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

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

June 7, 2021

Preprint of a paper in the *American Sociological Review*.
Published version: https://doi.org/10.1177%2F00031224211004187

Keywords: social statistics, research design, descriptive inference, causal inference, estimands

Abstract

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

(166 words)

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

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

[‡]Assistant Professor, Program in Quantitative Social Science and affiliate, Sociology, Dartmouth College, rebeccajohnson.io, rebecca.ann.johnson@dartmouth.edu

[§]Assistant Professor and Arthur H. Scribner Bicentennial Preceptor, 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

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

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

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

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

Stating both the theoretical estimand and the empirical estimand separately from the estimation strategy partitions the link between theory and evidence into three steps involving different modes of argument (Fig. 1). The distinction between the theoretical and empirical estimands is subtle but important: the former may involve unobservable quantities such as counterfactuals while the latter involves only observable data. Our full argument for the separate statement of the theoretical and empirical estimands appears in the section that introduces the empirical estimand. The choice of theoretical estimands requires substantive argument about the theory and goals, while the choice of empirical estimands requires

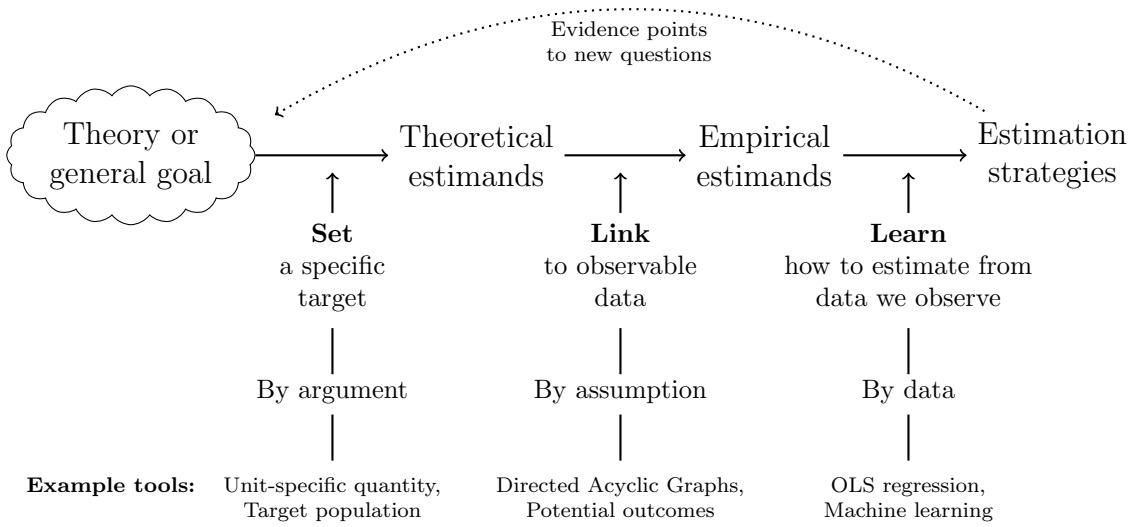


Fig. 1. Three critical choices in quantitative social science arguments. The first choice is the theoretical estimands, which set the targets of inference. Argument is required to link the theoretical estimands to the broader theory. The second choice is the empirical estimands, which link the targets to observable data. The connection requires substantive assumptions that can be formalized in Directed Acyclic Graphs. The third choice is the estimation strategies, which captures what we will actually do with data. We select estimation strategies based on the data.

conceptual argument about unobserved data. The choice of estimation strategies is distinct because it can be at least partially data-driven. Separating these steps helps researchers make principled choices, allows readers to evaluate claims, and enables the community to build on research findings. Too often, research papers involve pages of rich theory followed by pages of procedures applied to data, with a vague link between the two. The theoretical and empirical estimand fill the void by making precise both the theoretical quantity we would like to know and the empirical quantity that our procedures are most directly designed to approximate. Our most emphatic argument is that the field has much to gain from a precise statement of the true target of inquiry even if the assumptions required to estimate it hold only imperfectly and the empirical tools available are limited. Stating the goal allows the reader and the community to engage meaningfully with and build on the work.

The paper proceeds in several sections. We first introduce our framework, highlighting each step of our proposed research process: setting the theoretical estimand (2.1), linking

to an empirical estimand (2.2), and learning an estimate from data (2.3). Section 3 demonstrates the prevalence of the problems we address through a review of the 2018 volume of *ASR* and illustrates how our framework would transform quantitative research through two in-depth examples. Section 4 concludes by describing how estimands clarify methodological issues for analysts, readers, and the broader community.

2 Estimands Link Theory and Evidence

This section details each step in our proposed link between theory and evidence. Figure 2 presents the examples that we use to introduce this general framework.

2.1 The Theoretical Estimand: Set the Target

The most important step in quantitative empirical research is a clear statement of the research question. The clarity of the question is paramount because no single analysis can prove or undermine an entire sociological theory (Lieberson and Horwich, 2008). When estimating causal effects, for instance, one must state the population over which heterogeneous effects are averaged (Brand and Xie, 2010; Xie, 2013). When estimating associations, one must be clear about whether the target of inference is causal (Hernán, 2018), and if so to be clear about the hypothetical intervention at the core of the claim (Greiner and Rubin, 2011; Hernán et al., 2016; Morgan and Winship, 2015; Sen and Wasow, 2016). A lack of clarity can lead to a table of regression coefficients that are, at best, weakly informative about theory (Keele et al., 2020; Westreich and Greenland, 2013). Before you apply a statistical procedure, you have to define the thing you are trying to estimate or measure (Katz et al., 2020). Without the language to make a more precise statement, researchers find themselves constrained to questions stated in terms of regression coefficients. We join a long line of increasingly urgent calls to think beyond the constraints of regression as it is commonly practiced (Duncan 1984; Lieberson 1987; Abbott 1988; Freedman 1991; Berk 2004).

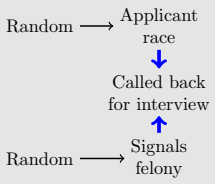
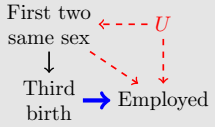
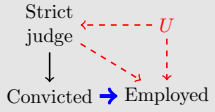
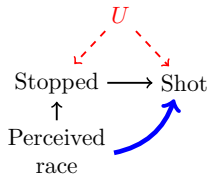
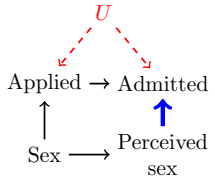
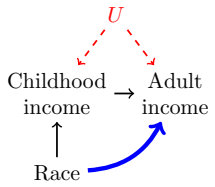
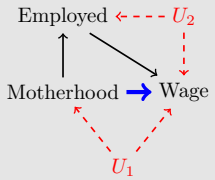
	Set the target: The theoretical estimand		Link to observables	Learn from data
	Unit-specific quantity	Target population of units	Identification	Estimation
Pager	Difference in whether application i would be called back if it signaled white with a felony vs. black without	Applications to jobs in Milwaukee		Logistic regression
Angrist and Evans	Difference in whether mother i would be employed if she had three vs. two children	Those who would have a third birth only if first two of the same sex		Two-stage least squares
Harding et al.	Difference in whether person i would be employed if convicted vs. if not	Those who would be convicted only under certain judges		Two-stage least squares
Fryer	Difference in whether person i would be stopped if perceived as black vs. white	Those stopped by police		Logistic regression
Bickel et al.	Difference in whether applicant i would be admitted if perceived as male vs. female	Applicants to Berkeley		Difference in proportions
Chetty et al.	Adult income that person i would be realized if childhood income took a particular value	U.S. population		OLS
Pal and Waldfogel	Wage that mother i would realize if she were an employed mother vs. an employed non-mother	U.S. civilian women ages 25–44 in March 2019		OLS Parametric g-formula

Fig. 2. Estimands are relevant to a broad range of social science studies. White boxes on the diagonal are the focus of the main text, but every study implicitly involves all four steps. Some steps (i.e. DAGs for identification, see Section 2.2) 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.

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

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\substack{\text{Mean over every } i \\ \text{among U.S. adults} \\ \text{(target population)}}} \underbrace{Y_i}_{\substack{\text{Whether each } i \\ \text{is employed} \\ \text{(unit-specific quantity)}}} \tag{2.1}$$

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

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\substack{\text{Mean over every } i \\ \text{among U.S. adults} \\ \\ \text{(target population)}}} \underbrace{\Big(\underbrace{Y_i(1)}_{\substack{\text{Employment if} \\ \text{enrolled in} \\ \text{job training}}} - \underbrace{Y_i(0)}_{\substack{\text{Employment if} \\ \text{not enrolled in} \\ \text{job training}}}\Big)}_{\text{(unit-specific quantity)}} \tag{2.2}$$

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

2.1.1 Specify the Unit-Specific Quantity

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

The unit-specific quantities in sociological research can be complex. For instance, Pager (2003) explores "the ways in which the effects of race and criminal record interact to produce new forms of labor market inequalities," (p. 938). The study navigates this difficult topic through a randomized design. Even before randomization, however, the real novelty of the study is the definition of the unit of analysis as an *application* rather than as a *person*. In the experiment, job postings were randomly assigned to receive applications from a white or black pair of applicants. Each member of the pair approached the employer at a different time to apply for the job posting, a combination which we call an application. For each posting, one application was randomly assigned to signal a felony conviction for possession of cocaine. For each application i, an outcome Y_i was observed: whether that application received a callback. Each application was thus randomized to one of the four treatment conditions captured by the 2×2 table in Figure 3 Panel A, each of which has a potential outcome Y_i(`Treatment Condition`). Taking the unit of analysis as the application rather than the person sidesteps problems that plague the study of race within a causal framework. It may be very difficult to disentangle race from individual identities in order to consider a counterfactual world in which a person signaled a different racial category

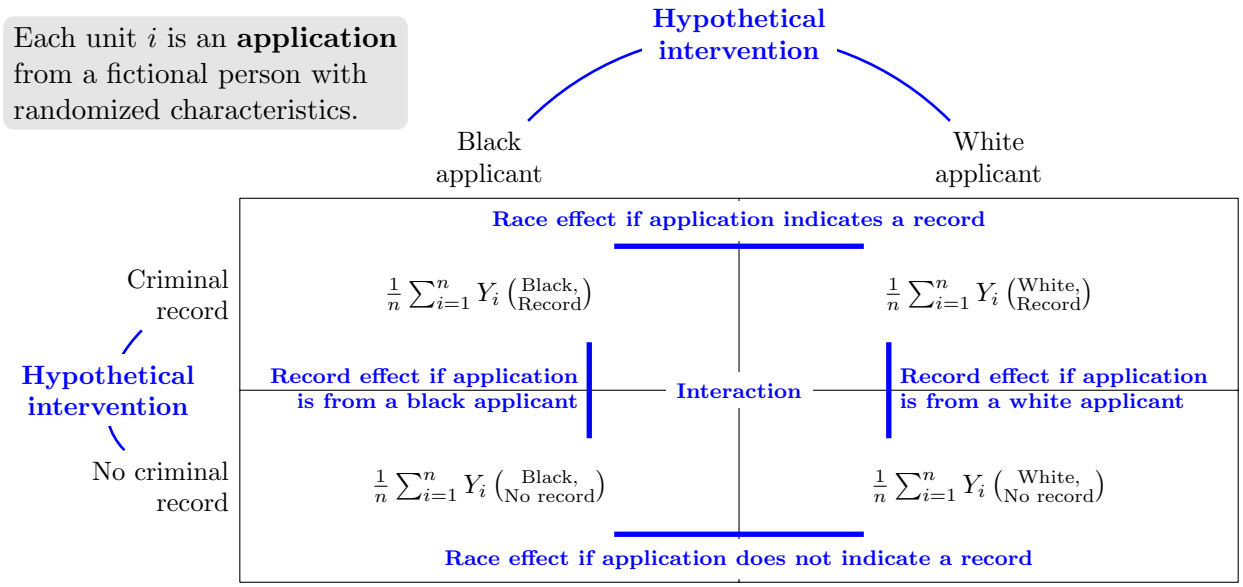
(Kohler-Hausmann, 2018). It is reasonably straightforward, however, to imagine an application signaling a different racial category (Sen and Wasow, 2016). Pager (2003) therefore makes progress by studying the application rather than the person as the unit of analysis. Randomization is made possible by the pivot in how the unit of analysis is defined.

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

The scientific insight of Pager (2003) contrasts with what could be learned in an observational study focused on a different unit-specific quantity (Fig. 3 Panel B). Suppose we defined the unit of analysis as a real flesh-and-blood applicant in an actual population of those applying for jobs. For real people (as opposed to applications), it is difficult to conceptualize all the things that would have to change in a world where a real person was counterfactually of another racial category—access to schooling, earlier experiences of discrimination, and inumerable opportunities that strengthen a resume. For these reasons, viewing race as a causal treatment may not be straightforward in a study where the unit of analysis is a person. Racial categories could instead denote two populations that differ in myriad ways due to systemic racism. Potential outcomes could be defined as a function of a criminal record *only*. The unit-specific quantity would involve two potential outcomes: the outcome each person would realize if they had a criminal record or if they did not. One could compare the causal effect of a criminal record across subpopulations of black and white applicants.

The colloquial term "moderation" could describe the research goal in both the observational design and the Pager (2003) design, but the meanings of the two estimands are distinct. The policy implications of the former would focus on preventing hiring decision makers from directly considering race and criminal histories when evaluating identical ap-

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



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

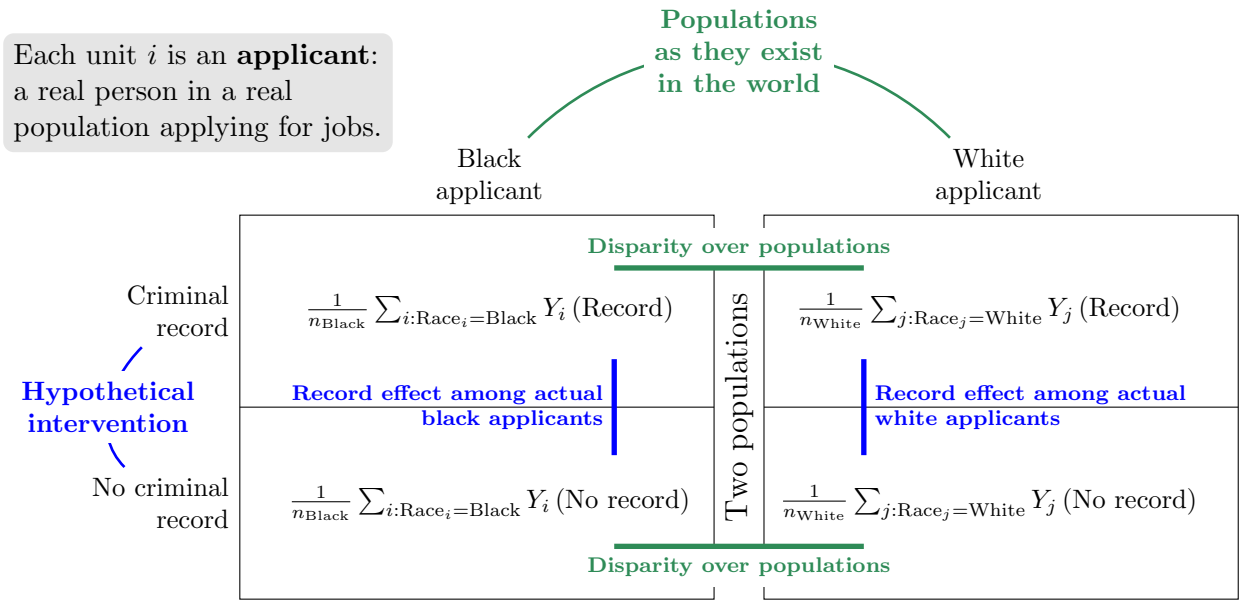


Fig. 3. Two estimands with different unit-specific quantities and different target populations. Both estimands could be termed the effect of a criminal record on the probability of a callback among black and white applicants, yet the two are quite different. A design targeting causal interaction (Pager, 2003) would randomly assign units (applicant-application pairs) to a cell of the 2×2 table that combines all values of both treatments. A design targeting effect heterogeneity would take applications in the real-world distribution for each subgroup and estimate the outcome they would realize if they signaled or did not signal a criminal record. Both estimands are of substantive interest.

plications. The policy implications of the latter would focus on how equalizing criminal histories (or signals of those histories) could reduce disparities across populations of actual job applicants of different racial categories. Both are worthwhile goals. Distinguishing them requires clarity about whether the unit-specific quantity is a potential outcome as a function of one or two treatments.

The unit-specific quantity provides an opportunity to clarify the causal component (if any) of our research claims. It invites us to be precise about the causes we are studying (e.g. a signal of race) and those that we are not (e.g. how racial disparities in access to education create differences in resumes). It allows us to be precise about the levels of the treatment being contrasted (e.g. a particular resume line about a felony conviction for possession of cocaine). Finally, precision about the unit-specific quantity facilitates the study of questions involving interventions to multiple variables; the causal interaction between race and a criminal record is only one example of many such questions (Table 1). An experimental protocol provides a perhaps unparalleled opportunity for clarity in these regards, but nothing prevents observational studies from aspiring to similar clarity. Even if the assumptions necessary to estimate the goal are doubtful, there is never a reason for the author's intention for the statistical evidence to be in doubt.

2.1.2 Define the Target Population

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

The target population is a widely-recognized issue in experiments. For example, Pager

Estimand name	Mathematical statement	DAG (Sec. 2.2)	Reference	Colloquial terms
Average treatment effect	$\frac{1}{n}\sum_i \Big(Y_i(d') - Y_i(d)\Big)$	$D \to Y$	Morgan and Winship (2015)	Effect
Conditional average treatment effect	$\frac{1}{n_x}\sum_{i:X_i=x} \Big(Y_i(d') - Y_i(d)\Big)$	$X \to D \to Y$, $X \to Y$	Athey and Imbens (2016)	Effect heterogeneity or moderation
Causal interaction	$\frac{1}{n}\sum_i \Big(\Big(Y_i(a',d') - Y_i(a',d)\Big) - \Big(Y_i(a,d') - Y_i(a,d)\Big)\Big)$	$A \to Y$, $D \to Y$	VanderWeele (2015)	Joint treatment effect
Controlled direct effect	$\frac{1}{n}\sum_i \Big(Y_i(d',m) - Y_i(d,m)\Big)$	$D \to M \to Y$, $D \to Y$	Acharya et al. (2016)	Mediation (Sec. 3.3)
Natural direct effect	$\frac{1}{n}\sum_i \Big(Y_i(d',M_i(d)) - Y_i(d,M_i(d))\Big)$	$D \to M \to Y$, $D \to Y$	Imai et al. (2011)	Mediation (Appendix B)
Effect of time-varying treatment	$\frac{1}{n}\sum_i \Big(Y_i(d_1',d_2') - Y_i(d_1,d_2)\Big)$	$D_1 \to D_2 \to Y$, $D_1 \to Y$	Wodtke et al. (2011)	Cumulative effect

Table 1. Unit-specific quantities defined in potential outcomes unlock many causal estimands for inquiry. Social scientists who define the research goal before moving to regression uncover more possible questions than those who confine themselves to regression parameters. This provides a non-exhaustive list of common causal estimands. The mathematical statement of each estimand involves counterfactuals—potential outcomes under unobserved treatment assignments—and is the parameter that the quantitative analysis would hope to estimate. The DAG depicts one potential set of identification assumptions to link unobservable quantities to observable data. 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).

(2003:965) explicitly notes that "one key limitation of the audit study design is its concentration on a single metropolitan area." The true target population may involve a larger set, such as all entry-level job openings in the U.S. If so, the researcher must argue that a particular piece of empirical evidence (an experiment in Milwaukee with entry-level job openings advertised in one newspaper and one website) is informative about that broader target population. Alternatively, one could define the target population more narrowly as entry-level job openings in Milwaukee. Then, the researcher must motivate not what Milwaukee tells us about other places but why we should care about Milwaukee specifically. A clear statement of the target population allows a researcher to clarify which approach they are taking.

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

Sometimes, one could defend the complier population as being of genuine theoretical in-

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

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

The target population clarifies debates about which methods are most credible. Some feel that econometric approaches like IV and regression discontinuity designs provide evidence that is too limited because the leap from the identified population to the full population of interest is too severe (Deaton, 2010; Deaton and Cartwright, 2018; Heckman and Urzua, 2010). Others argue that the causal identification problems in the full population are so

difficult that we are better off focusing on the subpopulation for whom we can identify an effect (Imbens, 2018, 2010; Samii, 2016). Both sides have fair points. A target population allows authors to navigate this tension directly, either by arguing that the subpopulation they identify is informative about the general population or is theoretically important in its own right.

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

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

2.2 Identification: Link to an Empirical Estimand

The same quality that makes a theoretical estimand liberating—it can involve unobservable data—also means that strong assumptions will be required to learn about that estimand from statistical procedures, which can only be applied to observable data. The second

step of our framework links the theoretical estimand to an empirical estimand: a target of inference that only involves observable data. That link can be formalized with tools like Directed Acyclic Graphs (DAGs) (Morgan and Winship, 2015; Pearl, 2009). Yet, despite decades of methodological advice to focus not only on technical fixes in regression but also on scientific issues like selection into treatment (Freedman, 1991), DAGs or other equivalent statements of conditional independence appear in only a minority of sociological studies. One reason causal assumptions are missing from research practice may be that authors believe their research goals are not causal and therefore lie outside the scope of problems for which those assumptions are needed. We argue that a clear statement of both the research goal (the theoretical estimand) and the concrete target of the statistical analysis (the empirical estimand) would clarify the identification assumptions are needed in a much wider range of questions. We introduce the idea of two estimands (theoretical and empirical) with a causal example before turning to more complex examples from demography and from the study of disparities.

2.2.1 Two estimands: An introduction with a causal effect

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

$$\underbrace{\tau = \frac{1}{n}\sum_{i=1}^{n}}_{\substack{\text{Mean}\\ \text{over } \textbf{all}\\ \text{applications}}} \left(\underbrace{Y_i\left(\substack{\text{White,}\\ \text{Record}}\right)}_{\substack{\text{Potential}\\ \text{outcome under}\\ \text{one condition}}} - \underbrace{Y_i\left(\substack{\text{Black,}\\ \text{No record}}\right)}_{\substack{\text{Potential}\\ \text{outcome under}\\ \text{another condition}}} \right) \tag{2.3}$$

For each unit, it is not possible to observe both potential outcomes, so the theoretical estimand τ is not an empirical quantity. The empirical estimand is the difference in the

observed outcomes between job applications actually assigned to each of these experimental conditions. This involve only observable quantities (no potential outcomes).

$$\theta = \underbrace{\frac{1}{n_{\mathrm{WR}}} \sum_{i \in \mathcal{S}_{\mathrm{WR}}} \overbrace{Y_i}^{\text{Factual outcomes}}}_{\substack{\text{Mean among} \\ \text{applications assigned} \\ \text{to the condition} \\ (\textbf{W}\text{hite, }\textbf{R}\text{ecord})}} - \underbrace{\frac{1}{n_{\mathrm{BN}}} \sum_{i \in \mathcal{S}_{\mathrm{BN}}} \overbrace{Y_i}^{\text{Factual outcomes}}}_{\substack{\text{Mean among} \\ \text{applications assigned} \\ \text{to the condition} \\ (\textbf{B}\text{lack, }\textbf{N}\text{o record})}} \tag{2.4}$$

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

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

Stating both estimands is important because there is no 1-1 mapping between a theoretical and empirical estimand. One could examine a particular empirical estimand—for example, the difference in the mean callbacks of black and white applicants in administrative records with no adjustment—which could correspond to theoretical estimands as diverse as a descriptive disparity or a causal effect. The argument of a research study has to clarify to which of many possible theoretical estimands they intend the empirical estimand to speak so that the reader can adequately evaluate the available evidence for the claim. This is especially true in more complex settings.

2.2.2 Setting 1: Demographic standardization

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

"would" happen in a "counterfactual" population. Without such an explicit statement, the goal is unclear.

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

2.2.3 Setting 2: Disparities in the presence of selection processes

Few sociology papers reason explicitly about identification assumptions. One reason may be that the objects of sociological inquiry appear on the surface to be descriptive sample quantities, which may be valid under weaker assumptions. Yet results which seem to be descriptive empirical regularities, or stylized facts (Hirschman, 2016), often take on a theoretical meaning only under identification assumptions. We review three examples (Table 2) with a common style. The authors cite a descriptive disparity—police shootings by race, graduate admissions by sex, and adult incomes by race—but control for a third variable that is a consequence of the demographic characteristic of interest. This produces problems from conditioning on a collider variable (Elwert and Winship, 2014). These examples highlight the need to state the theoretical estimand, the empirical estimand, and the identification assumptions under which the two are equal even when the target quantity may not appear to be causal at first glance. A precise statement of the theoretical estimand can inform the assumptions to identify that estimand.

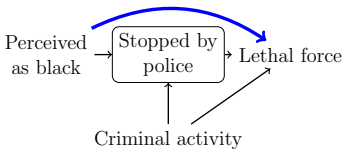
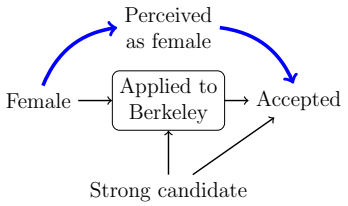
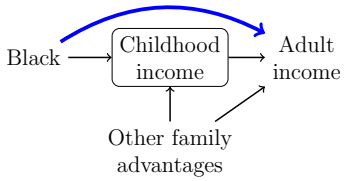
Study	Empirical Regularity	Misleading Conclusion	Directed Acyclic Graph
Fryer (2019)	Among those they stop, police shoot the same proportion of black individuals as white individuals.	Police do not discriminate against black individuals when using lethal force.	
Bickel et al. (1975)	Among those who apply, Berkeley departments admit a higher proportion of women than of men.	Admissions committees do not discriminate against women.	
Chetty et al. (2020)	Among those with equal childhood incomes, black and white women earn similar amounts as adults.	Equalizing childhood incomes would eliminate the racial gap in women's adult incomes.	

Table 2. Empirical regularities can be misleading without estimands. Each example reports an empirical regularity with a vague connection to a theoretical claim. The empirical regularity supports the misleading conclusion only under identification assumptions that the node at the bottom of each Directed Acyclic Graph (DAG, Pearl 2009) does not affect both the variable that the researchers hold constant (boxed) and the outcome (at right). We draw the Fryer (2019) example from a critique by Knox et al. (2020) which highlights this and other issues with the original paper. In the first row, equal use of lethal force against black individuals stopped by police may stem from the fact that being stopped is a collider: among those stopped, the behavior of blacks is likely to be less dangerous. In the second row, equal or higher acceptance rates among female candidates who apply to Berkeley could result because applying to Berkeley is a collider: among women, only the strong candidates apply. In the third row, childhood income is a collider: black families who overcome discrimination to attain incomes comparable to those of white families likely have other advantages that may contribute to their children's incomes in adulthood. When we state the theoretical and empirical estimands, the DAG makes clear that they are not equal and thus the descriptive quantity does not support the conclusions drawn.

Fryer (2019) examines police interactions by race in several administrative data sources. In records from New York City, the use of sub-lethal force was higher for blacks than non-blacks. Yet data from Houston on the most extreme form of force, police-involved shootings, showed no differences across racial groups. In both of these settings, the theoretical estimand (racial bias) is the difference in force if we intervene to change an officer's perception of an individual's race, averaged over those stopped by police. The empirical estimand is the difference in force used against black and white individuals who are involved in police interactions. Knox et al. (2020) highlight a key issue: the sample only includes those who interacted with police, either due to a stop or a 911 call, yet race affects whether these events occur (Table 2). If being black increases the risk of being stopped, then black individuals with a range of behaviors are stopped while only the most dangerous white individuals are stopped. Because the white individuals who are stopped are more dangerous than the black individuals who are stopped, an unbiased officer might actually use lethal force against whites at a *higher* rate among those who have been stopped. That is, equivalent rates are actually consistent with racial discrimination.[1]

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

In response to a comment by Durlauf and Heckman (2020), Fryer (2020) claims that he never sought to study racial bias, but only "racial differences" by repeatedly caveating the

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

results with the phrase "conditional on interaction," (Fryer 2020:1). He writes, "I am not sure how many more ways we would have needed to caveat our results to satisfy [Durlauf and Heckman]," (Fryer 2020:1). But caveats are exactly the problem. No one is well-served when methods make empirical evidence transparent (disparities in shooting conditional on a stop) but the theoretical quantity that motivates that evidence remains vague. Retreating from theoretical claims does not make the link between theory and evidence stronger. Rather, directly confronting the gap between a precise goal and the available evidence opens the door to transparent discussions and new tools to address that gap, as demonstrated by Knox et al. (2020). This is just one reason why it is essential to transparently state the study's true goals.

This problem is more general than the use of administrative data to study police bias. Bickel et al. (1975) study graduate admissions at Berkeley and discover that, although men are admitted at rates nine percentage points higher than women school-wide, women are admitted at higher rates than men within departments. The theoretical estimand is the difference in admission if we intervene to change a committee's perception of an applicant's sex. However, the empirical estimand—the disparity among students who actually apply to Berkeley—is not well situated to speak to that counterfactual. If, due to discrimination at the undergraduate level, many men apply to Berkeley but only the most qualified women apply to Berkeley, equal rates of admission among the men and women we observe could actually be consistent with sex-based discrimination against women.[2]

As a third example, Chetty et al. (2020) show that black and white women who are raised in families with similar incomes have similar earnings as adults. At face value, one might interpret this in terms of a theoretical estimand: if we intervened to equalize the childhood

[2]Bickel et al. (1975) explicitly assume away this problem: "in any given discipline male and female applicants do not differ in respect of their intelligence, skill, qualifications, promise, or other attribute deemed legitimately pertinent to their acceptance as students. It is precisely this assumption that makes the study of 'sex bias' meaningful, for if we did not hold it any differences in acceptance of applicants by sex could be attributed to differences in their qualifications, promise as scholars, and so on"(p. 398). While we applaud the explicitness of the assumption, it is questionable when discrimination affects decisions to apply for graduate school.

incomes of black and white women, the racial income gap in adulthood would disappear. Yet this would be misleading because family income is a consequence of both race and other family advantages; the black families who overcome discrimination to achieve incomes comparable to those of whites are likely to be advantaged in many other ways. In other words, childhood income is a collider variable (Table 2). The racial income gap in adulthood that would persist if we equalized childhood family incomes (a theoretical estimand, see Lundberg 2020) is likely to be different from empirical evidence about the racial gap in adult incomes among those observed with equal childhood incomes (an empirical estimand).

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

Table 2 formalizes these selection problems in causal DAGs (Pearl, 2009). Our framework aligns well with DAGs because they are nonparametric: they allow us to focus on one set of considerations (causal relationships) while delaying questions about the shape of statistical associations for the subsequent choice of an estimation strategy. We have argued here that identification assumptions that can be stated in DAGs apply to a wider set of sociological problems than applied researchers may think (including missing data problems for purely descriptive inferences). We refer readers to other pedagogical sources for an introduction to identification using DAGs (Morgan and Winship, 2015; Pearl and Mackenzie, 2018).

2.3 Estimation: Learn the Empirical Estimand from Data

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

For example, identification might lead to an empirical estimand θ that is the sample-average difference between the expected outcome among those treated $D = 1$ and among

those untreated $D = 0$ conditional on pre-treatment covariates $\vec{X}$.

$$\begin{array}{c} \text{Mean over} \\ \text{entire sample} \\ \searrow \end{array} \qquad \begin{array}{c} \text{Expected outcome among cases} \\ \text{with the covariate values } \vec{x}_i \text{ of unit } i \\ \text{who are factually treated } (D = 1) \\ \swarrow \end{array}$$

$$\begin{aligned} \theta = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \mathbf{E}(Y \mid \vec{X} = \vec{x}_i, D = 1) \\ - \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \mathbf{E}(Y \mid \vec{X} = \vec{x}_i, D = 0) \end{aligned} \tag{2.5}$$

$$\begin{array}{c} \nearrow \\ \text{Mean over} \\ \text{entire sample} \end{array} \qquad \begin{array}{c} \nwarrow \\ \text{Expected outcome among cases} \\ \text{with the covariate values } \vec{x}_i \text{ of unit } i \\ \text{who are factually untreated } (D = 0) \end{array}$$

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

$$\begin{array}{c} \text{Mean over} \\ \text{entire sample} \\ \searrow \end{array} \qquad \begin{array}{c} \text{Regression prediction} \\ \text{at observed covariate values } \vec{x}_i \\ \text{with treatment set to } D = 1 \\ \swarrow \end{array}$$

$$\begin{aligned} \hat{\theta} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \hat{\mathbf{E}}(Y \mid \vec{X} = \vec{x}_i, D = 1) \\ - \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \hat{\mathbf{E}}(Y \mid \vec{X} = \vec{x}_i, D = 0) \end{aligned} \tag{2.6}$$

$$\begin{array}{c} \nearrow \\ \text{Mean over} \\ \text{entire sample} \end{array} \qquad \begin{array}{c} \nwarrow \\ \text{Regression prediction} \\ \text{at observed covariate values } \vec{x}_i \\ \text{with treatment set to } D = 0 \end{array}$$

Estimation is tightly linked with prediction. Suppose we wanted to predict the outcome value Y for random units sampled from the subpopulation with covariates $\vec{X} = \vec{x}_i$ and treatment $D = 1$, and we fit our models on a different set of units sampled from the same population (e.g. we are not forecasting the future). For example, we might observe employment for a sample and want to predict whether a new person sampled from the same population who is 43 years old is employed. The prediction that would be as close as possible to the true values (in terms of expected squared error) is the conditional mean in the population, $\mathbf{E}(Y \mid \vec{X} = \vec{x}_i, D = 1)$. One could therefore select the estimator $\hat{\mathbf{E}}(Y \mid \vec{X} = \vec{x}_i, D = 1)$ out

of many candidate estimators by selecting the one that empirically minimizes out-of-sample mean squared prediction error. A researcher could consider the empirical mean among those observed with those covariates, the prediction of a regression model with or without interactions and squared terms, or some machine learning tool. Rather than arguing among these model specifications conceptually, we could decide among them empirically: the one that best estimates $\mathbf{E}(Y \mid \vec{X} = \vec{x}_i, D = 1)$ is the one that minimizes expected squared prediction error in out-of-sample cases. That procedure provides an empirical basis to adjudicate choices about functional forms.

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

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

might be small. Social science theory could suggest some amount of smoothness: we could expect employment among 43-year-olds to be similar to the employment of those who are 42 and 44. We might share information across these covariate values by averaging over all those ages 42–44, thus producing a slightly more precise estimator by drawing on our assumption of smoothness.

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

2.3.1 Concrete Estimation Example: The Family Gap in Pay

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

Our focus in this example is estimation. However, as argued throughout this paper, clear reasoning about estimation requires that we first define the theoretical and empirical esti-

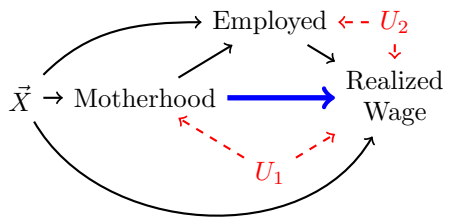


Fig. 4. Identification assumptions for the motherhood wage penalty. To identify the controlled direct effect of motherhood on the wages that would be realized under employment, one must assume that the covariates $\vec{X}$ are sufficient to block all confounding of motherhood and of employment. The red nodes U_1 and U_2 represent threats to identification.

mand. The original paper is not entirely clear; while the authors deploy "causal estimation techniques" (Pal and Waldfogel 2016:108), they also define the target quantity in a way that seems to appeal to a descriptive disparity between two populations, as "the differential in hourly wages between women with children and women without children," (Pal and Waldfogel 2016:104). We take the goal to be causal. However, we cannot simply define the goal as the average causal effect of motherhood on the wages of mothers, because wages are not defined for those who are not employed (Pal and Waldfogel 2016:109–110 acknowledge this complexity in a footnote). In order to target a well-defined unit-specific quantity for every mother in the population, we take the theoretical estimand to be the controlled direct effect of motherhood on the wages women would realize if they were employed, averaged over the population of mothers (Line 1 in Figure 5). We take the empirical estimand to be the descriptive gap between employed mothers and non-mothers conditional on covariates (Line 2 in Figure 5). The theoretical and empirical estimands are equal under the assumptions presented in Fig 4, which we emphasize includes a very strong assumption of no mediator-outcome confounding. Appendix D discusses alternative ways to frame the problem. Our focus here is on estimating the empirical estimand by using regression to predict the unknown conditional expectations (Line 3 in Figure 5). This estimation strategy is known as the parametric g-formula in biostatistics (Hernán and Robins, 2020, Ch. 13) and the imputation estimator in econometrics (e.g. Hahn 1998:321).

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

The imputation estimator unlocks new tools: the same procedure holds regardless of whether the algorithm used to predict the outcome variable is OLS regression, logistic regression for a binary outcome, or a machine learning strategy. In the OLS case, it simplifies back to a familiar result: the estimated treatment effect is the coefficient $\hat{\beta}$ on motherhood. However, that simplification is only possible under the (doubtful) no-interactions assumption that the treatment effect is the same value $\hat{\beta}$ at the covariate value $\vec{x}_i$ of every unit i. In other words, a coefficient $\hat{\beta}$ is only consistent for the average effect under special cases, such as when the effect is the same in every subgroup.[3] By motivating the estimation procedure as a tool for predicting conditional means rather than estimating coefficients, researchers open the door to alternative strategies that do not rely on the implausible assumption of linearity with no interactions.

In this example, we can actually conduct estimation without *any* assumptions about the functional form. The result holds even when we share no information across observations, imputing the expected wage within a covariate cell by the mean wage of those observed with

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

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

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

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\Big(\underbrace{Y_i(\texttt{Mother, Employed})}_{\text{if she were an employed mother}} \;\underset{\text{versus}}{-}\; \underbrace{Y_i(\texttt{Non-mother, Employed})}_{\text{if she were an employed non-mother}}\Big)$$

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

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

$$\theta = \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{\mathbf{E}\left(Y \,\middle|\, \begin{array}{lcl}\text{Motherhood} & = & \texttt{Mother,}\\ \text{Employment} & = & \texttt{Employed,}\\ \text{Covariates } \vec{X} & = & \texttt{Observed } \vec{x}_i\end{array}\right)}_{\text{actually are mothers}} \;\underset{\text{versus}}{-}\; \underbrace{\mathbf{E}\left(Y \,\middle|\, \begin{array}{lcl}\text{Motherhood} & = & \texttt{Non-mother,}\\ \text{Employment} & = & \texttt{Employed,}\\ \text{Covariates } \vec{X} & = & \texttt{Observed } \vec{x}_i\end{array}\right)}_{\text{actually are not mothers}}\right)$$

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

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

$$\underset{\substack{\uparrow\\ \text{estimate of}\\ \text{the estimand}}}{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{\hat{\mathbf{E}}\left(Y \,\middle|\, \begin{array}{lcl}\text{Motherhood} & = & \texttt{Mother,}\\ \text{Employment} & = & \texttt{Employed,}\\ \text{Covariates } \vec{X} & = & \texttt{Observed } \vec{x}_i\end{array}\right)}_{\substack{\uparrow\\ \text{estimated } \hat{Y}_i(\text{Mother})}} \;\underset{\text{versus}}{-}\; \underbrace{\hat{\mathbf{E}}\left(Y \,\middle|\, \begin{array}{lcl}\text{Motherhood} & = & \texttt{Non-mother,}\\ \text{Employment} & = & \texttt{Employed,}\\ \text{Covariates } \vec{X} & = & \texttt{Observed } \vec{x}_i\end{array}\right)}_{\substack{\uparrow\\ \text{estimated } \hat{Y}_i(\text{Non-mother})}}\right)$$

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

If that regression is **ordinary least squares**

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

$$\hat{\mathbf{E}}\left(Y \,\middle|\, \begin{array}{lcl}\text{Motherhood,}\\ \text{Employment} & = & \texttt{Employed,}\\ \text{Covariates } \vec{X} & = & \texttt{Observed } \vec{x}_i\end{array}\right)$$

$$= \begin{cases} \hat{\alpha} + \vec{x}_i'\hat{\vec{\gamma}} & \text{if Motherhood} = \texttt{Non-Mother}\\ \hat{\alpha} + \hat{\beta} + \vec{x}_i'\hat{\vec{\gamma}} & \text{if Motherhood} = \texttt{Mother}\end{cases}$$

By parametric approximation; Intercept; Coefficient on motherhood; Coefficients on other covariates

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left(\overbrace{\left(\hat{\alpha} + \hat{\beta} + \vec{x}_i'\hat{\vec{\gamma}}\right)}^{\hat{Y}_i(\text{Mother})} - \overbrace{\left(\hat{\alpha} + \vec{x}_i'\hat{\vec{\gamma}}\right)}^{\hat{Y}_i(\text{Non-mother})}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(\hat{\alpha} + \vec{x}_i'\hat{\vec{\gamma}} - \hat{\alpha} - \vec{x}_i'\hat{\vec{\gamma}}\right)}_{\substack{\text{Cancels because model}\\ \text{assumes no interactions}}} + \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}$$

$$= \hat{\beta} \quad \leftarrow \text{coefficient on motherhood}$$

If that regression is **anything else**

Does not simplify
(because the model is realistically complex)

$$\hat{\mathbf{E}}\left(Y \,\middle|\, \begin{array}{lcl}\text{Motherhood,}\\ \text{Employment} & = & \texttt{Employed,}\\ \text{Covariates } \vec{X} & = & \texttt{Observed } \vec{x}_i\end{array}\right)$$

$$= \hat{f}(\text{Motherhood}, \texttt{Employed}, \texttt{Observed } \vec{x}_i)$$

where $\hat{f}()$ is a prediction function.

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

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

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

Fig. 5. Estimands unlock new estimation tools. Once we state the estimand, we can use a predictive algorithm to impute unknown conditional means. It could be a linear model (e.g. OLS) or a nonlinear model (e.g. logit, random forest). The result equals a coefficient for linear models only. The empirical estimand guides the use of a realistically complex functional form.

exactly that set of covariates (far left of Fig. 6). So why would one assume a functional form, like a parametric model where the estimand is estimated by a coefficient (far right of Fig. 6)? In this case, the OLS model is clearly misspecified: it assumes that the family gap in pay does not vary by age (see lower right panel of Fig. 6), despite evidence in the other panels that the gap is larger in magnitude at younger ages. Yet one might prefer the OLS coefficient if the sample size were very small so that the stratification estimator at the left was infeasible or produced extremely uncertain estimates. Stating the estimand therefore does not preclude the use of a regression model as an approximation; rather, it provides a precise statement of the research goal so that we can begin to reason about the best empirical approximation. In a large sample, we may often be able to estimate by more credible assumptions than parametric models.

Using the stratification estimator weakens the *estimation* assumptions but cannot weaken the *identification* assumptions. For example, no estimator will get us out of making assumptions about unobserved confounding that affects both employment and wages (U_2 in Figure 4); those assumptions are out of scope for an algorithm because no algorithm can see the non-existent wages of the non-employed. While the choice of the best estimation method could be made using the data, we need subject matter knowledge to assess whether the identification assumptions are plausible.

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

The purpose of an estimation strategy is to estimate the empirical estimand. We would like valid procedures for statistical inference that produce, for example, a 95% confidence interval that actually would contain the empirical estimand in 95% of hypothetical samples. A valid interval is elusive in both parametric models and machine learning approaches. Parametric models (e.g. OLS) come with readily-available confidence intervals for the coefficient of interest. Those intervals would provide the expected coverage for the coefficient that would

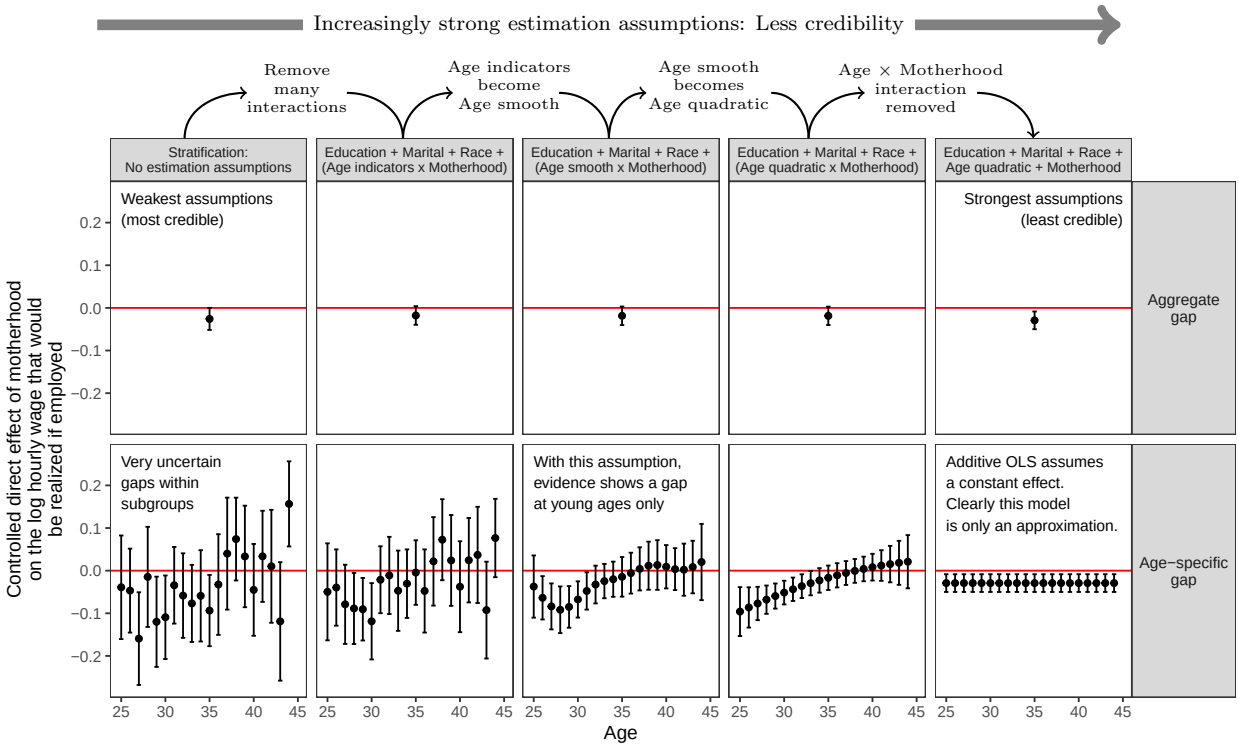


Fig. 6. A series of estimation strategies (columns) for two estimands (rows). Each estimand is the gap in log hourly wages between mothers and childless women, conditional on age, education, marital status, and race and aggregated over the covariate distribution of mothers. Estimands differ by aggregating over ages (top row) or not (bottom row). Estimation strategies range from weakest assumptions (left) to strongest assumptions (right). In the notation of the top titles, terms such as (Age indicators × Motherhood) represent an interaction and its lower-order terms. Provided that samples sizes are large enough to yield estimates that are sufficiently precise, one would prefer the estimation strategies to the left because they are more credible. Machine learning approaches such as the Generalized Additive Model (center column, Wood 2017) represent a middle ground between parametric models (OLS, far right) and nonparametric approaches (stratification, far left). Some findings, such as the population average gap (top row), are relatively invariant to the estimation strategy and can be defended under minimal estimation assumptions (far left). Other findings, such as the age-specific gap (bottom row), require modeling assumptions to achieve adequate precision. We suggest that the tendency to define estimands by a regression coefficient has prevented social scientists from recognizing setting when inference can proceed from more minimal estimation assumptions (at left). Instead of beginning from the right and moving left, we propose that researchers default to the left side and move right, motivating each choice to add an assumption. 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. All analyses make a common support restriction to the 98% of observations i such that both employed mothers and employed non-mothers are observed within the covariate stratum with $\vec{X} = \vec{x}_i$. Error bars are 95% confidence intervals calculated using replicate weights (Appendix D).

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

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

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

shadows and onto the page of the research paper, thereby facilitating arguments about these difficult issues.

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

2.4 Summary of Research Framework

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

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

Our contention is that greater attention to estimands could revolutionize substantive claims and reorient methodological guidance. A necessary condition for this argument is that current quantitative practice in sociology does not already explicitly or implicitly specify the theoretical and empirical estimands. To investigate this, we review the 2018 volume of the

American Sociological Review and show that we cannot consistently determine the theoretical estimand. We then turn to what it would mean for the field to reorient methodological choices around our framework. Unlike new statistical adjustments, changing the theoretical estimand can set the research on a completely different path, making it difficult to produce general statements about what would happen. Instead, we demonstrate in two specific examples: a descriptive example about the gender gap in college completion and a causal example about the effect of paternal incarceration on maternal depression.

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

Fig. 7 summarizes our review of all 32 articles using quantitative data in the 2018 volume of the *American Sociological Review*. The goal of this review was to assess whether our proposed framework merely introduces new terminology for existing practices: can we already translate standard summaries of quantitative analyses into unambiguous theoretical estimands involving unit-specific quantities aggregated over target populations even if they are not stated explicitly? Because the theoretical estimand links statistical analyses to theory, we considered not only the procedures applied to the data but also how the authors interpreted the procedures and results. Two of us read each paper and iterated to come to a joint assessment. Our determinations on each paper are summarized in Appendix H. For zero papers were we completely certain of both the unit-specific quantity and the target population. The fact that past research does not fit into our framework is unsurprising: we had not yet proposed this framework. Yet the conflicts between standard practice and our framework are nonetheless troubling: when there is disagreement about what the research goal is, it is difficult to adjudicate downstream debates about identification and estimation.

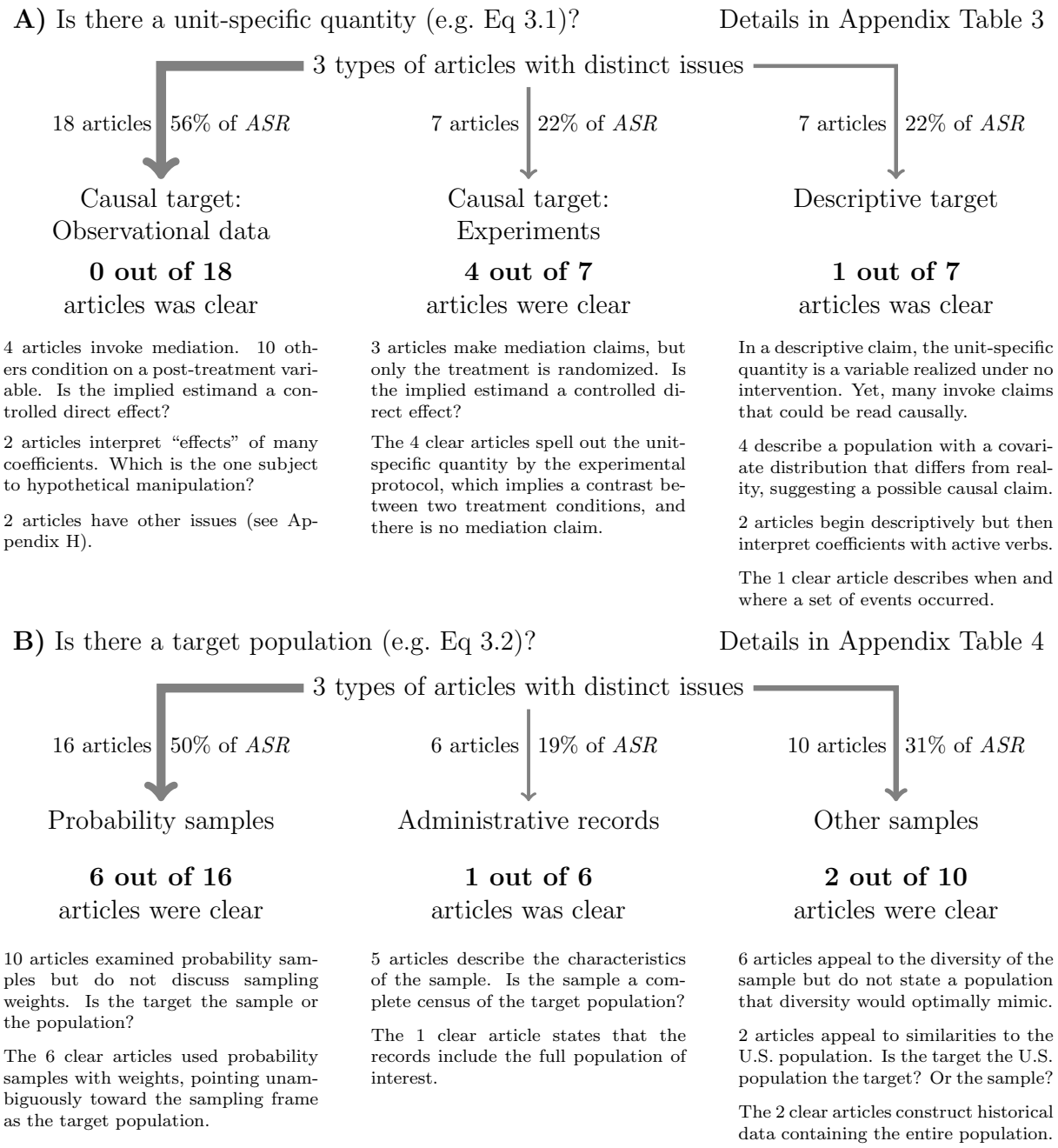


Fig. 7. Our methodological framework differs from standard practice. In our framework, all estimands are functions of actual outcomes Y_i (for descriptive estimands) or potential outcomes $Y_i(d)$ (for causal estimands), aggregated over well-defined populations of units indexed by i. Zero articles in the 2018 volume of *ASR* wrote the estimand this way. Some articles used sufficiently precise language that either the unit-specific quantity or the target population could be inferred unambiguously, rendering mathematical formalism superfluous. However, no article used language that was sufficiently precise for us to infer both the unit-specific quantity and the target population without some ambiguity. Each article is categorized in each panel above by the single issue we considered to be most apparent. Details are in Appendix H.

3.1.1 Unit-specific quantity

Our framework advocates the statement of unit-specific quantities as either realized random variables (for descriptive goals) or random variables that would be realized if one or more treatments were fixed to values they would not have otherwise taken (for causal goals). In our framework, every unit-specific quantity involves components like those in Eq. 3.1.

$$\begin{array}{c} \text{Unit-specific quantity:} \\ Y_i \quad \text{or} \quad Y_i(t) \\ \nearrow \qquad\qquad \nwarrow \\ \begin{array}{c}\textbf{Descriptive}\\ \text{Outcome as}\\ \text{it factually}\\ \text{exists}\end{array} \qquad \begin{array}{c}\textbf{Causal}\\ \text{Outcome if}\\ \text{assigned to}\\ \text{treatment value } t\end{array} \end{array} \tag{3.1}$$

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

We can confidently state the unit-specific quantity in the form of Eq. 3.1 in only 5 out of 32 papers (16%). Ambiguities in this quantity differ across three categories of papers. In studies drawing causal claims from observational data (56% of *ASR*), zero articles provide enough detail for us to be entirely confident of the intended intervention. Most of this category (78%) conducts an analysis that conditions on a post-treatment variable, often targeting a regression coefficient net of this variable. Four articles explicitly mention mediation; ten do not explicitly discuss mediation but nonetheless condition on a post-treatment variable. For all 14, we are unsure which of the many possible mediation estimands (Sec. 3.3) is the object of inquiry. Experiments (22% of *ASR*) involve a well-defined causal contrast for the treatment effect as specified by the experimental protocol. However, mediation claims with non-randomized mediators (appearing in 43% of such studies) are subject to the same issues common among observational studies (discussed in Sec. 3.3).

Descriptive studies constitute a minority of articles (22% of *ASR*). However, 6 of these 7 studies conduct at least one analysis reaching beyond pure description to claims that we consider to be at least implicitly causal. For example, Ciocca Eller and DiPrete (2018:1187) do not discuss causal effects or identification assumptions and yet examine disparities in college completion after "counterfactually shifting the dropout risk distribution of entering black students so that it more closely resembles the distribution for white students. In our framework, there are two types of claims: descriptive claims about unit-specific quantities in an observed population and causal claims about unit-specific quantities that would be realized if each individual were exposed to a hypothetical intervention. Both types of claims are important. Descriptive claims might include a comparison of rates between two groups—e.g., college completion rates for Black and White students. Causal claims involve a hypothetical causal intervention and identification assumptions. What does not fit in our framework is the middle ground: claims about what would happen under some `condition` (e.g., if Black students' dropout risk was similar to that of White students) which present a regression prediction but do not make causal assumptions explicit. The middle-ground claims only tell us how our model-specific predictions would change if we alter the input for `condition` in the model. Crucially, without identification assumptions and a hypothetical causal intervention, these are counterfactuals *of the model*, but they need not correspond to the effect of `condition` being realized in the world. Because the estimand does not describe the world as it is, or the world as it might be under a clearly stated intervention, we do not consider non-causal counterfactual estimands to provide a compelling link between the model and the theory.[4]

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

3.1.2 Target Population

In our framework, a theoretical estimand involves a precise statement of the target population about whom claims are made. A target population is a set of existing units (indexed by i) such that the theoretical estimand can be defined as some aggregation over that population, such as the average in Eq. 3.2.

$$\underset{\substack{\text{Mean over}\\ \text{some specified}\\ \text{population}}}{\frac{1}{n}\sum_{i=1}^{n}}\underset{\substack{\text{Of a quantity}\\ \text{defined for}\\ \text{each unit}}}{\begin{pmatrix}\text{Unit-Specific}\\ \text{Quantity}\end{pmatrix}_i} \tag{3.2}$$

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

We can confidently state the target population in only 9 out of 32 papers (28%). Half of *ASR* articles draw on probability samples, but 62% of those articles do not discuss how (if at all) survey weights are incorporated into the main analyses. It is then ambiguous whether the target population is the sample, the sampling frame, or some broader population. For instance, Liu (2018) uses data from Framingham, Massachusetts. Is this particular town of theoretical interest, or is the hope that it is informative about a broader population? With administrative records (19% of *ASR*), authors are often remarkably clear about who is in the records, but only 1 out of 6 articles (Font et al., 2018) explicitly states whether the set of units in the records is the entire population about which they seek to draw inference. Finally, other samples such as Amazon Mechanical Turk and datasets constructed by the author appear in 31% of *ASR*. Most of these studies defend the chosen sample on the grounds of its diversity. Yet diversity is only helpful if that diversity matches the diversity of some target population of interest, enabling weighted inferences to be valid for the target. What target population should our diverse sample mimic? It is difficult to assess things like confidence intervals and

significance tests (typically based on the idea of sampling from a target population) when authors have described the sample but left the target population unstated.

3.1.3 Summary of the review of *ASR*

As readers, we often ask "what is the estimand?" and cannot reverse-engineer an answer from published articles. In its purest form, our framework proposes mathematical precision to resolve these ambiguities. The causal contrast becomes clear when estimands are written as functions of unit-specific quantities: either actual outcomes Y_i for descriptive estimands or potential outcomes $Y_i(d)$ for causal estimands. Explicit aggregation over a well-defined set of units indexed by i leaves no ambiguity about the target population. But mathematical formalization is not absolutely necessary; we would be happy if authors stated the estimand in words with sufficient precision that we could translate the description to a particular estimand. This might be possible through description of an experimental protocol or a clear hypothetical intervention in an observational study, paired with a precise statement of the target population. What is troubling is our inability to unambiguously translate the description provided by authors into a precise research goal. As a result, it is difficult to know what was learned or to reason about methodological procedures by which it could be learned better. Productive scientific exchange is difficult when articles do not make clear what question was answered.

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

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

Buchmann and DiPrete (2006) summarize a gender reversal: while men historically com-

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

Suppose the researchers instead summarized a series of descriptive estimands: the probability of college completion as a function of gender, birth cohort, and parental characteristics, stated without any appeal to regression coefficients.

$$\tau(g,p,c) = \mathrm{P}\left(\overset{\substack{\text{Probability}\\ \text{of college}\\ \text{completion } Y}}{Y = 1} \;\middle|\; \begin{array}{l} \overset{\substack{\text{Among}\\ \text{those}\\ \text{with}}}{G = g} \\ C = c \\ P = p \end{array} \right) \tag{3.3}$$

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

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

into question the equality of τ and θ.

$$\theta(g,p,c) = \mathrm{P}\left(\overset{\substack{\text{Probability}\\\text{of college}\\\text{completion } Y}}{Y=1} \;\middle|\; \overset{\substack{\text{Among}\\\text{those}\\\text{with}}}{\begin{matrix} G=g \\ C=c \\ P=p \\ R=1 \end{matrix}}\right) \tag{3.4}$$

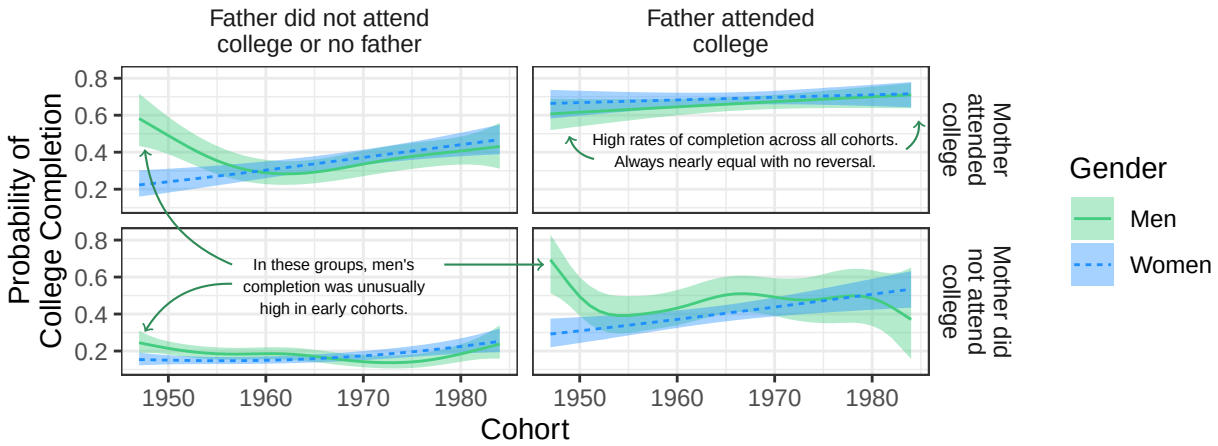
Gender g, birth cohort c, and parent characteristics p

and who are **alive and willing to respond** ($R=1$)

Fig. 8 summarizes these descriptive estimands with gender g represented by colors, parent characteristics p represented by the grid of plots, and cohorts c represented by the x-axis. We extend the series to include all data now available. Instead of dichotomizing birth cohorts at 1966 (the original specification), we use generalized additive models (GAMs, Wood 2017) to estimate smooth but flexible curves. In one respect, our result reproduces the original authors' claims: the disparity reversed the most among those for whom at least one parent did not attend college (panels other than the top right). This descriptive statement involves no appeal to effects defined as regression parameters: it is simply a description within subgroups.

Flexible descriptions can spark theoretical questions that would otherwise remain buried by parametric specifications. The original authors focus on why the gap closed, but our results raise a different question: why was college completion so common among men born in the early 1950s whose father did not attend college? One candidate explanation is that these are the same cohorts in which men were conscripted to serve in the Vietnam War (especially birth years 1950–1952, Angrist and Chen 2011). The Vietnam War could have produced especially high college completion rates for this cohort because veterans received scholarships upon their return under the GI Bill. The high completion rate could also arise in part from an identification problem: the theoretical estimand involves everyone born in

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



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

Fig. 8. Descriptive estimands are worthwhile goals. The language of "effects" common in multiple regression models can produce confusion. The figure is purely descriptive, presenting estimates of the mean within subgroups with no control variables. The only model serves to smooth over cohorts (Wood, 2017). The evidence base in the figure is analogous to the logistic regression model from Buchmann and DiPrete (2006, Table 2 Model 1): the predictors of that model define the subgroups in this plot (gender, cohort, and parent characteristics). It is clear that these subgroups produce an interesting description. It is not clear that this description, or the analogous logit model, allows one to attribute the reversal to any particular "effect." We propose that descriptive estimands—means within subgroups—can specify the research goal while avoiding the tendency to state results in vaguely-defined effects and attributions.

1950, but those killed in the war were not around decades later to complete the GSS survey. Perhaps the men who would not have completed college were disproportionately drafted and killed in the war, contributing to a gap between the theoretical and empirical estimand.

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

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

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

3.3.1 Example: Effects of Paternal Incarceration

Wildeman et al. (2012) estimates how incarceration of a child's father has collateral consequences across the family by increasing the probability that the child's mother will be depressed. Using a design that adjusts for observed sources of confounding, the authors report that paternal incarceration increases the log odds of maternal depression by 0.32. Our replication recovers a similar estimate of 0.28 (details in Appendix F).[5] We convert the model into an estimate of a nonparametric estimand: paternal incarceration increases the

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

probability of maternal depression by 4 percentage points (95% CI: -0.02, 0.10) on average. This illustrates a first advantage of an approach centered on estimands: it becomes clear that the effect size is small, although important nonetheless. The original authors proceed to a series of mediation claims, which we use to make broader points about causal mediation.

3.3.2 Direct Effects: Defining a Mediation Estimand

We focus on one mediator from Wildeman et al. (2012): paternal incarceration may cause the mother to reside with a new partner, which could in turn affect her depression. Claims about this mechanism would invoke counterfactuals for both the treatment (what if the father had not been incarcerated?) and the mediator (what if the mother had not repartnered?). The potential outcomes, $Y_i(d, m)$, are thus functions of both the treatment value d and the mediator value m. This definition of potential outcomes allows one to target many mediation estimands defined as contrasts over the outcomes that would realized at different values of d and m (Imai et al., 2011; Pearl, 2001). We focus on one particular set of mediation estimands—controlled direct effects—which can be estimated under more credible assumptions than other mediation estimands because they can be identified even in the presence of treatment-induced mediator-outcome confounding (Acharya et al., 2016). A controlled direct effect $\tau(m)$ compares the outcome under two different treatment values that would persist if we intervened to fix the treatment to a particular value m.

$$\underbrace{\tau(m)}_{\text{Controlled direct effect}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{Mean over sample}} \Big(\overbrace{Y_i(1, \underbrace{m}_{\text{if her repartnering was set to the value } m})}^{\text{Whether mother } i \text{ would be depressed if father } i \text{ was incarcerated}} \overset{\text{vs}}{-} \overbrace{Y_i(0, m)}^{\text{if father } i \text{ was } \textbf{not} \text{ incarcerated}} \Big) \tag{3.5}$$

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

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

Because mediation invokes counterfactual assignments of both the treatment and the mediator, it is necessary to adjust for variables that confound the assignment of both of these variables. Sociologists frequently discuss confounding of the treatment but almost never discuss confounding of the mediator. Wildeman et al. (2012, p. 222) "adjust for preexisting differences between mothers who have and have not experienced the incarceration of their child's father," but they do not say anything to address concerns that an unobserved variable U might affect the mediator (maternal repartnering) and the outcome (maternal depression, Fig. 9). This threat persists even in randomized experiments where the treatment (but not the mediator) is randomized (see Appendix Table 8 for examples from *ASR*). The assumption of no treatment-outcome or mediator-outcome confounding is often doubtful. We will nonetheless proceed under this assumption in order to further illustrate complexities of mediation that arise (at least implicitly) in many studies.

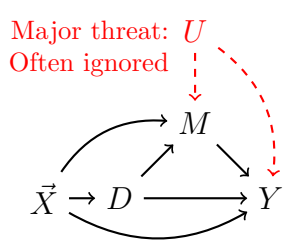


Fig. 9. Causal structure for mediation estimands. These require identifying the effect of the treatment D and the effect of the mediator M on the outcome Y. The unobserved mediator-outcome confounder, U, is a threat to inference.

3.3.4 Estimating Direct Effects

Due to our identification assumptions, direct effects can be estimated by the imputation estimator (Figure 5): fit a statistical model for Y as a function of $\{\vec{X}, D, M\}$, plug in the new values d and m for the variables D and M, predict $\hat{Y}_i(d, m)$ for each unit i, and average over units as specified by the estimand.

$$\underbrace{\hat{\tau}(m)}_{\substack{\text{Estimated}\\ \text{controlled}\\ \text{direct}\\ \text{effect}}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\substack{\text{Mean}\\ \text{over}\\ \text{sample}}} \Big(\overbrace{\hat{Y}_i(1, m)}^{\substack{\text{Predicted probability of depression for mother } i\\ \text{with father incarceration}\\ \text{set to the value 1}}} \overset{\text{vs}}{-} \overbrace{\hat{Y}_i(0, m)}^{\substack{\text{with father incarceration}\\ \text{set to the value 0}}} \Big) \tag{3.6}$$

$$\underbrace{\hat{Y}_i(1, m)}_{\text{and with her repartnering set to the value } m}$$

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

3.3.5 Results

The four plots in the upper left of Fig. 10 correspond to average potential outcomes: the proportion of mothers who would be depressed under each possible value of the treatment (father incarceration) and the mediator (mother having a new partner). The direct effects (right column) are the difference in the average potential outcomes across treatment conditions within a mediator condition. The direct effect to which this estimand corresponds is subtle. Paternal incarceration would *reduce* maternal depression by 1 percentage point under a subsequent intervention to repartner any mother who would not otherwise repartner. This is a direct effect because the intervention would remove the causal pathway through repartnering. Meanwhile, paternal incarceration would *increase* maternal depression by 5 percentage points under a subsequent intervention to prevent any mother from repartnering

(middle right). The "direct effect" is an ambiguous estimand until we state the value to which the mediator is fixed.

3.3.6 Mediation: Concluding Remarks

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

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

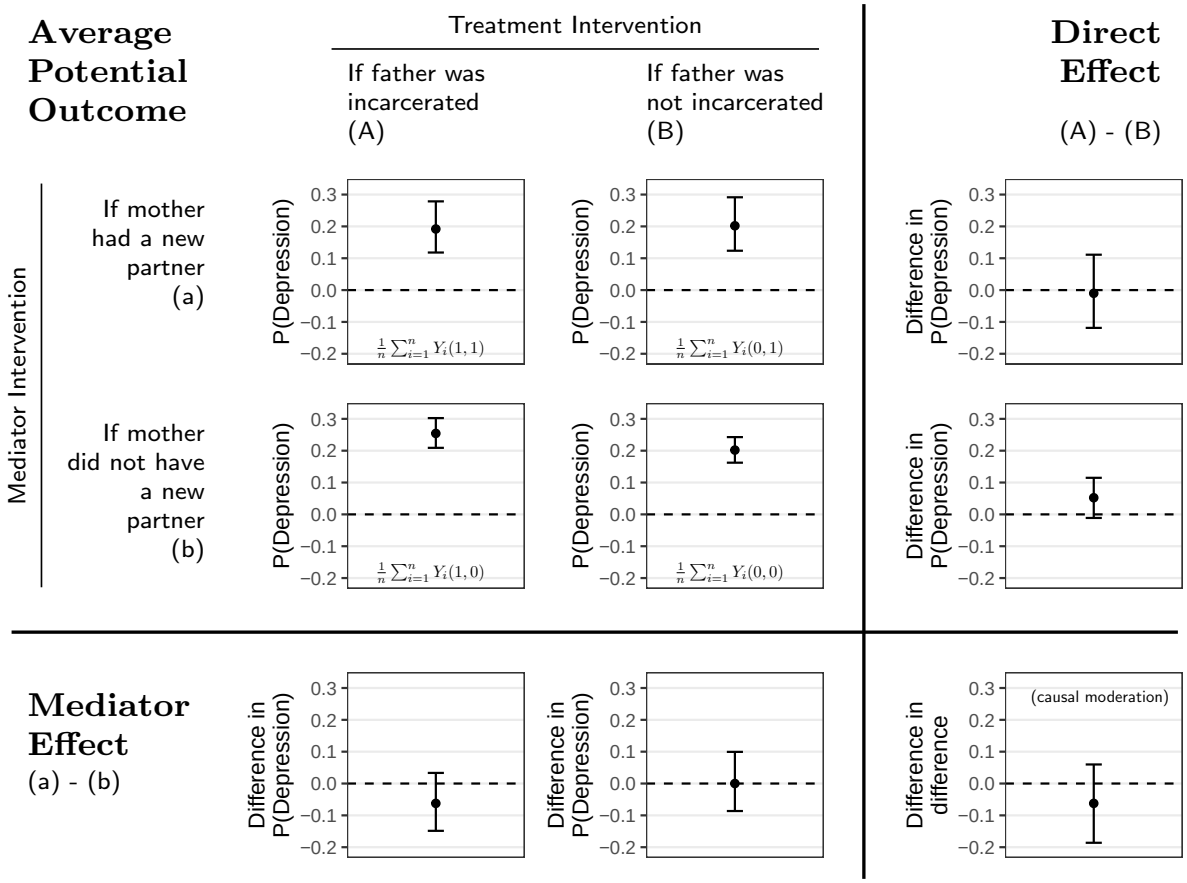


Fig. 10. Controlled direct effects involve an intervention to the treatment and the mediator. Figure explores the degree to which the effect of paternal incarceration on maternal depression operates through the mother residing with a new romantic partner. In a possibly counterfactual world in which a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

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

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

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

4.1 For Analysts: Estimands Ground Methodological Choices

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

For example, researchers estimating binary outcome models are often confused about whether interaction terms are needed and how to interpret them if so. The extensive methodological debate on this topic frames these issues within the terms of logit and probit models (Berry et al., 2010; Nagler, 1991; Rainey, 2016). Yet the problem is more fundamental: the researcher must begin by stating what they mean by the term "interaction." Fig. 11 illustrates that a treatment may multiply the probability of an outcome by a fixed amount (no interaction) while increasing that probability by an additive amount that depends on a pre-treatment covariate (interaction). Whether or not interaction is present depends on the estimand. No argument about a model can resolve this question; it is fundamentally a question about the research goal itself.

Estimands likewise offer guidance about which variables to include in a model. Mood (2010:67) warns that problems arise in logistic regression because "we can seldom include in a model all variables that affect an outcome." Breen et al. (2018:47) write that logit coefficient estimates "are lower bounds to the true or underlying coefficients unless all relevant covariates are included." But are these "true" coefficients even a well-defined estimand? What would it mean to include all relevant covariates? Our framework instead offers a straightforward answer: you must include the covariates needed to identify the estimand. For descriptive estimands, that might only be the predictors of interest.

Presentation of results can also be guided by estimands. If the estimand is a difference

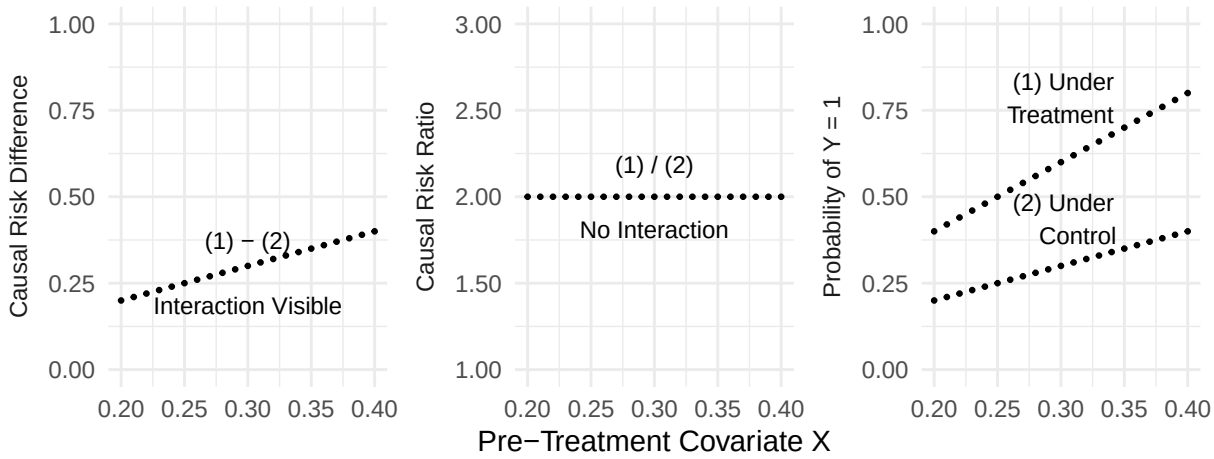


Fig. 11. Illustration: The presence or absence of interaction may depend on whether the estimand is a multiplicative or an additive effect. In this example, the treatment multiplies the probability of $Y = 1$ by a constant value (2.00) at all values of X. If the baseline outcome is a function of X, however, then a constant multiplicative effect implies an additive effect that is a function of X: the probability of $Y = 1$ in this example increases by a greater amount when X is greater. Whether interaction is present depends on the estimand. No amount of literature about the proper interpretation of coefficients in binary outcome models can help a researcher as long as that researcher's goal of assessing the presence or absence of interaction remains underspecified. Appendix G provides simulation details.

in probabilites, then the researcher would naturally present predicted probabilities rather than reporting coefficient estimates (Breen et al., 2018; Mize, 2019; Mood, 2010). Questions about which model to use, what assumptions are required, and how to present results all become clearer once the research goal has been stated.

4.2 For Readers: Estimands Clarify the Author's Claim

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

Readers often like to see that results are 'robust'. In the 2018 volume of the *American*

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

In general, robustness checks provide useful information only in the context of a well-defined target and clarity about the alternatives to which we are evaluating robustness. In contrast, robustness checks as currently applied treat all specifications as equally valid. Young and Holsteen (2017) provides tools to automate this procedure. Yet when an unguided search for robustness is taken to its logical extreme with thousands of specifications, it is impossible to defend each individual specification. The resulting benchmark for methodological rigor devolves into a requirement that sociologists report only the results that survive a test of methodological invariance: they are the same even if we target several different estimands through several different estimation strategies. This is not a requirement of a credible claim. Conversations about robustness would be more productive if they centered on how each check resolves a specific point of uncertainty in the link between theory and evidence.

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

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

in studies B and C?

Distinguishing between differences in estimation strategies versus differences in estimands is essential in the case of replication. Questions of replication often focus on questions about statistical power and hypothesis testing. Yet a replication can also fail because it targets a different estimand from the original study, as when the replication uses a different pool of experimental subjects.[6] A replication focused on the same estimand as the original study provides evidence about statistical issues like false positives. A replication focused on a different estimand provides evidence about theoretical issues regarding the generality of the phenomenon across settings. Specificity about the estimand of each paper is key to the advancement of general theories of social life that produce results that replicate across many studies in many distinct settings. For the field as a whole to grow, it would help if each paper's contribution to knowledge is stated precisely.

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

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

A renewed focus on estimands will be important as sociology navigates a methodological landscape that is changing rapidly. A pivot to new sources of 'big data' creates an ever-greater need for clarity about the gap between the theoretical goal and selection issues

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

that constrain the data available (as in Section 2.2). As sociologists increasingly engage with predictive tasks (Salganik et al., 2020), estimands will clarify key distinctions among different types of prediction: for cases from the same data generating process as the training data (standard prediction), for the outcome that would be realized under an intervention (causal prediction), or for future events that have not yet occurred (forecasting). Each setting corresponds to a different theoretical estimand and requires a different set of identification assumptions. Estimands can also improve the use of inductive measurement strategies such as latent class analysis for surveys, methods for text as data, and unsupervised machine learning. Explicit statements of the estimands in these settings would both clarify what these procedures are learning from data and what evidence would be necessary to contradict the finding. They also highlight the importance of choosing an estimand of interest in one sample and then evaluating the estimand on a second sample (Egami et al., 2018). Broadly, estimands will be key to whatever methods quantitative sociology develops in the coming years.

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

purpose precise: what is the estimand?

References

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 S. H. Chen 2011. Schooling and the vietnam-era gi bill: Evidence from the draft lottery. *American Economic Journal: Applied Economics*, 3(2):96–118.

Angrist, J. D. and 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.

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

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

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.

Berry, W. D., J. H. DeMeritt, and J. Esarey 2010. Testing for interaction in binary logit and probit models: is a product term essential? *American Journal of Political Science*, 54(1):248–266.

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.

Blitzstein, J. K. and J. Hwang 2019. *Introduction to Probability*. CRC Press.

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.

Buchmann, C. and T. A. DiPrete 2006. The growing female advantage in college completion: The role of family background and academic achievement. *American Sociological Review*, 71(4):515–541.

Buja, A., L. Brown, R. Berk, E. George, E. Pitkin, M. Traskin, K. Zhang, L. Zhao, et al. 2019. Models as approximations I: consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544.

Chetty, R., N. Hendren, M. R. Jones, and S. R. Porter 2020. Race and economic opportunity in the united states: An intergenerational perspective. *The Quarterly Journal of Economics*, 135(2):711–783.

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

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.

Duncan, O. D. 1984. *Notes on Social Measurement: Historical and Critical*. Russell Sage Foundation.

Durlauf, S. N. and J. J. Heckman 2020. An empirical analysis of racial differences in police use of force: A comment. *Journal of Political Economy*, 128(10).

DAmour, A., P. Ding, A. Feller, L. Lei, and J. Sekhon 2020. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*.

Egami, N., C. J. Fong, J. Grimmer, M. E. Roberts, and B. M. Stewart 2018. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*.

Elwert, F. and C. Winship 2010. Effect heterogeneity and bias in main-effects-only regression models. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, Pp. 327–36.

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. 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.

Fryer, R. G. 2020. An empirical analysis of racial differences in police use of force: A response. *Journal of Political Economy*, 128(10).

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

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.

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.

Katz, J. N., G. King, and E. Rosenblatt 2020. Theoretical foundations and empirical evaluations of partisan fairness in district-based democracies. *American Political Science Review*, 114(1):164–178.

Keele, L., R. T. Stevenson, and F. Elwert 2020. The causal interpretation of estimated associations in regression models. *Political Science Research and Methods*, 8(1):1–13.

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

Knox, D., W. Lowe, and J. Mummolo 2020. Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637.

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

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.

Liu, H. 2018. Social and genetic pathways in multigenerational transmission of educational attainment. *American Sociological Review*, 83(2):278–304.

Lundberg, I. 2020. The gap-closing estimand: A causal approach to study interventions that close disparities across social categories. `https://doi.org/10.31235/osf.io/gx4y3`. SocArXiv.

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

Mize, T. D. 2019. Best practices for estimating, interpreting, and presenting nonlinear interaction effects. *Sociological Science*, 6:81–117.

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.

Nagler, J. 1991. The effect of registration laws and education on U.S. voter turnout. *The American*

Political Science Review, Pp. 1393–1405.

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. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth conference on Uncertainty in Artificial Intelligence*, Pp. 411–420.

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.

Preston, S., P. Heuveline, and M. Guillot 2000. *Demography: Measuring and Modeling Population Processes*. Malden, MA: Blackwell Publishers.

Raftery, A. E. 1995. Bayesian model selection in social research. *Sociological Methodology*, 25:111–164.

Rainey, C. 2016. Compression and conditional effects: A product term is essential when using logistic regression to test for interaction. *Political Science Research and Methods*, 4(3):621–639.

Salganik, M. J., I. Lundberg, A. T. Kindel, 108 others, and S. McLanahan 2020. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403.

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

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.

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

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.

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.

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

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.

Wildeman, C., J. Schnittker, and K. Turney 2012. Despair by association? the mental health of mothers with children by recently incarcerated fathers. *American Sociological Review*, 77(2):216–243.

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. and K. Holsteen 2017. Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods and Research*, 46(1):3–40.

ONLINE APPENDICES

What is Your Estimand?

Defining the Target Quantity Connects Statistical Evidence to Theory

Ian Lundberg, Rebecca Johnson, and Brandon Stewart

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 1. While Section 3.3 discussed mediation, it touched on only one type of mediation estimand (a controlled direct effect). This section compares that estimand to a different mediation estimand: the natural direct effect. This example also provides an additional example of a widely-read sociological study which implicitly involves a mediation claim.

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

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

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

C Target population details in Angrist and Evans (1998)

To calculate the target population size in Angrist and Evans (1998), we rely on the descriptive statements in the text. Between 37 and 53 % of mothers ages 21–35 have two or more children (Table 1 in Angrist and Evans 1998:453), and the proportion of women having a third child was between 6 and 7 percentage points greater among those whose first two children were of the same sex compared with those whose first two children were of different sexes (bottom row of Table 3 in Angrist and Evans 1998:457). The size of the target population is therefore at most only 4 % (0.53×0.07) of all mothers ages 21–35.

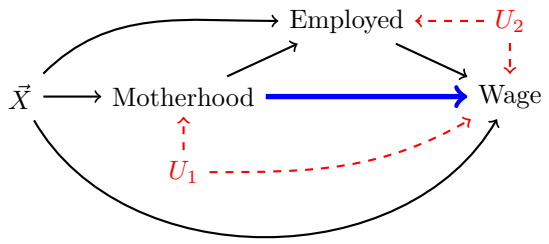
D Family gap in pay: Example details

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

Fig 12 Panel A presents the causal assumptions underlying the research question as it is presented in the main text. Because one cannot earn a wage unless one is employed, we assume that employment is a cause of wages. We focus on the direct effect of motherhood on the wages women would realize if they were employed (blue edge). Identifying this direct effect relies on two key assumptions. The first is that the covariates $\vec{X}$ (age, education, marital status, and race) are sufficient to block all backdoor paths between motherhood and wages. This assumption entails that there are no unobserved variables like U_1 that affect both motherhood and wages. The second assumption is that the same covariates $\vec{X}$ are sufficient to block all backdoor paths between employment and wages. This assumption entails that there are no unobserved variables like U_2 that affect both employment and wages. This latter assumption is often ignored in practice, but it is important because it is doubtful. The quality of one's resume, for instance, might both increase the probability of employment and increase the wage that one would realize if employed. The main text proceeds to estimation to illustrate that step of the process, but selection along variables like U_2 is a serious concern in any study for which the outcome variable (e.g. wage) is only defined for those taking a particular value of a mediator (e.g. employment).

Transparency about the assumed causal structure opens dialogues about alternative causal assumptions. For example, Elwert and Winship (2014:41–42) discuss the effect of motherhood on wages but assume that wages cause employment—the opposite of our assumption that employment causes wages. In that view, the outcome of interest is *offer wage*: the pay employers offer, after which individuals decide whether to accept the offer. In that framework, wages cause employment rather than the reverse. In the process of circulating drafts of this paper, we corresponded with Elwert and Winship and arrived at the possibility

A) Basic DAG illustrating key assumptions in the motherhood example



B) Detailed DAG with a more complex causal process

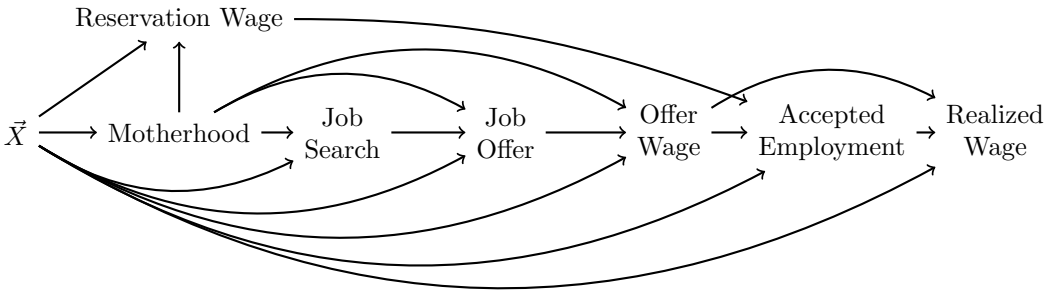


Fig. 12. DAGs showing causal assumptions for the motherhood wage penalty. Panel (A) shows the basic assumptions of our main analysis. We focus on the direct effect (blue edge) of motherhood on the wages that women would realize in a possibly counterfactual world in which they were employed. In this structure, we assume that employment causes wages: one first has to be employed in order to earn a wage. Identification requires that the covariates $\vec{X}$ (age, education, marital status, and race) block all backdoor paths between motherhood and wage (e.g. U_1 does not exist) and between employment and wages (e.g. U_2 does not exist). These assumptions are very strong; in particular, it seems likely that some unobserved variable U_2 affects both employment and wages. Panel (B) expands the DAG to a more detailed causal process: motherhood affects a woman's reservation wage (the lowest wage at which she would accept employment), her job search, whether she is offered a job, the wage she is offered, whether she accepts the offer, and the wage she realizes. In the assumptions of (B), the offer wage (an unobserved variable) confounds the relationship between accepting employment and the wage realized after negotiation, thereby contradicting the assumption of no U_2 in (A). The two sets of assumptions are therefore distinct. Broadly, a research framework centered on the estimand could lead the motherhood penalty literature to consider a more detailed set of questions about how motherhood affects all of these downstream labor market outcomes. We are grateful to Felix Elwert and Christopher Winship for suggesting the more detailed DAG.

that the true causal process may be a more complex structure as presented in Fig 12 Panel B. Motherhood affects whether women search for a job, whether they are offered a job, whether they accept that offer, and the wage they ultimately realize (possibly after negotiations). One way that motherhood shapes accepted employment may be by affecting women's reservation wages: the lowest wage level such that one would accept an offer of employment rather than remaining unemployed. This more complex set of causal assumptions illustrates how the literature on the motherhood wage penalty could ask many distinct research questions about effects on reservation wages, offer wages, and realized wages. Further, the process in Panel B illustrates another problem with the assumptions in Panel A: offer wages can play the role of U_2, by affecting employment and the wage one would actually realize in the job.

Our dialogue about the assumptions of Fig 12 Panels A and B illustrates the kind of scholarly communication that can happen in the framework we propose. If we had instead followed the standard approach and defined the goal as the coefficient on motherhood, it would have been much more difficult to have discussions about the causal process as the reader would be unable to discern which causal process we were working with. This is one of the main benefits that the field stands to gain from our approach. Some readers may agree with your causal assumptions and others may not, and making those assumptions explicit is a good step toward productive conversation about those issues. The main text maintains the simpler causal assumptions for the purpose of using this example to illustrate issues with estimation rather than identification.

Taking the goal to be the direct effect of motherhood on wages that would be realized under employment (Fig 12 Panel A), we next proceed to estimate that quantity. We draw data from the 2019 Annual Social and Economic Supplement (ASEC) of the March Current Population Survey, in the form supplied by the Integrated Public Use Microdata Series (IPUMS, Ruggles et al. 2019). We restrict the sample to women ages 25–44 (16,977 mothers, 8,183 non-mothers) who fall in the region of common support (16,659 mothers, 8,025 non-mothers, see next paragraph). To estimate the potential wage that would be realized if one

were an employed mother or an employed non-mother, we fit models on the subsample who worked for wages or salary in the last year (11,555 mothers, 6,277 non-mothers). To draw inference about the target population of mothers regardless of actual employment, we make predictions for all 16,659 mothers on the region of common support (discussed two paragraphs below). To construct the outcome (log hourly wage), we divide each individual's annual wage and salary income by the number of hours worked in the last year. To determine the number of hours worked, we multiply the report of the usual hours worked per week by the number of weeks worked in the last year. Because division can produce extreme values, we truncate hourly wages at the 1st and 99th percentile (unweighted). The outcome variable is the log of truncated hourly wages. We operationalize motherhood by whether the respondent reports any of their own children residing in the household. Following Pal and Waldfogel (2016), we include four other predictors: age, education (less than high school, high school, some college, college or more), marital status (married with spouse present, versus not), and race (white, black, other).

To calculate the proportion of observations in the region of common support, we separate the data into subgroups of age, education, marital status, and race. We note whether each subgroup is on the region of common support (at least one employed mother and one employed non-mother is observed) or not (no one is observed in at least one of these two conditions). We then aggregate over mothers and over non-mothers. The region of common support includes 98.1% of mothers and 98.1% of non-mothers. All analyses restrict to this region of common support. Common support is often a problem when the covariates have a very large number of values (DAmour et al., 2020). For example, if even one covariate is continuous (with a unique value for each observation), then common support will always be a problem and some additional assumptions will be needed.

To produce point estimates of the family gap in pay, we apply the parametric g-formula (Hernán and Robins, 2020, Ch. 13), an approach involving several steps. We first apply each algorithm (discussed in next paragraph) to learn the conditional mean of log wages

given all the predictors among employed women. Then, we predict the outcome for each mother (including those that are not employed) with her observed covariates, and with motherhood changed to zero and all other covariates left unchanged. For each mother, we difference the predictions to estimate the family gap in pay at her specific covariate set. Finally, we aggregate over all mothers by a weighted mean using the ASEC survey weights. To estimate the controlled direct effect of motherhood on the wages women would realize if they were employed, this procedure requires the identification assumptions that potential wages are independent of both motherhood and employment within subgroups of the observed covariates. Although this assumption is likely to be imperfect, we focus this example on issues of estimation and assume for the sake of illustration that the identification assumption holds.

The only difference across the estimators is the algorithm used to learn the conditional mean of log wages given predictors. These algorithms correspond to the columns of Figure 6. Column 1 shows results from a nonparametric stratification estimator which estimates the mean by the observed mean of individuals observed with each covariate set, weighted by the survey weights. Column 2 shows results from a linear regression of log wages on all predictors, with age included as a series of indicator variables as well as their interaction with motherhood. Column 3 shows results from a generalized additive model (GAM, Wood 2017) that adds an assumption that the association between age and log wages follows a smooth functional form, which is allowed to differ for mothers and non-mothers. The thin-plate spline applied follows the defaults of the `mgcv` package in R, with automatic smoothing penalty. Column 4 shows results from an OLS regression model that requires that the association between age and log wage follows a quadratic form interacted with motherhood. Column 5 shows results from an OLS regression model that maintains the quadratic form for age but removes the interaction between age and motherhood. In all of the algorithms, we apply the ASEC survey weights in order to prioritize fitting of the conditional expectation function in regions of the covariate space that are heavily weighted.

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

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

where $\hat{\theta}_r$ is each replicate-specific estimate for $r = 1, \ldots, 160$ 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 $\hat{\mathrm{V}}(\hat{\theta})$.

The specification with a quadratic for age interacted with motherhood achieves the best predictive performance. To arrive at this result, we sorted the sample by the ASEC survey weights and assigned observations to five folds systematically: the first set of five observations were assigned to folds 1–5, respectively, the second set of 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 predic-

tions for the held-out fold. After running through all five folds, we pooled all predictions and calculated the weighted mean squared error over the full sample. This procedure provides a valid estimate of out-of-sample predictive performance because each observation's prediction was estimated on a sample that excluded that observation. Because this entire procedure can be wrapped in a single function, we can also pass it through the same variance estimation procedure as the point estimates. The result suggests that the variance of the estimated mean squared errors is high, so that we are hesitant to conclude that the data strongly point toward one estimation approach over another, with the exception that nonparametric stratification has substantially worse predictive performance. Nonetheless, if one sought a purely data-driven approach to select only one estimator, the cross-validation approach would point toward the smooth age specification interacted with motherhood, with all other predictors entered additively.

E Buchmann and DiPrete (2006): Example details

Our replication of Buchmann and DiPrete (2006) focuses on one descriptive component of the paper: the trend across birth cohorts in college completion for men and women within subgroups of parents' education, which appears in Fig. 8 of this paper and corresponds to results presented in Table 2 of the original paper. Following the original authors, we analyze data from the General Social Survey. While they used data collected from 1972–2002, we extend through all data currently available (1972–2018). Following the original authors, we focus on white respondents ages 25–34 ($N = 10,601$). We restrict to the 1947–1984 birth cohorts so that every cohort can be observed at every age from 25–34 ($N = 8,679$). Because the original authors do not conduct multiple imputation, we restrict to observations without missing values on any of the variables ($N = 7,605$). Following the original authors, we define mother's education as a binary indicator of whether the mother attended college. Father's education is more complicated in the original authors' specification: their model includes

three categories (father attended college, father did not attend college, or there was no father) but for some interaction terms they combine fathers who did not attend college and those with no father into a single category. To avoid complexity, we use the binary variable definition throughout, categorizing fathers as either (1) attended college or (2) either did not attend college or are unknown.

We estimate the probability of college completion within subgroups of the covariates using a generalized additive logistic regression model (Wood, 2017), weighted by the GSS sampling weights. Our specification allows all interactions among gender, mother's education, father's education, and a smooth function of birth cohort. The appeal of this highly flexible specification is that in a very large sample it would converge toward estimating the proportion completing college within each subgroup (gender $\times$ father's education $\times$ mother's education $\times$ cohort) by the simple average of college completion within that subgroup. The smooth term for cohorts pools information from observations in nearby birth cohorts, but is much more flexible than a line or a binary indicator for the cohort exceeding a particular value (the original authors dichotomize at 1966). The values reported in Fig. 8 are point estimates and confidence intervals based on standard errors calculated by the `predict` function in the `mgcv` package in `R` (Wood, 2017).

F Wildeman et al. (2012): Example details

Our replication of Wildeman et al. (2012) is imperfect because replication code and the version of the data used by the original authors are unavailable. We are grateful to the authors for their correspondence with us, which made it possible for us to replicate the main results of the paper. The analyses use the Fragile Families and Child Wellbeing Study, a birth cohort study of 4,898 children born in large U.S. cities in 1998–2000, with an oversample of non-marital births (Reichman et al., 2001). The analyses restrict to the subsample for whom the mother completed interviews when the child was approximately 3 and 5 years

old and have non-zero weights from the mother survey at age 5 ($N = 3,770$) and for whom the outcome variables (maternal depression and life satisfaction, where the latter is not examined in this replication) were non-missing ($N = 3,760$). Following the main specification of the original authors, we further restrict to those for whom the father had a history of incarceration at some point up through the age 3 interview ($N = 1,595$). Analyses use the city weights provided by the survey. The target population is therefore mothers of children born in the 20 sampled U.S. cities in 1998–2000, for whom the father was incarcerated at some point up to child age 3. Following the original authors, we can think of this as a population at risk of paternal incarceration between child ages 3 and 5.

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

We multiply impute missing value of all predictors using the `Amelia` package in `R` (Honaker et al., 2011). While the original authors conduct multiple imputation by chained equations, we use Amelia because it is faster, thereby allowing us to conduct analyses on 100 imputed datasets, as compared with the 20 imputations used by the original authors. More imputations may improve the performance of the procedures, for which statistical guarantees are asymptotic in the number of imputations. Most variables are missing for less than 15% of observations. One variable missing in 25% of cases is whether the mother reports that either of her biological parents ever experienced a two-week period of feeling depressed, down

in the dumps, or blue; this is sometimes missing because the mother has no knowledge of one of her own parents. Variables involving father reports are also often missing because survey nonresponse and attrition were higher for fathers, some of whom never participate in a survey. Father's foreign-born status and age are missing in 22% of observations, and paternal impulsivity is missing in 40% of observations because this variable is based on a scale administered at child age 1 to fathers in only 18 of the 20 sample cities. An argument can be made that the model would be more transparent by excluding these variables with high rates of missingness, but we include them to maintain maximal comparability to the specification of the original authors.

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

G Risk ratios and risk difference: An example of interaction estimands that are distinct

Researchers often explore how a treatment effect varies as a function of pre-treatment covariates. Fig. 11 illustrates one setting in which the presence or absence of an interaction depends on the estimand. This section provides details about that simulation.

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

$$X_i \sim \text{Discrete Uniform}(\{.20, .21, \ldots, .40\}) \tag{G.1}$$

$$D_i \sim \text{Bernoulli}(0.50) \tag{G.2}$$

$$Y_i(0) \sim \text{Bernoulli}(X_i) \tag{G.3}$$

$$Y_i(1) \sim \text{Bernoulli}(2X_i) \tag{G.4}$$

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

outcomes under each treatment.

$$\text{Causal Risk Difference}(x) = \underbrace{\frac{1}{n_x}\sum_{i:X_i=x} Y_i(1)}_{\substack{\text{Mean potential outcome}\\ \text{under treatment among}\\ \text{those with } X_i = x}} - \underbrace{\frac{1}{n_x}\sum_{i:X_i=x} Y_i(0)}_{\substack{\text{Mean potential outcome}\\ \text{under control among}\\ \text{those with } X_i = x}} \tag{G.5}$$

$$\text{Causal Risk Ratio}(x) = \frac{\frac{1}{n_x}\sum_{i:X_i=x} Y_i(1)}{\frac{1}{n_x}\sum_{i:X_i=x} Y_i(0)} \quad \leftarrow \quad \text{ratio of the means above} \tag{G.6}$$

We might say that interaction is present if the causal effect of the treatment is a function of the pre-treatment covariate x. However, the conclusion could depend on what estimand is meant by "effect". Interaction may be present for the risk difference but not for the risk ratio (Fig. 11). This example illustrates the broader point that the language commonly used to state sociological goals (e.g. effect heterogeneity, interaction) is insufficiently precise.

H Review of *ASR* 2018

We reviewed the 2018 volume of *ASR* to determine if our framework was redundant: can we already unambiguously translate research goals into theoretical estimands defined as unit-specific quantities aggregated over target populations? Table 3 and 4 state our coding of the unit-specific quantity and target population, respectively, for all 32 articles published in the 2018 volume of the *American Sociological Review*. Because the theoretical estimand is directly related to how statistical results are interpreted with respect to theory, we considered not only the statistical procedures applied to the data but also the ways the authors interpreted those procedures. Justifications for the codings are provided for the unit-specific quantity in Tables 5–10 and for the target population in Tables 11–14. We arrived at these codings after two of the authors read every paper and iterated to come to agreement about how it should be coded.

Author	Type of article	Difficulty translating the unit-specific quantity to our framework	Details in
Liu	Causal observational	Underspecified mediation	Table 5
Wedow et al.	Causal observational	Underspecified mediation	Table 5
Barber et al.	Causal observational	Underspecified mediation	Table 5
Desmond and Travis	Causal observational	Underspecified mediation	Table 5
McDonnell and King	Causal observational	Post-treatment controls	Table 6
Mun and Jung	Causal observational	Post-treatment controls	Table 6
Font et al.	Causal observational	Post-treatment controls	Table 6
Kadivar	Causal observational	Post-treatment controls	Table 6
Schneider et al.	Causal observational	Post-treatment controls	Table 6
Inanc	Causal observational	Post-treatment controls	Table 6
Gauchat and Andrews	Causal observational	Post-treatment controls	Table 6
Gutierrez	Causal observational	Post-treatment controls	Table 6
Villarreal and Tamborini	Causal observational	Post-treatment controls	Table 6
Ludwig and Brüderl	Causal observational	Post-treatment controls	Table 6
Barr et al.	Causal observational	Bundled treatment	Table 7
Wilmers	Causal observational	Linear continuous treatment	Table 7
Goldstein	Causal observational	Many treatments	Table 7
Horowitz	Causal observational	Many treatments	Table 7
Schilke and Rossman	Causal experiment	Underspecified mediation	Table 8
Simpson et al.	Causal experiment	Underspecified mediation	Table 8
Quadlin	Causal experiment	Underspecified mediation	Table 8
Hahl et al.	Causal experiment	Ok	
Mize and Manago	Causal experiment	Ok	
Weisshaar	Causal experiment	Ok	
Flores and Schachter	Causal experiment	Ok	
Ferguson and Koning	Descriptive	Describes another population	Table 9
Ciocca Eller and DiPrete	Descriptive	Describes another population	Table 9
Guinea-Martin et al.	Descriptive	Describes another population	Table 9
Bloome et al.	Descriptive	Describes another population	Table 9
Offer and Fischer	Descriptive	Causal verbs	Table 10
Freeland and Hoey	Descriptive	Causal verbs	Table 10
Mullins et al.	Descriptive	Ok	

Table 3. Review of unit-specific quantities. Systematic coding of individual papers from *ASR* 2018. Table 5–10 justify the coding of each paper.

Author	Type of article	Difficulty translating the target population to our framework	Details in
Barr et al.	Probability sample	Does not discuss weights	Table 11
Flores and Schachter	Probability sample	Does not discuss weights	Table 11
Gauchat and Andrews	Probability sample	Does not discuss weights	Table 11
Guinea-Martin et al.	Probability sample	Does not discuss weights	Table 11
Gutierrez	Probability sample	Does not discuss weights	Table 11
Horowitz	Probability sample	Does not discuss weights	Table 11
Inanc	Probability sample	Does not discuss weights	Table 11
Liu	Probability sample	Does not discuss weights	Table 11
Schneider et al.	Probability sample	Does not discuss weights	Table 11
Wedow et al.	Probability sample	Does not discuss weights	Table 11
Barber et al.	Probability sample	Ok	
Bloome et al.	Probability sample	Ok	
Ciocca Eller and DiPrete	Probability sample	Ok	
Desmond and Travis	Probability sample	Ok	
Ludwig and Brüderl	Probability sample	Ok	
Villarreal and Tamborini	Probability sample	Ok	
Ferguson and Koning	Administrative records	Are records the full population?	Table 12
Goldstein	Administrative records	Are records the full population?	Table 12
McDonnell and King	Administrative records	Are records the full population?	Table 12
Mun and Jung	Administrative records	Are records the full population?	Table 12
Wilmers	Administrative records	Are records the full population?	Table 12
Font et al.	Administrative records	Ok	
Freeland and Hoey	Other sample	Diversity of sample	Table 13
Mize and Manago	Other sample	Diversity of sample	Table 13
Offer and Fischer	Other sample	Diversity of sample	Table 13
Quadlin	Other sample	Diversity of sample	Table 13
Simpson et al.	Other sample	Diversity of sample	Table 13
Weisshaar	Other sample	Diversity of sample	Table 13
Hahl et al.	Other sample	Similarity to the U.S. population	Table 14
Schilke and Rossman	Other sample	Similarity to the U.S. population	Table 14
Kadivar	Other sample	Ok	
Mullins et al.	Other sample	Ok	

Table 4. Review of target populations. Systematic coding of individual papers from *ASR* 2018. Table 11–14 justify the coding of each paper.

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

Table 5. Difficulty translating the unit-specific quantity to our framework: Underspecified mediation in observational studies (4 articles). As discussed in Sec. 3.3, we interpret any claim about a "direct effect" or an effect "net of" or "mediated by" something as a causal claim about the effect of a treatment on an outcome in a possibly counterfactual world where the mediator was set to a specified value. In our framework, direct effects therefore require a statement of the value at which the mediator is fixed by a hypothetical intervention. The magnitude of the direct effect may be different at different values of the mediator.

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

Table 6. Difficulty translating the unit-specific quantity to our framework: Post-treatment conditioning (10 articles). These studies potentially condition on a post-treatment variable; they either adjust for (1) a covariate that temporally follows a treatment, (2) a covariate measured in the same time period as the treatment and is potentially caused by it, or (3) are not clear enough about the temporal and causal ordering of covariates and treatments to determine. Is the target a mediation claim about a direct effect? If so, then making the target explicit would facilitate clear reasoning about the goal and the identification assumptions required to estimate the causal effect of the mediator. If not, then the authors have made an identification error by conditioning on a post-treatment variable. With these papers, we are unsure which has occurred.

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

Table 7. Difficulty translating the unit-specific quantity to our framework: Other issues (4 articles). These issues include continuous treatment variables estimated with a linear functional form for which the precise contrast between values of the treatment is not clear, as well as studies with many predictors of which it is not clear which are viewed as causal treatments. In each case, either (1) we are unsure which variables are the treatments that define unit-specific counterfactual quantities as opposed to distinguishing separate populations of units, (2) we are unsure of the specific treatment values over which the causal effect is contrasted, or (3) we are unsure of both.

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

Table 8. Difficulty translating the unit-specific quantity to our framework: Underspecified mediation in experiments (3 articles). Randomization of treatment assignment in an experiment makes the unit-specific quantity unambiguous for the total effect: the difference between the potential outcomes under two treatment conditions. The direct effect net of a non-randomized mediator, however, suffers from the same issues common in observational studies targeting mediation estimands (Table 5). As discussed in Sec. 3.3, we interpret any claim about a "direct effect" or an effect "net of" or "mediated by" something as a causal claim about the effect of a treatment on an outcome in a possibly counterfactual world where the mediator was set to a specified value. In our framework, direct effects therefore require a statement of the value at which the mediator is fixed by a hypothetical intervention. Further, even in experiments with randomized treatments, mediation claims rely on assumptions to identify the causal effect of the mediator.

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

Table 9. Difficulty translating the unit-specific quantity to our framework: Inference to non-existent populations (4 articles). Some studies reweight a population so that a predictor follows a different distribution from that observed in the factual population. In practice, many such articles formally appeal to non-causal counterfactuals (or more informally what 'would have happened'). For instance, researchers might investigate how outcomes would change for group A if that group's characteristics were made more similar to group B (e.g., how would male students' rates of college entry change if one shifted their actual high school achievement to the counterfactual high school achievement equivalent to female students?). Our definition of the theoretical estimand rules out these settings because they don't describe the state of the world where such data would arise and thus cannot adjudicate between alternative adjustment strategies. These studies could be brought into our framework by either describing an existing population or invoking unit-specific causal counterfactuals under a well-defined intervention.

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

Table 10. Difficulty translating the unit-specific quantity to our framework: Causal verbs attributed to predictors (2 articles). In these articles, the primary component of the paper is purely descriptive, yet at least one analysis or interpretation pairs predictor variables with causal verbs that point toward at least some interest in a causal claim. If interpretations give an active role to a variable as (for example) shaping, leading to, pressuring, or shifting some outcome, those are claims that we interpret as implicitly implying a counterfactual defined over that variable: the outcome would have been different if the variable took a different value. In our framework, any counterfactual statement is ambiguous unless it is clearly tied to a causal claim about unit-specific outcomes that would be realized under a well-specified intervention. Alternately, the authors could interpret the results in solely predictive, non-causal terms, and be clear about how the purely descriptive quantity is relevant to theory.

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

Table 11. Difficulty translating the target population to our framework: Ambiguity about survey weights (9 articles). Is the sample of theoretical interest in itself, or do the authors seek to draw inference to a broader target population? If the target is the sample, weights are unneeded. If the target is the sampling frame, weights are needed. If the target is broader than the sampling frame, additional argument is needed about how the estimates are informative about that broader population. When studies do not discuss the use or lack of use of survey weights in the main analysis, we cannot unambiguously say what the target population is. That hinders methodological choices: the correct use of weights cannot be determined when the target population is ambiguous.

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

Table 12. Difficulty translating the target population to our framework: Are administrative records a census of the entire target population? (5 articles). We are often unsure whether the population that appears in administrative records is the population of substantive interest, or whether authors seek to reach further to a broader population. Whether the administrative records contain the entire target population is also necessary to guide methodological choices such as the construction of standard errors.

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

Table 13. Difficulty translating the target population to our framework: Appeals to diversity (5 articles). Articles with non-probability samples often appeal to the diversity of cases represented in the sample. However, diversity is only helpful if it makes the sample more informative about some diverse target population. In these studies, we are unsure what the target population is. Knowing the target population would inform the assumptions required to draw inference about this population and could direct methodological choices to construct and weight the sample to yield better estimates. While the target population remains unstated it is difficult to reason about these methodological choices.

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

Table 14. Difficulty translating the target population to our framework: Similarities to the U.S. population (2 articles). Unlike studies that simply appeal to diversity, these studies additionally appeal to similarities of the characteristics of those in the sample to the characteristics of the U.S. population. Is the target then the U.S. population? If so, one could imagine reweighting the sample to yield better estimates for this target population. If the target is only the sample, and this is just a way of describing the sample, then reweighting would be inappropriate. Appropriate methodological guidance depends on the target population.

References

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

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.

Barber, J. S., Y. Kusunoki, H. H. Gatny, and J. Budnick 2018. The dynamics of intimate partner violence and the risk of pregnancy during the transition to adulthood. *American sociological review*, 83(5):1020–1047.

Barr, A. B., L. G. Simons, R. L. Simons, S. R. Beach, and R. A. Philibert 2018. Sharing the burden of the transition to adulthood: African american young adults transition challenges and their mothers health risk. *American Sociological Review*, 83(1):143–172.

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.

Buchmann, C. and T. A. DiPrete 2006. The growing female advantage in college completion: The role of family background and academic achievement. *American Sociological Review*, 71(4):515–541.

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

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

DAmour, A., P. Ding, A. Feller, L. Lei, and J. Sekhon 2020. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*.

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

Ferguson, J.-P. and R. Koning 2018. Firm turnover and the return of racial establishment segregation. *American Sociological Review*, 83(3):445–474.

Flores, R. D. and A. Schachter 2018. Who are the illegals? the social construction of illegality in the united states. *American Sociological Review*, 83(5):839–868.

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.

Freeland, R. E. and J. Hoey 2018. The structure of deference: Modeling occupational status using affect control theory. *American Sociological Review*, 83(2):243–277.

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

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.

Guinea-Martin, D., R. Mora, and J. Ruiz-Castillo 2018. The evolution of gender segregation over the life course. *American Sociological Review*, 83(5):983–1019.

Gutierrez, C. M. 2018. The institutional determinants of health insurance: Moving away from labor market, marriage, and family attachments under the ACA. *American Sociological Review*, 83(6):1144–1170.

Hahl, O., 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.

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

Honaker, J., G. King, M. Blackwell, et al. 2011. Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47.

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

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.

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.

Kadivar, M. A. 2018. Mass mobilization and the durability of new democracies. *American Sociological Review*, 83(2):390–417.

Liu, H. 2018. Social and genetic pathways in multigenerational transmission of educational attainment. *American Sociological Review*, 83(2):278–304.

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.

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.

Mullins, D. A., D. Hoyer, C. Collins, T. Currie, K. Feeney, P. François, P. E. Savage, H. Whitehouse, and P. Turchin 2018. A systematic assessment of axial age proposals using global comparative historical evidence. *American Sociological Review*, 83(3):596–626.

Mun, E. and J. Jung 2018. Policy generosity, employer heterogeneity, and womens employment opportunities: The welfare state paradox reexamined. *American Sociological Review*, 83(3):508–535.

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.

Offer, S. and C. S. Fischer 2018. Difficult people: Who is perceived to be demanding in

personal networks and why are they there? *American sociological review*, 83(1):111–142.

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.

Quadlin, N. 2018. The mark of a womans record: Gender and academic performance in hiring. *American Sociological Review*, 83(2):331–360.

Reichman, N. E., J. O. Teitler, I. Garfinkel, and S. S. McLanahan 2001. Fragile Families: Sample and design. *Children and Youth Services Review*, 23(4-5):303–326.

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.

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.

Simpson, B., A. Harrell, D. Melamed, N. Heiserman, and D. V. Negraia 2018. The roots of reciprocity: Gratitude and reputation in generalized exchange systems. *American Sociological Review*, 83(1):88–110.

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

Wedow, R., M. Zacher, B. M. Huibregtse, K. Mullan Harris, B. W. Domingue, and J. D. Boardman 2018. Education, smoking, and cohort change: Forwarding a multidimensional theory of the environmental moderation of genetic effects. *American Sociological Review*, 83(4):802–832.

Weisshaar, K. 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.

Wildeman, C., J. Schnittker, and K. Turney 2012. Despair by association? the mental health of mothers with children by recently incarcerated fathers. *American Sociological Review*, 77(2):216–243.

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.

Wood, S. N. 2017. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.