Setting the Target: Precise Estimands and the Gap Between Theory and Empirics [*]

Ian Lundberg[†] Rebecca Johnson[‡] Brandon M. Stewart[§]

January 7, 2020

Keywords: social statistics, research design, quantitative methods, causal inference, estimands

Abstract

The link between theory and quantitative empirical evidence is a longstanding hurdle in sociological research. Ambiguity about the role that statistical evidence plays in an argument may produce misleading conclusions and poor methodological practice. This ambiguity could be reduced if researchers would state the theoretical estimand—the central quantity at the core of a given paper—in precise language. Our approach envisions three choices in the research process: (1) choice of a theoretical estimand, which will be informative for theory, (2) choice of an empirical estimand, which is informative about the theoretical estimand under some identification assumptions, and (3) choice of an estimation strategy to learn the empirical estimand from data. Key advantages of this approach include improved clarity on the object of interest, transparency about how empirical evidence contributes to knowledge of that quantity, and the ability to easily plug in new statistical tools for estimation.

(145 words)

[*] Replication code is posted on Github [link] and will be posted on Dataverse once the paper is forthcoming in a journal. For helpful discussions and feedback relevant to this project, we thank Dalton Conley, Matt Desmond, Adam Goldstein, Justin Grimmer, Tod Hamilton, Erin Hartman, Daniel Karell, Gary King, Sarah Mustillo, Matt Salganik, and members of the Stewart Lab. 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.

[†] Ph.D. Candidate, Department of Sociology and Office of Population Research, Princeton University, ianlundberg.org, ilundberg@princeton.edu

[‡] Ph.D. Candidate, Department of Sociology and Office of Population Research, Princeton University, https://scholar.princeton.edu/rebeccajohnson/, raj2@princeton.edu

[§] Assistant Professor, Department of Sociology and Office of Population Research, Princeton University, brandonstewart.org, bms4@princeton.edu. 149 Wallace Hall, Princeton University, Princeton, NJ 08540.

1 Introduction

Methodological choices in sociology are dominated by concerns about model specification. In the 2019 volume of the *American Sociological Review*, two papers use the exact phrase "results are robust to alternative specifications" (Homan, 2019; Travis, 2019). Others echo the sentiment, writing that "findings are consistent across a number of additional model specifications" (Legewie and Fagan, 2019), that results hold "for a variety of different model specifications" (Chmielewski, 2019), and that the finding "is substantively and statistically indistinguishable across these different specifications" (Light and Thomas, 2019). Another study footnotes steps taken "to rule out misspecification" (Wang et al., 2019). A comment claims that the original study "suffers from misspecification" (Auspurg et al., 2019). These papers share a commitment to finding a properly-specified, 'correct' regression model. Yet, if we are honest, all empirical models suffer from misspecification. This manuscript provides a framework for evaluating quantitative work under the baseline assumption that models are simply tools that help us approximate unknown quantities of interest and cannot be, in a general sense, correct.

Quantitative contributions in sociology often equate the research goal with one or more coefficients of a regression model. The most troubling implication of this perspective combined with a worldview in which all models are misspecified is that the research goal itself is ill-defined. Because we view the model as only an approximation to a more complex reality, we argue that any coefficient in that model is at best an approximation to some target quantity that can be defined apart from the model. The first step to sound methodological choices, then, is to state this target quantity—the estimand—separately from the method used to approximate it. Stating an estimand guides subsequent methodological choices and aids transparent communication among authors, reviewers, and readers.

An estimand is some function (e.g. a mean) of the population distribution of some unit-specific quantity (such as a variable Y_i or a unit-specific causal effect τ_i). Examples of estimands include the unemployment rate or the population-average effect of job training on employment. These estimands are generally unstated, with methods sections discussing the tools—survey designs, experimental comparisons, regression specifications—that are best viewed as estimation strategies used to approximate the estimand. A separate statement of the estimand clarifies the goal of the analysis and facilitates reasoning about the sense in which a particular regression (or other estimation procedure) is a reasonable approximation. It also frees researchers to consider quantities not easily captured by a single regression coefficient and clarifies how black-box machine learning algorithm can be used to estimate conventional quantities of interest. In contrast, vague goals and buried assumptions at best hinder scholarly conversation and at worst can sow deeply misleading conclusions, as we show in three studies about demographic disparities police shootings, graduate admissions, and income (Section 2.2).

We propose a standard under which quantitative work partitions the link between theory and evidence into three steps, each of which involves different kinds of approximations that can be defended by different types of argument (Fig 1). (1) Choice of a *theoretical estimand*, which is the quantity that would most advance theory if it were known. The theoretical estimand summarizes the population distribution of some quantity τ_i specific to each unit i. Some τ_i may not be observable, such as individual causal effects. (2) Choice of an *empirical estimand* that might be learned from the data we observe, such as the mean difference between those in a treatment and control group. The empirical and theoretical estimands are linked by assumptions about the data we do not directly observe (e.g. random assignment). (3) Choice of an *estimation strategy* to learn the empirical estimand, such as a regression model.

The three steps involve different kinds of argument. The choice of a theoretical estimand requires substantive argument about the broader theory and goals of inference, the choice of an empirical estimand requires conceptual argument about unobserved and unobservable data, and the choice of an estimation strategy can be entirely grounded in the data. Separat-

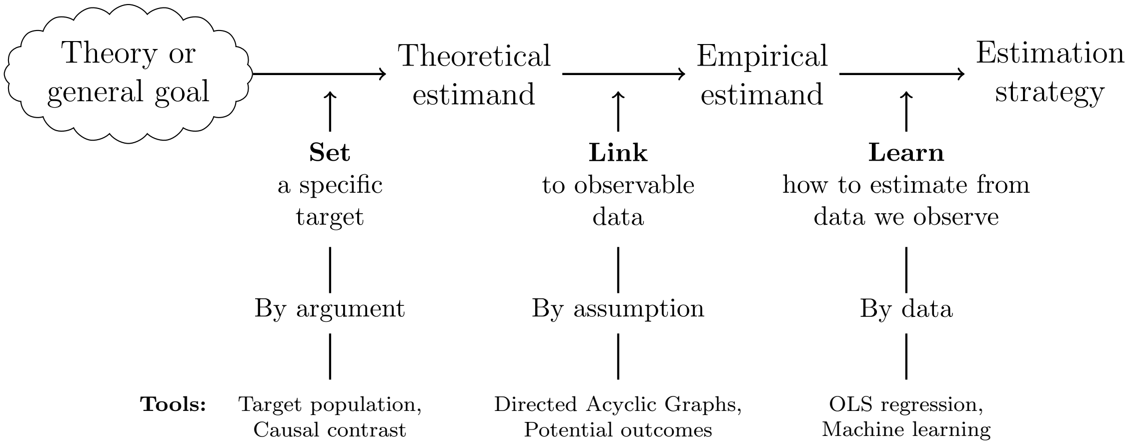


Fig. 1. Three critical choices in quantitative social science arguments. The first choice is the theoretical estimand, which sets the target of inference. Argument is required to link the theoretical estimand to the broader theory. The second choice is the empirical estimand, which links the target to observable data. The connection requires substantive assumptions that can be formalized in Directed Acyclic Graphs. The third choice is the estimation strategy, which captures what we will actually do with data. We can learn a good estimation strategy by empirical considerations with the data available through a machine learning approach.

ing these steps would help authors make principled methodological choices, help reviewers pinpoint the source of their concerns, and help readers understand exactly how new studies extend what is known from prior work.

We are not the first sociologists to argue for a perspective in which models are viewed as approximations. Prior scholars have emphasized that the search for a correct model holds social science to an unattainable standard from the natural sciences (Lieberson, 1987), criticized the tendency of sociologists to rely on an assumed general linear reality (Abbott, 1988), highlighted unwarranted methodological focus on the specification of regression models (Freedman, 1991), and written entire textbooks that view the model as an approximation (Aronow and Miller, 2019; Berk, 2004). Yet, somehow these critiques have not fully permeated research practice which continues to rely on research goals equated with regression coefficients. Our hope is to gain traction on this problem by providing a three-step process for research that places the estimand at the center, thus giving applied researchers the tools to no longer frame the goal of their research as testing a regression coefficient.[1]

The paper proceeds in several sections. We first introduce our overall framework, highlighting each step of our proposed research process: setting the theoretical estimand (2.1), linking to an empirical estimand (2.2), and learning an estimate from data (2.3). Each section shows how the framework clarifies specific issues in published studies that range from debates about how to design audit studies (Section 2.1.1) to debates about how to measure racially-biased policing (Section 2.2). We defer connections to prior work until Section 3, by which point we have laid the necessary technical foundation for a thorough discussion. Section 4 highlights three reasons estimands are essential in all quantitative work: they ground methodological choices, add clarity to the search for model robustness, and partition the role of evidence in social science. Finally, in Section 5, we raise and reply to possible objections before concluding.

2 Estimands Link Theory and Evidence

This section details each step in our proposed link between theory and evidence, referencing a series of real settings from published research (Table 1) in which clarity about the theoretical estimand, empirical estimand, and estimation strategy may clarify some of the methodological choices in published research.

2.1 The Theoretical Estimand: Set the Target

The first step in any research process—and the step we advocate most strongly—is a precise definition of the research goal. Many research goals can be stated in a straightforward form: as the mean (or other summary) over some population of units of a unit-specific quantity

[1] Past accounts have already problematized goals stated by regression specifications. Our framework offers a new solution. For instance, Abbott (1988) proposes three tools for transcending a general linear reality: demographic methods, sequence analysis, and network methods. Yet researchers using any of these methods still need to state a target quantity that motivates the approach.

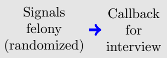
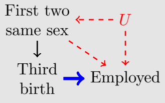
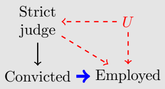
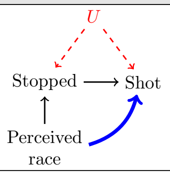
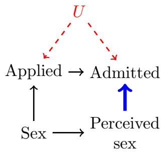
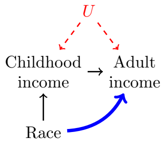
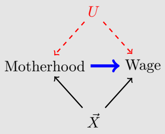
	Set the target		Link to observables	Learn from data
	Causal contrast	Target population	Identification	Estimation
Pager	Signals felony conviction	Applications to jobs in Milwaukee	 Signals felony (randomized) Callback for interview	Mean difference
Angrist and Evans	Third birth	Those who would have a third birth only if first two of the same sex		Two-stage least squares
Harding et al.	Convicted	Those who would be convicted only under certain judges		Two-stage least squares
Fryer	Perceived race	Those stopped by police		Mean difference
Bickel et al.	Perceived sex	Applicants to Berkeley		Mean difference
Chetty et al.	Childhood income	U.S. population		OLS
Pal and Waldfogel	Motherhood	U.S. civilian women ages 25–44 in March 2019		OLS Parametric *g*-formula

Table 1. 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) 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.

subscripted by i. For instance, we might study the employment rate among U.S. adults.

$$\frac{1}{\substack{\text{Size of U.S.}\\\text{adult population}}} \sum_{\substack{i \text{ in U.S.}\\\text{adult population}}} \text{Employed}_i \tag{2.1}$$

Stating the estimand explicitly as a mean clarifies the population at stake, in this case including the entire U.S. adult population rather than only those in the labor force. One can also write causal research goals in similar notation. For instance, how would the probability of employment differ if we enrolled a randomly chosen individual in job training, as compared with not enrolling them in job training? We can define this causal goal using potential outcomes notation (Imbens and Rubin, 2015)[2] as the difference in the potential employment each person would realize if enrolled in job training—denoted Y_i(Job training)—versus if they did not—denoted Y_i(No job training).

$$\frac{1}{\substack{\text{Size of U.S.}\\\text{adult population}}} \sum_{\substack{i \text{ in U.S.}\\\text{adult population}}} \left(\underbrace{\text{Employed}_i(\text{Job training})}_{\substack{\text{Employment if received}\\\text{job training}}} - \underbrace{\text{Employed}_i(\text{No job training})}_{\substack{\text{Employment if did not receive}\\\text{job training}}} \right) \tag{2.2}$$

Just like a descriptive estimand (the employment rate) we have expressed the causal estimand (the effect of job training) in notation that invokes a sum over individual-specific quantities. Stating the theoretical estimand in this form clarifies two critical components: (1) the causal contrast at the core of the claim (if any) and (2) the target population over which the unit-specific quantity is aggregated.

[2] Both potential outcomes (Imbens and Rubin, 2015) and structural causal models (Pearl, 2009) are useful for defining the estimands. Each has its own strengths. It is often easier to state the target population in the potential outcomes notation, but the language of structural causal models is sometimes easier for reasoning about identification assumptions. We use both frameworks at points in this paper and we do not take a strong position to prioritize either one over the other.

2.1.1 Specify the Causal Contrast

For causal estimands, a first-order concern is the causal contrast at the core of the claim. In the job training example, a well-designed study would provide specific details about the type of program deployed, such as the training curriculum and length of the program. In observational studies in sociology, it can be challenging to be precise about the specific intervention under consideration. In studies of the effect of a college degree (Brand and Xie, 2010; Hout, 1988), would the ideal experiment be to assign some students randomly to enroll in college? To complete college? Perhaps results would be different if assigned to a selective college (Zhou, 2019) or to completion of a particular graduate degree (Torche, 2011). These difficulties may be even more acute in studies of aggregate scales, such as neighborhood disadvantage (Wodtke et al., 2011). Estimands about different causal contrasts provide us with distinct information about the world.

The causal contrasts in sociological theory may seem insurmountably complex. In these cases, it can be helpful to break off one concrete aspect of a theory to produce a well-defined causal contrast. For example, Pager (2003) sets out to explore a complex general theory: “the ways in which the effects of race and criminal record interact to produce new forms of labor market inequalities,” (Pager 2003:938). Both race and a criminal record are complex, multifaceted constructs that affect numerous domains of life, making them difficult to manipulate (Kohler-Hausmann, 2018). To make progress on this problem, it can help to translate broad theories of racial inequality into specific testable contrasts (Sen and Wasow, 2016). Pager (2003) provides an excellent example of this procedure. The study randomly assigned job postings to receive applications from a white or black pair of applicants. The pair of applicants physically approached one employer and applied for the same job. Within each pair, one applicant was randomly assigned to signal a felony conviction for possession of cocaine.[3] The unit of analysis in this design (indexed by i) is the (job opening × applicant)

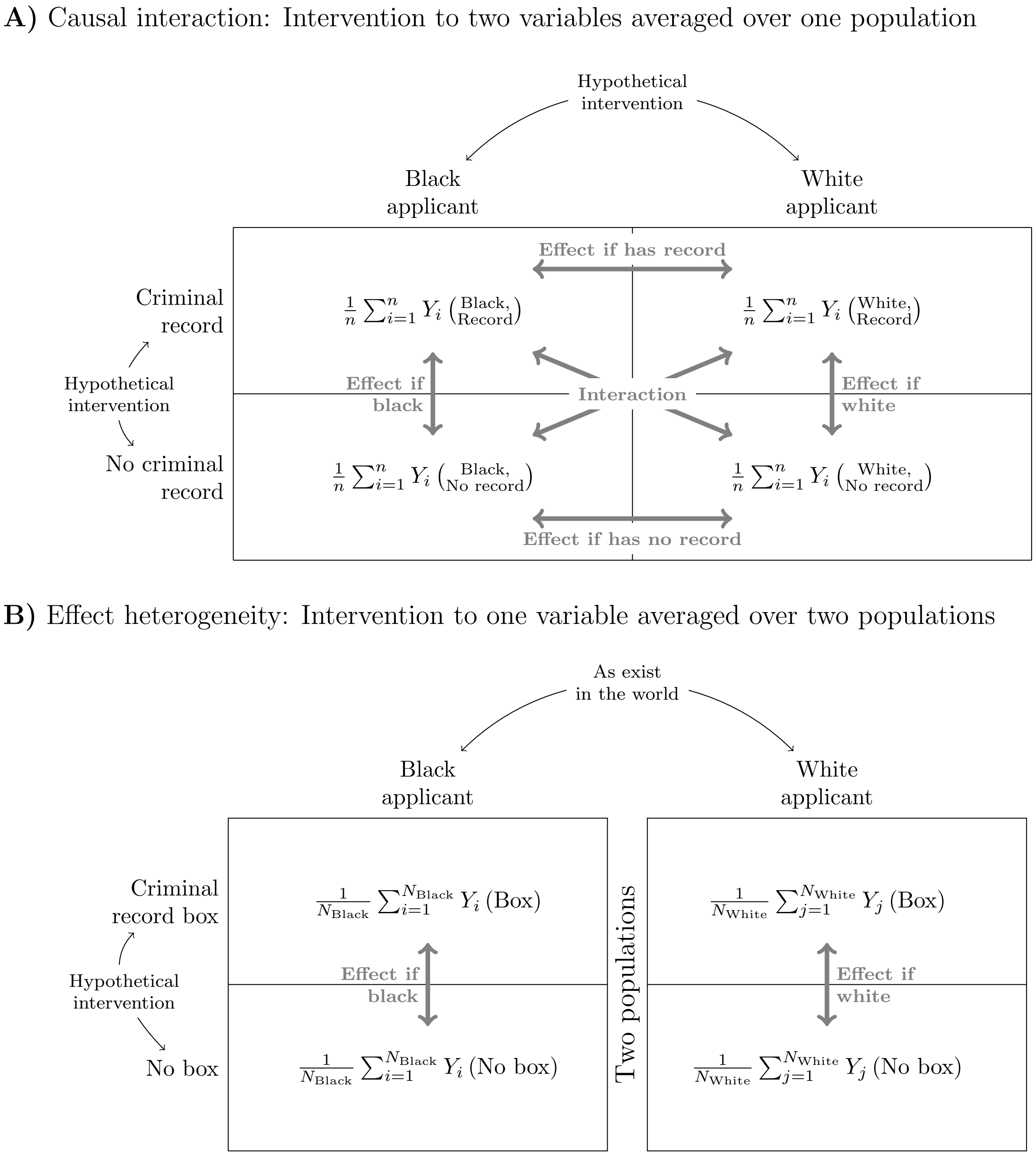
[3] The conviction in the audit study by Pager (2003) was for possession of cocaine with an intent to distribute, with 18 months of prison time, as signaled by prison work experience on the resume, a parole officer as a reference, and often a checkbox on the application in which the employer asked the applicant

pair. Each unit is randomly assigned to one of the four treatment conditions captured by the 2 × 2 table in Figure 2 Panel A, with each potential outcome stated as a function of two interventions (race and criminal record). Although the general theory involves broad constructs such as race, the causal contrast in its most precise form involves a more limited estimand: the difference between the outcome for one specific pair of research assistants (labeled black) as compared with another specific pair of research assistants (labeled white) over random assignment to signal a criminal record in a particular way versus not.

The design of Pager (2003) highlights a second aspect of causal contrasts: they can involve interventions to multiple variables. One of the most striking findings of the study directly involves the joint intervention. The probability of being called back for an interview was lower for a black applicant without a criminal record than for a white applicant with a criminal record. Because both treatments are randomized, they are evaluated over the same distribution of job types with the same set of other applicant characteristics. Thus, the lower callback rates cannot be attributed to any source other than the particular research assistant (white or black) sent to apply for the job and the particular criminal record condition that they signaled. An estimand that involves a joint manipulation of two variables is often termed causal interaction.

One well-stated estimand can lay the groundwork for future work targeting other estimands. Pager's 2003 result may lead us to believe that racial inequality in hiring could be partially (but not completely) closed by interventions to ban the box on job applications asking about criminal records. A researcher embedded within a company could conduct a follow-up experiment to explore this possibility. This experiment may be unethical, but the thought experiment helps to highlight core ideas (or alternatively might motivate an observational study). The researcher might randomly assign actual job applicants to receive an application form with or without a box requesting information about their criminal background. In this design, black and white applicants for the jobs would differ as they do in the

whether they had been convicted of a crime.



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

Hypothetical intervention
Black applicant
White applicant
Effect if has record
Criminal record
$\frac{1}{n}\sum_{i=1}^{n} Y_i\binom{\text{Black,}}{\text{Record}}$
$\frac{1}{n}\sum_{i=1}^{n} Y_i\binom{\text{White,}}{\text{Record}}$
Hypothetical intervention
Effect if black
Interaction
Effect if white
No criminal record
$\frac{1}{n}\sum_{i=1}^{n} Y_i\binom{\text{Black,}}{\text{No record}}$
$\frac{1}{n}\sum_{i=1}^{n} Y_i\binom{\text{White,}}{\text{No record}}$
Effect if has no record

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

As exist in the world
Black applicant
White applicant
Criminal record box
$\frac{1}{N_{\text{Black}}}\sum_{i=1}^{N_{\text{Black}}} Y_i\,(\text{Box})$
$\frac{1}{N_{\text{White}}}\sum_{j=1}^{N_{\text{White}}} Y_j\,(\text{Box})$
Hypothetical intervention
Effect if black
Two populations
Effect if white
No box
$\frac{1}{N_{\text{Black}}}\sum_{i=1}^{N_{\text{Black}}} Y_i\,(\text{No box})$
$\frac{1}{N_{\text{White}}}\sum_{j=1}^{N_{\text{White}}} Y_j\,(\text{No box})$

Fig. 2. Two estimands that involve different interventions and different target populations. Both estimands could be termed the effect of a criminal record an 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 randomly assign only one treatment: whether a box on the application requests information about the applicant's criminal record. Both estimands are of substantive interest.

population of job applicants: race would not be subject to manipulation. In this way, the operationalization of race may more fully capture this complex construct, at the cost that the "effect" of race would not be identified. Yet, the experiment could tell us the differential effect of the criminal record box on the probability of a callback for these two populations. This study would target an estimand, often termed effect heterogeneity, in which the manipulation is applied to one variable (the box on the application) while a moderating variable (applicant race) is not subject to manipulation. This follow-up study would be less able to speak to the effect of some component of race, but it would be more able to speak to the potential policy consequences of a ban-the-box initiative applied in a real-world setting with the actual distribution of application types by race.

To highlight the different information provided by these two estimands, imagine that the Pager (2003) study was repeated today and the outcome was quite similar across all four experimental conditions. This would not suggest that race of individual is unrelated to hiring because many consequences of race are contained in other components of hiring. Similarly, a company should not look at the results of the ban the box survey showing that asking for information about criminal records is worse for black applicants, and conclude that the problem would be addressed by declining to solicit information on applicant race. Policy implications require clarity about which variable is manipulated.

While the distinction between causal interaction and effect heterogeneity is reasonably straightforward in experiments, it can be woefully murky in observational studies. Suppose a researcher analyzes administrative records of job applications with a regression model including control variables, race, criminal record, and the interaction term between race and a criminal record. The researcher might make a hypothesis about this interaction term. But to what causal contrast does this interaction speak? Does the researcher mean to approximate some manipulation to criminal records? To race? To both? Different estimands require different identifying assumptions (see Section 2.2). Stating the estimand in terms of an intervention, even if only hypothetical, would clarify the aims of such a study in a way that is not possible by stating a regression specification alone.

Making the causal contrast explicit is also worthwhile for sociology because it opens the door to numerous new research goals. Causal interaction is only one of many estimands that involve interventions to more than one variable (we provide examples in Table 2), and past scholarship has touched on these estimands only indirectly insofar as they can be related to the parameters of the regression models that dominate sociological research. We suggest that a renewed attention to the estimand may create opportunities for research goals that one may never have considered when thinking in terms of regression coefficients alone. Further, we suggest that arguments that make the hypothetical intervention as precise as possible can lend clarity about the goal that a study aims to accomplish.

2.1.2 Define the Target Population

The importance of reasoning about the target population is also well-known in experiments, since the process of designing the experiment forces the researcher to be clear about who is eligible for treatment. In the example above, (Pager 2003:965) explicitly notes that "one key limitation of the audit study design is its concentration on a single metropolitan area." Yet the target population is often quite limited in observational studies as well, even when the data are drawn from a large and diverse population. A heavy reliance on research goals equated with regression coefficients has, in our view, led to a degree of complacency about the target population in non-experimental studies. Authors write about the "effect of X" as though the effect is a constant value that applies equally to all units in the population. Yet the notion that a coefficient would apply uniformly across the population is implausible, as prior authors have noted. Xie (2013) calls for "recognition of inherent

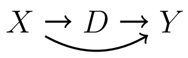
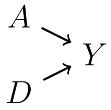
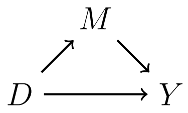
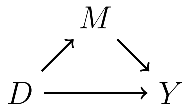
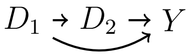
Estimand	Mathematical statement	DAG (Sec. 2.2)	Reference	Colloquial terms
Average treatment effect	$\frac{1}{n}\sum_i Y_i(d') - Y_i(d)$	$D \rightarrow Y$	Morgan and Winship (2015)	Effect
Conditional average treatment effect	$\frac{1}{n_x}\sum_{i:X_i=x}(Y_i(d') - Y_i(d))$		Athey and Imbens (2016)	Effect heterogeneity or moderation
Causal interaction	$\frac{1}{n}\sum_i \left(\left(Y_i(a',d') - Y_i(a',d)\right) - \left(Y_i(a,d') - Y_i(a,d)\right)\right)$		VanderWeele (2015)	Joint treatment effect
Controlled direct effect	$\frac{1}{n}\sum_i \left(Y_i(d',m) - Y_i(d,m)\right)$		Acharya et al. (2016)	Mediation (Appendix B)
Natural direct effect	$\frac{1}{n}\sum_i \left(Y_i(d',M_i(d)) - Y_i(d,M_i(d))\right)$		Imai et al. (2011)	Mediation (Appendix B)
Effect of dynamic treatment regime	$\frac{1}{n}\sum_i Y_i(\vec{d'}) - Y_i(\vec{d})$		Wodtke et al. (2011)	Cumulative effect

Table 2. Many causal estimands are potential targets of social science 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 estimands. Y indicates the outcome, D indicates the treatment, M indicates a mediator, $\vec{X}$ indicates pre-treatment covariates, capital letters indicate random variables, and lower case letters indicate fixed values. Controlled direct effects and other mediation-based estimands appear in sociology, though not always labeled as such (Appendix B).

individual-level heterogeneity" in causal inference, and Morgan and Winship (2015:47) posit that statement of the characteristics of 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).

The target population can be selected based in part on the population about which one can credibly draw conclusions. For example, experimentalists rarely claim inference over an entire population, but instead over the units in the study or the sampling frame from which they are drawn. In observational settings, the population to which results can credibly generalize is more elusive. One estimation approach for which the target population has been the subject of substantial discussion is instrumental variables (IV). Under IV, authors identify causal effects by relying on an instrument: a variable assumed to affect the outcome only through the treatment. The population to which IV estimates generalize is limited to a subpopulation, often termed compliers, whose treatment status is causally affected by the instrument (Angrist et al., 1996; Imbens and Angrist, 1994). IV analyses are uninformative about the effect among non-compliers, for whom the instrument has no effect on the treatment. The complier perspective affords researchers an opportunity to clarify the subpopulation to whom estimates apply.

Angrist and Evans (1998) study the effect of a third birth on women's employment by exploiting the fact that 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 resulting estimates capture the effect of a third birth among those who would have a third birth if and only if their first two children are of the same sex. The authors provide enough information for the reader to conclude that the target population is extremely small: 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. If another study used a different empirical strategy and came to a different estimate, both could be correct; the difference might be attributable to a difference in the estimand rather than the validity of the empirical approach.

In some examples, the LATE may be the estimand of greatest interest. 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. Because careful statement of the estimand has become standard in instrumental variables analysis, the authors are very clear about the target of inference: the effect among offenders who, had they faced a more lenient judge, would have been sentenced to probation rather than prison (Harding et al. 2018:67). This estimand may be especially useful for substantive theory: these are individuals whose sentences might plausibly change if judges were encouraged to be more lenient in sentencing. The relevant population for policy may be similar to the population about which the instrumental variables strategy is informative. Stating the target population should not feel to researchers like an extra limitation they must acknowledge; it is instead an opportunity to argue transparently for why the target population of the study is the one about which we would most like to know for theory or policy.

Clarity about estimands entails a tradeoff between the desired effect of interest and the tractable identification problem (more in the next section). Instrumental variables are an example of a broad class of approaches, including regression discontinuity designs and experiments, for which the causal effect averages over a very specific subpopulation. There is a debate within the econometrics literature between those who feel that these specific estimands are too limiting (Deaton, 2010; Deaton and Cartwright, 2018; Heckman and Urzua, 2010) and others who argue that they are at least as informative as less credible alternatives (Aronow and Samii, 2016; Imbens, 2018, 2010). We simply note that while observational data can appear to offer generalizability to the full population, we can only draw credible inferences where we have common support (comparable treatment and control units). 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. Many settings with observational data have common support on only a small subset of the population.

To summarize, the theoretical estimand is a chance for researchers to define the aim of the study, and potentially to argue that a quantity which is empirically tractable may also be a quantity of substantive interest. A clear statement of the goal—in terms of an average over some well-defined population—may provide greater clarity in this regard than approaches that simply equate a theory with a particular regression coefficient.

2.2 Identification: Link to an Empirical Estimand

When the theoretical estimand is causal, we never get to observe the potential outcomes of a unit under at least one level of the treatment. No amount of big data can reveal these potential outcomes without assumptions. Only by assumption can we convert a causal research goal to an empirical estimand: a statistical quantity that can be learned from data. In concrete terms, even millions of administrative records on job applicants and their felony conviction status would not reveal the causal effect of a felony conviction on callbacks. What makes the Pager (2003) study compelling is that 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—is highly plausible because of the design decision to randomize both signals. The mean difference in callback rates among those with and without this signal is a statistical quantity involving entirely observable data, yet it gains causal meaning because of the identification assumption. Similar assumptions are required for causal inference in observational settings, although in those settings their plausibility may be in greater doubt.

Although the term "identification" is most often used for causal estimands, non-causal estimands may also involve data we do not observe. For example, researchers often rely on probability samples to draw inferences about population means, such as the employment rate, yet every survey suffers from non-response. Some people refuse to answer the survey entirely, and others refuse to answer a key question of interest. Making inference to the original target population from a survey with non-response requires an identification assumption that the probability of employment is equal among responders (who we see) and non-responders (who we do not see). Even if we use administrative records like tax returns or unemployment filings, there are individuals missing from the records for reasons related to their employment status. Because they involve data we do not have, identification assumptions must be defended on conceptual rather than empirical grounds. In the survey setting, we would have to reason about the risk that unemployed individuals may be too busy searching for work to take the survey or might refuse to answer the employment question out of embarrasment about the answer. Identification arguments about missing data are essential to credible and transparent research.

Despite the emphasis on identification in standard textbooks (Morgan and Winship, 2015), few sociology papers reason explicitly about the assumptions required to deal with the missing data problems that arise in causal inference or in descriptive population inference. One reason for this discrepancy may be that the objects of sociological inquiry appear on the surface to be descriptive sample quantities, and thus valid without 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 3) 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 third variable is sometimes termed a collider variable because it is the consequence of the variable of interest and also the consequence of other unobserved variables; in a Directed Acyclic Graphs, arrows "collide" at this variable (Elwert and Winship 2014 provide an accessible introduction).[4] These examples highlight the need to state the theoretical estimand, the empirical estimand, and the identification assumptions under which the two are equal. The need for clarity about these steps holds even when the target quantity may not appear to be causal at first glance. Our general point is that a precise statement of the theoretical estimand can guide one toward the need for causal 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. It is clear in both of these settings that 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 stopped. In a well-argued critique, Knox et al. (2019) highlight a number of issues that hinder inference about discrimination from these data, among which one is especially relevant to the present discussion: the sample only includes those who are stopped by police, and being stopped is a consequence of race (Table 3). 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. Once a stop occurs, an unbiased officer might actually use lethal force against whites at a *higher* rate than against blacks, since the whites who are stopped are more dangerous than the blacks who are stopped. Equivalent rates are actually consistent with racial discrimination.[5]

The core empirical fact has not changed; it is the same probability of a police-involved shooting given race of the suspect in the sample. The meaning of that empirical fact,

[4] In our examples about demographic disparities, the problem is especially bad in two of the three cases because no data are available for one value of the collider: we only see application decisions among those who apply, and we only see whether someone is shot among those who are stopped by police. Knox et al. (2019) discuss this problem in greater depth.

[5] To be clear, Fryer (2019) discusses issues of sample selection. In fact he titles an entire top-level section "A note on potential selection into police data sets." His approach is to control for the measures he has available including precinct and officer characteristics, but this assessment cannot adjudicate selection on unmeasured police bias in stops.

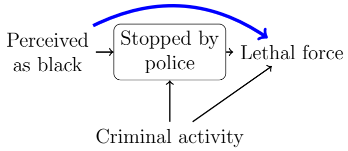
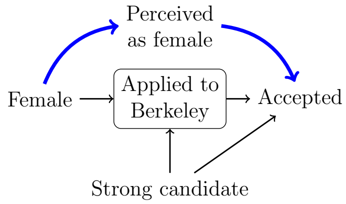
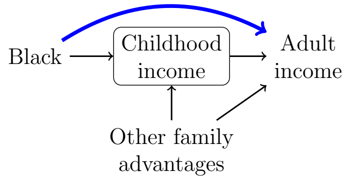
Study	**Empirical regularity**	**Misleading conclusion**	**Directed Acyclic Graph**
Fryer (2019)	Among those they stop, police shoot the same proportion of black individuals as white individuals.	Police do not discriminate against black individuals when using lethal force.	 Perceived as black Stopped by police Lethal force Criminal activity
Bickel et al. (1975)	Among those who apply, Berkeley departments admit a higher proportion of women than of men.	Committees do not discriminate against women.	 Perceived as female Female Applied to Berkeley Accepted Strong candidate
Chetty et al. (2018)	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.	 Black Childhood income Adult income Other family advantages

Table 3. 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. (2019) 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.

however, has changed quite dramatically if we accept the assumption that being stopped by police 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 that clarify the jump from the observed association to a theoretical conclusion about racial bias.

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, within departments, women are admitted at higher rates than men.[6] The theoretical estimand may be the difference in admission if we intervene to change a committee's perception of an applicant's sex. In this estimand, perceived sex would be independent of qualifications. The empirical data, however, are limited to the men and women who actually apply to Berkeley. Bickel et al. (1975) assume away this problem by stating that results require that men and women are equally qualified.[7] Yet this assumption seems quite doubtful. Due to discrimination at the undergraduate level, it is likely that many men apply to Berkeley but only the most qualified women apply to Berkeley. Thus, equal rates of admission among the men and women we observe could actually be consistent with sex-based discrimination against women. Because sex and qualifications both cause application to Berkeley, a simple comparison among those who apply to Berkeley can be misleading.

As a third example, Chetty et al. (2018) show that black and white women who are raised in families with similar incomes have similar earnings as adults. At face value, one might

[6] Bickel et al. (1975) is often cited as an example of either Simpson's Paradox or the ecological fallacy: in 1973, 44% of the 8,442 men who applied to graduate school at Berkeley were admitted compared to only 35% of 4,321 women (school-wide difference) but, if one looked at rates *within* departments, women were admitted at higher rates in most departments.

[7] Bickel et al. (1975:398) acknowledge their assumption: "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 hold it any differences in acceptance of applicants by sex could be attributed to differences in their qualifications, promise as scholars, and so on." This assumption is questionable in an environment where discrimination occurs prior to selection into applying for graduate school.

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 3). The racial income gap in adulthood that would persist if we equalized childhood family incomes (a theoretical estimand) 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 each of these cases, what appears to be a descriptive empirical regularity is commonly interpreted in light of a particular causal effect: the effect of being stopped by police on being shot, the effect of applying to Berkeley on being accepted, and the effect of childhood income on adult income. Justification of this interpretation requires an argument that reaches beyond the data to theoretical assumptions about how the data come to be and about causal relationships among variables in the data. No set of assumptions is ever perfect, yet formalizing the assumptions allows us to pinpoint the weaknesses of any particular link between a theoretical and empirical estimand. Making causal assumptions explicit is therefore essential to productive communication about why a particular empirical result is suggestive of a substantive claim. Yet doing so is only possible once the theoretical claim (the estimand) has been stated.

Table 3 reasons about these issues using Directed Acyclic Graphs (DAGs), a visual tool to depict an assumed set of causal relationships among variables (Pearl, 2009).[8] Just as our first step involves argument about the research goal separately from any regression coefficients, DAGs allow us to reason about what assumptions we need to link data to that goal rather than reasoning about a regression coefficient. In particular, nonparametric identification

[8] Directed Acyclic Graphs (DAGs) may seem similar to linear structural equation models (LSEMs), which have a long history in sociology (e.g. Blau and Duncan 1967). The two share common roots but are not the same. In an LSEM, by virtue of linearity and additivity assumptions, a number on each edge can summarize the magnitude and direction of the effect.

allows us to focus on one set of considerations (causal relationships) while delaying questions about the shape of statistical associations for the subsequent choice of an estimation strategy. The aim of this paper is to highlight that the identification assumptions that can be stated in DAGs may apply to a wider set of sociological problems than applied researchers may think; we refer readers to other pedagogical sources for an introduction to identification using DAGs (Morgan and Winship, 2015; Pearl and Mackenzie, 2018).

2.3 Estimation: Learn the Empirical Estimand from Data

Identification leads us to an empirical estimand θ, such as the proportion employed among those willing to answer a survey. This goalpost becomes the target for the next step in which we turn data into an estimate through the estimation strategy. Sociologists overwhelmingly select estimation strategies that equate the target parameter θ with a coefficient β from a regression model, but this is unnecessarily restrictive. With the estimand defined separately from any regression, it becomes possible to select an estimation strategy that proceeds under weaker (more credible) assumptions about the form of the relationship between predictors and the outcome. In particular, an empirical estimand opens the door to estimation strategies that involve two steps: a prediction algorithm to predict the outcome as a function of the covariates and an aggregation strategy that converts predictions to an estimate of the estimand.

To illustrate this possibility, we conduct an empirical exercise inspired by Pal and Waldfogel's (2016) examination of the family gap in pay: the log hourly wage gap between mothers and non-mothers conditional on age, education, race, and marital status. Following those authors, we analyze data from the Annual Social and Economic Supplement of the March Current Population Survey. To focus on the main issues, we defer details about the sample and estimation procedures to Appendix C. We focus on the most recent data collected in 2019, thereby updating the original results with the most current evidence. We select this example because of the large sample and the small number of covariates involved, which allow us to illustrate key points. Our conclusion bolsters the claims of the original authors, showing that their conclusions hold under milder assumptions than those maintained in the original paper. Defining the estimand as the coefficient on motherhood in a particular regression commits one to estimation by that regression; defining the estimand separately from the regression allows estimation under assumptions that are more credible.

The empirical estimand (Eq. 2.3) involves two components. The difference term within the parentheses compares the mean log hourly wages $\bar{Y}$ of mothers and non-mothers, within subgroups of the covariates $\vec{X}$. To aggregate these subgroup-specific estimands to a population-average estimand, the equation sums over all subgroups $\vec{X} = \vec{x}$ weighted by the prevalence of that covariate subgroup among mothers.

$$\theta = \underbrace{\sum_{\vec{x}} \text{P}(\vec{X} = \vec{x} \mid \text{Mother})}_{\text{Aggregated across subgroups}} \underbrace{\left(\bar{Y}_{\text{Mothers with } \vec{X}=\vec{x}} - \bar{Y}_{\text{Nonmothers with } \vec{X}=\vec{x}}\right)}_{\text{Mean difference within subgroups}} \tag{2.3}$$

One straightforward estimator for θ is a nonparametric stratification estimator, which we argue is an important conceptual starting place for estimation. Nonparametric stratification proceeds as follows. First, split the data into subgroups that have the same values for all covariates $\vec{X}$. One subgroup, for instance, would be white women at age 35 with a college degree. Second, calculate the subgroup-specific mean difference in the outcome $\bar{Y}$, such as the difference between the pay of mothers and non-mothers in this group. Third, aggregate across subgroups according to the observed distribution of the target population chosen in step two. In short, nonparametric stratification replaces all unknown quantities with model-free estimates of those quantities by weighted means, denoted by hats as in $\hat{\bar{Y}}_{\text{Mothers with } \vec{X}=\vec{x}}$.

$$\hat{\theta} = \sum_{\vec{x}} \hat{\text{P}}(\vec{X} = \vec{x} \mid \text{Mother}) \left(\hat{\bar{Y}}_{\text{Mothers with } \vec{X}=\vec{x}} - \hat{\bar{Y}}_{\text{Nonmothers with } \vec{X}=\vec{x}}\right) \tag{2.4}$$

The nonparametric stratification estimator is appealing because it proceeds entirely through weighted means; it does not assume any functional form. It corresponds to a simple prediction function: to predict for a new observation with $\vec{X} = \vec{x}$, take the mean of all previous cases observed with $\vec{X} = \vec{x}$. This estimation strategy imposes no assumptions on the shape of the relationship between $\vec{X}$ and Y beyond the discretization of age into years. For instance, because we calculate one mean for 25-year-olds and a different mean for 30-year-olds, this estimator does not get bogged down in debates about the proper shape of the relationship between age and earnings.

The nonparametric stratification estimator highlights a concern with estimation of all kinds—common support. The estimator breaks down if there exists a subgroup of $\vec{X}$ in which mothers are observed but non-mothers are not. This common support requirement can be viewed as a feature or a bug; it rules out the risk of extrapolation to covariate values not observed in the data at the cost of limiting the target population. In the family gap in pay example, the sample size is sufficiently large that 99 % of mothers fall in a covariate subgroup where non-mothers are also observed. In all analyses, we focus our attention to this region of common support as the target population. By beginning any estimation exercise from a default starting place that considers nonparametric stratification, it becomes easier to recognize the common support problems that plague research in settings when the covariates have a very large number of values (D'Amour et al., 2017). 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. For instance, we assume in this example that even though age is continuous, it can be adequately represented by the whole-number years in which the data are reported.

In current practice, authors more often assume a parametric model for the outcome Y as a function of all the predictors. For instance, we might assume a linear relationship between mean log hourly wages Y and the predictor variables, with no interactions.

$$\mathbf{E}(Y \mid \text{Motherhood}, \vec{X}) \underbrace{=}_{\substack{\text{By linear} \\ \text{approximation}}} \alpha + \beta(\text{Mother}) + \vec{X}'\vec{\gamma} \tag{2.5}$$

This model relies on an approximation which is unable to capture interactions or nonlinearities, so it will always estimate a log wage gap between mothers and non-mothers that is a constant value β regardless of the values of the other variables $\vec{X}$. As a result, the estimand θ is (by the model assumptions) equal to the coefficient β. To set up a more general procedure, we can arrive at this result by plugging in the predicted values from the regression model to produce an estimate of θ.

$$\hat{\theta}_{\text{OLS}} = \sum_{\vec{x}} \hat{\text{P}}(\vec{X} = \vec{x} \mid \text{Mother})\Bigg(\underbrace{\Big(\hat{\alpha} + \hat{\beta} + \vec{x}'\hat{\vec{\gamma}}\Big)}_{\substack{\text{Predicted } Y \text{ at } \vec{X} = \vec{x} \\ \text{if one is a mother}}} - \underbrace{\Big(\hat{\alpha} + \vec{x}'\hat{\vec{\gamma}}\Big)}_{\substack{\text{Predicted } Y \text{ at } \vec{X} = \vec{x} \\ \text{if one is not a mother}}} \Bigg) \tag{2.6}$$

$$= \sum_{\vec{x}} \hat{\text{P}}(\vec{X} = \vec{x} \mid \text{Mother})\hat{\beta} \tag{2.7}$$

$$= \hat{\beta} \tag{2.8}$$

This general strategy—assume a parametric model for the outcome and impute the outcome under different treatment conditions—is well-known in biostatistics as the parametric g-formula (Hernán and Robins, 2020, Ch. 13) and has been called an imputation estimator in econometrics (Hahn 1998:321, Abadie and Imbens 2006:241, Abadie and Imbens 2011:3).

In the more general case, one could apply any prediction function $f(\text{Motherhood}, \vec{X})$, such as a black-box machine learning algorithm, and use this function to predict the expected outcome terms in the empirical estimand. For instance, $f()$ could be an OLS model with additional interaction terms, a Generalized Additive Model (Wood, 2017) with a nonlinear smooth term for age, or a random forest that can capture arbitrary interactions.

$$\hat{\theta}_{\text{ML}} = \sum_{\vec{x}} \hat{\text{P}}(\vec{X} = \vec{x} \mid \text{Mother})\Bigg(\underbrace{\hat{f}\Big(\text{Mother}, \vec{x}\Big)}_{\substack{\text{Predicted } Y \text{ at } \vec{X} = \vec{x} \\ \text{if one is a mother}}} - \underbrace{\hat{f}\Big(\text{Nonmother}, \vec{x}\Big)}_{\substack{\text{Predicted } Y \text{ at } \vec{X} = \vec{x} \\ \text{if one is not a mother}}} \Bigg) \tag{2.9}$$

Even if it is impossible to summarize the prediction function $f()$ by a small number of parameters, the aggregate quantity $\hat{\theta}_{\text{ML}}$ can be derived from those predictions. Thus, a statement of the estimand is an essential first step to the application of black-box algorithms to social science questions by clarifying that the regression coefficient itself is not of inherent interest.

More flexible modeling approaches, such as nonparametric stratification or machine learning, are appealing because they can yield substantive conclusions under weaker (more credible) functional form assumptions than parametric models. Nonparametric stratification captures all nonlinearities and interactions in the data without requiring the researcher to specify them. Many machine learning algorithms seek to approximate this flexibility automatically while sharing some information to improve the precision of estimates, freeing the researcher from choices such as whether to include interactions or squared terms. There are also reasons to be cautious about machine learning approaches. For example, the statistical theory needed to place standard errors around estimates does not exist for all machine learning estimators. This is an active area of research (e.g. Wager and Athey 2018 develop inference for random forests) and may be overcome in part through computational techniques like bootstrapping, though this can be computationally demanding with a complex estimator. For the sake of argument, we will temporarily suspend these concerns and highlight a fact that can be demonstrated primarily with point estimates: whether the substantive conclusions change is an empirical question that will differ across specific cases.

Fig 3 shows empirical results for the family gap in pay. When the estimand is the gap aggregated over the covariate distribution of all mothers, the substantive size and precision of the estimated gap are similar across estimation strategies. Mothers have lower log wages than non-mothers, regardless of the strategy by which we adjust for covariates.[9] Claims about the aggregate gap need not rely on the assumptions of an additive model with a quadratic term for age (the specification of Pal and Waldfogel 2016), because the gap is evident even with estimation by nonparametric stratification. Allowing all nonlinearities

[9] Statistical significance of the aggregate family gap in pay differs across estimation approaches; the gap is significant at the 0.05 level under the stratification and no-interactions approaches only. We do not emphasize this difference because the substantive distinction between p-values just above and just below 0.05 is not meaningful (Gelman and Stern, 2006)

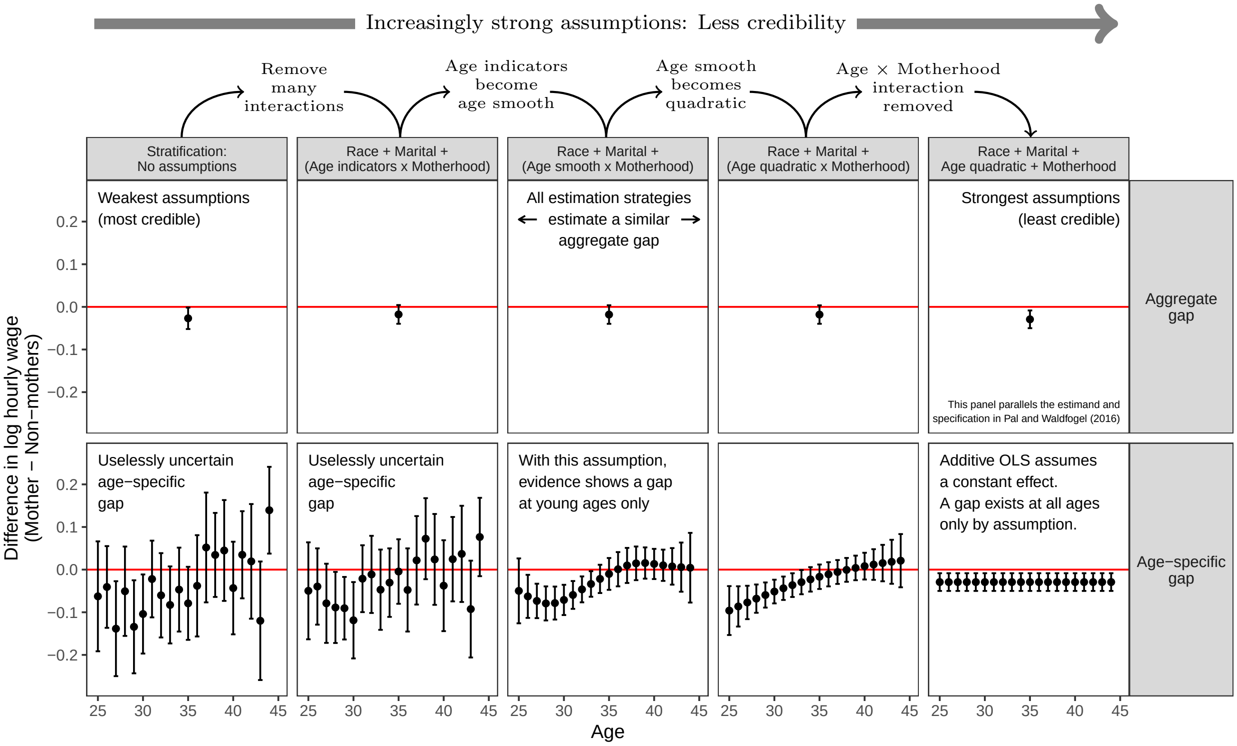


Fig. 3. A series of estimation strategies (columns) for two estimands (rows). Each estimand is the gap in log hourly wages between mothers and childless women, conditional on age, education, race, and marital status and aggregated over the covariate distribution of mothers. Estimands differ by aggregating 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 × Motherhood) represent an interaction and its lower-order terms. Provided that estimates 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 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 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. For instance, instead of defaulting to an additive model and motivating any included interactions, one could default to a fully interactive model and motivate why some interactions are omitted. Data come from the 2019 Annual Social and Economic Supplement of the March Current Population Survey. Error bars are 95% confidence intervals calculated using replicate weights.

and interactions does not change the substantive result, removing the need for any checks for robustness to alternative specifications. On the other hand, suppose we were interested in the family gap in pay within subgroups of age (bottom row). The result differs when we apply the same estimation strategies to this estimand. Nonparametric stratification yields estimates that are uselessly uncertain because the number of observations is small at each age. Estimates only become sufficiently precise once we assume that the only interaction is between motherhood and age and that the shape of the association with age is smooth (middle column). Under these assumptions, the data provide evidence of a family gap in pay that differs by age: young mothers earn less than young non-mothers, but older mothers' wages are similar to those of older non-mothers. The extremely strong assumption of an OLS model with no interactions (right column) obscures this fact because the model specification invokes an additive approximation that is guaranteed to miss any variation in the gap by covariates.

Examining both rows of Fig 3 together highlights what we mean when we say that parametric models are best viewed as parsimonious approximations. The OLS model in the lower right is clearly misspecified: it estimates a family gap in pay that is too close to zero younger ages and too far from zero at older ages. Yet these errors cancel out so that the aggregate gap (top right, equivalent to the coefficient on motherhood) is approximately equal to the nonparametric stratification result (top left), which can be viewed as a reasonable benchmark because the sample size is large enough to arrive at a precise estimate with no modeling assumptions. It is therefore possible that the regression coefficients commonly applied in studies are reasonable approximations to the nonparametric answer that one would get by applying nonparametric stratification to a very large sample. Yet, we see no reason to expect this approximation to perform well in general. We therefore suggest that a statement of the estimand separately from the estimation strategy is important because it lays the groundwork for researchers to begin with the most flexible model specification possible before moving to more restrictive assumptions as needed.

If one enumerates a series of candidate estimation strategies, empirical evidence can help to select among these estimators. In our example, the empirical estimand involves a conditional expectation function for an outcome (log wage) given a specified set of predictors. We could estimate this conditional expectation function by some estimated prediction function $\hat{f}()$, choosing the functional form among a sequence of candidates as the one that achieves the best predictive performance (lowest mean squared error) when predicting new observations. This decision rule is principled because the true conditional expectation function (if it were known) would achieve the best mean squared error. To illustrate, we conduct cross validation in the ASEC data. We split the data into five equally-sized folds, repeatedly fit the model on four folds and predict for the remaining fold, and then aggregate all squared prediction errors by the weighted mean using survey weights. The resulting mean squared error is highest (worst performance) for nonparametric stratification (0.334) and comparable for all other approaches (roughly 0.313). If we followed a rule of selecting the estimation strategy with the best mean squared prediction error, we would select the OLS model assuming a quadratic age form interacted with motherhood.

Caution is warranted, however, when selecting a prediction rule based on mean squared error because our aim is not prediction. Mean squared error ($\frac{1}{n}\sum_i(Y_i - \hat{Y}_i)^2$) is the optimal metric if we need to learn the entire conditional expectation function accurately. For an aggregate estimand θ such as the family gap in pay, we would rather know our error for that estimand, $(\hat{\theta} - \theta)^2$. If we were able to learn the conditional expectation function perfectly we would know θ perfectly, but the best estimator of the conditional expectation function is not necessarily the best estimator of the aggregate quantity θ. The optimal estimator depends on the estimand, as is visually apparent in this setting: nonparametric stratification is potentially optimal for the aggregate gap but is clearly suboptimal for the age-specific gap and is known by the cross-validation exercise above to be suboptimal for the full conditional expectation function. By stating the estimand, researchers take the first step necessary to derive a decision rule that converts out-of-sample predictive performance into a metric of the performance of any given estimator for a particular estimand (e.g. Athey and Imbens 2016). An entire literature exists in biostatistics to provide automated two-step procedures to modify prediction functions to target particular aggregate estimands (Van der Laan and Rose, 2011). Because this literature is still developing, we hesitate to provide direct guidance on the optimal rule to select among a set of estimators. Future research into the optimal decision rule for a variety of estimands is needed; the first step to this research is for social scientists to define the estimands for which statisticians can develop improved estimators.

Choosing an estimation strategy is not an easy task. Empirical evidence such as out-of-sample performance can help, but a heavy reliance on out-of-sample performance must acknowledge that mean squared prediction error and mean squared error for the estimand may not be equivalent. Regardless of the strategy one uses to select an estimator, the entire process of evaluating estimators can only begin once one states the estimand separately from the estimation strategy. Stating the research goal as a regression coefficient locks one into this particular estimation strategy and forces one to fight a losing battle to defend a regression specification that everyone knows to be imperfect. Given that techniques to approximate statistical parameters are likely to continue improving in the future, social scientists have much to gain from defining their estimand separately from regression models in order to tap into these growing opportunities.

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. This strategy may involve a machine learning algorithm. Together, these three steps make a clear linkage between theory and empirical evidence in which each step can involve a principled choice.

3 Relationship to Prior Work

Our call for explicit statement of an estimand and the assumptions by which we learn about it from data is closely related to the shift toward "causal empiricism" in economics and political science (Samii, 2016) and statistical perspectives that begin from an assumption the model is misspecified (Aronow and Miller, 2019; Berk et al., 2019; Buja et al., 2019) or that back out an effective estimand from an estimation strategy (D'Amour and Airoldi, 2017). Each step of our proposal bears some relationship to prior methodological contributions.

Our first step (choosing the theoretical estimand) relates to arguments about how empirical evidence relates to broad sociological theories. Lieberson and Horwich (2008) argue against a widely-held expectation that a single study can prove or undermine an entire sociological theory. Instead, they advocate that each theory be made elaborate with many predictions evaluated by separate teams in a variety of research designs, similar to the way advances are made in the natural sciences. Our paper provides a language to promote this communal view of science: we can collectively develop general theory most effectively when each study is precise about the theoretical estimand at the core of its empirical contribution, thereby facilitating constructive dialogue across studies. This call relates to Hernàn's (2018) argument that researchers in public health should define their estimand in explicitly causal terms, thereby highlighting what the study aspires to learn for theory and policy. Many methodological papers encourage researchers to be more explicit about the hypothetical intervention their study aims to approximate (Greiner and Rubin, 2011; Hernán et al., 2016; Sen and Wasow, 2016). Those who do not state the goal risk the "table 2 fallacy": causal interpretation of more than one coefficient in a way that is unwarranted (Westreich and Greenland, 2013) or requires implausibly strong assumptions (Keele et al., 2019). Finally, theoretical estimands relate to discussions of the population over which heterogeneous treatment effects are aggregated (Brand and Xie, 2010; Xie, 2013). Across these arguments, the common thread is a call for researchers to be precise about how their research goal relates to general theory, a task that enters our framework in the step of defining the theoretical estimand.

Our second step (choosing an empirical estimand and linking to the theoretical estimand by identification assumptions) has also been the subject of substantial scholarship. Freedman (1991) argues that social scientists are overly dependent on regression modeling and technical fixes, when in fact they ought to pay more attention to the scientific issues at play in a given study (e.g. selection into treatment). The identification link between the theoretical estimand and empirical estimand provides an opportunity for researchers to focus directly on these scientific issues, building an argument for the importance of an empirical quantity separately from technical considerations about how to learn that quantity. Identification has entered the sociological literature in the form of DAGs and potential outcomes used to specify the conditions under which causal inferences can be drawn from observational data (Morgan and Winship, 2015; Pearl, 2009; Pearl and Mackenzie, 2018). This technique has been increasingly adopted in sociology journals (Elwert and Winship, 2014; Sharkey and Elwert, 2011; Wodtke et al., 2016, 2011). *Sociological Science* has gone so far as to require that "manuscripts that offer causal claims based on empirical analyses should clearly state the assumptions required to warrant causal interpretations."[10] While we join those arguing for the more widespread use of identification, our framework also highlights that identification requires a well-specified goal in the form of a theoretical estimand.

Our third step (choosing an estimation strategy) relates to debates about the robustness of sociological results to modeling errors (Young, 2009) and to new calls for the incorporation of machine learning into sociological research (Molina and Garip, 2019). The estimation step epitomizes one specific setting in which sociologists may be particularly well-advised to incorporate predictive checks into their methods (Watts, 2014). It is not clear that prediction

[10] The *Sociological Science* submission guidelines that specify the need to state assumptions are online: https://www.sociologicalscience.com/for-authors/submission-guidelines/

can aid in definition of the theoretical estimand or defense of identification assumptions, but it certainly can play a role in selecting an estimator with good properties (our third step).

Although each of our three choices—the theoretical estimand, the empirical estimand, and the estimation strategy—has been the subject of substantial research, we found very few examples in which sociologists (or social scientists in general) are explicit about all three components. Authors who use matching or inverse probability of treatment weighting tend to be very clear about the target population and required identification assumptions but use logistic regression to estimate the probability of treatment with almost no discussion of the estimation assumptions involved in specifying the treatment as the inverse logit of an additive, linear function of predictors (e.g. Brand and Xie 2010). Young (2018) rightly points toward a crisis in science about model uncertainty arising from questionable models used for estimation, but the estimand in question is still stated as a regression coefficient. Molina and Garip (2019) highlight the promise of machine learning approaches for estimation but without a clear link to the target of estimation. In fact, the word estimand is not in common use in sociology.[11] This is not to say that the target of inference is never discussed, but we believe that using the term itself might increase the focus on the target of inference as an object of interest.

In short, prior literature points toward various pitfalls in the link between theory and empirics, and this paper unifies these concerns under a common framework involving three distinct steps.

4 Estimands are Essential in All Quantitative Work

Bringing methodological choices under the umbrella of estimands yields three key benefits: estimands ground methodological choices, add clarity to the search for model robustness,

[11] The word "estimand" has never appeared in the main text of an article in the *American Sociological Review*. According to a Google Scholar search it has, as of August 2019, appeared twice in footnotes in *American Sociological Review* (Wodtke et al., 2011; Zhou, 2019). It has appeared once (in a comment) in the *American Journal of Sociology* (Luo et al., 2016). The excellent textbook by Morgan and Winship (2015) discusses treatment effects and target populations but never uses the word "estimand."

and partition the role of evidence in social science.

4.1 Estimands Ground Methodological Choices

A framework involving all three steps is needed for a simple reason: the estimand (step 1) guides all subsequent methodological choices about identification (step 2) and estimation (step 3). It is difficult (or impossible) to reason about these methodological choices when the goal is not clear. A lack of clarity about estimands is therefore a core problem in sociological methodology. Specific debates at present address only individual symptoms. For binary outcomes, methodologists encourage reporting predicted probabilities instead of logistic regression coefficients (Breen et al., 2018; Mood, 2010). Yet this is fundamentally an issue of clarity about the estimand: the reason researchers report coefficients that are difficult to interpret is because they have not been clear from the start that the estimand of greatest interest is a probability. Methodologists have clarified that those who study social mobility while conditioning on consequences of social origins can produce misleading results (Zhou, 2019). Yet clarity about the estimand (the causal effect of social origins) would have highlighted the severe challenges of identifying this estimand and the dangers of post-treatment conditioning, possibly moving the discipline toward estimation approaches to resolve some of these problems (Zhou and Wodtke, 2019) decades earlier. Debates could be clearer and advances more rapid if we recognized the underlying disease: a lack of clarity about the estimand.

Separating the steps can pinpoint methodological criticism to the precise place at which it applies. Critiques in the first two steps require conceptual arguments. One might criticize a theoretical estimand by arguing that the quantity studied is only marginally informative for theory or policy (Deaton, 2010). One might criticize an identification strategy by arguing that the decision to condition on certain variables has undermined the link between empirical evidence and the theoretical goal (Elwert and Winship, 2014). Critiques of the third step (estimation) may instead hinge on empirical evidence. If a machine learning algorithm achieves lower mean squared prediction error in holdout data than an OLS regression, this provides evidence that the algorithm is a better estimation strategy for the conditional mean than the OLS regression.

Separating these critiques is important because the best evidence in each type of critique differs.

4.2 Estimands Add Clarity to the Search for Model Robustness

In the 2018 volume of the *American Sociological Review*, at least 20 papers reported a robustness check (Appendix A). Our framework asks: to what do we want our models to be robust? Some forms of robustness focus on the theoretical estimand (e.g. a different outcome), others focus on a different identification strategy (e.g. a different conditioning set), and still others focus on the estimation strategy (e.g. a different functional form). These forms of robustness are very different. When applying a robustness check within our framework, one would defend the view of the world under which each particular specification would be meaningful. In contrast, robustness checks as currently applied treat all specifications as equally valid. Young (2018) 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 would devolve 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.

Statement of the estimand can help clarify when robustness checks provide the most useful information. Robustness across causal contrasts and target populations may provide useful context for our theoretical understanding. Robustness across conditioning sets of variables 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 can provide useful information about our evidence, but only in the context of a well-defined target and clarity about the alternatives to which we are evaluating robustness.

4.3 Estimands Partition the Role of Evidence in Social Science

Although there have long been concerns about the accuracy of published results (Ioannidis, 2005; Leamer, 1983) recent attempts to replicate psychology studies in mass have given these concerns new fuel (Camerer et al., 2018; Open Science Collaboration, 2015). Scholars pin the blame on a host of different culprits: the abuse of p-values (Wasserstein et al., 2016), researcher degrees of freedom (Simmons et al., 2011), a failure to properly make causal inferences (Samii, 2016), and a host of other questionable practices. Sociology has not been immune to these critiques, with scholars highlighting concerns over transparency (Freese, 2007; Freese and Peterson, 2017), model uncertainty/misspecification (Winship and Western, 2016; Young, 2009, 2018), abuse of p-values (Gerber and Malhotra, 2008), and causal inference (Elwert and Winship, 2014; Morgan and Winship, 2015). All of these critiques of methodological practice are concerned with how we judge a quantitative finding and the way in which knowledge is built cumulatively within the field, yet they miss a key element of the problem: it is hard to judge empirical evidence when the theoretical target goes unstated.

Differences in estimation strategies versus estimands are especially essential in the case of replication, which may fail because the original study was a statistical fluke or because the replication actually targets a different estimand. When replication studies are conducted on a sample drawn from a different population (i.e. a different panel survey or a pool of experimental subjects at a different university), the replication targets a different theoretical estimand. The theoretical estimand provides language to clarify the difference.[12] It still may undermine the importance of the original study to know that the result does not transport

[12] Freese and Peterson (2017) also discuss replication as a setting where the target parameter may be different: replications vary in their degree of similarity to the original study. Framed 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.

to a new setting, but those who pursue replications should at least be aware that their argument may be about contextual differences in estimands rather than about the validity of the estimates themselves. Beyond the domain of replication, estimands clarify when the results of papers diverge because they have different research goals, different strategies to estimate those goals, or simply different data. This may promote constructive dialogue across papers, both methodological and substantive.

5 Objections

Because this proposal may invite controversy, we briefly respond here to several potential objections: (1) the proposed framework is superfluous because it reflects existing best practice, (2) the proposed framework is damaging because it narrows the scope of sociology, and (3) the proposed framework sets an unachievable standard.

5.1 Objection 1: The Proposal is Superfluous

Our proposal to foreground the target of inference (estimand) is a call for clarity about the relationship between theory and evidence. One might reasonably object that no one is arguing in favor of *less* clarity and thus, the proposal is itself superfluous.[13] Sociologists value clarity, but this is not being achieved in the vast majority of current empirical work. 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) are impossible to answer without reference to an objective. Common practice remains to define that target in terms of a regression coefficient, and then debate with reviewers about how best to specify that regression. Our contribution is to highlight how explicit estimands stated apart from regression coefficients can avoid these problems by clarifying the goals of authors, thus providing grounds to discuss what the most

[13] Our imaginary interlocutor who argues that our proposal is superfluous is inspired by Lizardo’s (2012) critique of analytical sociology.

useful methodological choices would be.

5.2 Objection 2: The Proposal Narrows Inquiry

The opposite objection might also be made, that the proposal so narrows sociological inquiry that to adopt it would damage the discipline. Social phenomena are complicated and theories of complex processes (e.g. the persistence of racial inequality in the U.S.) are not easily reducible to a few estimands. Our goal is not to reduce the richness of theory, but rather to clarify the piece of a given theory to which a particular quantitative analysis is able to speak. There will always be pieces of a theory that cannot be empirically tested. Clarifying those boundaries can lead to more productive exchange between scholars. Importantly, we also do not intend to place a disproportionate focus on causal estimands. Sociological research requires descriptive facts such as the distribution of incomes in the United States or the racial wealth gap, all of which can be stated as theoretical estimands. Description is an important goal in itself which should be prioritized (Gerring, 2012). Failing to recognize descriptive goals as explicitly descriptive can also lead to misleading results.

Evidence for important questions is likely to be imperfect, and we should not restrict ourselves to settings where identification assumptions are unimpeachable. The turn toward "causal empiricism" (Samii, 2016) in political science and economics has led toward a focus on questions that can be answered by randomized experiments, instrumental variables, regression discontinuities, and other techniques for which identification assumptions are considered especially credible. Critics argue that the adoption of these techniques is like losing one's keys in a parking lot and then searching only under the streetlight because that is where it is possible to see; they focus on a very narrow scope of questions for a very specific subpopulation by virtue of a focus on only the questions that can be answered well. The narrowness of the questions at stake leads to answers that are of little use in policymaking or theoretical development (Deaton, 2010; Heckman and Urzua, 2010).

Economists also tend to focus on treatments, such as school expenditures, that one can easily imagine manipulating in an experiment. Sociologists often examine treatments for which an experimental manipulation is more difficult to imagine. If cultural capital is defined as embodied in one's habitus as a set of durable dispositions composed of many components, it can be difficult to conceptualize the causal contrast of what would happen if one's habitus were different. One path forward in this setting is to break complex constructs down into smaller components for which a well-defined causal contrast is more straightforward. For instance, Khan (2010) discusses various elements that compose higher cultural capital among students at an elite boarding school. These include the ease with which students navigate interactions with authority figures and physical attractiveness. While neither may be especially amenable to a real-world intervention, they are sufficiently concrete that we can reason about the hypothetical outcome that might be realized if these specific attributes were different. It may then be possible to aggregate many specific causal effects into a broader theory about a complex construct like cultural capital.

To avoid falling into the trap of narrow questions, sociologists will need to maintain a commitment to ask big questions even if it takes creativity to define the causal contrast and even if the answers that are possible are not definitive. 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. Obfuscation of the true goals does not make an argument more compelling.

5.3 Objection 3: The Proposal Sets an Unattainable Standard

The final objection is that the goals we are proposing are simply unattainable. Critics may worry that we never understand the world well enough to draw a DAG, and therefore that causal inference from observational data should never be attempted. Even those who take this position are likely to believe that observational evidence can suggest a causal effect, or that it may reveal hypotheses that should be tested in experiments. Unfortunately, if our knowledge of the causal DAG is truly zero, no observational evidence can inform a causal hypothesis. By making the causal goal of research explicit and precise, we gain the ability to state formally the reasons why a descriptive finding points toward a causal claim: because the required DAG is at least mildly plausible. By clarifying our assumptions, we draw attention to areas where other scholars might disagree. Far from leading us to turn a blind eye to the problems of causal inference, precision brings these issues to the fore so that our vague intuitions can be made precise.

Our framework begins with the premise that it is possible for researchers to choose an estimand. Selection of this estimand is the most consequential point of the research process. A second version of the 'unattainable standard' objection, is that certain subfields may lack the theoretical closure necessary to find a set of reviewers willing 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 the debate we ought to be having—focusing on what quantities are most important to theory rather than talking past each other about methodological choices most appropriate for studying different things.

6 Conclusion

Sociology stands out from other social sciences for its richness of theory and ambitious attempts to answer questions of pressing social relevance even when doing so is difficult. Yet the methods sections of sociology papers often leave the reader unsure of the goal of estimation (the estimand) and its relevance to the theory to be tested. Quantitative sociologists sell themselves short when pages of theoretical development are followed by a table of regression coefficients that, when examined more closely, are only loosely connected to the theory the author is trying to develop by a model everyone knows is misspecified. At best this creates an uncomfortable ambiguity about the author's intentions and at worst can lead to deeply misleading conclusions.

Bridging the gap between theory and quantitative evidence starts with a clear statement of the target of inquiry, the theoretical estimand. We advocate a three step research process which involves (1) choosing a theoretical estimand, (2) choosing an empirical estimand which is informative about the theoretical estimand under a set of identification assumptions, and (3) choosing an estimation strategy to learn that empirical estimand. The key advantages of this approach include improved clarity and access to cutting-edge tools in identification and estimation.

We do not expect this advice to be easy to adopt. We have ourselves published research in which the estimand is not entirely clear. Being specific about the estimand takes time and effort. It requires the discomfort of facing the reality that our empirical evidence may inform a smaller portion of our theoretical process or target population than we would like. However, being clear about both the aspirations and limitations of quantitative evidence is an essential part of enabling the field to build knowledge collaboratively. Being clear about the estimand will not solve all of science's problems. If we want to hit a target, though, it is helpful to know where we should aim.

References

Abadie, A. and G. W. Imbens 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A. and G. W. Imbens 2011. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.

Abbott, A. 1988. Transcending general linear reality. *Sociological Theory*, 6(2):169–186.

Acharya, A., M. Blackwell, and M. Sen 2016. Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review*, 110(3):512–529.

Angrist, J. D. and W. N. Evans 1998. Children and their parents' labor supply: Evidence from exogenous variation in family size. *The American Economic Review*, 88(3):450–477.

Angrist, J. D., G. W. Imbens, and D. B. Rubin 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Aronow, P. M. and B. T. Miller 2019. *Foundations of Agnostic Statistics.* Cambridge University Press.

Aronow, P. M. and C. Samii 2016. Does regression produce representative estimates of causal effects? *American Journal of Political Science*, 60(1):250–267.

Athey, S. and G. Imbens 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Auspurg, K., J. Brüderl, and T. Wöhler 2019. Does immigration reduce the support for welfare spending? a cautionary tale on spatial panel data analysis. *American Sociological Review*, 84(4):754–763.

Berk, R., A. Buja, L. Brown, E. George, A. K. Kuchibhotla, W. Su, and L. Shazo 2019. Assumption lean regression. *The American Statistician*, (Online first.).

Berk, R. A. 2004. *Regression analysis: A constructive critique.* Sage.

Bickel, P. J., E. A. Hammel, and J. W. O'Connell 1975. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398–404.

Blau, P. M. and O. D. Duncan 1967. *The American Occupational Structure.* New York: Wiley.

Bloome, D., S. Dyer, and X. Zhou 2018. Educational inequality, educational expansion, and intergenerational income persistence in the United States. *American Sociological Review*, 83(6):1215–1253.

Brand, J. E. and Y. Xie 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., K. B. Karlson, and A. Holm 2018. Interpreting and understanding logits, probits, and other nonlinear probability models. *Annual Review of Sociology*, 44:39–54.

Buja, A., R. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, K. Zhan, and L. Zhao 2019. Models as approximations, part I: A conspiracy of nonlinearity and random regressors in linear regression. *arXiv preprint arXiv:1404.1578.*

Camerer, C. F., A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, et al. 2018. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637.

Chetty, R., N. Hendren, M. R. Jones, and S. R. Porter 2018. Race and economic opportunity in the United States: An intergenerational perspective. Technical report, National Bureau of Economic Research.

Chmielewski, A. K. 2019. The global increase in the socioeconomic achievement gap, 1964 to 2015. *American Sociological Review*, 84(3):517–544.

Ciocca Eller, C. and T. A. DiPrete 2018. The paradox of persistence: Explaining the black-white gap in bachelors degree completion. *American Sociological Review*, 83(6):1171–1214.

D'Amour, A. and E. Airoldi 2017. The effective estimand. `https://www.alexdamour.com/content/effective_estimand_prearxiv.pdf`. Unpublished manuscript.

D'Amour, A., P. Ding, A. Feller, L. Lei, and J. Sekhon 2017. Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*.

Deaton, A. 2010. Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48(2):424–55.

Deaton, A. and N. Cartwright 2018. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2–21.

Desmond, M. and A. Travis 2018. Political consequences of survival strategies among the urban poor. *American Sociological Review*, 83(5):869–896.

Elwert, F. and C. Winship 2014. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40:31–53.

Firebaugh, G., C. Warner, and M. Massoglia 2013. Fixed effects, random effects, and hybrid models for causal analysis. In *Handbook of Causal Analysis for Social Research*, Pp. 113–132. Springer.

Font, S. A., L. M. Berger, M. Cancian, and J. L. Noyes 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.

Freese, J. 2007. Replication standards for quantitative social science: Why not sociology? *Sociological Methods and Research*, 36(2):153–172.

Freese, J. and D. Peterson 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.

Gauchat, G. and K. T. Andrews 2018. The cultural-cognitive mapping of scientific professions. *American Sociological Review*, 83(3):567–595.

Gelman, A. and H. Stern 2006. The difference between significant and not significant is not itself statistically significant. *The American Statistician*, 60(4):328–331.

Gerber, A. S. and N. Malhotra 2008. Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods and Research*, 37(1):3–30.

Gerring, J. 2012. Mere description. *British Journal of Political Science*, 42(4):721–746.

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.

Greiner, D. J. and D. B. Rubin 2011. Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93(3):775–785.

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., M. Kim, and E. W. Zuckerman Sivan 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.

Harding, D. J., J. D. Morenoff, A. P. Nguyen, and S. D. Bushway 2018. Imprisonment and labor market outcomes: Evidence from a natural experiment. *American Journal of Sociology*, 124(1):49–110.

Heckman, J. J. and S. Urzua 2010. Comparing IV with structural models: What simple IV can and cannot identify. *Journal of Econometrics*, 156(1):27–37.

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 J. M. Robins 2020. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC.

Hernán, M. A., B. C. Sauer, S. Hernández-Díaz, R. Platt, and I. Shrier 2016. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology*, 79:70–75.

Hirschman, D. 2016. Stylized facts in the social sciences. *Sociological Science*, 3:604–626.

Homan, P. 2019. Structural sexism and health in the United States: A new perspective on health inequality and the gender system. *American Sociological Review*, 84(3):486–516.

Horowitz, J. 2018. Relative education and the advantage of a college degree. *American Sociological Review*, 83(4):771–801.

Hout, M. 1988. More universalism, less structural mobility: The american occupational structure in the 1980s. *American Journal of Sociology*, 93(6):1358–1400.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto 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. 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 J. D. Angrist 1994. Identification and estimation of local average treatment effects. *Econometrica: Journal of the Econometric Society*, Pp. 467–475.

Imbens, G. W. and D. B. Rubin 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.

Ioannidis, J. P. 2005. Why most published research findings are false. *PLoS medicine*, 2(8):e124.

Kadivar, M. A. 2018. Mass mobilization and the durability of new democracies. *American Socio-*

logical Review, 83(2):390–417.

Keele, L., R. T. Stevenson, and F. Elwert 2019. The causal interpretation of estimated associations in regression models. *Political Science Research and Methods*, Pp. 1–13.

Khan, S. R. 2010. *Privilege: The Making of an Adolescent Elite at St. Paul's School.* Princeton University Press.

Knox, D., W. Lowe, and J. Mummolo 2019. Administrative records mask racially biased policing. Available at SSRN: http://dx.doi.org/10.2139/ssrn.3336338.

Kohler-Hausmann, I. 2018. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Northwestern University Law Review*, 113:1163.

Leamer, E. E. 1983. Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43.

Legewie, J. and J. Fagan 2019. Aggressive policing and the educational performance of minority youth. *American Sociological Review*, 84(2):220–247.

Lieberson, S. 1987. *Making It Count: The Improvement of Social Research and Theory.* University of California Press.

Lieberson, S. and J. Horwich 2008. Implication analysis: A pragmatic proposal for linking theory and data in the social sciences. *Sociological Methodology*, 38(1):1–50.

Light, M. T. and J. T. Thomas 2019. Segregation and violence reconsidered: do whites benefit from residential segregation? *American Sociological Review*, 84(4):690–725.

Lizardo, O. 2012. Analytical sociology's superfluous revolution. *Sociologica: Italian Online Sociological Review*, 1.

Ludwig, V. and J. Brüderl 2018. Is there a male marital wage premium? New evidence from the United States. *American Sociological Review*, 83(4):744–770.

Luo, L., J. Hodges, C. Winship, and D. Powers 2016. The sensitivity of the intrinsic estimator to coding schemes: Comment on Yang, Schulhofer-Wohl, Fu, and Land. *American Journal of Sociology*, 122(3):930–961.

McDonnell, M.-H. and B. G. King 2018. Order in the court: How firm status and reputation shape the outcomes of employment discrimination suits. *American Sociological Review*, 83(1):61–87.

Mize, T. D. and B. Manago 2018. Precarious sexuality: How men and women are differentially categorized for similar sexual behavior. *American Sociological Review*, 83(2):305–330.

Molina, M. and F. Garip 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. and C. Winship 2015. *Counterfactuals and Causal Inference.* Cambridge University Press.

Mustillo, S. A., O. A. Lizardo, and R. M. McVeigh 2018. Editors comment: A few guidelines for quantitative submissions. *American Sociological Review*, 83(6):1281–1283.

Open Science Collaboration 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Pager, D. 2003. The mark of a criminal record. *American Journal of Sociology*, 108(5):937–975.

Pal, I. and J. Waldfogel 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. 2009. *Causality.* Cambridge University Press.

Pearl, J. and D. Mackenzie 2018. *The Book of Why: The New Science of Cause and Effect.* Basic Books.

Quadlin, N. 2018. The mark of a womans 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.

Ruggles, S., S. Flood, R. Goeken, J. Grover, M. Erin, J. Pacas, and M. Sobek 2019. *IPUMS USA: Version 9.0 [dataset].* Minneapolis, MN: IPUMS.

Samii, C. 2016. Causal empiricism in quantitative research. *The Journal of Politics*, 78(3):941–955.

Schilke, O. and G. Rossman 2018. It's only wrong if it's transactional: Moral perceptions of obfuscated exchange. *American Sociological Review*, 83(6):1079–1107.

Schneider, D., O. P. Hastings, and J. LaBriola 2018. Income inequality and class divides in parental investments. *American Sociological Review*, 83(3):475–507.

Sen, M. and O. Wasow 2016. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19:499–522.

Sharkey, P. and F. Elwert 2011. The legacy of disadvantage: Multigenerational neighborhood effects on cognitive ability. *American Journal of Sociology*, 116(6):1934–81.

Simmons, J. P., L. D. Nelson, and U. Simonsohn 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366.

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.

Travis, A. 2019. The organization of neglect: Limited liability companies and housing disinvestment. *American Sociological Review*, 84(1):142–170.

Van der Laan, M. J. and S. Rose 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.

Villarreal, A. and C. R. Tamborini 2018. Immigrants' economic assimilation: Evidence from longitudinal earnings records. *American Sociological Review*, 83(4):686–715.

Wager, S. and S. Athey 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wang, D. J., H. Rao, and S. A. Soule 2019. Crossing categorical boundaries: A study of diversification by social movement organizations. *American Sociological Review*, 84(3):420–458.

Wasserstein, R. L., N. A. Lazar, et al. 2016. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133.

Watts, D. J. 2014. Common sense and sociological explanations. *American Journal of Sociology*, 120(2):313–351.

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 S. Greenland 2013. The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, 177(4):292–298.

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 B. Western 2016. Multicollinearity and model misspecification. *Sociological Science*, 3(27):627–649.

Wodtke, G. T., F. Elwert, and D. J. Harding 2016. Neighborhood effect heterogeneity by family income and developmental period. *American Journal of Sociology*, 121(4):1168–1222.

Wodtke, G. T., D. J. Harding, and F. Elwert 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. 2017. *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC.

Xie, Y. 2013. Population heterogeneity and causal inference. *Proceedings of the National Academy of Sciences*, 110(16):6262–6268.

Young, C. 2009. Model uncertainty in sociological research: an application to religion and economic growth. *American Sociological Review*, 74(3):380–397.

Young, C. 2018. Model uncertainty and the crisis in science. *Socius*, 4:1–7.

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

Zhou, X. and G. T. Wodtke 2019. A regression-with-residuals method for estimating controlled direct effects. *Political Analysis*, 27(3):360–369.

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

B Mediation involves interventions to two variables

This section briefly expands on the mediation estimands listed in Table 2. 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.

$$\text{CDE}(0) = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(1,0) - Y_i(0,0)\Big) \tag{B.1}$$

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.

$$\text{CDE}(1) = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(1,1) - Y_i(0,1)\Big) \tag{B.2}$$

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 $M_i(0)$.

$$\text{NDE} = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(1,M_i(0)) - Y_i(0,M_i(0))\Big) \tag{B.3}$$

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 is quantity as advised in other commentaries (Mustillo et al., 2018). This fact reinforces the importance of stating the causal contrast in precise terms at the start of the analysis.

C Family gap in pay: Example details

This appendix provides details about our estimation illustration of the family gap in pay. We draw data from the 2019 Annual Social and Economic Supplement (ASEC) of the March Current Population Survey ($N = 25,610$), 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 who were employed in the last year and were not self-employed ($N = 18,075$). 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 three other predictors: 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 mothers in the region of common support, we separate the data into subgroups of age, education, race, and marital status. Within each subgroup for which at least one mother is observed, we note whether any non-mothers are observed. We then aggregate over the weighted distribution of mothers across subgroups to summarize the proportion of mothers who fall in a covariate subgroup where at least one mother is observed (99%). All analyses restrict to this region of common support.

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. Then, we predict the outcome for each mother 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.

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 3. 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`,

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

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 $\hat{\text{V}}(\hat{\theta})$.

In the main text, we note that 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 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 quadratic age specification interacted with motherhood, with all other predictors entered additively.