-supplier relations affect US workers' wages, 1978 to 2014	" β_1 is the effect of an increasing share of suppliers' revenue coming from dominant buyers (x_{it})," (p. 219). For a continuous treatment x_{it} , a well-defined unit-specific quantity would involve a causal contrast between two specific values of the treatment, such as the effect of a shift from 50% to 60% of revenue coming from dominant buyers. The "effect" defined as the regression coefficient is only well-defined if the parametric linearity assumption holds: if a fixed increase in the treatment has the same effect regardless of the initial value of the treatment. The causal contrast—a general "effect"—is therefore defined only under the parametric modeling assumptions.
Goldstein	The social ecology of speculation: Community organization and non-occupancy investment in the US housing bubble	The author hypothesizes about the effects of several variables on non-occupancy investment (NOI)—for instance, a greater share of community-based organizations or a lower rate of residential vacancies being associated with less NOI. Yet the regressions examine the effects of these variables simultaneously. For instance, Model 2 examines the effects of annual appreciation and a regulatory index simultaneously; model 4 adds five more variables of interest. These models cannot simultaneously identify the causal effects of all the variables except in very special circumstances, such as when the variables do not affect each other.
Horowitz	Relative education and the advantage of a college degree	"To test the relative education and skill-biased technological change hypotheses, I report three sets of independent variables: variables for cohort-level educational attainment, individual-level educational attainment, and an interaction between the two to show how the returns to education change as educational attainment increases across cohorts," (p. 780). Although often phrased in association terms, the claims are causal. For instance, "greater education generally leads to more skill usage, but this effect weakens when more people earn college degrees," (p. 785). But which of these many variables is subject to a hypothetical intervention? A single model cannot simultaneously identify the causal effects of all the variables except in very special circumstances, such as when the variables do not affect each other.

Table A.5: Difficulty translating the unit-specific quantity to our framework: Other issues (4 articles). These issues include continuous treatment variables estimated with a linear functional form for which the precise contrast between values of the treatment is not clear, as well as studies with many predictors of which it is not clear which are viewed as causal treatments. In each case, either (1) we are unsure which variables are the treatments that define unit-specific counterfactual quantities as opposed to distinguishing separate populations of units, (2) we are unsure of the specific treatment values over which the causal effect is contrasted, or (3) we are unsure of both.

Authors	Title	Underspecified mediation claim
Schilke and Rossman	It's only wrong if it's transactional: Moral perceptions of obfuscated exchange	A hypothesis stated by the authors exemplifies an underspecified mediation estimand: "Perceived attributional opacity mediates the negative effect of structural obfuscation (versus taboo quid pro quo) on moral disapproval," (p. 1088)." Although structural obfuscation is randomized, perceived attributional opacity is not. Mediation depends on identifying the causal effect of the mediator. Further, the degree to which opacity mediates this effect may depend on the value to which opacity is hypothetically set.
Simpson et al.	The roots of reciprocity: Gratitude and reputation in generalized exchange systems	The authors write that "feelings of gratitude partially mediated the relationship between first-mover and second-mover giving" (p. 103). The direct effect of first-mover giving on second-mover giving may depend on the second-mover's specific value of gratitude, however. The value to which the mediator is hypothetically fixed needs to be specified. In addition, the authors do not discuss the assumptions required to identify the causal effect of the gratitude mediator. They do randomize the timing of the mediator's measurement—either before or after the second-movers decision—to reduce priming effects, but randomizing the timing of measurement does not randomize the value of the gratitude mediator. There may remain other unmeasured respondent attributes—e.g., prosociality—that confound the relationship between the mediator and outcome).
Quadlin	The mark of a woman's record: Gender and academic performance in hiring	"The effect of high achievement is no longer significant in this model, implying that perceptions of competence and commitment account for the relationship between achievement and callbacks.," (p. 347). The size of the direct effect may depend on the value to which perceptions of competence and commitment are hypothetically fixed, however. Further, the causal effect of perceived competence and commitment is not identified by randomizing the treatment—high achievement—only.

Table A.6: Difficulty translating the unit-specific quantity to our framework: Underspecified mediation in experiments (3 articles). Randomization of treatment assignment in an experiment makes the unit-specific quantity unambiguous for the total effect: the difference between the potential outcomes under two treatment conditions. The direct effect net of a non-randomized mediator, however, suffers from the same issues common in observational studies targeting mediation estimands (Table A.3). As discussed in Sec. 2.3.3, we interpret any claim about a "direct effect" or an effect "net of" or "mediated by" something as a causal claim about the effect of a treatment on an outcome in a possibly counterfactual world where the mediator was set to a specified value. In our framework, direct effects therefore require a statement of the value at which the mediator is fixed by a hypothetical intervention. Further, even in experiments with randomized treatments, mediation claims rely on assumptions to identify the causal effect of the mediator.

Authors	Title	Inference to a non-existent population without explicit reference to an intervention.
Ferguson and Koning	Firm turnover and the return of racial establishment segregation	The authors’ statement of their aims suggests a goal of describing segregation in a population: “this type of decomposition is necessary to describe, explain, and evaluate countervailing trends like those we explore here,” (p. 446). However, they describe populations different from the actual population in ways that require careful definition and assumptions. We propose that an explicitly causal framing could clarify claims such as the following: “...the two types of segregation were nearly equal contributors to total employment segregation in the United States. At that time, either reallocating workers across occupations within establishments or reallocating workers across establishments without altering occupational segregation would have had roughly the same impact on total segregation,” (p. 456-457). This “would” statement appeals to a counterfactual situation; in our framework, any counterfactual situation involves a causal claim about unit-specific outcomes that would be realized under a well-specified intervention.
Ciocca Eller and DiPrete	The paradox of persistence: explaining the black-white gap in bachelor’s degree completion	“One statistical approach we use across all steps, however, is Fairlie’s (2005) decomposition technique. This technique uses logistic regression and counterfactual substitution of coefficient values to assess the difference in outcomes between two groups. It specifically isolates the proportion of the overall difference that is accounted for by group differences in covariates versus group differences in coefficients. We use black students as our reference group for the Fairlie decomposition, as well as randomized variable inputs to ensure robustness of results, although the outcomes are similar regardless of whether we use black, white, or pooled students as the reference group and regardless of whether or not we insert variables randomly. We also compute Fairlie decompositions using the polychoric factors described in the previous section, rather than individual variables, to increase interpretability; both strategies produce comparable results,” (p. 1181). The resulting quantities border on causal interpretations. For example, they describe the racial gap in college dropout after “counterfactually shifting the dropout risk distribution of entering black students so that it more closely resembles the distribution for white students,” (p. 1187). In our framework, a counterfactual situation involves a causal claim about unit-specific outcomes that would be realized under a well-specified intervention.
Guinea-Martin et al.	The evolution of gender segregation over the life course	“In other words, in a hypothetical situation where occupations are completely integrated, 44.4 percent of gender segregation (14.5 points) would remain at this age due to the contributions of economic and time-related segregation—which amount to 17.1 and 27.3 percent of overall gender segregation, respectively (see Table 9),” (p. 1008). This “would” statement appeals to a counterfactual situation; in our framework, a counterfactual situation involves a causal claim about unit-specific outcomes that would be realized under a well-specified intervention.
Bloome et al.	Educational inequality, educational expansion, and intergenerational income persistence in the United States	While the authors note “our model-based approach is descriptive” (p. 1225), and draw a distinction between their approach and “an alternative approach focused on identifying the causal effects of education on income persistence,” they then try to infer what the between-generation correlation in incomes would have been if there had not been a simultaneous educational expansion: “But if the only changes across cohorts were rising educational inequality and growing educational returns, we predict the correlation would have increased .06 points on the correlation scale, or about 14 percent of the baseline persistence level of .43,” (p. 1230). This “would” statement appeals to a counterfactual situation; in our framework, the theoretical estimand invoking a counterfactual situation must involve a causal claim about unit-specific outcomes that would be realized under a well-specified intervention.

Table A.7: Difficulty translating the unit-specific quantity to our framework: Inference to non-existent populations (4 articles).

Some studies reweight a population so that a predictor follows a different distribution from that observed in the factual population. In practice, many such articles formally appeal to non-causal counterfactuals (or more informally what ‘would have happened’). For instance, researchers might investigate how outcomes would change for group A if that group’s characteristics were made more similar to group B (e.g., how would male students’ rates of college entry change if one shifted their actual high school achievement to the counterfactual high school achievement equivalent to female students?). Our definition of the theoretical estimand rules out these settings because they don’t describe the state of the world where such data would arise and thus cannot adjudicate between alternative adjustment strategies. These studies could be brought into our framework by either describing an existing population or invoking unit-specific causal counterfactuals under a well-defined intervention.

Authors	Title	Causal verbs attributed to variables
Offer and Fischer	Difficult people: Who is perceived to be demanding in personal networks and why are they there?	“The major contribution of this study is our examination of the different types of constraints that may pressure people to interact with others whom they would otherwise prefer to avoid or disengage from,” (p. 133). The causal verb “pressure” suggests a causal effect of these constraints on interactions.
Freeland and Hoey	The structure of deference: Modeling occupational status using affect control theory	Primarily describes a new occupational deference score and establishes descriptive construct validity. However, some subsequent claims use causal language. For instance, the authors describe how they investigate the occupational deference scores’ criterion validity, examining the relationship between the scores and various job-related measures of subjective well-being. While the aim of the criterion validity analysis is descriptive—“to determine whether deference scores empirically perform as theoretically predicted, by displaying statistically significant effects net of other variables in the right direction” (p. 260)—the interpretations of the results lean toward causal claims. For example, the phrases “leads to” and “increase” in the following discussion have causal connotations: “if deference leads to greater satisfaction at work, it should also lead to greater respect outside the work environment, and thus should increase general happiness” (p. 260).

Table A.8: Difficulty translating the unit-specific quantity to our framework: Causal verbs attributed to predictors (2 articles). In these articles, the primary component of the paper is purely descriptive, yet at least one analysis or interpretation pairs predictor variables with causal verbs that point toward at least some interest in a causal claim. If interpretations give an active role to a variable as (for example) shaping, leading to, pressuring, or shifting some outcome, those are claims that we interpret as implicitly implying a counterfactual defined over that variable: the outcome would have been different if the variable took a different value. In our framework, any counterfactual statement is ambiguous unless it is clearly tied to a causal claim about unit-specific outcomes that would be realized under a well-specified intervention. Alternately, the authors could interpret the results in solely predictive, non-causal terms, and be clear about how the purely descriptive quantity is relevant to theory.

Authors	Title	Probability samples for which weights are not discussed
Liu	Social and genetic pathways in multigenerational transmission of educational attainment	Uses the Framingham Heart Study and the Health and Retirement Study but does not discuss weights or whether inference beyond these samples, especially Framingham, is relevant to the theoretical claim.
Schneider et al.	Income inequality and class divides in parental investments	Uses the Consumer Expenditure Survey and the American Heritage Time Use Survey but does not discuss survey weights.
Inanc	Unemployment, temporary work, and subjective well-being: the gendered effect of spousal labor market insecurity	Uses the British Household Panel Study but does not discuss weights.
Gauchat and Andrews	The cultural-cognitive mapping of scientific professions	Uses the National Science Foundation's Science and Technology Survey (2004) and the version that was embedded in the GSS (2006; 2012) but does not discuss survey weights.
Horowitz	Relative education and the advantage of a college degree	Uses the Current Population Survey but does not discuss weights.
Wedow et al.	Education, smoking, and cohort change: Forwarding a multidimensional theory of the environmental moderation of genetic effects	Uses the Health and Retirement Study and the National Longitudinal Study of Adolescent to Adult Health but does not discuss weights.
Flores and Schachter	Who are the "illegals"? The social construction of illegality in the United States	"GfK recruits respondents by beginning with an addressed-based sampling frame based on a USPS database of all addresses in the United States. Respondents are then randomly selected and invited to join the GfK database using probability-based sampling methods," (p. 846). The authors do not discuss weights, but it is difficult to believe that this sampling procedure produces an equal-probability sample.
Guinea-Martin et al.	The evolution of gender segregation over the life course	This study uses a weighted decomposition, but we interpret these weights as the prevalence of demographic groups in the sample rather than involving the survey weights.
Gutierrez	The institutional determinants of health insurance: Moving away from labor market, marriage, and family attachments under the ACA	Uses the National Survey on Drug Use and Health and weights the descriptive statistics using the survey weights (Table 2), but does not discuss the weights in the regression specification.
Barr et al.	Sharing the burden of the transition to adulthood: African American young adults' transition challenges and their mothers' health risk	Data are from the Family and Community Health Study. The authors never report using the survey weights. The online documentation for the study does not clarify if it is a probability sample or if weights are available.

Table A.9: Difficulty translating the target population to our framework: Ambiguity about survey weights (9 articles). Is the sample of theoretical interest in itself, or do the authors seek to draw inference to a broader target population? If the target is the sample, weights are unneeded. If the target is the sampling frame, weights are needed. If the target is broader than the sampling frame, additional argument is needed about how the estimates are informative about that broader population. When studies do not discuss the use or lack of use of survey weights in the main analysis, we cannot unambiguously say what the target population is. That hinders methodological choices: the correct use of weights cannot be determined when the target population is ambiguous.

Authors	Title	Administrative records: Are these a census of the target population or not?
McDonnell and King	Order in the court: How firm status and reputation shape the outcomes of employment discrimination suits	“We empirically analyze a unique database of more than 500 employment discrimination suits brought between 1998 and 2008,” (p. 61), which correspond to discrimination suits with a jury verdict against the 826 companies surveyed for the <i>Fortune 1000</i> status ratings. The authors note that analyzing these firms “does limit the scope of our findings to firms that are fairly large and visible, which are likely to have the kind of well established reputations and status orderings that are perceptible to lay audience members” (p. 68). The sample is described clearly, but are these large and visible firms the entire target population or is the target population some broader category of firms, as suggested by the authors’ use of standard errors?
Goldstein	The social ecology of speculation: Community organization and non-occupancy investment in the US housing bubble	“I assess the hypotheses by applying panel regression techniques to an annual city-level dataset covering the years 2000 to 2006. The sample is composed of 1,565 census places, which are nested within 475 counties...The sample is drawn from all places in the contiguous United States with population over 20,000 circa 2007 (n = 1,999). Censoring in the 2005 to 2007 ACS (to protect confidentiality in small areas) reduced the sample by 341. Unavailable housing price data further reduced the sample by 93 places. This leaves an analytical sample of 1,565 places” (p. 1118). Is the target population all places, and the omission of small areas is a limitation? Or is the target population all places exceeding a particular population? This is an example where the sample is described clearly, but it is not clear whether the sample contains the entire target population.
Wilmer	Wage stagnation and buyer power: How buyer-supplier relations affect US workers’ wages, 1978 to 2014	“I focus on publicly traded corporations...Publicly traded U.S. companies account for 37 percent and 30 percent of total U.S. employment in 1978 and 2014, respectively. These firms tend to be large and well-capitalized, and they are relatively concentrated in finance and manufacturing industries. In robustness tests, I assess whether the results found for publicly traded firms are biased by this skewed industry composition,” (p. 217-218). But if the target population is publicly traded firms, then how could the composition of those firms induce a bias? There is only a bias if the target population is all firms. For this reason, the target is unclear.
Ferguson and Koning	Firm turnover and the return of racial establishment segregation	“To measure workplace segregation, we leverage the population of EEO-1 establishment surveys gathered by the Equal Employment Opportunity Commission (EEOC) over more than four decades, from the early 1970s until recently. These surveys provide annual data on workforce racial composition for every large private-sector establishment in the U.S. economy,” (p. 446)...“In 2007, the EEOC made several important changes to the EEO-1 form and reporting process. They began collecting data once again from establishments whose size is below the mandatory reporting threshold. Participation in this new program is not mandatory for small establishments, however, so patterns in those data are not necessarily representative of national trends. We exclude these smaller establishments from our analyses” (p. 454). The note about excluding small establishments that are below the size requirements that trigger mandatory reporting suggests that the author’s target is large firms subject to mandated EEO reporting. Is the entire target population contained in the data?
Mun and Jung	Policy generosity, employer heterogeneity, and Women’s employment opportunities: The welfare state paradox reexamined	“Although it is not a representative sample of the entire population of firms, it includes most of the prominent firms that lead the Japanese business community,” (p. 516). Is the target population this sample of prominent firms, or is it the entire population of firms?

Table A.10: Difficulty translating the target population to our framework: Are administrative records a census of the entire target population? (5 articles). We are often unsure whether the population that appears in administrative records is the population of substantive interest, or whether authors seek to reach further to a broader population. Whether the administrative records contain the entire target population is also necessary to guide methodological choices such as the construction of standard errors.

Authors	Title	Other samples: Defenses that appeal to diversity
Simpson et al.	The roots of reciprocity: Gratitude and reputation in generalized exchange systems.	“Although not representative of the general population, MTurk samples are substantially more diverse than many other types of convenience samples,” (p. 101). About what target population is this diverse sample intended to be informative?
Offer and Fischer	Difficult people: Who is perceived to be demanding in personal networks and why are they there?	“Drawing on a survey of over 1,100 diverse respondents,” (p. 111)...“Our study uses a larger and more diverse sample and expands what we know about these relationships...” (p. 113). About what target population is this diverse sample intended to be informative?
Mize and Manago	Precarious sexuality: How men and women are differentially categorized for similar sexual behavior	“We recruited participants from Amazon Mechanical Turk (mTurk) for Studies 2A and 2B to obtain a diverse sample of U.S. adults,” (p. 317). Is the target population all U.S. adults, or just this sample? The fact that the sample is diverse does not imply that it yields reliable estimates for the target population of U.S. adults.
Weisshaar	From opt out to blocked out: The challenges for labor market re-entry after family-related employment lapses	“The job listings were sampled from 50 major metropolitan areas in the United States, allowing for a range of labor market contexts,” (p. 45). Is the target population all major metropolitan areas? What is the definition of “major”? Capturing a diversity of labor market contexts does not in itself define the target population for which the sample may be informative.
Freeland and Hoey	The structure of deference: Modeling occupational status using affect control theory	“We extracted a representative list of occupations,” (p. 250). The procedure to select these roughly 300 occupations involves sampling high, medium, and low-income populations from the major occupational groupings. But is the target to learn, from this sample of occupations, about an association over the unweighted population of all occupations or, for instance, a population of occupations weighted by their size? This is unclear because the procedure does not appeal to a clear target population of occupations, but rather defends the selection of certain occupations based on sampling from diverse categories.
Quadlin	The mark of a woman’s record: Gender and academic performance in hiring	The author appeals to diversity: “To maximize the overall sample size and enhance the geographic diversity of the sample, applications were submitted to positions in major metropolitan areas corresponding to five regions of the United States,” (p. 338). The author restricts to full-time job openings within a 30-mile radius, excludes “entry-level” or “general” jobs, and retained only the most recent job posted by each employer. We agree that the sample is diverse, yet the target population is difficult to define. Is it the population of full-time job openings, weighted so that each employer is given equal weight, in major metropolitan areas (where “major” is undefined) of the U.S. within 30 miles of the city center, which are not entry-level or general jobs? And does the target population only include jobs posted on the chosen job-search website, or does the author want to argue that results are informative about all jobs? One must either argue that the sample is of genuine theoretical interest or how it is informative about a broader population. Clarity would enable future studies to extend the results through designs intended to generalize to a target population of more general interest.

Table A.11: Difficulty translating the target population to our framework: Appeals to diversity (5 articles).

Articles with non-probability samples often appeal to the diversity of cases represented in the sample. However, diversity is only helpful if it makes the sample more informative about some diverse target population. In these studies, we are unsure what the target population is. Knowing the target population would inform the assumptions required to draw inference about this population and could direct methodological choices to construct and weight the sample to yield better estimates. While the target population remains unstated it is difficult to reason about these methodological choices.

Authors	Title	Other samples: Defenses that appeal to similarity to the U.S. population
Hahl et al.	The authentic appeal of the lying demagogue: Proclaiming the deeper truth about political illegitimacy.	“MTurk has been used widely in experimental research and has been found to provide a subject pool that is slightly more educated and technologically savvy than the national average,” (p. 10).
Schilke and Rossman	It’s only wrong if it’s transactional: Moral perceptions of obfuscated exchange	“Although AMT [Amazon Mechanical Turk] respondents are not perfectly representative of the U.S. population (e.g., they are more politically liberal on average), they do feature substantial demographic diversity and are thus often preferable to college student samples, especially for research concerning political issues,” (p. 1090).

Table A.12: Difficulty translating the target population to our framework: Similarities to the U.S. population (2 articles). Unlike studies that simply appeal to diversity, these studies additionally appeal to similarities of the characteristics of those in the sample to the characteristics of the U.S. population. Is the target then the U.S. population? If so, one could imagine reweighting the sample to yield better estimates for this target population. If the target is only the sample, and this is just a way of describing the sample, then reweighting would be inappropriate. Appropriate methodological guidance depends on the target population.

Appendix B

Supplementary materials for Chapter 3

B.1 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.2.

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.

$$\theta_x(t) = \mathbf{E}(Y(t) \mid X = x) \tag{B.1.1}$$

By the law of iterated expectation

$$= \mathbf{E} \left(\mathbf{E} \left(Y(t) \mid X = x, \vec{L} \right) \mid X = x \right) \tag{B.1.2}$$

By conditional mean independence

$$= \mathbf{E} \left(\mathbf{E} \left(Y(t) \mid X = x, \vec{L}, T = t \right) \mid X = x \right) \tag{B.1.3}$$

By consistency

$$= \mathbf{E} \left(\mathbf{E} \left(Y \mid X = x, \vec{L}, T = t \right) \mid X = x \right) \tag{B.1.4}$$

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.

$$\theta_x(\pi) = \mathbf{E} \left(\sum_t \pi(t, \vec{L}) Y(t) \mid X = x \right) \quad (\text{B.1.5})$$

By the law of iterated expectation

$$= \mathbf{E} \left(\sum_t \mathbf{E} \left(\pi(t, \vec{L}) Y(t) \mid X = x, \vec{L} \right) \mid X = x \right) \quad (\text{B.1.6})$$

Pull constant out of the inner $\mathbf{E}$

$$= \mathbf{E} \left(\sum_t \pi(t, \vec{L}) \mathbf{E} \left(Y(t) \mid X = x, \vec{L} \right) \mid X = x \right) \quad (\text{B.1.7})$$

By conditional mean independence

$$= \mathbf{E} \left(\sum_t \pi(t, \vec{L}) \mathbf{E} \left(Y(t) \mid X = x, \vec{L}, T = t \right) \mid X = x \right) \quad (\text{B.1.8})$$

By consistency

$$= \mathbf{E} \left(\sum_t \pi(t, \vec{L}) \mathbf{E} \left(Y \mid X = x, \vec{L}, T = t \right) \mid X = x \right) \quad (\text{B.1.9})$$

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.2 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.2.1 defines the notation to be used throughout the proofs. Section B.2.2 proves that estimation with predicted outcomes is consistent if the outcome prediction function is consistent. Section B.2.3 proves that estimation with predicted treatment probabilities is consistent if the treatment prediction function is consistent. Sections B.2.4 and B.2.5 show that doubly-robust estimation is consistent if the prediction function is consistent for outcomes or treatment probabilities, respectively.

B.2.1 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.

$$P(\text{Event} \mid \text{Condition}) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i: \text{Condition}_i = \text{True}} (\text{Event}_i = \text{True}) \quad (\text{B.2.1})$$

$$E(\text{Outcome} \mid \text{Condition}) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i: \text{Condition}_i = \text{True}} \text{Outcome}_i \quad (\text{B.2.2})$$

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

$$\frac{1}{n} \sum_{i: \text{Condition}_i = \text{True}} \text{Outcome}_i \rightarrow E(\text{Outcome} \mid \text{Condition}) \quad (\text{B.2.3})$$

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 .

$$\eta_x = \sum_{i: X_i=x} \eta_i, \quad \eta_{x\vec{\ell}} = \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}} \eta_i, \quad \eta_{x\vec{\ell}t} = \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \eta_i \quad (\text{B.2.4})$$

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

$$\hat{g}(t, x, \vec{\ell}) \xrightarrow{P} g(t, x, \vec{\ell}) = \mathbf{E} \left(Y \mid T = t, X = x, \vec{L} = \vec{\ell} \right) \quad \forall t, x, \vec{\ell} \quad (\text{B.2.5})$$

or the treatment prediction function,

$$\hat{m}(t, x, \vec{\ell}) \xrightarrow{P} m(t, x, \vec{\ell}) = \mathbf{P} \left(T = t \mid X = x, \vec{L} = \vec{\ell} \right) \quad \forall t, x, \vec{\ell} \quad (\text{B.2.6})$$

B.2.2 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.

$$\hat{\theta}_x^{\text{Outcome}}(\pi) = \frac{1}{\eta_x} \sum_{i: X_i=x} \eta_i \sum_t \hat{\pi}(t, \vec{L}_i) \hat{g}(t, X_i, \vec{L}_i) \quad (\text{B.2.7})$$

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

$$= \frac{1}{\eta_x} \sum_{\vec{\ell}} \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}} \eta_i \sum_t \hat{\pi}(t, \vec{L}_i) \hat{g}(t, X_i, \vec{L}_i) \quad (\text{B.2.8})$$

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

$$= \sum_{\vec{\ell}} \frac{\eta_{x\vec{\ell}}}{\eta_x} \sum_t \frac{1}{\eta_{x\vec{\ell}}} \underbrace{\sum_{i: X_i=x, \vec{L}_i=\vec{\ell}} \eta_i \hat{\pi}(t, \vec{L}_i) \hat{g}(t, X_i, \vec{L}_i)}_{\text{Weighted sample mean of a constant}} \quad (\text{B.2.9})$$

$$= \sum_{\vec{\ell}} \underbrace{\frac{\eta_{x\vec{\ell}}}{\eta_x}}_{\rightarrow \mathbf{P}(\vec{L}=\vec{\ell} \mid X=x)} \sum_t \underbrace{\hat{\pi}(t, \vec{\ell})}_{\rightarrow \pi(t, \vec{\ell})} \underbrace{\hat{g}(t, x, \vec{\ell})}_{\rightarrow g(t, x, \vec{\ell})} \quad (\text{B.2.10})$$

$$\rightarrow \sum_{\vec{\ell}} \mathbf{P}(\vec{L} = \vec{\ell} \mid X = x) \sum_t \pi(t, \vec{\ell}) g(t, x, \vec{\ell}) \quad (\text{B.2.11})$$

Replace $g(\cdot)$ with its definition

$$\rightarrow \sum_{\vec{\ell}} \mathbf{P}(\vec{L} = \vec{\ell} \mid X = x) \sum_t \pi(t, \vec{\ell}) \mathbf{E}(Y \mid T = t, X = x, \vec{L} = \vec{\ell}) \quad (\text{B.2.12})$$

Invoke causal identification assumptions (B.2.13)

$$\rightarrow \sum_{\vec{\ell}} \mathbf{P}(\vec{L} = \vec{\ell} \mid X = x) \sum_t \pi(t, \vec{\ell}) \mathbf{E}(Y(t) \mid X = x, \vec{L} = \vec{\ell}) \quad (\text{B.2.14})$$

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

$$\rightarrow \mathbf{E} \left(\sum_t \pi(t, \vec{L}) Y(t) \mid X = x \right) \quad (\text{B.2.15})$$

$$\rightarrow \theta_x(\pi) \quad (\text{B.2.16})$$

B.2.3 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 $m()$ is a correctly-specified logistic regression and if $\pi(t, \vec{\ell})$ is $m()$ 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.

$$\hat{\theta}_x^{\text{Treatment}} = \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)} Y_i \quad (\text{B.2.17})$$

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

$$= \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{\vec{\ell}} \sum_t \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \eta_i Y_i \quad (\text{B.2.18})$$

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

$$= \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{\vec{\ell}} \sum_t \frac{\eta_{x\vec{\ell}t}}{\eta_{x\vec{\ell}t}} \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \underbrace{\frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})}}_{\text{Constant within inner sum}} \eta_i Y_i \quad (\text{B.2.19})$$

Pull the constant out of the inner sum and rearrange

$$= \sum_{\vec{\ell}} \sum_t \underbrace{\frac{\eta_{x\vec{\ell}t}}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}}}_{\substack{\rightarrow \text{P}(\vec{L}=\vec{\ell}, T=t | X=x) \\ \text{Proof in Eq B.2.25--B.2.31}}} \underbrace{\frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})}}_{\substack{\rightarrow \frac{\pi(t, \vec{\ell})}{\text{P}(T=t, | X=x, \vec{L}=\vec{\ell})}}} \underbrace{\frac{1}{\eta_{x\vec{\ell}t}} \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \eta_i Y_i}_{\rightarrow \text{E}(Y | X=x, \vec{L}=\vec{\ell}, T=t)} \quad (\text{B.2.20})$$

$$\rightarrow \sum_{\vec{\ell}} \sum_t \underbrace{\frac{\text{P}(\vec{L}=\vec{\ell}, T=t | X=x)}{\text{P}(T=t, | X=x, \vec{L}=\vec{\ell})}}_{\substack{= \text{P}(\vec{L}=\vec{\ell} | X=x) \\ \text{Proof in Eq. B.2.32--B.2.33}}} \pi(t, \vec{\ell}) \text{E}(Y | X=x, \vec{L}=\vec{\ell}, T=t) \quad (\text{B.2.21})$$

Invoke causal identification assumptions for the expectation

$$\rightarrow \sum_{\vec{\ell}} \sum_t \text{P}(\vec{L}=\vec{\ell} | X=x) \pi(t, \vec{\ell}) \text{E}(Y(t) | X=x, \vec{L}=\vec{\ell}) \quad (\text{B.2.22})$$

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

$$\rightarrow \text{E} \left(\sum_t \pi(t, \vec{L}) Y(t) | X=x \right) \quad (\text{B.2.23})$$

$$\rightarrow \theta_x(\pi) \quad (\text{B.2.24})$$

Eq B.2.20 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.

$$\frac{\eta_{x\vec{\ell}t}}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} = \frac{\eta_{x\vec{\ell}t}}{\sum_{\vec{\ell}} \sum_t \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \eta_i \frac{\hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \quad (\text{B.2.25})$$

Pull out constants from the sum

$$= \frac{\eta_{x\vec{\ell}t}}{\sum_{\vec{\ell}} \sum_t \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \eta_i} \quad (\text{B.2.26})$$

Simplify the inner-most sum

$$= \frac{\eta_{x\vec{\ell}t}}{\sum_{\vec{\ell}} \sum_t \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \eta_{x\vec{\ell}t}} \quad (\text{B.2.27})$$

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

$$= \frac{\frac{1}{\eta_x} \eta_{x\vec{\ell}t}}{\sum_{\vec{\ell}} \frac{\eta_{x\vec{\ell}}}{\eta_x} \sum_t \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \frac{1}{\eta_{x\vec{\ell}}} \eta_{x\vec{\ell}t}} \quad (\text{B.2.28})$$

Replace estimates with their probability limit parameters

$$\rightarrow \frac{P(\vec{L} = \vec{\ell}, T = t \mid X = x)}{\sum_{\vec{\ell}} P(\vec{L} = \vec{\ell} \mid X = x) \sum_t \frac{\pi(t, \vec{\ell})}{P(T=t \mid X=x, \vec{L}=\vec{\ell})} P(T = t \mid X = x, \vec{L} = \vec{\ell})} \quad (\text{B.2.29})$$

Cancel terms in the internal sum in the denominator

$$\rightarrow \frac{P(\vec{L} = \vec{\ell}, T = t \mid X = x)}{\underbrace{\sum_{\vec{\ell}} P(\vec{L} = \vec{\ell} \mid X = x)}_{=1} \underbrace{\sum_t \pi(t, \vec{\ell})}_{=1}} \quad (\text{B.2.30})$$

$$= P(\vec{L} = \vec{\ell}, T = t \mid X = x) \quad (\text{B.2.31})$$

Eq. B.2.21 above posits that a ratio involving conditional probabilities has a simpler form. That is proven below by expanding the numerator and then canceling terms.

$$\frac{P(\vec{L} = \vec{\ell}, T = t \mid X = x)}{P(T = t, \mid X = x, \vec{L} = \vec{\ell})} = \frac{P(T = t \mid X = x, \vec{L} = \vec{\ell}) P(\vec{L} = \vec{\ell} \mid X = x)}{P(T = t, \mid X = x, \vec{L} = \vec{\ell})} \quad (\text{B.2.32})$$

$$= P(\vec{L} = \vec{\ell} \mid X = x) \quad (\text{B.2.33})$$

B.2.4 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,

$$\hat{\theta}_x^{\text{Robust}}(\pi) = \hat{\theta}_x^{\text{Outcome}}(\pi) - \underbrace{\frac{1}{\sum_{i: X_i=x} \hat{w}_i} \sum_{i: X_i=x} \hat{w}_i \left(\hat{g}(T_i, X_i, \vec{L}_i) - Y_i \right)}_{\text{Correction } \hat{\delta}_x^{\text{Outcome}}(\pi)} \quad (\text{B.2.34})$$

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.2.2 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 $\hat{\delta}_x^{\text{Outcome}}(\pi) \rightarrow 0$.

$$\hat{\delta}_x^{\text{Outcome}}(\pi) = \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, X_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)} \left(\hat{g}(T_i, X_i, \vec{L}_i) - Y_i \right) \quad (\text{B.2.35})$$

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

$$= \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{\vec{\ell}} \sum_t \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \eta_i \left(\hat{g}(T_i, X_i, \vec{L}_i) - Y_i \right) \quad (\text{B.2.36})$$

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

$$\frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{\vec{\ell}} \sum_t \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \eta_i \left(\hat{g}(T_i, X_i, \vec{L}_i) - Y_i \right) \quad (\text{B.2.37})$$

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

$$= \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{\vec{\ell}} \sum_t \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \eta_{x\vec{\ell}t} \underbrace{\frac{1}{\eta_{x\vec{\ell}t}} \sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \eta_i \left(\hat{g}(T_i, X_i, \vec{L}_i) - Y_i \right)}_{\substack{\rightarrow 0 \\ \text{Weighted mean of outcome prediction errors} \\ \text{within sample stratum } x, \vec{\ell}, t}} \quad (\text{B.2.38})$$

$$\rightarrow 0 \quad (\text{B.2.39})$$

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

B.2.5 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,

$$\hat{\theta}_x^{\text{Robust}}(\pi) = \hat{\theta}_x^{\text{Outcome}}(\pi) - \frac{1}{\sum_{i: X_i=x} \hat{w}_i} \sum_{i: X_i=x} \hat{w}_i \left(\hat{g}(T_i, X_i, \vec{L}_i) - Y_i \right) \quad (\text{B.2.40})$$

$$\begin{aligned} &= \underbrace{\frac{1}{\sum_{i: X_i=x} \hat{w}_i} \sum_{i: X_i=x} \hat{w}_i Y_i}_{=\hat{\theta}_x^{\text{Treatment}}(\pi)} + \underbrace{\hat{\theta}_x^{\text{Outcome}}(\pi) - \frac{1}{\sum_{i: X_i=x} \hat{w}_i} \sum_{i: X_i=x} \hat{w}_i \hat{g}(T_i, X_i, \vec{L}_i)}_{\text{Correction } \hat{\delta}_x^{\text{Treatment}}(\pi)} \\ &\quad (\text{B.2.41}) \end{aligned}$$

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.2.3 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 $\hat{\delta}_x^{\text{Treatment}}(\pi) \rightarrow 0$.

$$\hat{\delta}_x^{\text{Treatment}}(\pi) = \hat{\theta}_x^{\text{Outcome}}(\pi) - \frac{1}{\sum_{i: X_i=x} \hat{w}_i} \sum_{i: X_i=x} \hat{w}_i \hat{g}(T_i, X_i, \vec{L}_i) \quad (\text{B.2.42})$$

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

$$\begin{aligned} &= \frac{1}{\eta_x} \sum_{i: X_i=x} \eta_i \sum_t \hat{\pi}(t, \vec{L}_i) \hat{g}(t, X_i, \vec{L}_i) \\ &\quad - \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, X_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, X_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)} \hat{g}(T_i, X_i, \vec{L}_i) \quad (\text{B.2.43}) \end{aligned}$$

(continued on next page)

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

$$\begin{aligned}
&= \sum_{\vec{\ell}} \left(\frac{1}{\eta_x} \underbrace{\sum_{i: X_i=x, \vec{L}_i=\vec{\ell}} \eta_i \overbrace{\sum_t \hat{\pi}(t, \vec{\ell}) \hat{g}(t, x, \vec{\ell})}^{\text{Constant with respect to } i}}_{=\eta_{x\vec{\ell}} \sum_t \hat{\pi}(t, \vec{\ell}) \hat{g}(t, x, \vec{\ell})} \right. \\
&\quad \left. - \frac{1}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, X_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}} \sum_t \underbrace{\sum_{i: X_i=x, \vec{L}_i=\vec{\ell}, T_i=t} \eta_i \overbrace{\frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \hat{g}(t, x, \vec{\ell})}^{\text{Constant with respect to } i}}_{=\eta_{x\vec{\ell}} t \frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})} \hat{g}(t, x, \vec{\ell})} \right) \quad (\text{B.2.44})
\end{aligned}$$

Simplify and note limits of each component,

with $\tilde{g}() \neq g()$ denoting the incorrect limit of $\hat{g}()$

$$\begin{aligned}
&= \sum_{\vec{\ell}} \left(\underbrace{\frac{\eta_{x\vec{\ell}}}{\eta_x}}_{\rightarrow \text{P}(\vec{L}=\vec{\ell}|X=x)} \sum_t \underbrace{\hat{\pi}(t, \vec{\ell})}_{\rightarrow \pi(t, \vec{\ell})} \underbrace{\hat{g}(t, x, \vec{\ell})}_{\rightarrow \tilde{g}(t, x, \vec{\ell})} - \sum_t \underbrace{\frac{\eta_{x\vec{\ell}} t}{\sum_{i: X_i=x} \frac{\eta_i \hat{\pi}(T_i, X_i, \vec{L}_i)}{\hat{m}(T_i, X_i, \vec{L}_i)}}}_{\substack{\rightarrow \text{P}(\vec{L}=\vec{\ell}, T=t|X=x) \\ \text{by Eq. B.2.25--B.2.31}}} \underbrace{\frac{\hat{\pi}(t, \vec{\ell})}{\hat{m}(t, x, \vec{\ell})}}_{\rightarrow \frac{\pi(t, \vec{\ell})}{m(t, x, \vec{\ell})}} \underbrace{\hat{g}(t, x, \vec{\ell})}_{\rightarrow \tilde{g}(t, x, \vec{\ell})} \right) \quad (\text{B.2.45})
\end{aligned}$$

$$\begin{aligned}
&\rightarrow \sum_{\vec{\ell}} \left(\text{P}(\vec{L}=\vec{\ell} | X=x) \sum_t \pi(t, \vec{\ell}) \tilde{g}(t, x, \vec{\ell}) \right. \\
&\quad \left. - \sum_t \underbrace{\text{P}(\vec{L}=\vec{\ell}, T=t | X=x)}_{=m(t, x, \vec{\ell}) \text{P}(\vec{L}=\vec{\ell}|X=x)} \frac{\pi(t, \vec{\ell})}{m(t, x, \vec{\ell})} \tilde{g}(t, x, \vec{\ell}) \right) \quad (\text{B.2.46})
\end{aligned}$$

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

$$\rightarrow \sum_{\vec{\ell}} \text{P}(\vec{L}=\vec{\ell} | X=x) \underbrace{\left(\sum_t \pi(t, \vec{\ell}) \tilde{g}(t, x, \vec{\ell}) - \sum_t \pi(t, \vec{\ell}) \tilde{g}(t, x, \vec{\ell}) \right)}_{=0} \quad (\text{B.2.47})$$

$$\rightarrow 0 \quad (\text{B.2.48})$$

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.

B.3 Variance estimation with complex 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.

B.4 Simulations

Three simulations illustrate the properties of the estimators. The first simulation (Fig 3.4 in the main text) illustrates how the doubly-robust estimator is approximately unbiased if the prediction function for either outcomes or for treatment probabilities 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 3.5 in the main text) shows that cross fitting improves on double robustness in a setting with random forests used for estimation.

B.4.1 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 B.1 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).

$$X \rightarrow L \overset{\curvearrowright}{\rightarrow} T \rightarrow Y$$

Figure B.1: Causal structure for the 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 B.2. The gap-defining category X is a binary indicator.

$$X \sim \text{Bernoulli}(.5) \tag{B.4.1}$$

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.

$$L \sim \begin{cases} \text{Uniform}(-1, 0) & \text{if } X = 0 \\ \text{Uniform}(0, 1) & \text{if } X = 1 \end{cases} \tag{B.4.2}$$

The true probability $m(1, x, \vec{\ell})$ of treatment $T = 1$ is quadratic in log odds (Fig B.2).

$$m(1, x, \ell) \equiv \text{logit}^{-1}(-2 + 4L^2) \tag{B.4.3}$$

$$T \sim \text{Bernoulli}(m(1, x, \ell)) \tag{B.4.4}$$

The outcome conditional mean $g(t, x, \ell)$ involves an interaction between L and T (Fig B.2).

$$g(t, x, \ell) \equiv 4 + 4L + 4T - 2TL \tag{B.4.5}$$

$$Y \sim \text{Normal}(\text{Mean} = g(t, x, \ell), \text{SD} = 1) \tag{B.4.6}$$

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 .

$$g(t, x, \ell) = \begin{cases} \beta_0 + \beta_1 t + \beta_2 \ell + \beta_3 t\ell & \text{correct specification} \\ \eta_0 + \eta_1 t + \eta_2 \ell & \text{incorrect specification} \end{cases} \quad (\text{B.4.7})$$

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

$$m(t, x, \ell) = \begin{cases} \text{logit}^{-1}(\lambda_0 + \lambda_1 \ell + \lambda_2 \ell^2) & \text{correct specification} \\ \text{logit}^{-1}(\kappa_0 + \kappa_1 \ell) & \text{incorrect specification} \end{cases} \quad (\text{B.4.8})$$

Fig B.2 illustrates the true data generating process and the approximations to which the incorrect specifications converge at large sample sizes. Fig 3.4 (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.

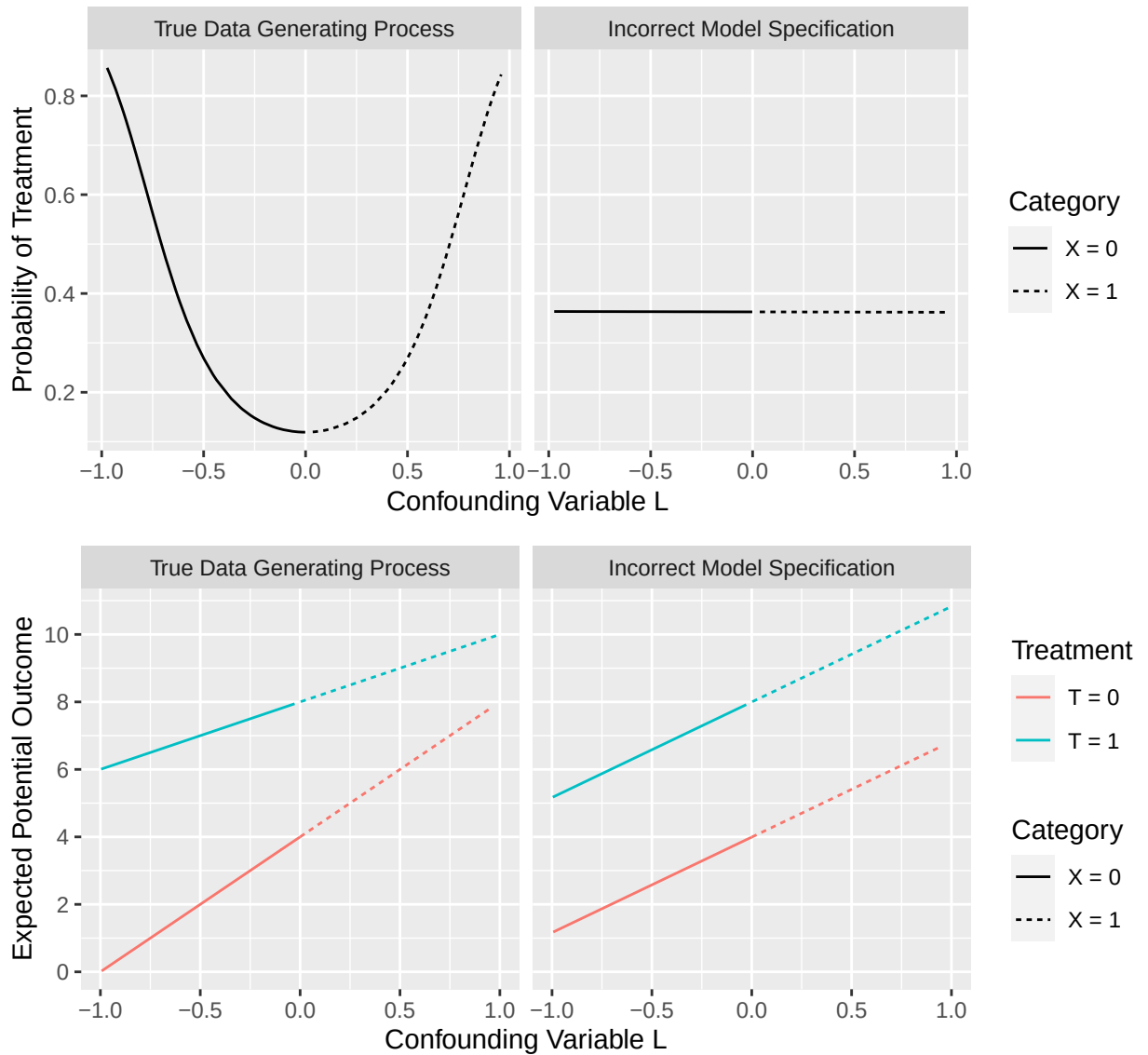
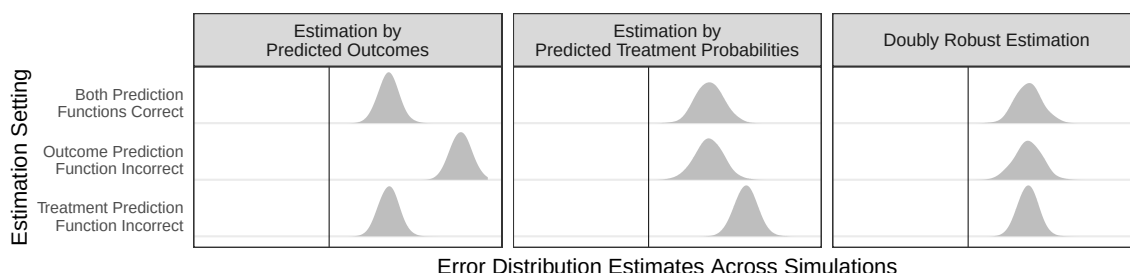


Figure B.2: Data generating process and functional form misspecifications to illustrate double robustness. Results of this simulation are shown in Fig 3.4 in the main text. The two categories X (solid and dashed lines) have disjoint support on the continuous confounding variable L . The log odds of treatment 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 3.4).

B.4.2 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 B.3 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 with unequal but known probabilities, analyzing it as a simple random sample produces bias for all estimators.



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

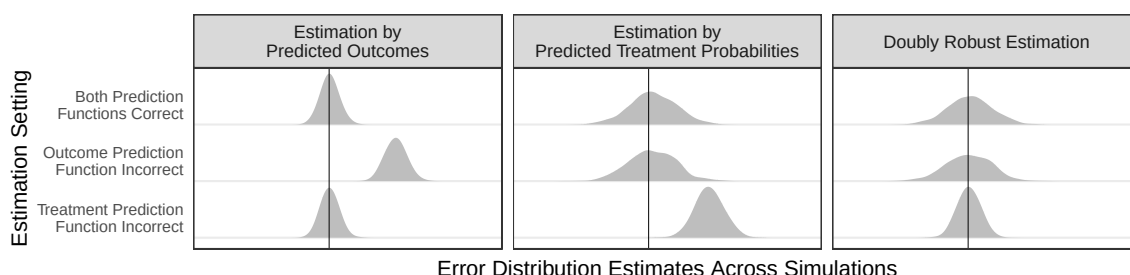


Figure B.3: Simulation showing estimation with sampling weights. In this simulation, units are sampled from the population with known but unequal probabilities of inclusion. Simulation details are provided in Sec. B.4.2.

B.4.3 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 B.4. 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 .

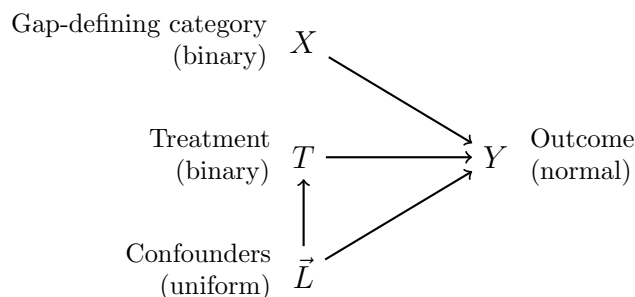


Figure B.4: Causal structure for cross fitting simulation.

Mathematical details of the simulation are given in Eq. B.4.9–B.4.14. 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 $X = 0$.

$$X \sim \text{Bernoulli}(.5) \quad (\text{B.4.9})$$

$$L_1, \dots, L_{10} \sim \text{Uniform}(-1, 1) \quad (\text{B.4.10})$$

$$m(1, X, \vec{L}) = \text{logit}^{-1}(.3L_1 + \dots + .3L_{10}) \quad (\text{B.4.11})$$

$$T \sim \text{Bernoulli}\left(m(1, X, \vec{L})\right) \quad (\text{B.4.12})$$

$$g(T, X, \vec{L}) = \begin{cases} L_1 + \dots + L_{10} + T & \text{if } X = 1 \\ L_1 + \dots + L_{10} - T & \text{if } X = 0 \end{cases} \quad (\text{B.4.13})$$

$$Y \sim \text{Normal}\left(\text{Mean} = g(T, X, \vec{L}), \text{SD} = 10\right) \quad (\text{B.4.14})$$

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 $n = \{250, 500, 750, 1000, 1250, 1500, 2000, 3000, 5000, 10000\}$, 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. B.4.9–B.4.14, the true gap-closing estimand under assignment to $T = 1$ is $\tau = 2$. Fig 3.5 in the main text reports root mean squared error (RMSE) of each estimator at each sample size.

$$\text{RMSE} = \sqrt{\frac{1}{1000} \sum_{r=1}^{1000} (\hat{\tau}_{1,0}^r(1) - \tau_{1,0}(1))^2} \quad (\text{B.4.15})$$

As is visible in Fig 3.5, RMSE converges to zero more quickly for doubly-robust estimation than for an approach with only predicted outcomes, and more quickly with cross fitting than in a single sample.

B.5 Details of the empirical example

B.5.1 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 (Beller, 2009; Jonsson et al., 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 B.3).

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.

B.5.2 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.

$$\text{logit}\left(m\left(t, x, \vec{\ell}\right)\right) = \text{logit}\left(\text{P}\left(T = t \mid X = x, \vec{L} = \vec{\ell}\right)\right) \quad (\text{B.5.1})$$

$$\underbrace{=}_{\text{By parametric approximation}} \alpha_m + x\beta_m + \vec{\ell}'\vec{\gamma}_m \quad (\text{B.5.2})$$

$$g\left(t, x, \vec{\ell}\right) = \text{E}\left(Y \mid T = t, X = x, \vec{L} = \vec{\ell}\right) \quad (\text{B.5.3})$$

$$\underbrace{=}_{\text{By parametric approximation}} \alpha_g + \underbrace{x\beta_g + t\lambda_g + (xt)\nu_g}_{\text{Category, treatment, and interaction}} + \underbrace{\vec{\ell}'\vec{\gamma}_g}_{\text{Pre-treatment covariates}} \quad (\text{B.5.4})$$

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.

B.5.3 Example: Specification of $\pi()$

Fig 3.6 reports estimates for two fixed and two stochastic assignment rules.

$$\text{Assign a professional occupation: } \pi(\text{Professional}, \vec{\ell}) = 1 \quad (\text{B.5.5})$$

$$\text{Assign a working-class occupation: } \pi(\text{Professional}, \vec{\ell}) = 0 \quad (\text{B.5.6})$$

$$\text{Assign a random occupational class: } \pi(\text{Professional}, \vec{\ell}) = P(T = \text{Professional}) \quad (\text{B.5.7})$$

$$\text{Assign randomly within covariates: } \pi(\text{Professional}, \vec{\ell}) = P(T = \text{Professional} \mid \vec{L} = \vec{\ell}) \quad (\text{B.5.8})$$

For the stochastic assignment rule (Eq. B.5.7 and B.5.8), $\pi()$ must be learned from the data. Because there is no double robustness to misspecifying this function, I estimate nonparametrically. For the rule in Eq. B.5.7, I estimate by the weighted sample mean in a professional occupation. For the rule in Eq. B.5.8, I estimate by the weighted sample mean within each stratum of the covariates.

B.5.4 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.

B.5.5 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).

$$\text{Bias}\left(\hat{\tau}_{1,0}(1)\right) = \overbrace{\mathbf{E}\left(\hat{\tau}_{1,0}(1) - \tau_{1,0}(1)\right)}^{\text{Bias of gap-closing estimator under treatment value 1}} \quad (\text{B.5.9})$$

$$= \mathbf{E}\left(\overbrace{\left(\hat{\theta}_1(1) - \hat{\theta}_0(1)\right)}^{\text{Difference in post-intervention mean estimates}} - \overbrace{\left(\theta_1(1) - \theta_0(1)\right)}^{\text{Difference in post-intervention mean true values}}\right) \quad (\text{B.5.10})$$

$$= \underbrace{\mathbf{E}\left(\hat{\theta}_1(1) - \theta_1(1)\right)}_{\substack{\text{Bias}(\hat{\theta}_1(1)) \\ \text{Bias of post-intervention mean} \\ \text{estimator for units in category 1}}} - \underbrace{\mathbf{E}\left(\hat{\theta}_0(1) - \theta_0(1)\right)}_{\substack{\text{Bias}(\hat{\theta}_0(1)) \\ \text{Bias of post-intervention mean} \\ \text{estimator for units in category 0}}} \quad (\text{B.5.11})$$

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.

B.6 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	37.20	37.31
N		1,374	10,954

Table B.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)	0.00 (0.04)	0.46 (0.09)
Treated (respondent holds a professional occupation)	0.38 (0.03)	
Treated $\times$ professional origin	0.10 (0.07)	
Age	0.02 (0.00)	0.02 (0.01)
Highest degree (less than high school omitted)		
High school	0.46 (0.03)	1.79 (1.12)
Junior college	0.62 (0.04)	2.72 (1.13)
Bachelor's	0.87 (0.04)	3.66 (1.12)
Graduate	1.06 (0.05)	4.83 (1.12)
Race (white omitted)		
Black	-0.02 (0.03)	-0.39 (0.16)
Other	-0.15 (0.04)	0.38 (0.13)
Female	-0.71 (0.02)	-0.78 (0.08)
Intercept	9.40 (0.08)	-5.06 (1.17)

Table B.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 3.6.

	Not conditional on treatment	Conditional on treatment
Professional class origin (father figure held a professional occupation)	0.05 (0.03)	0.02 (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)	-0.14 (0.04)
Female	-0.74 (0.02)	-0.71 (0.02)
Intercept	9.41 (0.08)	9.40 (0.08)

Table B.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 ($0.05 - 0.02 = 0.03$). This procedure does not estimate a gap-closing estimand (see Sec. 3.7.1).

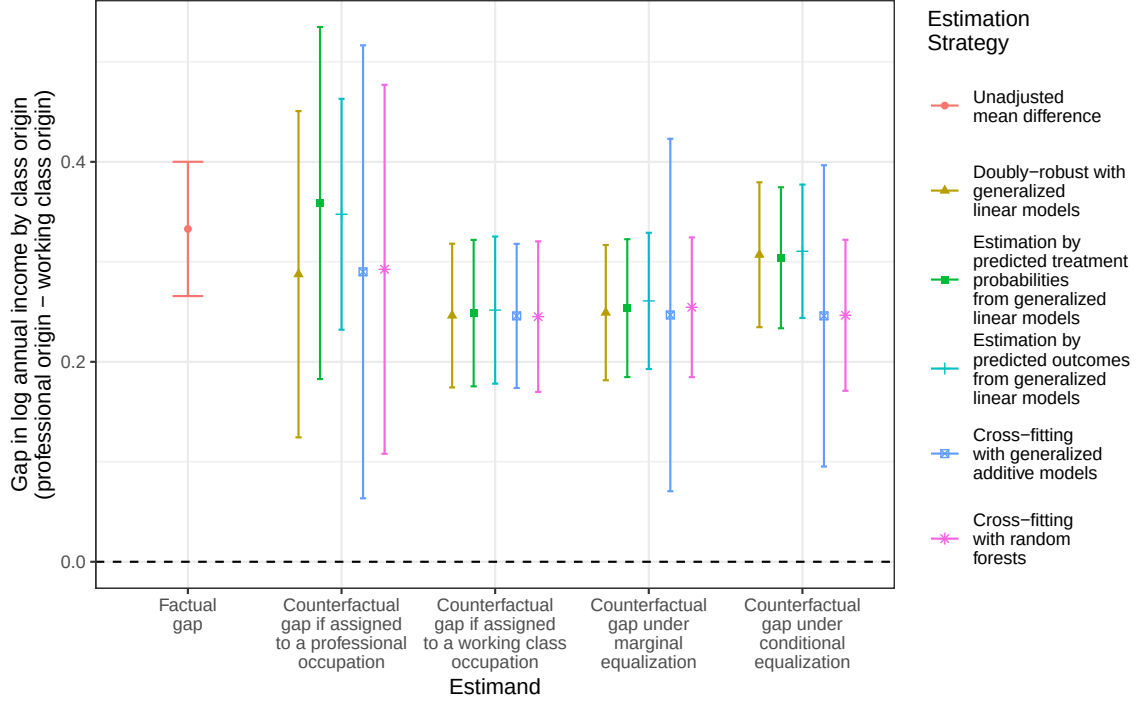


Figure B.5: Gap-closing estimands by alternative estimators. Estimation by predicted outcomes follows Eq. 3.5.3. Estimation by predicted treatment probabilities reweights observed outcomes for treated units to estimate the population average potential outcome as in Eq. 3.5.5. Doubly robust estimation combines the two and is consistent if either is consistent (Eq. 3.5.6). 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. 3.5.4 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 B.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 C

Supplementary materials for Chapter 4

C.1 Weighting of the CPS-ASEC

The CPS-ASEC samples people with unequal probabilities. All analyses in the main text weight by the ASECWT weighting variable designed for the ASEC, with inference calculated from variances estimated with replicate weights (Appendix C.6). Supplemental analyses show similar results with the LINKFW1YWT weight provided by IPUMS for the subsample that links forward one year (Flood et al., 2020); because replicate weights are not available for this weight so that the appropriate construction of confidence intervals is not clear, results with that weight are provided in Fig C.9 Panel E. Because the U.S. population grows over time, I allow the weight variable to place greater weight on the more recent years proportional to the relatively larger size of the population (see Fig C.7).

In addition to weighting population averages, the analyses also weight the generalized additive models (Wood, 2017) used to estimate conditional outcomes given predictors. The reason to weight the GAM is because it is only an approximation to a conditional mean function that is likely to be more complex; in fact, there are likely to be omitted interactions that the additive aspect of the GAM will miss. The aim is for the GAM to provide a useful and parsimonious approximation to the more complex conditional mean function. In order to prioritize the quality of that approximation in regions of the covariate space that are more heavily weighted by the sampling weights (where the predictions will matter most), the GAM estimation is weighted.

C.2 Methods for descriptive scatterplots (Fig 4.1)

Because of the large number of occupations, some occupations are sparsely populated. The descriptive occupation-level analyses in Fig 4.1 use a partial-pooling strategy to improve precision. I first score each occupation on two dimensions: (1) the racial/ethnic composition of those employed in the occupation and (2) the proportion of those employed in the occupation this year to report a work-limiting disability next year. I then estimate occupation-level summaries with bivariate person-level regression models for the outcome of interest (e.g. onset of work-limiting disability) as a function of occupation random intercepts, thereby partially pooling information across occupations and improving efficiency by pulling estimates (particularly for small occupations) toward the overall mean. I repeat this process for each occupational summary of interest: the proportion with onset of work-limiting disability next year, the proportion non-Hispanic Black, the proportion Hispanic, and the proportion identifying with other racial categories (non-Hispanic white omitted). The scatterplots visualize occupation-level onset of work-limiting disability plotted against occupation-level racial composition. The best-fit line is a linear regression through the estimated points (occupations), with each occupation weighted proportional to size.

C.3 Threats to causal identification

Causal identification from observational data requires assumptions (Fig 4.5A), and those assumptions are never perfect. This section discusses two threats to identification, both of which derive in part from the fact that lagged survey reports of health coarsely capture a latent construct. Both threats are mitigated in part by the adjustment for several measures of lagged health in the analyses: adjustment for

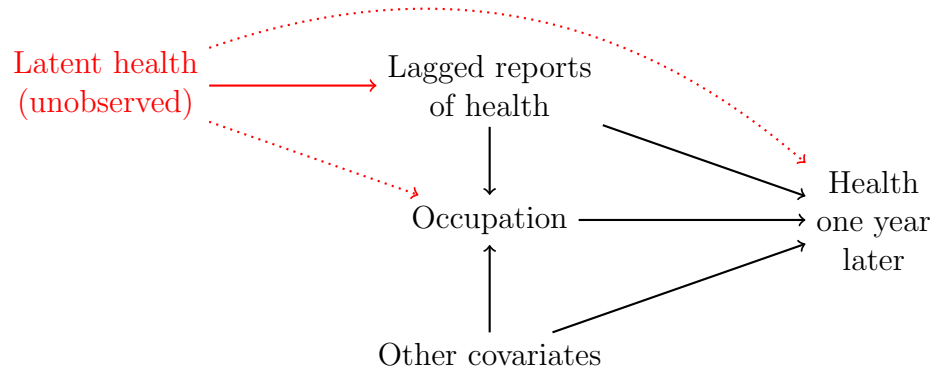


Figure C.1: Threat to causal identification: Coarse measurement of lagged health. The lagged health reports are likely to be coarse measures of latent health. To the degree that uncaptured components of latent health directly affect health outcomes and the occupation one holds (dotted edges), identification will be imperfect. This risk is mitigated to the degree that the lagged reports of health capture the relevant aspects of latent health that affect occupations and affect the health outcome, so that the dotted edges may be relatively unimportant.

lagged self-rated health and restriction of the sample to those who never quit a job for health reasons and who do not initially report a work-limiting disability (the lagged outcome).

Despite these adjustments, latent health remains a possible confounder (Fig C.1). If it were possible to know a true continuous and multi-dimensional construct of latent health, it is likely that latent health would cause occupation in year t and the health outcome in year $t + 1$ through paths not fully captured by the measures of lagged health. Because health is intrinsically difficult to measure, this threat is unavoidable. Nonetheless, adjustment for lagged reports of health may substantially reduce the magnitude of the bias, because those lagged reports may capture the primary pathways through which latent health affects the treatment (occupation) and the outcome (health one year later).

A second threat arises from the possibility that lagged health is a pre-treatment collider variable (Fig. C.2). Because the CPS panel can only be linked over two years, I observe occupation in years t and $t + 1$ but not in year $t - 1$. It is likely

that occupation in year $t - 1$ causes occupation in year t . If occupations cause health outcomes (the main claim of this paper), then occupation in year $t - 1$ affects lagged reports of health from year t . Lagged health is a collider variable because it is also a consequence of latent health conditions. If those latent health conditions cause health one year later through pathways other than the observed measures of health, then there is a danger of M -bias: holding lagged reports of health constant induces an association between lagged health and the occupation held at $t - 1$, thus opening a backdoor path which creates confounding (Greenland, 2003; Greenland et al., 1999). Intuitively, suppose that in order to have no work-limiting disability at year t (one of the lagged reports of health) you either have to hold a safe occupation or you have to have very good latent health. Among those with no work-limiting disability at year t , those in the most dangerous occupations would then have the best latent health. That induced association could cause the results to underestimate the contribution of occupational segregation to health, because the most dangerous jobs are filled with the people who are healthiest in unobserved ways.

Despite the possibility of inducing some M -bias, adjustment for lagged reports of health is worthwhile because it mitigates the most serious threat to identification: that lagged health affects future health and affects the occupation one holds. Further, each of these threats is substantially diminished to the degree that the lagged reports of health capture the aspect of latent health most relevant to the outcome variable. Because the lagged reports of health include a lagged dependent variable, the threats may be reasonably small. For these reasons, estimates are useful despite these limitations.

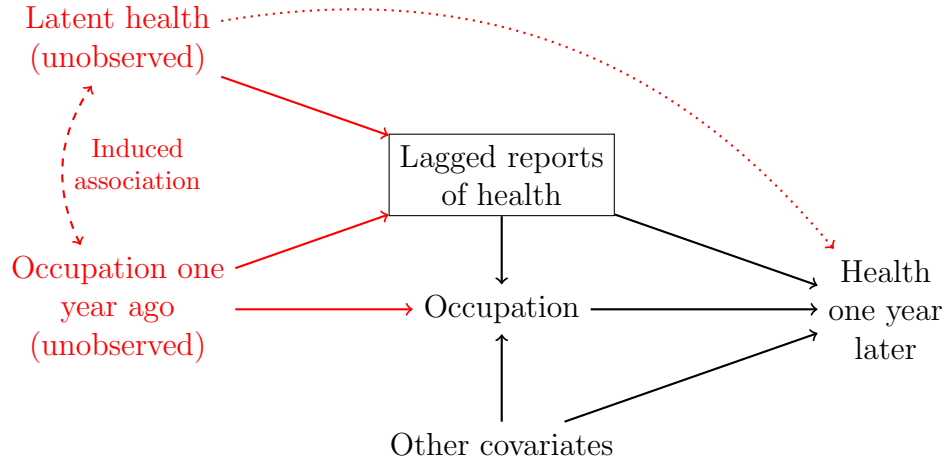


Figure C.2: Threat to causal identification: *M*-bias. Lagged health is a collider variable because it is likely to be a consequence of both latent health and the occupation one a year ago. Conditioning on lagged health induces an association between those causes. This opens a backdoor path (dashed) that would confound estimates of the causal effect of occupation on health one year later. This risk is mitigated to the degree that the lagged reports of health capture the relevant aspects of latent health that affect the health outcome, so that the dotted edge may be relatively unimportant.

C.4 Details of how prediction functions are specified and used to estimate the main results

This section provides details about the methods applied to the CPS-ASEC to yield the main estimates of this paper. Under the identifying assumptions in the DAG presented in Fig 4.5A, the observed data are informative about the unobserved data. In particular, we can predict the probability of disability $\hat{y}_i(d)$ that each person i would experience if counterfactually exposed to occupation d , based on the health outcomes actually observed among those who share their race and covariate values $\vec{\ell}$ and are factually exposed to occupation d . The task of a statistical prediction function is simply to predict health outcomes as a function of the category (race), all covariates (e.g. education), and occupation. Because racial disparities are central to the research question, I estimate separate models on subsamples defined by racial

categories. This means that the relationship between the predictors and the outcome is allowed to vary by race. The model a Generalized Additive Models (GAM, Wood 2017) specified as by Eq. C.4.1.

$$\begin{aligned}
& \mathbf{E}(\text{Onset of Work-Limiting Disability} \mid \text{Occupation}_i, \text{Covariates}_i, \text{Race}_i = r) \\
&= \underbrace{\alpha_{\text{Occupation}_i}^r}_{\text{Occupation-specific intercept (see C.4.2)}} + \beta_1^r(\text{Education}_i = \text{Less than High School}) \\
&\quad + \beta_2^r(\text{Education}_i = \text{Some College}) + \beta_3^r(\text{Education}_i = \text{College}) \\
&\quad + \beta_4^r(\text{Self-Rated Health}_i = \text{Poor}) + \beta_5^r(\text{Self-Rated Health}_i = \text{Fair}) \\
&\quad + \beta_6^r(\text{Self-Rated Health}_i = \text{Good}) + \beta_7^r(\text{Self-Rated Health}_i = \text{Very Good}) \\
&\quad + \beta_8^r(\text{Sex}_i = \text{Female}) + \beta_9^r(\text{Foreign-Born}_i = \text{True}) \\
&\quad + \underbrace{\beta_{10}^r(\text{Year}_i \geq 2014)}_{\substack{\text{Intercept shift when the} \\ \text{survey structure changed} \\ \text{(see Appendix C.5)}}} + \underbrace{s_{\text{Year}}^r(\text{Year}_i) + s_{\text{Age}}^r(\text{Age}_i)}_{\substack{\text{Smooth functions of year and age} \\ \text{(thin-plate splines)}}} \tag{C.4.1}
\end{aligned}$$

Every term has an r superscript because the model is estimated separately by race. The smooth terms $s()$ are thin-plate splines (Wood, 2003) to allow a smooth but possibly non-linear functional form for continuous predictors. The indicator for $(\text{Year}_i \geq 2014)$ allows a discontinuity at 2014 because of a change in questionnaire design at that point (Appendix C.5). As discussed in Section 4.4, the model implicitly conditions on lagged reports of no work-limiting disability (a lagged dependent variable) and never quitting a job for health reasons by restricting the sample along these variables.

An important choice is how to estimate the occupation-specific intercepts α_d^r . Estimates could be very uncertain if I estimated an unregularized coefficient on an indicator variable for each occupation, because in some occupations the number of people observed is small. A standard multilevel model would improve precision by regularizing the occupation-specific intercept α_d^r toward the grand mean. But the en-

tire motivation for the model is that occupational hazards may vary systematically as a function of racial composition. I therefore model the occupation-specific intercepts as functions of the racial composition of each occupation, denoted π_{rd} for each racial category r .

$$\begin{array}{ccccccc}
 \text{Occupation-specific intercept} & & \text{Proportion non-Hispanic Black in occupation } d & & \text{Proportion Hispanic in occupation } d & & \text{Proportion other in occupation } d \\
 \downarrow & & \downarrow & & \downarrow & & \downarrow \\
 \alpha_d^r = s_{\text{Black}}^r(\pi_{\text{Non-Hispanic Black},d}) + s_{\text{Hispanic}}^r(\pi_{\text{Hispanic},d}) + s_{\text{Other}}^r(\pi_{\text{Other},d}) + \delta_d^r & & & & & & \\
 & \nwarrow & & \nearrow & & & \nearrow \\
 & \text{Thin-plate splines for smooth nonlinear functions} & & & & & \text{Idiosyncratic component} \\
 & \text{to predict outcomes for individuals in racial category } r & & & & & \text{regularized toward} \\
 & \text{given the racial composition of their occupations} & & & & & \text{the grand mean} \\
 & & & & & & \sim \mathcal{N}(\delta_0^r, \sigma_{\delta^r}^2)
 \end{array}
 \tag{C.4.2}$$

In this specification, the thin-plate splines are regularized toward lines and only the δ_d^r component is regularized toward the grand mean. The advantage is that occupations with a given racial composition are all regularized toward each other. Thus, regularization will lead to over-estimation of the outcome in some occupations at a given racial composition and under-estimation the outcome in other occupations with that composition. But the counterfactual assignment rule will up-weight or down-weight all occupations with a given racial composition in a similar way. As a result, the biases may counterbalance more favorably than in a standard multilevel model specification. It is also important to be mindful that this is entirely a finite-sample consideration: as the sample size with each occupation grows, the idiosyncratic component δ_d^r would come to dominate the occupation-specific estimate so that the chosen method of regularization declines in importance.

The only additional hurdle is that the occupation-specific racial compositions π_{rd} are not known and must be estimated from the data. I therefore employ a two-

step procedure: I first estimate the racial composition of each occupation by random intercept models (Eq. C.4.3) and then use the estimated $\hat{\pi}_{rd}$ in estimation of the outcome model.

$$\pi_{rd} \sim \mathcal{N}(\bar{\pi}_r, \sigma_r^2) \quad (\text{C.4.3})$$

Probability that a person in occupation d is of racial category r

Population proportion in race r

Variance across occupations

The random intercepts for occupational composition (Eq. C.4.3) are similar to simply taking the proportion in race r among those observed in occupation d . The only difference is that the estimate is regularized toward the population proportion $\bar{\pi}_r$ in race r . This helps for small occupations, where the racial composition is especially uncertain. The amount of shrinkage σ_r^2 is chosen empirically by the `mgcv` package (Wood, 2017). The analysis captures statistical uncertainty at both steps because uncertainty is estimated by replication methods in which the entire procedure is carried out on repeated samples (Sec. C.6).

With the estimated prediction function, I then impute the counterfactual probability of work-limiting disability onset $\hat{y}_i(d)$ for each person i in each occupation d (Fig 4.5C). Then, I average over a counterfactual assignment rule for occupations, weighting each potential outcome $\hat{y}_i(d)$ by the prevalence of the occupation d among people with the education level of person i (Fig 4.5E). Finally, I aggregate the person-specific estimates over people in each subpopulation defined by race and ethnicity, using survey weights (Fig 4.5E).

C.4.1 Doubly-robust estimates

The main text presents models from an approach centered on predicted outcomes. An alternative approach (Fig C.9 Panel I) is a doubly robust estimator: build predictions

for the probability of assignment to each occupation (analogous to a propensity score) and correct the outcome predictions by their estimated bias (the weighted mean error with inverse probability of treatment weights). Lundberg (2021) discusses this estimator in greater depth. The benefit of a doubly robust estimator is that it is consistent if either the treatment model or the outcome model is correctly specified.

The treatment in the present paper is occupation, with 451 categories. Doubly robust estimators can have high variance when inverse probability of treatment weights are highly variable, which especially occurs when some probabilities of treatment are very near zero (Kang and Schafer, 2007). This means that doubly robust estimation may produce poor estimates for outcomes in extremely rare occupations. To avoid this problem, I focus the doubly robust estimator on a different counterfactual assignment rule: assign occupations from among the 192 largest occupations that collectively contain 90% of the population of people.

With this simplification, it is still difficult to estimate the factual probability of holding each of the 192 relevant occupations as a function of all observed confounders. Fitting a multinomial logistic regression model is infeasible with so many outcome categories. I therefore break the problem into two parts. I first estimate the probability of holding an occupation in each of the nine major occupational categories in the Census occupational classification system, which are indicated by the first of the three-digit detailed occupation code. Within each major classification, I then estimate the probability of holding each detailed occupation among the occupations in that major category (between 9 and 33 occupations). The probability of holding an occupation given covariates is the product of the probability of the major category and the probability of the detailed occupation given the major category. I estimate the multinomial models with the `multinom` function in the `nnet` package in R (Venables and Ripley, 2002), entering all covariates additively with linear terms for continuous

covariates, and with estimation carried out separately for each racial category as in the outcome model specification.

Given the estimated factual probability of treatment, one can learn about the outcomes under the counterfactual assignment rule by weighting each observation by a weight that is the product of the survey weight, the counterfactual probability of being assigned to the factual occupation, and the inverse of the factual probability of the factual occupation (Lundberg, 2021). Using these weights, I de-bias the estimated outcome predictions by subtracting the weighted mean outcome residual from the outcome predictions. Results (Fig C.9 Panel I) are substantively similar to the outcome modeling estimates presented in the main text.

C.5 Measurement issues with work-limiting disability in the CPS-ASEC

There are two important limitations to the operationalization of work-limiting disability in the CPS-ASEC: changes in the survey instrument over time and limitations in the definition of disability for a self report.

C.5.1 Changes in the CPS-ASEC survey instrument

Prior to 2014, the CPS-ASEC instrument was relatively constant over time. The household respondent answered a work-limiting disability question for each household member: “(Do you/Does anyone in the household) have a health problem or disability which prevents (you/them) from working or which limits the kind or amount of work (you/they) can do?” Two important changes occurred after that, one in 2014 and one in 2016.

In 2014, the wording of this question remained the same but other aspects of the CPS-ASEC underwent a major redesign.¹ All respondents answered new questions about health insurance, and 3/8 of the sample received income questions in a new format. After 2014, the CPS-ASEC followed the new format for all respondents. Although the work-limiting disability question did not change, Fig C.3 Panel A shows that the reported prevalence of work-limiting disability rose substantially under the new questionnaire structure. One reason for this increase might be the change in the order of the questionnaire. In the traditional questionnaire, the module immediately preceding work-limiting disability was about Social Security survivor benefits. In the redesigned questionnaire, the module immediately preceding work-limiting disability was about Supplemental Security Income (SSI), a Social Security program that provides transfers to individuals with disabilities (Czajka et al. 2015:5). Asking someone to report their receipt of SSI may prime them to report a work-limiting disability at a higher rate in the immediately subsequent questionnaire module.

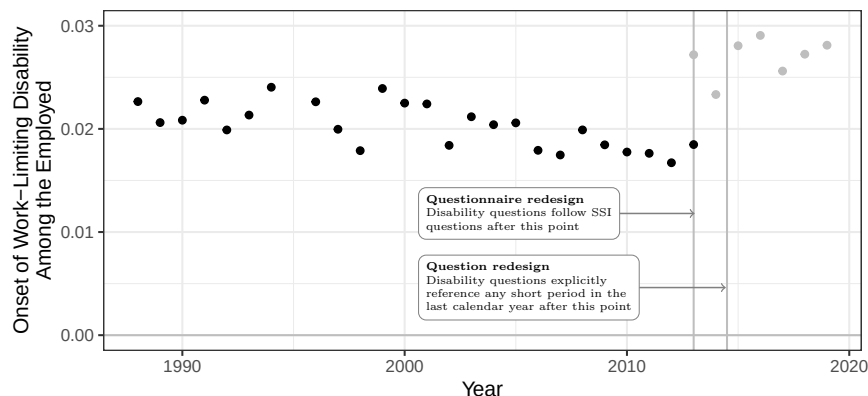
In 2016, the wording of the work-limiting disability question changed to focus explicitly on work-limiting disability in the past calendar year, with an added prompt about limitations that were in place “even for a short time.”² As illustrated in Figure C.3 Panel A, this wording change produced no apparent change in the proportion reporting a work-limiting disability.

Despite both of these changes, the *Black-white disparity* in work-limiting disability remains remarkably consistent across the years 2005–2020 (Fig C.3 Panel B). The main analyses therefore pool over all years. Fig C.9 Panel G shows that the results are similar if subsetting to the years preceding either change, and Panel H shows a similar result if subsetting to the 2014–2020 period in which one or both changes were in effect.

¹Details of the CPS-ASEC 2014 redesign are available in a BLS memo: https://cps.ipums.org/cps/resources/other_docs/ASEC_Redesign.pdf

²Full question wordings are available at https://cps.ipums.org/cps-action/variables/DISABWRK#questionnaire_text_section.

A) Reports of a work-limiting disability increased when the questionnaire changed



B) The key disparity of interest remained roughly constant

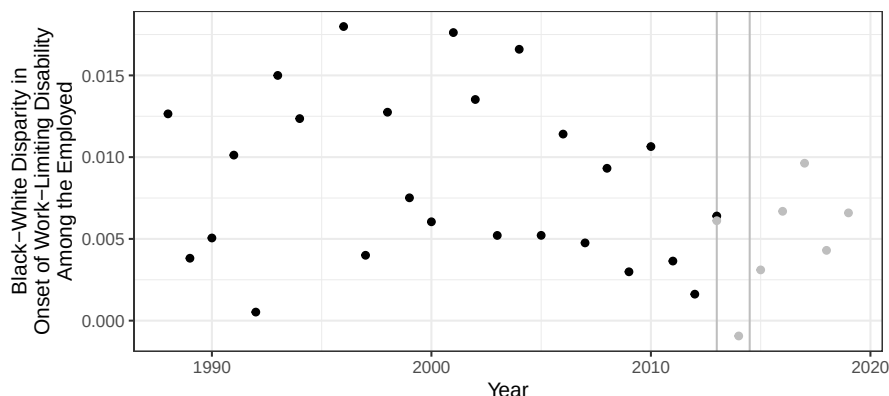


Figure C.3: Changes to the survey instrument affected levels but not disparities in reported work-limiting disability. For consistency with other plots, the x -axis here refers to the first year of observation, but the outcome is measured in the second year of observation. Therefore, CPS questionnaire changes in 2014 and before 2016 correspond to vertical lines placed at 2013 and before 2015 in this plot. In 2014, the CPS questionnaire was redesigned and modules were re-ordered so that the disability questions immediately followed the Supplemental Security Income questions instead of the Survivor Benefit questions. The redesigned questionnaire was given to 3/8 of the sample (gray dot on the vertical line), while 5/8 of the sample received the original questionnaire (black dot on the vertical line). Before 2014, everyone answered the traditional questionnaire (black dots) and after 2014 everyone answered the redesigned questionnaire (gray dots). Beginning in 2016, the question about “a health problem or a disability which prevents him/her from working or which limits the kind or amount of work” was changed to reference the past calendar year and to include the prompt for conditions lasting “even for a short time.” The 2014 redesign increased reported disability, while the 2016 question change appears to have had little effect on reported disability. The Black-white disparity in disability onset is roughly constant over both changes to the survey instrument. The main analyses therefore pool over both formats of the question, with alternative specifications separating at the more consequential update (2014) reported in Fig C.9 Panels G and H.

C.5.2 Validity in the meaning of the disability question

Self-reported work-limiting disability faces the same problems of survey measurement that affect any question: it captures disabilities that limit work only to the extent that respondents understand them as such. Hale (2001) critiques the CPS work-limiting disability question for defining disabilities in terms of work limitations rather than in terms of specific medical impairments. In an evaluation that compares the CPS work-limiting disability question to a broader definition of disability from a series of questions in the National Health Interview Survey, Burkhauser et al. (2002) show that the work-limiting disability question estimates a smaller population with a disability but that employment trends in both populations are similar. Maestas et al. (2019) collect new data to show that those who report a workplace accommodation for health reasons may still answer “no” to a question about whether they have a work-limiting disability. One should therefore understand a work-limiting disability as one particular operationalization of a broader concept of disability. It is possible that self-perceptions of disability deviate from how a medical professional would diagnose a disability, and those deviations may differ across individuals. Yet, one’s own perception that one has a work-limiting disability may be a worthy object of study in its own right; it is real for the person involved even if it is distinct from the diagnosis a medical professional would provide. A further advantage of accepting these self-reports as worthy of study is the ability to use the very large samples of the CPS with detailed occupational information linked over a 12-month interval.

C.6 Statistical inference by replicate weights

Statistical inference (e.g. standard errors and confidence intervals) involves a claim about how an estimator would vary if applied to hypothetical new samples drawn from the sampling frame by the same procedure used to generate the observed sam-

ple. Standard procedures assume a simple random sample and apply only if that is how the data were generated. The CPS is not a simple random sample. Instead, geographic Primary Sampling Units (PSUs) are selected within geographic strata, and then geographic Ultimate Sampling Units (USUs, clusters of households) are selected within PSUs by systematic sampling after sorting USUs systematically by labor market characteristics. Further, selection occurs with unequal probabilities in order to yield accurate estimates in subpopulations (e.g. state-specific unemployment estimates). This complex multistage cluster design renders standard inference procedures inappropriate. Further, as the analyst I cannot directly observe the strata, PSU, or USUs because these geographic identifiers are hidden in the data files for privacy reasons.

Inference therefore relies on a set of replicate weights constructed by the Bureau of Labor Statistics. Fig C.4 summarizes the sampling process of the CPS and the procedure by which the BLS constructs the replicate weights (Bureau of Labor Statistics 2006 Ch. 14, Fay and Train 1995). Fig C.5 illustrates the mechanical process by which the replicate weights yield a variance estimate: by conducting the estimation algorithm with each set of replicate weights in turn and estimating variation across the sets of replicate weights, one can simulate how the estimator would vary across repeated samples drawn from the CPS design. Because this entire procedure is computational, it is possible to estimate sampling variability by replicate weights even for potentially complicated estimation algorithms that involve regression, prediction, and aggregation to counterfactual summaries.

An advantage of the replicate weight procedure is that it applies equally well to any estimation algorithm, even one that is complex. Within each replicate, I estimate the predictive model for the outcome, impute unobserved counterfactuals, take averages, and store an estimate, thereby simulating variability across repeated samples drawn by the sampling design of the CPS.

Sampling Process	Replicate Process
Begin with the sampling frame: All residential addresses in the U.S.	Begin with the sample: Chosen households
Define geographic Primary Sampling Units (PSUs) within two types of geographic strata.	PSUs and strata are already defined.
Non-Self-Representing Strata (NSR): Each small-population PSUs is placed in a stratum with others. Self-Representing Strata (SR): Each large-population PSU is placed in a stratum alone.	For NSR, replicate the selection of PSUs. Combine pairs of NSRs into pseudostrata. Multiply sampling weight by a replicate factor of 1.5 and 0.5. More complex procedures are used for triplets and to adjust for unequal strata populations For SR, replicate the selection of USUs. Assign USUs into pairs by adjacency in the systematic sampling sort. Conduct an analogous reweighting procedure.
Select Ultimate Sampling Units (USUs) (clusters of households) in each PSU	All USUs remain, but with new replicate weights per above.
For privacy reasons, these steps are conducted by the Bureau of Labor Statistics and are not visible to researchers.	
Interview those households Estimate $\hat{\theta}$ with weights $\vec{w}$	Analyze those households Estimate $\hat{\theta}_r^*$ with each replicate weight $\vec{w}_r$
The variance of an estimator is how it would vary across repeated CPS samples drawn this way	The estimated variance is a function of variation across the 160 replicates: $\hat{V}(\hat{\theta}) = \frac{4}{160} \sum_{r=1}^{160} \left(\hat{\theta}_r^* - \hat{\theta} \right)^2$ ↑ Note: The factor of 4 is needed because the replicates inflate and deflate weights without completely zeroing out any weights.

Figure C.4: Replicate weights mimic repetitions of the CPS sampling process. Illustration is based on the BLS description (Bureau of Labor Statistics, 2006; Fay and Train, 1995). Because the CPS is not a simple random sample, standard approaches to statistical inference are inappropriate. Replicate weights simulate variation that comes from selection of geographic Primary Sampling Units (PSUs) within strata. For PSUs selected with probability one, replicate weights mimic variability within strata. Although the weight construction procedure is described by the BLS, as the researcher I only have access to the generated weights $\vec{w}$ and replicate weights $\{\vec{w}_r\}_{r=1}^{160}$. Inference in this paper proceeds by conducting the entire estimation algorithm repeatedly with each set of replicate weights, calculating the variance of the point estimate, and constructing a confidence interval by a normal approximation.

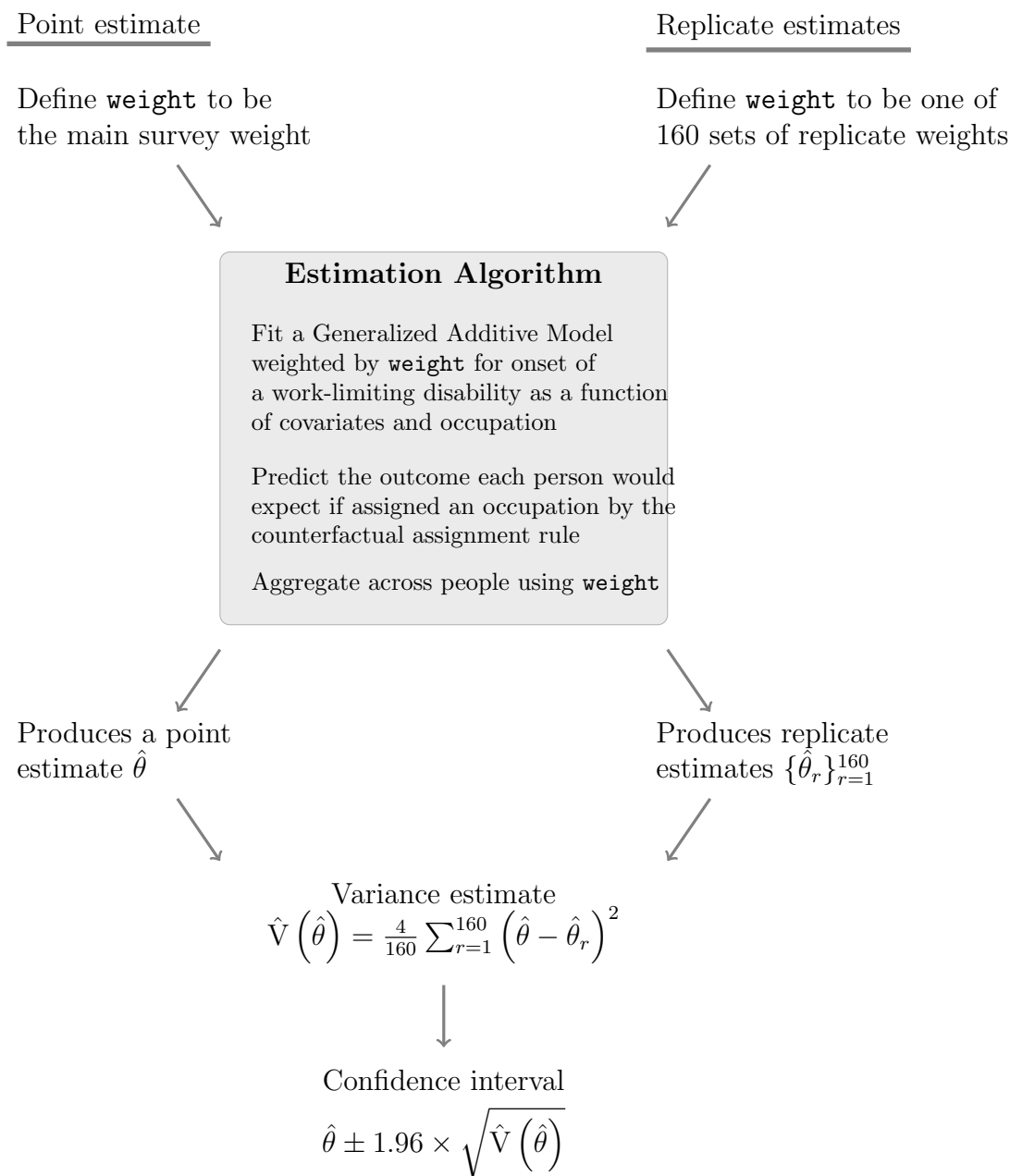


Figure C.5: Procedure to construct confidence intervals with replicate weights. The entire estimation procedure—regression, imputation, and aggregation—can be placed within an algorithm that accepts a weight as the input and returns an estimate. Passing the main survey weight through that algorithm yields a point estimate. Passing the replicate weights through the algorithm yields a set of 160 replicate estimates. Pooling across these estimates yields an estimated variance. By a normal approximation, the point estimate and estimated variance produce a 95% confidence interval.

C.7 Supplemental Figures

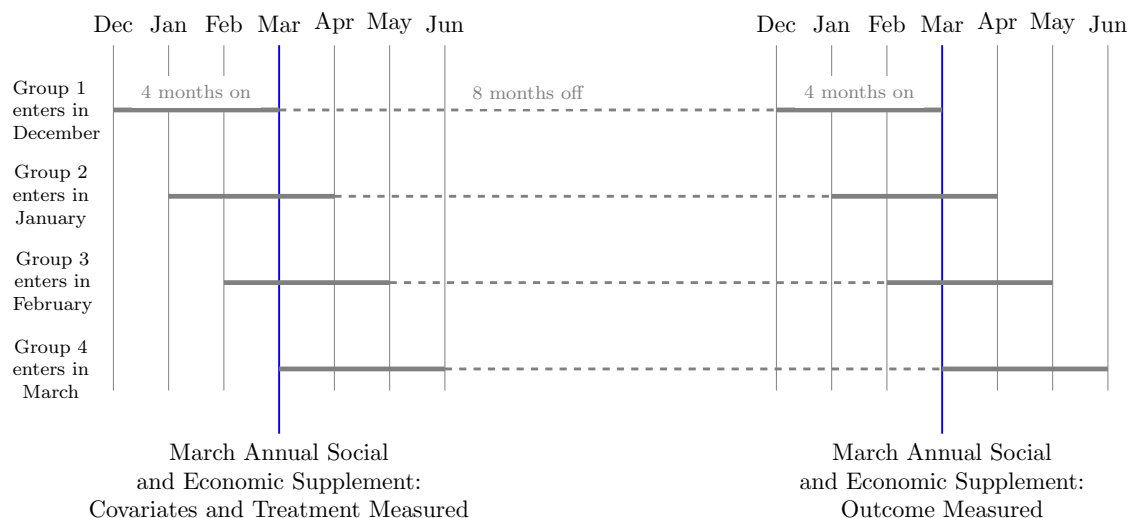


Figure C.6: Panel structure of the Current Population Survey. Households are sampled and assigned to rotation groups, which enter the panel on different months throughout the year. Each rotation group is included in the monthly survey for four months, excluded for eight months, and then included again for four more months. A basic survey is administered each month with the primary goal of tracking the unemployment rate. The Annual Social and Economic Supplement is a large survey administered each March. Because of the panel structure, households that enter the CPS panel in December to March are included in the ASEC survey for two adjacent years. The ASEC also includes oversamples which follow a different panel structure and are not linkable across years. For example, since 2001 the ASEC has included oversamples of low-income households drawn from those who enter the CPS in other months, with the goal of providing estimates relevant to evaluations of the State Children’s Health Insurance Program. Because the oversamples are not linkable across years, they are excluded from the linked data analyses in this paper. See Flood and Pacas (2017) for more information.

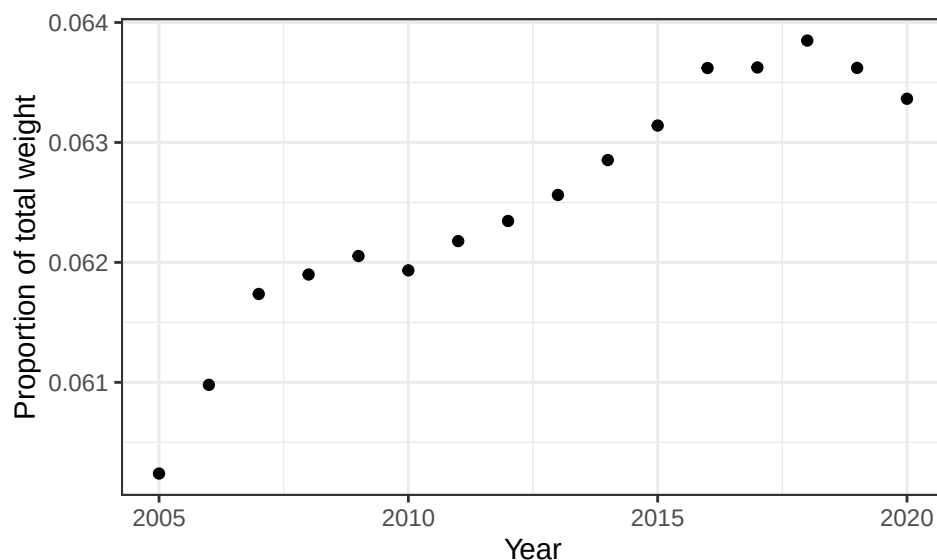


Figure C.7: The total weight on the sample increases over time. This is because the U.S. population ages 25–60 increases in size over time; when drawing inference about this population, the more recent years are weighed more heavily because they contain a larger share of the total population.

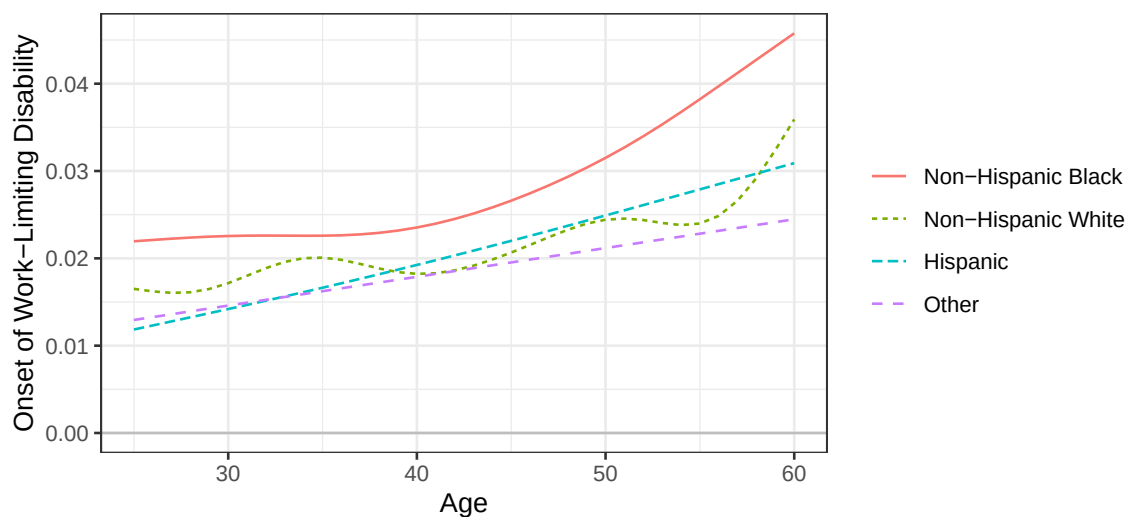


Figure C.8: Conditional association between age and work-limiting disability. These smooth terms are visualized here to build intuition for the Generalized Additive Model (GAM, Wood 2017) predicting the onset of a work-limiting disability as a function of all predictors. The association is free to differ by race because models are estimated separately by race, and the smooth curve is not restricted to any parametric form (e.g. quadratic). These curves have no causal interpretation; they depict the population average conditional association between age and the outcome.

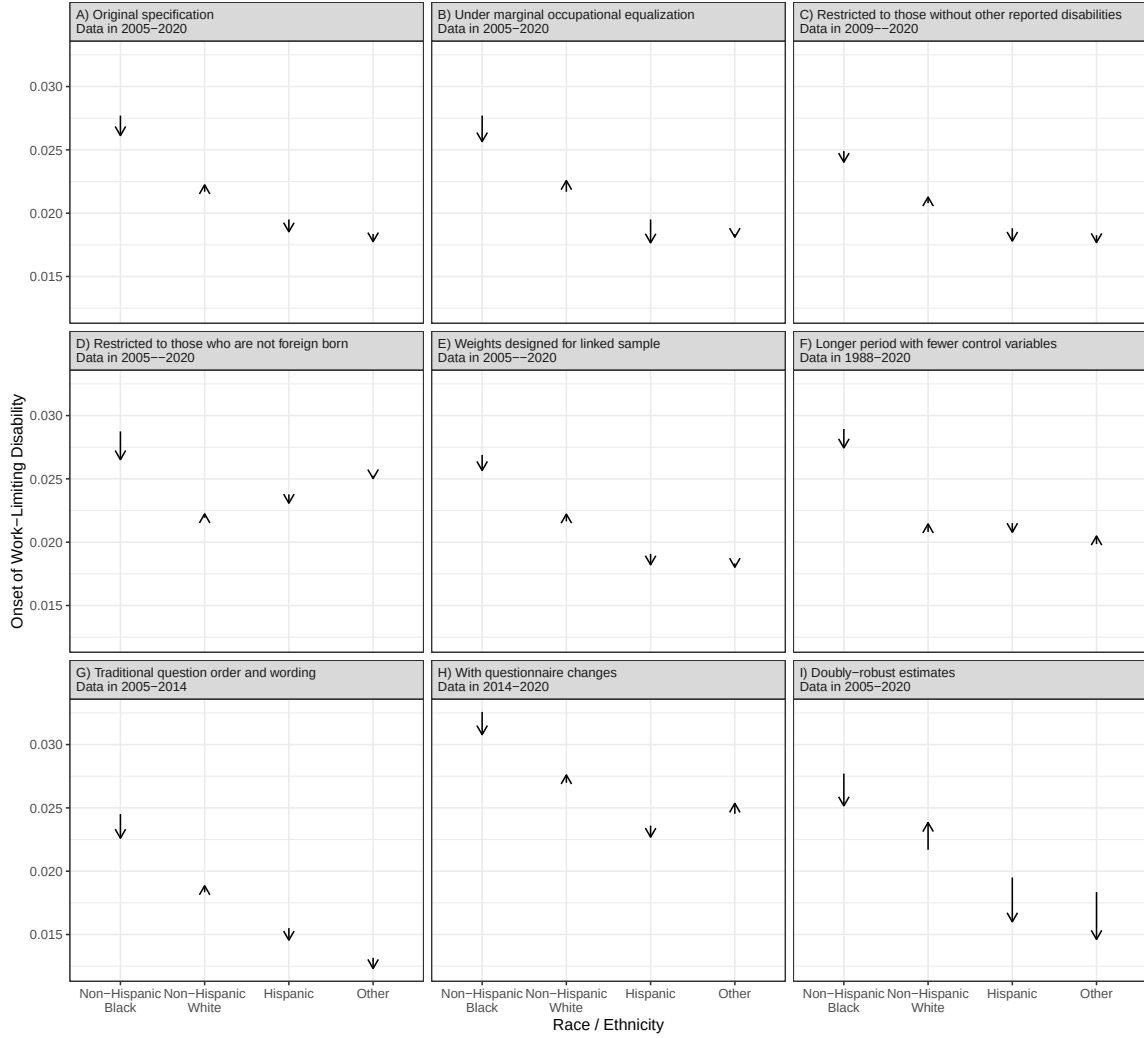


Figure C.9: Alternative specifications for the main result. Each arrow goes from the factual prevalence of work-limiting disability to the counterfactual prevalence under equalization of occupation assignments within education. In each panel, the reported years correspond to year t in which covariates and occupation are measured; the onset of a work-limiting disability is in the following year $t + 1$. Panel A is the specification from the main text ($N = 246,658$). Panel B shows results for marginal equalization: assign occupations by their population prevalence for all individuals regardless of education. Panel C uses additional data available in 2009–2020 to subset to those who report never leaving a job for health reasons and who report no difficulty with hearing, vision, remembering, walking or climbing stairs, performing basic activities outside the home along, or taking care of personal needs ($N = 165,961$). Panel D subsets to those who are not foreign-born ($N = 208,256$). Panel E uses the LNKFW1YWT weights provided by IPUMS for the subsample that links across years instead of the standard ASECWT weight ($N = 246,658$), which cannot be used in the main analyses because replicate versions of that weight are not available. Panel F provides point estimates for the full 1988–2019 period in which work-limiting disability was measured by the CPS, at the cost of not adjusting for foreign-born status or lagged self-rated health ($N = 531,896$). Panel G subsets to those who received the original version of the disability question in both year t and year $t + 1$ ($N = 156,247$). Panel H subsets to those who received the new version of the disability question in both year t and year $t + 1$, where the new version additionally prompted for disabilities lasting a short time ($N = 71,351$). Panel I shows results with a doubly-robust estimator for a slightly modified assignment rule (see Appendix C.4.1).

Bibliography

- Abbott, A. (1988). Transcending general linear reality. *Sociological Theory*, 6(2):169–186.
- Acharya, A., Blackwell, M., and Sen, M. (2016). Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review*, 110(3):512–529.
- Ahonen, E. Q., Fujishiro, K., Cunningham, T., and Flynn, M. (2018). Work as an inclusive part of population health inequities research and prevention. *American Journal of Public Health*, 108(3):306–311.
- An, W. and Glynn, A. N. (2019). Treatment effect deviation as an alternative to Blinder-Oaxaca decomposition for studying social inequality. *Sociological Methods & Research*, Online first.
- Angrist, J. D. and Chen, S. H. (2011). Schooling and the vietnam-era gi bill: Evidence from the draft lottery. *American Economic Journal: Applied Economics*, 3(2):96–118.
- Angrist, J. D. and Evans, W. N. (1998). Children and their parents’ labor supply: Evidence from exogenous variation in family size. *The American Economic Review*, 88(3):450–477.
- Aronow, P. M. and Miller, B. T. (2019). *Foundations of Agnostic Statistics*. Cambridge University Press.
- Aronow, P. M., Samii, C., 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.
- Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.
- Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.
- Banzhaf, S., Ma, L., and Timmins, C. (2019). Environmental justice: The economics of race, place, and pollution. *Journal of Economic Perspectives*, 33(1):185–208.
- Barber, J. S., Kusunoki, Y., Gatny, H. H., and Budnick, J. (2018). The dynamics of intimate partner violence and the risk of pregnancy during the transition to adulthood. *American Sociological Review*, 83(5):1020–1047.
- Barr, A. B., Simons, L. G., Simons, R. L., Beach, S. R., and Philibert, R. A. (2018). Sharing the burden of the transition to adulthood: African american young adults’ transition challenges and their mothers’ health risk. *American Sociological Review*, 83(1):143–172.
- Bearman, P. S., Moody, J., and Stovel, K. (2004). Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, 110(1):44–91.
- Beller, E. (2009). Bringing intergenerational social mobility research into the twenty-first century: Why mothers matter. *American Sociological Review*, 74(4):507–528.

- Berk, R., Buja, A., Brown, L., George, E., Kuchibhotla, A. K., Su, W., and Shazo, L. (2021). Assumption lean regression. *The American Statistician*, 75(1):76–84.
- Berk, R. A. (2004). *Regression Analysis: A Constructive Critique*. Sage.
- Berry, W. D., DeMeritt, J. H., and Esarey, J. (2010). Testing for interaction in binary logit and probit models: is a product term essential? *American Journal of Political Science*, 54(1):248–266.
- Bertrand, M. and Mullainathan, S. (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*, pages 647–671.
- Bickel, P. J., Hammel, E. A., and O’Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398–404.
- Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865.
- Blau, P. M. and Duncan, O. D. (1967). *The American Occupational Structure*. New York: Wiley.
- Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, pages 436–455.
- Blitzstein, J. K. and Hwang, J. (2019). *Introduction to Probability*. CRC Press.
- Bloome, D., Dyer, S., and Zhou, X. (2018). Educational inequality, educational expansion, and intergenerational income persistence in the United States. *American Sociological Review*, 83(6):1215–1253.
- Brand, J. E. and Xie, Y. (2010). Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. *American Sociological Review*, 75(2):273–302.
- Breen, R., Karlson, K. B., and Holm, A. (2018). Interpreting and understanding logits, probits, and other nonlinear probability models. *Annual Review of Sociology*, 44:39–54.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- Brucker, D. L., Mitra, S., Chaitoo, N., and Mauro, J. (2015). More likely to be poor whatever the measure: Working-age persons with disabilities in the United States. *Social Science Quarterly*, 96(1):273–296.
- Buchanan, S., Vossenas, P., Krause, N., Moriarty, J., Frumin, E., Shimek, J. A. M., Mirer, F., Orris, P., and Punnett, L. (2010). Occupational injury disparities in the U.S. hotel industry. *American Journal of Industrial Medicine*, 53(2):116–125.
- Buchmann, C. and DiPrete, T. A. (2006). The growing female advantage in college completion: The role of family background and academic achievement. *American Sociological Review*, 71(4):515–541.
- Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., Zhao, L., et al. (2019). Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544.
- Bureau of Labor Statistics (2006). *Design and Methodology: Current Population Survey Technical Paper 66*. U.S. Department of Labor.
- Bureau of Labor Statistics (2020). Employer-reported workplace injuries and illnesses–2019. *U.S. Department of Labor News Release*, 20–2030.

- Burgard, S. A. and Lin, K. Y. (2013). Bad jobs, bad health? how work and working conditions contribute to health disparities. *American Behavioral Scientist*, 57(8):1105–1127.
- Burkhauser, R. V., Butler, J., Kim, Y.-W., and Weathers, R. R. (1999). The importance of accommodation on the timing of disability insurance applications: Results from the survey of disability and work and the Health and Retirement Study. *Journal of Human Resources*, pages 589–611.
- Burkhauser, R. V., Daly, M. C., Houtenville, A. J., and Nargis, N. (2002). Self-reported work-limitation data: What they can and cannot tell us. *Demography*, 39(3):541–555.
- Cherniack, M. (1986). *The Hawk’s Nest Incident: America’s Worst Industrial Disaster*. New Haven, CT: Yale University Press.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.
- Chetty, R., Hendren, N., Jones, M. R., and Porter, S. R. (2020). Race and economic opportunity in the united states: An intergenerational perspective. *The Quarterly Journal of Economics*, 135(2):711–783.
- Chetty, R., Hendren, N., Kline, P., Saez, E., and Turner, N. (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 DiPrete, T. A. (2018). The paradox of persistence: Explaining the black-white gap in bachelor’s degree completion. *American Sociological Review*, 83(6):1171–1214.
- Coleman, J. S. (1990). *Foundations of Social Theory*. Harvard University Press.
- Czajka, J. L., Rosso, R., et al. (2015). Redesign of the income questions in the Current Population Survey Annual Social and Economic Supplement: Further analysis of the 2014 split-sample test. *Mathematica Policy Research*.
- Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48(2):424–55.
- Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2–21.
- Desmond, M. and Travis, A. (2018). Political consequences of survival strategies among the urban poor. *American Sociological Review*, 83(5):869–896.
- Du Bois, W. E. (1899). *The Philadelphia Negro: A Social Study*. Philadelphia: University of Pennsylvania Press.
- Dudík, M., Erhan, D., Langford, J., Li, L., et al. (2014). Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511.
- Duncan, O. D. (1984). *Notes on Social Measurement: Historical and Critical*. Russell Sage Foundation.
- Durlauf, S. N. and Heckman, J. J. (2020). An empirical analysis of racial differences in police use of force: A comment. *Journal of Political Economy*, 128(10).
- D’Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2020). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*.

- Edelmann, A., Wolff, T., Montagne, D., and Bail, C. A. (2020). Computational social science and sociology. *Annual Review of Sociology*, 46:61–81.
- Edwards, R. and Smith, S. M. (2020). Job market remains tight in 2019, as the unemployment rate falls to its lowest level since 1969. *Bureau of Labor Statistics: Monthly Labor Review*, April.
- Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC press.
- Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., and Stewart, B. M. (2018). How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*.
- Elwert, F. and Winship, C. (2010). Effect heterogeneity and bias in main-effects-only regression models. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, pages 327–36.
- Elwert, F. and Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40:31–53.
- Erikson, R., Goldthorpe, J. H., and Portocarero, L. (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.
- Fay, R. E. and Train, G. F. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. In *Proceedings of the Section on Government Statistics, American Statistical Association, Alexandria, VA*, pages 154–159.
- Ferguson, J.-P. and Koning, R. (2018). Firm turnover and the return of racial establishment segregation. *American Sociological Review*, 83(3):445–474.
- Firebaugh, G., Warner, C., and Massoglia, M. (2013). Fixed effects, random effects, and hybrid models for causal analysis. In *Handbook of Causal Analysis for Social Research*, pages 113–132. Springer.
- Fleischer, N. L., Tiesman, H. M., Sumitani, J., Mize, T., Amarnath, K. K., Bayakly, A. R., and Murphy, M. W. (2013). Public health impact of heat-related illness among migrant farmworkers. *American Journal of Preventive Medicine*, 44(3):199–206.
- Flood, S., King, M., Rodgers, R., Ruggles, S., and Warren, J. R. (2020). *Integrated Public Use Microdata Series, Current Population Survey: Version 8.0 [dataset]*. Minneapolis, MN: IPUMS.
- Flood, S. M. and Pacas, J. D. (2017). Using the Annual Social and Economic Supplement as part of a Current Population Survey Panel. *Journal of Economic and Social Measurement*, 42(3-4):225–248.
- Flores, R. D. and Schachter, A. (2018). Who are the “illegals”? The social construction of illegality in the united states. *American Sociological Review*, 83(5):839–868.
- Font, S. A., Berger, L. M., Cancian, M., and Noyes, J. L. (2018). Permanency and the educational and economic attainment of former foster children in early adulthood. *American Sociological Review*, 83(4):716–743.
- Freedman, D. A. (1991). Statistical models and shoe leather. *Sociological Methodology*, 21:291–313.

- Freeland, R. E. and Hoey, J. (2018). The structure of deference: Modeling occupational status using affect control theory. *American Sociological Review*, 83(2):243–277.
- Freese, J. and Peterson, D. (2017). Replication in social science. *Annual Review of Sociology*, 43:147–165.
- Fryer, R. G. (2019). An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, 127(3):1210–1261.
- Fryer, R. G. (2020). An empirical analysis of racial differences in police use of force: A response. *Journal of Political Economy*, 128(10).
- Fujishiro, K., Hajat, A., Landsbergis, P. A., Meyer, J. D., Schreiner, P. J., and Kaufman, J. D. (2017). Explaining racial/ethnic differences in all-cause mortality in the multi-ethnic study of atherosclerosis (mesa): Substantive complexity and hazardous working conditions as mediating factors. *SSM-Population Health*, 3:497–505.
- Gauchat, G. and Andrews, K. T. (2018). The cultural-cognitive mapping of scientific professions. *American Sociological Review*, 83(3):567–595.
- Geronimus, A. T. (1992). The weathering hypothesis and the health of african-american women and infants: Evidence and speculations. *Ethnicity & Disease*, 2(3):207–221.
- Geronimus, A. T., Hicken, M., Keene, D., and Bound, J. (2006). “Weathering” and age patterns of allostatic load scores among blacks and whites in the United States. *American Journal of Public Health*, 96(5):826–833.
- Glynn, A. N. and Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1):36–56.
- Goldstein, A. (2018). The social ecology of speculation: Community organization and non-occupancy investment in the U.S. housing bubble. *American Sociological Review*, 83(6):1108–1143.
- Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306.
- Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.
- Greiner, D. J. and Rubin, D. B. (2011). Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93(3):775–785.
- Guinea-Martin, D., Mora, R., and Ruiz-Castillo, J. (2018). The evolution of gender segregation over the life course. *American Sociological Review*, 83(5):983–1019.
- Gutierrez, C. M. (2018). The institutional determinants of health insurance: Moving away from labor market, marriage, and family attachments under the ACA. *American Sociological Review*, 83(6):1144–1170.
- Hahl, O., Kim, M., and Zuckerman Sivan, E. W. (2018). The authentic appeal of the lying demagogue: Proclaiming the deeper truth about political illegitimacy. *American Sociological Review*, 83(1):1–33.
- Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.
- Hale, T. W. (2001). The lack of a disability measure in today’s current population survey. *Monthly Labor Review*, 124:38.

- Harding, D. J., Morenoff, J. D., Nguyen, A. P., and Bushway, S. D. (2018). Imprisonment and labor market outcomes: Evidence from a natural experiment. *American Journal of Sociology*, 124(1):49–110.
- Harknett, K., Schneider, D., and Wolfe, R. (2020). Losing sleep over work scheduling? The relationship between work schedules and sleep quality for service sector workers. *SSM-Population Health*, 12:100681.
- Heckman, J. J. and Urzua, S. (2010). Comparing IV with structural models: What simple IV can and cannot identify. *Journal of Econometrics*, 156(1):27–37.
- Hedström, P. and Bearman, P. (2009). *The Oxford Handbook of Analytical Sociology: Oxford University Press, 2009*. Oxford University Press.
- Hernán, M. A. (2016). Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*, 26(10):674–680.
- Hernán, M. A. (2018). The c-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health*, 108(5):616–619.
- Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.
- Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R., and Shrier, I. (2016a). Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology*, 79:70–75.
- Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R., and Shrier, I. (2016b). Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology*, 79:70–75.
- Heymann, J. (2003). *Global Inequalities at Work: Work’s Impact on the Health of Individuals, Families, and Societies*. Oxford University Press.
- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.
- Hirschman, D. (2016). Stylized facts in the social sciences. *Sociological Science*, 3:604–626.
- Hofman, J. M., Sharma, A., and Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, 355(6324):486–488.
- 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*, pages 93–109.
- Honaker, J., King, G., Blackwell, M., et al. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47.
- Horowitz, J. (2018). Relative education and the advantage of a college degree. *American Sociological Review*, 83(4):771–801.
- Horvitz, D. G. and Thompson, D. J. (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.

- Hout, M. (2012). Social and economic returns to college education in the United States. *Annual Review of Sociology*, 38:379–400.
- Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(4):765–789.
- Imbens, G. (2018). Comments on understanding and misunderstanding randomized controlled trials: A commentary on Cartwright and Deaton. *Social Science & Medicine*.
- Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.
- Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, 48(2):399–423.
- Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica: Journal of the Econometric Society*, pages 467–475.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- Inanc, H. (2018). Unemployment, temporary work, and subjective well-being: The gendered effect of spousal labor market insecurity. *American Sociological Review*, 83(3):536–566.
- 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 Arah, O. A. (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 VanderWeele, T. J. (2018). Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology*, 29(6):825–835.
- Jajtner, K. M., Mitra, S., Fountain, C., Nichols, A., et al. (2020). Rising income inequality through a disability lens: Trends in the United States 1981–2018. *Social Indicators Research*, pages 1–34.
- Jonsson, J. O., Grusky, D. B., Di Carlo, M., Pollak, R., and Brinton, M. C. (2009). Microclass mobility: Social reproduction in four countries. *American Journal of Sociology*, 114(4):977–1036.
- Juon, H.-S., Hong, A., Pimpinelli, M., Rojulpote, M., McIntire, R., and Barta, J. A. (2021). Racial disparities in occupational risks and lung cancer incidence: Analysis of the national lung screening trial. *Preventive Medicine*, 143:106355.
- Kadivar, M. A. (2018). Mass mobilization and the durability of new democracies. *American Sociological Review*, 83(2):390–417.
- Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

- Katz, J. N., King, G., and Rosenblatt, E. (2020). Theoretical foundations and empirical evaluations of partisan fairness in district-based democracies. *American Political Science Review*, 114(1):164–178.
- Kaufman, J. S. (2008). Epidemiologic analysis of racial/ethnic disparities: some fundamental issues and a cautionary example. *Social Science & Medicine*, 66(8):1659–1669.
- Keele, L., Stevenson, R. T., and Elwert, F. (2020). The causal interpretation of estimated associations in regression models. *Political Science Research and Methods*, 8(1):1–13.
- Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (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 Bryan, B. (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.
- Knox, D., Lowe, W., and Mummolo, J. (2020). Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637.
- 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 Rao, J. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, pages 1010–1019.
- Krieger, N. (2001). A glossary for social epidemiology. *Journal of Epidemiology & Community Health*, 55(10):693–700.
- Laurison, D. and Friedman, S. (2016). The class pay gap in higher professional and managerial occupations. *American Sociological Review*, 81(4):668–695.
- Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., et al. (2020). Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062.
- Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.
- Lewis, T. T., Cogburn, C. D., and Williams, D. R. (2015). Self-reported experiences of discrimination and health: scientific advances, ongoing controversies, and emerging issues. *Annual Review of Clinical Psychology*, 11:407–440.
- Liebertson, S. (1987). *Making It Count: The Improvement of Social Research and Theory*. University of California Press.
- Liebertson, S. and Horwich, J. (2008). Implication analysis: A pragmatic proposal for linking theory and data in the social sciences. *Sociological Methodology*, 38(1):1–50.
- Link, B. G. and Phelan, J. (1995). Social conditions as fundamental causes of disease. *Journal of Health and Social Behavior*, pages 80–94.
- Lipscomb, H. J., Loomis, D., McDonald, M. A., Argue, R. A., and Wing, S. (2006). A conceptual model of work and health disparities in the United States. *International Journal of Health Services*, 36(1):25–50.

- Liu, H. (2018). Social and genetic pathways in multigenerational transmission of educational attainment. *American Sociological Review*, 83(2):278–304.
- Lowry, S. J., Blecker, H., Camp, J., De Castro, B., Hecker, S., Arbabi, S., Traven, N., and Seixas, N. S. (2010). Possibilities and challenges in occupational injury surveillance of day laborers. *American Journal of Industrial Medicine*, 53(2):126–134.
- Lucas, S. R. (2001). Effectively maintained inequality: Education transitions, track mobility, and social background effects. *American Journal of Sociology*, 106(6):1642–1690.
- Ludwig, V. and Brüderl, J. (2018). Is there a male marital wage premium? New evidence from the United States. *American Sociological Review*, 83(4):744–770.
- 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.
- Lundberg, I. (2021). The gap-closing estimand: A causal approach to study interventions that close disparities across social categories. SocArXiv manuscript: <https://doi.org/10.31235/osf.io/gx4y3>.
- Lundberg, I., Johnson, R., and Stewart, B. M. (2021a). What is your estimand? defining the target quantity connects statistical evidence to theory. <https://osf.io/preprints/socarxiv/ba67n/>.
- Lundberg, I., Johnson, R., and Stewart, B. M. (2021b). What is your estimand? Defining the target quantity connects statistical evidence to theory. SocArXiv manuscript: <https://doi.org/10.31235/osf.io/ba67n>.
- Maestas, N., Mullen, K. J., and Rennane, S. (2019). Unmet need for workplace accommodation. *Journal of Policy Analysis and Management*, 38(4):1004–1027.
- Marmot, M. and Wilkinson, R. (2005). *Social Determinants of Health*. Oxford University Press.
- McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403.
- McDonnell, M.-H. and King, B. G. (2018). Order in the court: How firm status and reputation shape the outcomes of employment discrimination suits. *American Sociological Review*, 83(1):61–87.
- McEwen, B. S. (1998). Protective and damaging effects of stress mediators. *New England Journal of Medicine*, 338(3):171–179.
- McIntosh, K., Moss, E., Nunn, R., and Shambaugh, J. (2020). Examining the black-white wealth gap. *Washington DC: Brookings Institutes*.
- Merkin, S. S., Basurto-Dávila, R., Karlamangla, A., Bird, C. E., Lurie, N., Escarce, J., and Seeman, T. (2009). Neighborhoods and cumulative biological risk profiles by race/ethnicity in a national sample of us adults: Nhanes iii. *Annals of Epidemiology*, 19(3):194–201.
- Meyer, B. D. and Mok, W. K. (2019). Disability, earnings, income and consumption. *Journal of Public Economics*, 171:51–69.

- Mize, T. D. (2016). Sexual orientation in the labor market. *American Sociological Review*, 81(6):1132–1160.
- Mize, T. D. (2019). Best practices for estimating, interpreting, and presenting non-linear interaction effects. *Sociological Science*, 6:81–117.
- Mize, T. D. and Manago, B. (2018). Precarious sexuality: How men and women are differentially categorized for similar sexual behavior. *American Sociological Review*, 83(2):305–330.
- Moen, P., Kelly, E. L., Tranby, E., and Huang, Q. (2011). Changing work, changing health: can real work-time flexibility promote health behaviors and well-being? *Journal of Health and Social Behavior*, 52(4):404–429.
- Molina, M. and Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology*, 45:27–45.
- Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, 26(1):67–82.
- Morgan, S. L. (2017). A coding of social class for the general social survey. *Chicago, Illinois: GSS Methodological Report No. 125*.
- Morgan, S. L. and Winship, C. (2015). *Counterfactuals and Causal Inference*. Cambridge University Press.
- Muñoz Díaz, I. and van der Laan, M. (2012). Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549.
- Mullins, D. A., Hoyer, D., Collins, C., Currie, T., Feeney, K., François, P., Savage, P. E., Whitehouse, H., and Turchin, P. (2018). A systematic assessment of “axial age” proposals using global comparative historical evidence. *American Sociological Review*, 83(3):596–626.
- Mun, E. and Jung, J. (2018). Policy generosity, employer heterogeneity, and women’s employment opportunities: The welfare state paradox reexamined. *American Sociological Review*, 83(3):508–535.
- Murphy, S. A., van der Laan, M. J., Robins, J. M., and Group, C. P. P. R. (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423.
- Mustillo, S. A., Lizardo, O. A., and McVeigh, R. M. (2018). Editors’ comment: A few guidelines for quantitative submissions. *American Sociological Review*, 83(6):1281–1283.
- Nagler, J. (1991). The effect of registration laws and education on U.S. voter turnout. *The American Political Science Review*, pages 1393–1405.
- Naimi, A. I., Schnitzer, M. E., Moodie, E. E., and Bodnar, L. M. (2016). Mediation analysis for health disparities research. *American Journal of Epidemiology*, 184(4):315–324.
- Nguyen, T. Q., Schmid, I., and Stuart, E. A. (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*, pages 693–709.

- Offer, S. and Fischer, C. S. (2018). Difficult people: Who is perceived to be demanding in personal networks and why are they there? *American Sociological Review*, 83(1):111–142.
- Ólafsdóttir, S. and Beckfield, J. (2020). Health inequalities: What do we know and what do we need to know? *Social Science & Medicine*, 267:113575–113575.
- Omi, M. and Winant, H. (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.
- Pais, J. (2014). Cumulative structural disadvantage and racial health disparities: The pathways of childhood socioeconomic influence. *Demography*, 51(5):1729–1753.
- Pal, I. and Waldfogel, J. (2016). The family gap in pay: New evidence for 1967 to 2013. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 2(4):104–127.
- 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*, pages 411–420.
- Pearl, J. (2009). *Causality*. Cambridge University Press.
- Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.
- Petersen, T. and Morgan, L. A. (1995). Separate and unequal: Occupation-establishment sex segregation and the gender wage gap. *American Journal of Sociology*, 101(2):329–365.
- Pfeffer, F. T. and Hertel, F. R. (2015). How has educational expansion shaped social mobility trends in the United States? *Social Forces*, 94(1):143–180.
- Phelan, J. C. and Link, B. G. (2015). Is racism a fundamental cause of inequalities in health? *Annual Review of Sociology*, 41:311–330.
- Preston, S., Heuveline, P., and Guillot, M. (2000). *Demography: Measuring and Modeling Population Processes*. Malden, MA: Blackwell Publishers.
- Prins, S. J., McKetta, S., Platt, J., Muntaner, C., Keyes, K. M., and Bates, L. M. (2019). Mental illness, drinking, and the social division and structure of labor in the United States: 2003-2015. *American Journal of Industrial Medicine*, 62(2):131–144.
- Quadlin, N. (2018). The mark of a woman’s record: Gender and academic performance in hiring. *American Sociological Review*, 83(2):331–360.
- Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25:111–164.
- Raftery, A. E. and Hout, M. (1993). Maximally maintained inequality: Expansion, reform, and opportunity in Irish education, 1921-75. *Sociology of Education*, pages 41–62.
- Rainey, C. (2016). Compression and conditional effects: A product term is essential when using logistic regression to test for interaction. *Political Science Research and Methods*, 4(3):621–639.
- Rao, J. N. and Wu, C. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401):231–241.

- Reichman, N. E., Teitler, J. O., Garfinkel, I., and McLanahan, S. S. (2001). Fragile Families: Sample and design. *Children and Youth Services Review*, 23(4-5):303–326.
- Rivera Drew, J. A., Flood, S., and Warren, J. R. (2014). Making full use of the longitudinal design of the Current Population Survey: Methods for linking records across 16 months. *Journal of Economic and Social Measurement*, 39(3):121–144.
- 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., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (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*, pages 70–82.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(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., Díaz, I., Rosenblum, M., and Stuart, E. A. (2014). Estimating population treatment effects from a survey subsample. *American Journal of Epidemiology*, 180(7):737–748.
- Ruggles, S., Flood, S., Goeken, R., Grover, J., Erin, M., Pacas, J., and Sobek, M. (2019). *IPUMS USA: Version 9.0 [dataset]*. Minneapolis, MN: IPUMS.
- Rust, K. F. and Rao, J. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5(3):283–310.
- Salganik, M. J., Lundberg, I., Kindel, A. T., 108 others, and McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403.
- Samii, C. (2016). Causal empiricism in quantitative research. *The Journal of Politics*, 78(3):941–955.
- Saperstein, A. and Penner, A. M. (2012). Racial fluidity and inequality in the United States. *American Journal of Sociology*, 118(3):676–727.
- Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.
- Schilke, O. and Rossman, G. (2018). It’s only wrong if it’s transactional: Moral perceptions of obfuscated exchange. *American Sociological Review*, 83(6):1079–1107.
- Schneider, D., Hastings, O. P., and LaBriola, J. (2018). Income inequality and class divides in parental investments. *American Sociological Review*, 83(3):475–507.
- Schulte, P. A., Pana-Cryan, R., Schnorr, T., Schill, A. L., Guerin, R., Felknor, S., and Wagner, G. R. (2017). An approach to assess the burden of work-related injury, disease, and distress. *American Journal of Public Health*, 107(7):1051–1057.

- Seabury, S. A., Terp, S., and Boden, L. I. (2017). Racial and ethnic differences in the frequency of workplace injuries and prevalence of work-related disability. *Health Affairs*, 36(2):266–273.
- Sen, M. and Wasow, O. (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.
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.
- Simpson, B., Harrell, A., Melamed, D., Heiserman, N., and Negraia, D. V. (2018). The roots of reciprocity: Gratitude and reputation in generalized exchange systems. *American Sociological Review*, 83(1):88–110.
- Smith, T. W., Davern, M., Freese, J., and Morgan, S. (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](http://gssdataexplorer.norc.uchicago.edu).
- Stainback, K. and Tomaskovic-Devey, D. (2012). *Documenting Desegregation: Racial and Gender Segregation in Private Sector Employment Since the Civil Rights Act*. Russell Sage Foundation.
- Stanbury, M. and Rosenman, K. D. (2014). Occupational health disparities: A state public health-based approach. *American Journal of Industrial Medicine*, 57(5):596–604.
- Storer, A., Schneider, D., and Harknett, K. (2020). What explains racial/ethnic inequality in job quality in the service sector? *American Sociological Review*, 85(4):537–572.
- Tomaskovic-Devey, D., Zimmer, C., Stainback, K., Robinson, C., Taylor, T., and McTague, T. (2006). Documenting desegregation: Segregation in American workplaces by race, ethnicity, and sex, 1966–2003. *American Sociological Review*, 71(4):565–588.
- 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.
- Torrats-Espinosa, G. (2021). Using machine learning to estimate the effect of racial segregation on COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, 118(7).
- 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 Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media.
- VanderWeele, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

- VanderWeele, T. and Robinson, W. (2014). On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology*, 25(4):473–484.
- Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.
- Villarreal, A. and Tamborini, C. R. (2018). Immigrants’ economic assimilation: Evidence from longitudinal earnings records. *American Sociological Review*, 83(4):686–715.
- Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.
- Watts, D. J. (2014). Common sense and sociological explanations. *American Journal of Sociology*, 120(2):313–351.
- Wedow, R., Zacher, M., Huibregtse, B. M., Mullan Harris, K., Domingue, B. W., and Boardman, J. D. (2018). Education, smoking, and cohort change: Forwarding a multidimensional theory of the environmental moderation of genetic effects. *American Sociological Review*, 83(4):802–832.
- Weisshaar, K. (2017). Publish and perish? An assessment of gender gaps in promotion to tenure in academia. *Social Forces*, 96(2):529–560.
- Weisshaar, K. (2018). From opt out to blocked out: The challenges for labor market re-entry after family-related employment lapses. *American Sociological Review*, 83(1):34–60.
- Western, B. (2006). *Punishment and Inequality in America*. Russell Sage Foundation.
- Westreich, D. and Greenland, S. (2013). The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, 177(4):292–298.
- Wildeman, C., Schnittker, J., and Turney, K. (2012). Despair by association? the mental health of mothers with children by recently incarcerated fathers. *American Sociological Review*, 77(2):216–243.
- Williams, D. R., Lawrence, J. A., and Davis, B. A. (2019). Racism and health: Evidence and needed research. *Annual Review of Public Health*, 40:105–125.
- Williams, D. R. and Mohammed, S. A. (2009). Discrimination and racial disparities in health: Evidence and needed research. *Journal of Behavioral Medicine*, 32(1):20–47.
- Wilmers, N. (2018). Wage stagnation and buyer power: How buyer-supplier relations affect U.S. workers’ wages, 1978 to 2014. *American Sociological Review*, 83(2):213–242.
- Winship, C. and Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods & Research*, 23(2):230–257.
- Wodtke, G. T., Harding, D. J., and Elwert, F. (2011). Neighborhood effects in temporal perspective: The impact of long-term exposure to concentrated disadvantage on high school graduation. *American Sociological Review*, 76(5):713–736.
- Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

- Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.
- Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(i01).
- Xie, Y. (2013). Population heterogeneity and causal inference. *Proceedings of the National Academy of Sciences*, 110(16):6262–6268.
- Young, C. and Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods and Research*, 46(1):3–40.
- Zhang, J. and Bareinboim, E. (2018). Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 2037–2045.
- 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 Wodtke, G. T. (2019). A regression-with-residuals method for estimating controlled direct effects. *Political Analysis*, 27(3):360–369.