American Sociological Association

American Sociological Review
2021, Vol. 86(3) 532–565

DOI:10.1177/00031224211004187
journals.sagepub.com/home/asr

SAGE

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

Ian Lundberg,[a] Rebecca Johnson,[b] and Brandon M. Stewart[a]

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.

Keywords

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

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. Furthermore, a goal defined entirely within a model bears a connection to theory that is questionable at best. This article presents a methodological framework for quantitative social science in which a precise statement of the

[a]Princeton University
[b]Dartmouth College

Corresponding Author:
Brandon M. Stewart, Department of Sociology and Office of Population Research, Princeton University, 149 Wallace Hall, Princeton, NJ 08540
Email: bms4@princeton.edu

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 of more than one variable (whether person *i* would be employed if they received job training and childcare), 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 or what do we aggregate that unit-specific quantity? The unit-specific quantity and target population combine to define the theoretical estimand: the thing we would like to know if we had data for the full population in all factual or counterfactual worlds of interest. A paper may have multiple theoretical estimands.

Each theoretical estimand is linked to an *empirical estimand* involving only observable quantities (e.g., a difference in means in a population) by assumptions about the relationship between the data we observe and the data we do not. These identification assumptions can be conveyed in a Directed Acyclic Graph (DAG). Finally, one chooses an *estimation strategy* to learn the empirical estimand (e.g., a regression model). We use 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 (Figure 1). The distinction between the theoretical and empirical estimands is subtle but important: the former may involve unobservable quantities such as counterfactuals, whereas 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; the choice of empirical estimands requires conceptual argument about unobserved data. The choice of estimation strategies is distinct because it can be at least partially data-driven. Separating these steps helps researchers make principled choices, allows readers to evaluate claims, and enables the community to build on research findings.

Too often, research papers involve pages of rich theory followed by pages of procedures applied to data, with a vague link between the two. The theoretical and empirical estimand fill the void by precisely stating 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

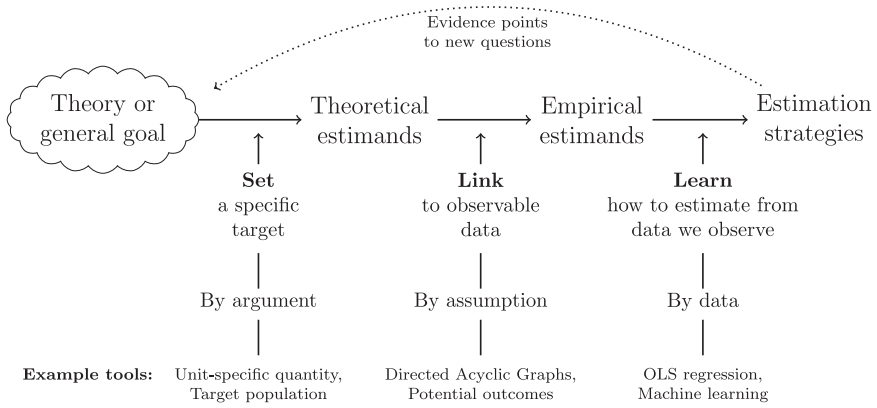


Figure 1. Three Critical Choices in Quantitative Social Science Arguments
Note: The first choice is the theoretical estimands, which set the targets of inference. Argument is required to link the theoretical estimands to the broader theory. The second choice is the empirical estimands, which link the 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.

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.

We first introduce our framework, highlighting each step of our proposed research process: setting the theoretical estimand, linking to an empirical estimand, and learning an estimate from data. Figure 2 presents the examples we use throughout these three sections. We then demonstrate the prevalence of the problems we address through a review of the 2018 volume of *American Sociological Review* (*ASR*), and we illustrate how our framework would transform quantitative research through two in-depth examples. We conclude by describing how estimands clarify methodological issues for analysts, readers, and the broader community.

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, 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, Stevenson, and Elwert 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, King, and Rosenblatt 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 (Abbott 1988; Berk 2004; Duncan 1984; Freedman 1991; Lieberson 1987).

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

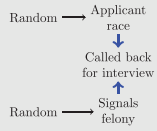
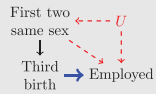
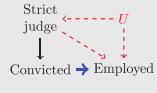
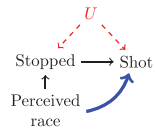
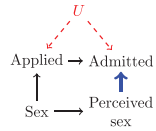
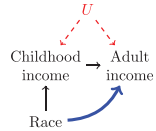
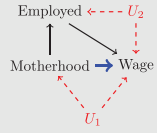
	Set the target: The theoretical estimand		Link to observables	Learn from data
	Unit-specific quantity	Target population of units	Identification	Estimation
Pager	Difference in whether application i would be called back if it signaled White with a felony vs. Black without	Applications to jobs in Milwaukee		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

Figure 2. Estimands Are Relevant to a Broad Range of Social Science Studies
Note: White boxes on the diagonal are the focus of the main text, but every study implicitly involves all four steps. Some steps (e.g., DAGs for identification) are simplified to fit in the table. In the identification step, thick arrows represent the causal effect at the center of the paper and dashed edges represent threats to identification.

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:

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

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 $Y_i(1)$—versus if they did not—denoted $Y_i(0)$:

$$\frac{1}{n}\sum_{i=1}^{n}\Big(Y_i(1) - Y_i(0)\Big) \quad (2)$$

Mean over every i among U.S. adults (target population); Employment if enrolled in job training; Employment if not enrolled in job training (unit-specific quantity)

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.

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 (2003:938) explores "the ways in which the effects of race and criminal record interact to produce new forms of labor market inequalities." 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 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 difficult to disentangle race from individual identities 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.

Pager's (2003) scientific insight contrasts with what could be learned in an observational

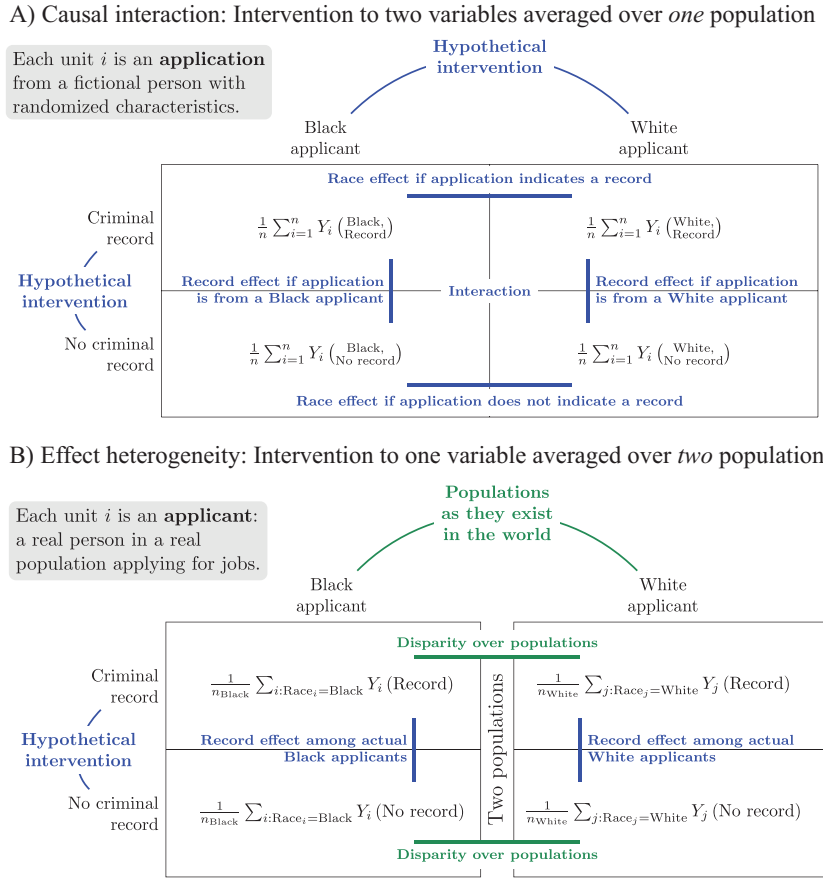


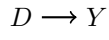
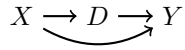
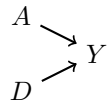
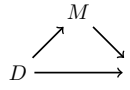
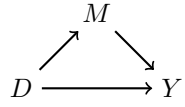
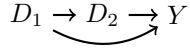
Figure 3. Two Estimands with Different Unit-Specific Quantities and Different Target Populations

Note: 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.

study focused on a different unit-specific quantity (Figure 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 innumerable opportunities that strengthen a résumé. 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

Table 1. Unit-Specific Quantities Defined in Potential Outcomes Unlock Many Causal Estimands for Inquiry

Estimand name	Mathematical statement	DAG	Reference	Colloquial terms
Average treatment effect	$\frac{1}{n}\sum_i \left(Y_i(d') - Y_i(d)\right)$		Morgan and Winship (2015)	Effect
Conditional average treatment effect	$\frac{1}{n_x}\sum_{i:X_i=x} \left(Y_i(d') - Y_i(d)\right)$		Athey and Imbens (2016)	Effect heterogeneity or moderation
Causal interaction	$\frac{1}{n}\sum_i \left(\left(Y_i(a',d') - Y_i(a',d)\right) - \left(Y_i(a,d') - Y_i(a,d)\right)\right)$		Vanderweele (2015)	Joint treatment effect
Controlled direct effect	$\frac{1}{n}\sum_i \left(Y_i(d',m) - Y_i(d,m)\right)$		Acharya et al. (2016)	Mediation (Illustrations: Example 2)
Natural direct effect	$\frac{1}{n}\sum_i \left(Y_i(d', M_i(d)) - Y_i(d, M_i(d))\right)$		Imai et al. (2011)	Mediation (Part B of the Online Supplement)
Effect of time-varying treatment	$\frac{1}{n}\sum_i \left(Y_i(d_1', d_2') - Y_i(d_1, d_2)\right)$		Wodtke et al. (2011)	Cumulative effect

Note: Social scientists who define the research goal before moving to regression uncover more possible questions than those who confine themselves to regression parameters. The table 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 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, *X* indicates pre-treatment covariates, capital letters indicate random variables, and lowercase letters indicate fixed values. Controlled direct effects and other mediation-based estimands appear in sociology, although not always labeled as such (see Part B of the Online Supplement).

would focus on preventing hiring decision-makers from directly considering race and criminal histories when evaluating identical applications. The policy implications of the latter would focus on how equalizing criminal histories (or signals of those histories) could reduce disparities across populations of actual job applicants of different racial categories. Both are worthwhile goals. Distinguishing them requires clarity about whether the unit-specific quantity is a potential outcome as a function of one or two treatments.

The unit-specific quantity provides an opportunity to clarify the causal component (if any) of our research claims. It invites us to be precise about the causes we are studying (e.g., a signal of race) and those we are not (e.g., how racial disparities in access to education create differences in résumés). It allows us to be precise about the levels of the treatment being contrasted (e.g., a particular résumé 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 (see 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 obscured.

Define the Target Population

The second building block of a theoretical estimand is the target population: the set of

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

The target population is a widely-recognized issue in experiments. For example, Pager (2003:965) explicitly notes that "one key limitation of the audit study design is its concentration on a single metropolitan area." The true target population may be broader, such as all entry-level job openings in the United States. 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 (IV) designs. For example, Angrist and Evans (1998) examine the effect of having three versus two children on women's employment under an 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 women 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 percent of all mothers (see Part C of the online supplement). 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 percent of mothers is informative about all mothers. Setting the target population would clarify which tack the authors are taking.

Sometimes, one could defend the complier population as being of genuine theoretical interest in itself. Harding and colleagues (2018) estimate the effect of prison on labor market outcomes by leveraging random variation in judges' propensities to sentence people convicted of felonies to probation instead of prison. The target population is offenders who would have been sentenced to probation rather than prison if they had 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 no untreated units (e.g., those on probation or not). Common support problems leave researchers 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 authors 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 Urzúa 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 2010, 2018; 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 that it 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.

IDENTIFICATION: LINK TO AN EMPIRICAL ESTIMAND

The same quality that makes a theoretical estimand liberating—it can involve unobservable data—also means 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 these 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 that 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.

Theoretical and Empirical 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 versus a Black applicant without a criminal record:

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i\left(\substack{\text{White,}\\\text{Record}}\right) - Y_i\left(\substack{\text{Black,}\\\text{No record}}\right)\right) \quad (3)$$

Mean over **all** applications

Potential outcome under one condition

Potential outcome under another condition

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 involves only observable quantities (no potential outcomes).

$$\theta = \frac{1}{n_{\text{WR}}}\sum_{i\in\mathcal{S}_{\text{WR}}} Y_i - \frac{1}{n_{\text{BN}}}\sum_{i\in\mathcal{S}_{\text{BN}}} Y_i \quad (4)$$

Factual outcomes

Mean among applications assigned to the condition (**W**hite, **R**ecord)

Mean among applications assigned to the condition (**B**lack, **N**o record)

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 our statistical procedures are designed to recover.

Stating both estimands is important because there is no one-to-one 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. Authors need to clarify to which of many possible theoretical estimands they intend the empirical estimand to speak, so the reader can adequately evaluate the available evidence for the claim. This is especially true in more complex settings.

Additional Setting 1: Demographic Standardization

Consider standardized mortality rates. We might compare the age-specific mortality rate (e.g., deaths per thousand among people age 50 to 54) in Mexico and the United States. A demographer might then aggregate age-specific estimates to a summary statement:

the mortality rate in the United States compared with the mortality rate in the Mexican population aggregated over the age distribution of the United States (Preston, Heuveline, and Guillot 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 United States? 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 United States. 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.

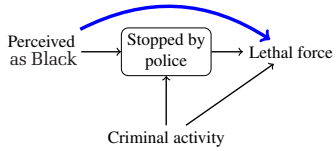
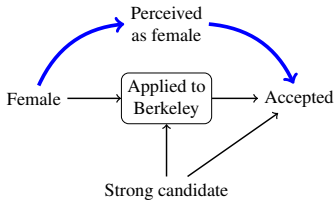
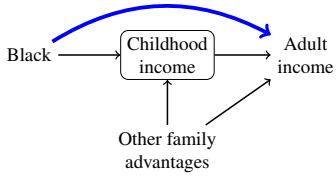
Sociologists fall prey to the same problem when, for example, they deploy Kitagawa-Blinder-Oaxaca decompositions (Kitagawa 1955) and related methods to discuss what would happen in counterfactual populations in which covariates took different values (e.g., Ciocca Eller and DiPrete 2018; Mize 2016; Storer, Schneider, and Harknett 2020). Like a standardized mortality rate, the methods used in these 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.

Additional Setting 2: Disparities in the Presence of Selection Processes

Few sociology papers explicitly state and support their 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 that 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.

Fryer (2019) examines police interactions by race in several administrative data sources. In records from New York City, the use of sublethal force was higher for Black than for non-Black individuals. 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 people stopped by police. The empirical estimand is the difference in force used against Black and White individuals who are involved in police interactions. Knox, Lowe, and Mummolo (2020) highlight a key issue: the sample only includes people who interacted with

Table 2. Empirical Regularities Can Be Misleading without Estimands

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.	

Note: 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 and colleagues (2020) that 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 Black individuals 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 they are not equal and thus the descriptive quantity does not support the conclusions drawn.

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 whereas 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 White individuals 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 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 (2020:4003) claims that he never sought to study racial bias, but only "racial differences" by repeatedly caveating the results with the phrase "conditional on interaction." Fryer (2020:4003) writes, "I am not sure how many more ways we would have needed to caveat our results to satisfy [Durlauf and Heckman]." 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 and colleagues (2020). This is just one reason why it is essential to transparently state a study's true goals.

This problem is more general than the use of administrative data to study police bias. Bickel, Hammel, and O'Connell (1975) study graduate admissions at Berkeley and discover that, although men are admitted at rates 9 percentage points higher than women schoolwide, 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, 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 and colleagues (2020) show that Black and White women who are raised in families with similar incomes have similar earnings as adults. At face value, one might interpret this in terms of a theoretical estimand: if we intervened to equalize the childhood incomes of Black and White women, the racial income gap in adulthood would disappear. Yet this would be misleading because family income is a consequence of both race and other family advantages; the Black families who overcome discrimination to achieve incomes comparable to those of White families 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 individuals 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 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).

ESTIMATION: LEARN THE EMPIRICAL ESTIMAND FROM DATA

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

Mean over entire sample; Expected outcome among cases with the covariate values $\vec{x}_i$ of unit i who are factually treated ($D = 1$)

$$\theta = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}(Y \mid \vec{X} = \vec{x}_i, D = 1) - \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}(Y \mid \vec{X} = \vec{x}_i, D = 0) \quad (5)$$

Mean over entire sample; Expected outcome among cases with the covariate values $\vec{x}_i$ of unit i who are factually untreated ($D = 0$)

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., $\hat{\mathbf{E}}(Y|\vec{X} = \vec{x}_i, D = 1)$) wherever it appears.

Mean over entire sample; Regression prediction at observed covariate values $\vec{x}_i$ with treatment set to $D = 1$

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{E}}(Y \mid \vec{X} = \vec{x}_i, D = 1) - \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{E}}(Y \mid \vec{X} = \vec{x}_i, D = 0) \quad (6)$$

Mean over entire sample; Regression prediction at observed covariate values $\vec{x}_i$ with treatment set to $D = 0$

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, $\mathbf{E}(Y|\vec{X} = \vec{x}_i, D = 1)$. One could therefore select the estimator $\hat{\mathbf{E}}(Y|\vec{X} = \vec{x}_i, D = 1)$ 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 $\mathbf{E}(Y|\vec{X} = \vec{x}_i, D = 1)$ 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 percent 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 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 individuals who are 42 and 44. We might share information across these covariate values by averaging over everyone age 42 to 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 a 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.

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 (details are in Part D of the online supplement). 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,

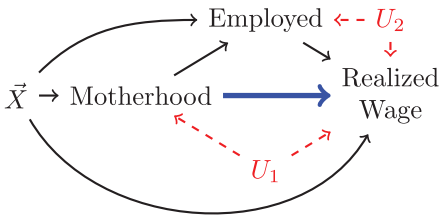


Figure 4. Identification Assumptions for the Motherhood Wage Penalty
Note: To identify the controlled direct effect of motherhood on the wages that would be realized under employment, one must assume the covariates $\vec{X}$ are sufficient to block all confounding of motherhood and of employment. The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

showing 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 article, clear reasoning about estimation requires that we first define the theoretical and empirical estimand. The original paper is not entirely clear: the authors deploy "causal estimation techniques" (Pal and Waldfogel 2016:108), but 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 individuals who are not employed (Pal and Waldfogel [2016:109–110] acknowledge this complexity in a footnote). 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 Figure 4, which we emphasize includes a very strong assumption of no mediator-outcome confounding. Part D of the online supplement 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, Blackwell, and Sen 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 $\widehat{\beta}$ on motherhood. However, that simplification is only possible under the (doubtful) no-interactions assumption that the treatment effect is the same value $\widehat{\beta}$ at the covariate value $\vec{x}_i$ of

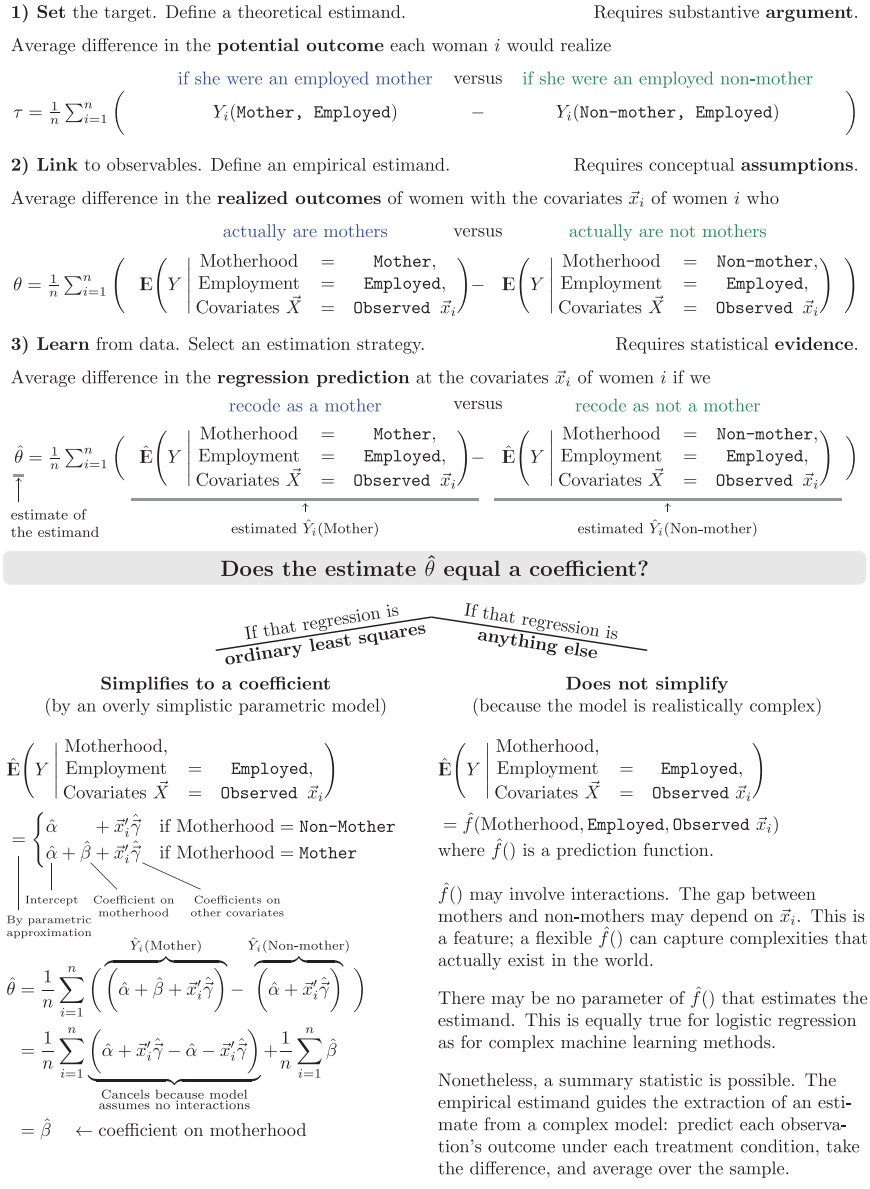


Figure 5. Estimands Unlock New Estimation Tools
Note: Once we state the estimand, we can use a predictive algorithm to impute unknown conditional means. It could be a linear model (e.g., OLS) or a nonlinear model (e.g., logit, random forest). The result equals a coefficient for linear models only. The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

every unit *i*. In other words, a coefficient $\widehat{\beta}$ 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,

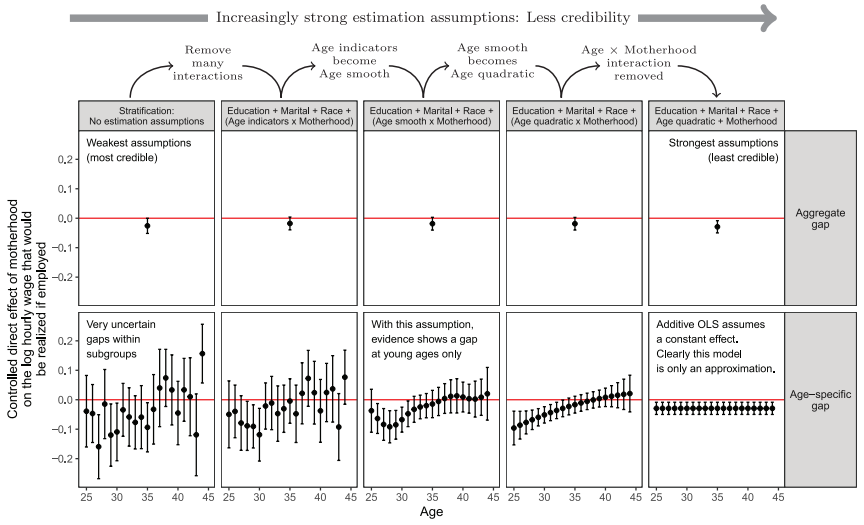


Figure 6. A Series of Estimation Strategies (columns) for Two Estimands (rows)
Note: Each estimand is the gap in log hourly wages between mothers and childless women, conditional on age, education, marital status, and race and aggregated over the covariate distribution of mothers. Estimands differ by 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 sample 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 the tendency to define estimands by a regression coefficient has prevented social scientists from recognizing settings when inference can proceed from more minimal estimation assumptions (at left). Instead of beginning from the right and moving left, we propose 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 percent 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 percent confidence intervals calculated using replicate weights (see Part D of the online supplement).

imputing the expected wage within a covariate cell by the mean wage of those observed with exactly that set of covariates (far left of Figure 6). So why would one assume a functional form, like a parametric model where the estimand is estimated by a coefficient (far right of Figure 6)? In this case, the OLS model is clearly misspecified: it assumes the family gap in pay does not vary by age (see lower right panel of Figure 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 we can begin to reason about the best empirical approximation. In a large sample, we may often be able to estimate the empirical estimand 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 nonexistent wages of the nonemployed. The choice of the best estimation method could be made using the data, but we need subject matter knowledge to assess whether the identification assumptions are plausible.

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 percent confidence interval that actually would contain the empirical estimand in 95 percent of hypothetical samples. A valid interval is elusive in both parametric models and machine learning approaches. Parametric models (e.g., OLS) come with readily-available confidence intervals for the coefficient of interest. Those intervals would provide the expected coverage for the coefficient that would be estimated if the regression were estimated on the full population. But that coefficient is not the empirical estimand: if the functional form assumed by the model is a poor approximation to the truth, then a confidence interval for the coefficient may have very poor coverage for the empirical estimand. It is therefore easy to produce a confidence interval for a coefficient, but the properties of that interval with respect to the estimand rely on the assumed functional form, which may be questionable. Flexible machine learning approaches avoid this problem by learning the functional form from the data. Yet they face a different problem: the statistical theory to place standard errors around machine learning estimates can be lacking. This is an area of active research (e.g., Wager and Athey 2018) and can sometimes be overcome by computational approaches such as bootstrapping. Part D of the online supplement 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 it can be difficult to summarize how the model extrapolates to accomplish this task, a problem that can be particularly acute in high-dimensional data (D'Amour et al. 2021). 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.

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.

STATISTICAL PRACTICE IN A TOP JOURNAL DOES NOT FOLLOW OUR FRAMEWORK

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.

Figure 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 Part H of the online supplement. We were completely certain of both the unit-specific quantity and the target population in zero papers. 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.

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 Equation 7:

Unit-specific quantity:

$$Y_i \quad \text{or} \quad Y_i(t) \tag{7}$$

Descriptive Outcome as it factually exists

Causal Outcome if assigned to treatment value t

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.

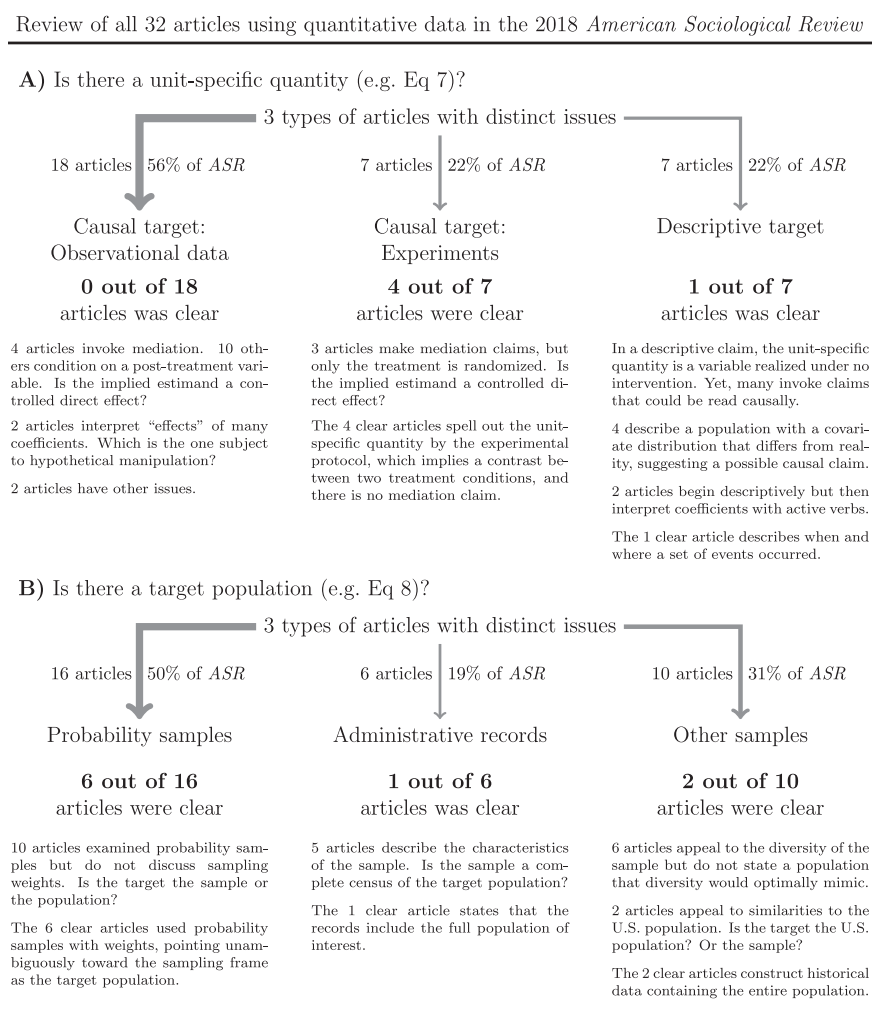


Figure 7. Our Methodological Framework Differs from Standard Practice
Note: In our framework, all estimands are functions of actual outcomes Y_i (for descriptive estimands) or potential outcomes $Y_i(d)$ (for causal estimands), aggregated over a well-defined population of units indexed by *i*. In the 2018 volume of *ASR*, no articles 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 the panels above by the single error we considered to be most apparent. Details are in Part H of the online supplement.

We can confidently state the unit-specific quantity in the form of Equation 7 in only 5 out of 32 papers (16 percent). Ambiguities in this quantity differ across three categories of papers. In studies drawing causal claims from observational data (56 percent of *ASR*), zero articles provide enough detail for us to be entirely confident of the intended intervention. Most of this category (78 percent) 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 is the object of inquiry. Experiments (22 percent 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 percent of such studies) are subject to the same issues common among observational studies.

Descriptive studies constitute a minority of articles (22 percent of *ASR*). However, 6 of these 7 studies conduct at least one analysis reaching beyond pure description to claims 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—for example, 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) that 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 the 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 the 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]

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 Equation 8:

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\substack{\text{Mean over}\\ \text{some specified}\\ \text{population}}} \underbrace{\left(\text{Unit-Specific Quantity}\right)_i}_{\substack{\text{Of a quantity}\\ \text{defined for}\\ \text{each unit}}} \tag{8}$$

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 percent). Half of *ASR* articles draw on probability samples, but 62 percent 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 percent 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 percent of *ASR* articles. 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.

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 sociology has 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 section uses specific examples to show how the proposed framework could transform quantitative sociology.

ILLUSTRATIONS: HOW ESTIMANDS IMPROVE PRACTICE

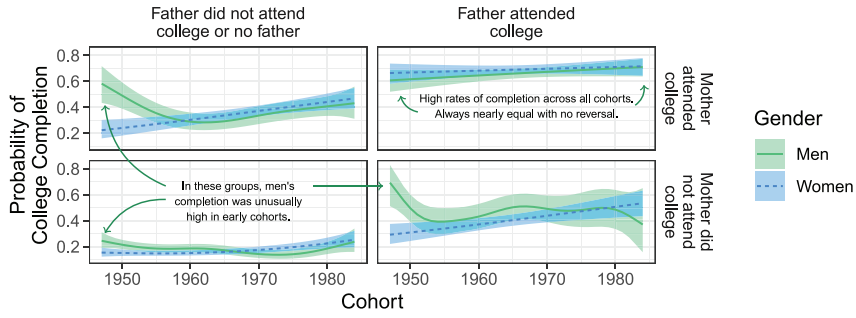
Unlike new statistical adjustments, changing the theoretical estimand can set the research on a completely different path, making it difficult to produce general statements about what would happen to results in the field. Instead, we demonstrate changes using two specific examples: a descriptive example about the gender gap in college completion and a causal example about the effect of paternal incarceration on maternal depression.

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

Buchmann and DiPrete (2006) summarize a gender reversal: whereas men historically completed college degrees at higher rates than women, the disparity reversed over the second half of the twentieth 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). The statement evokes something more meaningful, but 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:

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



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

Figure 8. Descriptive Estimands Are Worthwhile Goals; the Language of "Effects" Common in Multiple Regression Models Can Produce Confusion
Note: 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 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.

Probability of college completion Y | Among those with | Gender g, birth cohort c, and parent characteristics p

$$\tau(g,p,c) = \mathbf{P}\left(Y = 1 \,\middle|\, \begin{matrix} G = g \\ C = c \\ P = p \end{matrix}\right) \tag{9}$$

Probability of college completion Y | Among those with | Gender g, birth cohort c, and parent characteristics p

$$\theta(g,p,c) = \mathbf{P}\left(Y = 1 \,\middle|\, \begin{matrix} G = g \\ C = c \\ P = p \\ R = 1 \end{matrix}\right) \tag{10}$$

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

Each person has a unit-specific realized outcome indicating whether or not that person completed college. 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 1947 to 1984 and observed at ages 25 to 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:

Figure 8 summarizes these descriptive estimands with gender g represented by solid and dashed lines, 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 individuals 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 fathers 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 to 1952; Angrist and Chen 2011). The Vietnam War could have produced especially high college completion rates for this cohort because veterans received scholarships upon their return under the GI Bill. The high completion rate could also arise in part from an identification problem: the theoretical estimand involves everyone born in 1950, but individuals killed in the war were not around decades later to complete the GSS survey. Perhaps the men who would not have completed college were disproportionately drafted and killed in the war, contributing to a gap between the theoretical and empirical estimand.

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

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

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

Wildeman, Schnittker, and Turney (2012) estimate 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 .32. Our replication recovers a similar estimate of .28 (details in Part F of the online supplement).[5] We convert the model into an estimate of a nonparametric estimand: paternal incarceration increases the probability of maternal depression by 4 percentage points (95 percent CI: –.02, .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.

We focus on one mediator from Wildeman and colleagues (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 be realized at different values of d and m (Imai et al. 2011; Pearl 2001). We focus on one particular set

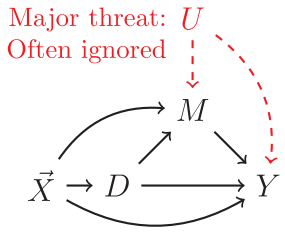


Figure 9. Causal Structure for Mediation Estimands
Note: These require identifying the effect of the treatment D and the effect of the mediator M on the outcome Y. The unobserved mediator-outcome confounder, U, is a threat to inference.

of mediation estimands—controlled direct effects—that 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:

$$\tau(m) = \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i(1,m) - Y_i(0,m)\Big) \quad (11)$$

Whether mother i would be depressed if father i was incarcerated vs if father i was **not** incarcerated

Controlled direct effect; Mean over sample; if her repartnering was set to the value m

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.

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 and colleagues (2012:222) "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, see Figure 9). This threat persists even in randomized experiments where the treatment (but not the mediator) is randomized (for examples from *ASR*, see Table 8 in the online supplement). The assumption of no treatment-outcome or mediator-outcome confounding is often doubtful. We will nonetheless proceed under this assumption to further illustrate complexities of mediation that arise (at least implicitly) in many studies.

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 $\widehat{Y}_i(d,m)$ for each unit i, and average over units as specified by the estimand:

$$\hat{\tau}(m) = \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{Y}_i(1,m) - \hat{Y}_i(0,m)\Big) \quad (12)$$

Predicted probability of depression for mother i with father incarceration set to the value 1 vs with father incarceration set to the value 0

Estimated controlled direct effect; Mean over sample; 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.

The four plots in the upper left of Figure 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

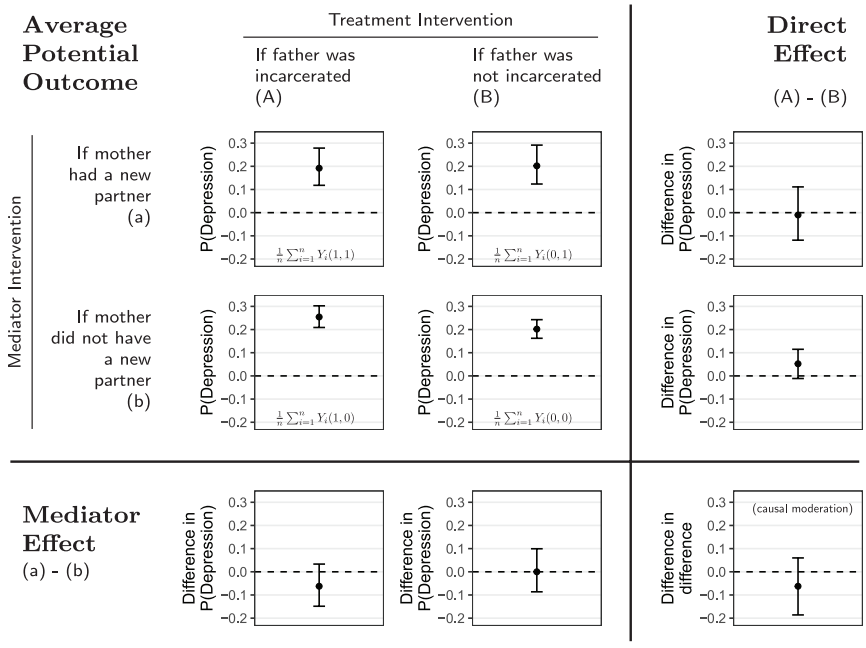


Figure 10. Controlled Direct Effects Involve an Intervention to the Treatment and the Mediator

Note: The 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 .025 and .975 quantiles of 10,000 simulated draws calculated by 100 likelihood-based samples from parameters estimated in each of 100 multiply-imputed datasets.

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. 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.

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—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 and colleagues (2012:233) "consider all the mechanisms simultaneously, and the relationship is reduced by approximately half." 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.

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 that involves (1) choosing one or more theoretical estimands, (2) choosing empirical estimands that 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 people 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.

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, Karlson, and Holm 2018), or 'Should I use fixed or random effects?' (Firebaugh, Warner, and Massoglia 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, DeMeritt, and Esarey 2010; Nagler 1991; Rainey 2016). Yet the problem is more fundamental: researchers must begin by stating what they mean by the term "interaction." Figure 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

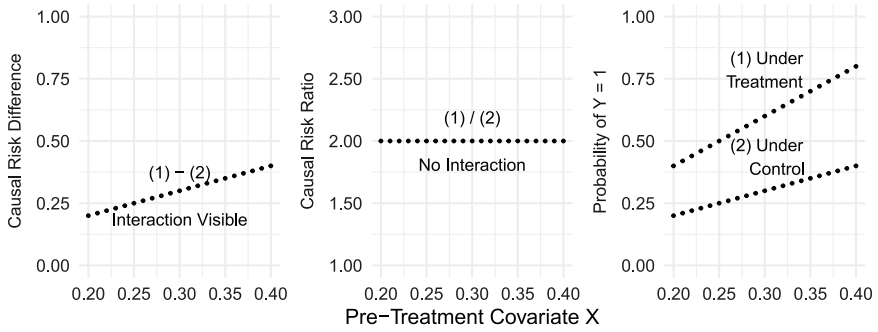


Figure 11. Illustration: The Presence or Absence of Interaction May Depend on Whether the Estimand Is a Multiplicative or an Additive Effect

Note: 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. Part G of the online supplement provides simulation details.

regression because "we can seldom include in a model all variables that affect an outcome." Breen and colleagues (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 in probabilities, then the researcher would naturally present predicted probabilities rather than reporting coefficient estimates (Breen et al. 2018; Mize 2019; Mood 2010). Questions about which model to use, what assumptions are required, and how to present results all become clearer once the research goal has been stated.

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 (see Part A of the online supplement). 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 sets that credibly identify the causal effect. Robustness across estimation strategies is only important among methods that 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. Yet, robustness checks as currently applied treat all specifications as equally valid. Young and Holsteen (2017) provide 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.

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 to 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 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.

Concluding Remarks: Estimands Prepare Us for the Future of Quantitative Sociology

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

Acknowledgments

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. Special thanks to Simone Zhang for many useful comments and excellent collaboration on related projects.

Funding

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.

Data Note

Replication code is available on Dataverse: https://doi.org/10.7910/DVN/ASGOVU.

ORCID iDs

Ian Lundberg https://orcid.org/0000-0002-1909-2270
Rebecca Johnson https://orcid.org/0000-0003-2475-6622
Brandon M. Stewart https://orcid.org/0000-0002-7657-3089

Notes

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.
2. Bickel and colleagues (1975:398) 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." We applaud the explicitness of the assumption, but it is questionable when discrimination affects decisions to apply for graduate school.
3. When the effect is not constant, $\widehat{\beta}$ 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.
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.
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 percent confidence interval (–.14 to .74) contains zero.
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 (e.g., 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.

References

Abbott, Andrew. 1988. "Transcending General Linear Reality." *Sociological Theory* 6(2):169–86.

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining Causal Findings without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110(3):512–29.

Angrist, Joshua D., and Stacey 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, Joshua D., and William N. Evans. 1998. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *The American Economic Review* 88(3):450–77.

Aronow, Peter M., and Benjamin T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge, UK: Cambridge University Press.

Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113(27):7353–60.

Berk, Richard A. 2004. *Regression Analysis: A Constructive Critique*. Thousand Oaks, CA: Sage.

Berk, Richard, Andreas Buja, Lawrence Brown, Edward George, Arun K. Kuchibhotla, Weijie Su, and Linda Shazo. 2021. "Assumption Lean Regression." *The American Statistician* 75(1):76–84.

Berry, William D., Jacqueline H. R. DeMeritt, and Justin Esarey. 2010. "Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?" *American Journal of Political Science* 54(1):248–66.

Bickel, Peter J., Eugene A. Hammel, and J. William O'Connell. 1975. "Sex Bias in Graduate Admissions: Data from Berkeley." *Science* 187(4175):398–404.

Blitzstein, Joseph K., and Jessica Hwang. 2019. *Introduction to Probability*, 2nd ed. Boca Raton, FL: CRC Press.

Brand, Jennie E., and Yu 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, Richard, Kristian B. Karlson, and Anders Holm. 2018. "Interpreting and Understanding Logits, Probits, and Other Nonlinear Probability Models." *Annual Review of Sociology* 44:39–54.

Buchmann, Claudia, and Thomas A. DiPrete. 2006. "The Growing Female Advantage in College Completion: The Role of Family Background and Academic Achievement." *American Sociological Review* 71(4):515–41.

Buja, Anders, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. 2019. "Models as Approximations I: Consequences Illustrated with Linear Regression." *Statistical Science* 34(4):523–44.

Chetty, Raj, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter. 2020. "Race and Economic Opportunity in the United States: An Intergenerational Perspective." *The Quarterly Journal of Economics* 135(2):711–83.

Ciocca Eller, Christina, and Thomas A. DiPrete. 2018. "The Paradox of Persistence: Explaining the Black-White Gap in Bachelor's Degree Completion." *American Sociological Review* 83(6):1171–1214.

D'Amour, Alexander, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. 2021. "Overlap in Observational Studies with High-Dimensional Covariates." *Journal of Econometrics* 221(2):644–54.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48(2):424–55.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210:2–21.

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

Durlauf, Steven N., and James J. Heckman. 2020. "An Empirical Analysis of Racial Differences in Police Use of Force: A Comment." *Journal of Political Economy* 128(10):3998–4002.

Egami, Nakoi, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2018. "How to Make Causal Inferences Using Texts." arXiv preprint arXiv:1802.02163.

Elwert, Felix, and Christopher Winship. 2010. "Effect Heterogeneity and Bias in Main-Effects-Only Regression Models." Pp. 327–36 in *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, edited by R. Dechter, H. Geffner, and J. Y. Hapern. Rickmansworth, UK: College Publications.

Elwert, Felix, and Christopher Winship. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology* 40:31–53.

Firebaugh, Glenn, Cody Warner, and Michael Massoglia. 2013. "Fixed Effects, Random Effects, and Hybrid Models for Causal Analysis." Pp. 113–32 in *Handbook of Causal Analysis for Social Research*, edited by S. Morgan. Dordrecht, Netherlands: Springer.

Font, Sarah A., Lawrence M. Berger, Maria Cancian, and Jennifer L. Noyes. 2018. "Permanency and the Educational and Economic Attainment of Former Foster Children in Early Adulthood." *American Sociological Review* 83(4):716–43.

Freedman, David A. 1991. "Statistical Models and Shoe Leather." *Sociological Methodology* 21:291–313.

Freese, Jeremy, and David Peterson. 2017. "Replication in Social Science." *Annual Review of Sociology* 43:147–65.

Fryer, Roland G. 2019. "An Empirical Analysis of Racial Differences in Police Use of Force." *Journal of Political Economy* 127(3):1210–61.

Fryer, Roland G. 2020. "An Empirical Analysis of Racial Differences in Police Use of Force: A Response." *Journal of Political Economy* 128(10):4003–4008.

Greiner, D. James, and Donald B. Rubin. 2011. "Causal Effects of Perceived Immutable Characteristics." *Review of Economics and Statistics* 93(3):775–85.

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66(2):315–31.

Harding, David J., Jeffrey D. Morenoff, Anh P. Nguyen, and Shawn D. Bushway. 2018. "Imprisonment and Labor Market Outcomes: Evidence from a Natural Experiment." *American Journal of Sociology* 124(1):49–110.

Heckman, James J., and Sergio Urzúa. 2010. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." *Journal of Econometrics* 156(1):27–37.

Hernán, Miguel A. 2018. "The C-Word: Scientific Euphemisms Do Not Improve Causal Inference from Observational Data." *American Journal of Public Health* 108(5):616–19.

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

Hernán, Miguel A., Brian C. Sauer, Sonia Hernández-Díaz, Robert Platt, and Ian 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, Daniel. 2016. "Stylized Facts in the Social Sciences." *Sociological Science* 3:604–626.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(4):765–89.

Imbens, Guido W. 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzúa (2009)." *Journal of Economic Literature* 48(2):399–423.

Imbens, Guido. 2018. "Understanding and Misunderstanding Randomized Controlled Trials: A Commentary on Cartwright and Deaton." *Social Science & Medicine* 210:50–52.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2):467–75.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, UK: Cambridge University Press.

Katz, Jonathan N., Gary King, and Elizabeth Rosenblatt. 2020. "Theoretical Foundations and Empirical Evaluations of Partisan Fairness in District-Based Democracies." *American Political Science Review* 114(1):164–78.

Keele, Luke, Randolph T. Stevenson, and Felix Elwert. 2020. "The Causal Interpretation of Estimated Associations in Regression Models." *Political Science Research and Methods* 8(1):1–13.

Kitagawa, Evelyn M. 1955. "Components of a Difference between Two Rates." *Journal of the American Statistical Association* 50(272):1168–94.

Knox, Dean, Will Lowe, and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." *American Political Science Review* 114(3):619–37.

Kohler-Hausmann, Issa. 2018. "Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination." *Northwestern University Law Review* 113(5):1163–1227.

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

Lieberson, Stanley, and Joel Horwich. 2008. "Implication Analysis: A Pragmatic Proposal for Linking Theory and Data in the Social Sciences." *Sociological Methodology* 38(1):1–50.

Liu, Hexuan. 2018. "Social and Genetic Pathways in Multigenerational Transmission of Educational Attainment." *American Sociological Review* 83(2): 278–304.

Lundberg, Ian. 2020. "The Gap-Closing Estimand: A Causal Approach to Study Interventions That Close Disparities across Social Categories." SocArXiv (https://doi.org/10.31235/osf.io/gx4y3).

Mize, Trenton D. 2016. "Sexual Orientation in the Labor Market." *American Sociological Review* 81(6):1132–60.

Mize, Trenton D. 2019. "Best Practices for Estimating, Interpreting, and Presenting Nonlinear Interaction Effects." *Sociological Science* 6:81–117.

Molina, Mario, and Filiz Garip. 2019. "Machine Learning for Sociology." *Annual Review of Sociology* 45:27–45.

Mood, Carina. 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, Stephen L., and Christopher Winship. 2015. *Counterfactuals and Causal Inference*. Cambridge, UK: Cambridge University Press.

Nagler, Jonathan. 1991. "The Effect of Registration Laws and Education on U.S. Voter Turnout." *The American Political Science Review* 85(4):1393–1405.

Pager, Devah. 2003. "The Mark of a Criminal Record." *American Journal of Sociology* 108(5):937–75.

Pal, Ipshita, and Jane 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, Judea. 2001. "Direct and Indirect Effects." Pp. 411–20 in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge, UK: Cambridge University Press.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Preston, Samuel, Patrick Heuveline, and Michel Guillot. 2000. *Demography: Measuring and Modeling Population Processes*. Malden, MA: Blackwell Publishers.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111–64.

Rainey, Carlisle. 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):621–39.

Salganik, Matthew J., Ian Lundberg, Alex Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, et al. 2020. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." *Proceedings of the National Academy of Sciences* 117(15):8398–8403.

Samii, Cyrus. 2016. "Causal Empiricism in Quantitative Research." *The Journal of Politics* 78(3):941–55.

Sen, Maya, and Omar 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, Adam, Daniel Schneider, and Kristen Harknett. 2020. "What Explains Racial/Ethnic Inequality in Job Quality in the Service Sector?" *American Sociological Review* 85(4):537–72.

Van der Laan, Mark J., and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Berlin: Springer Science & Business Media.

VanderWeele, Tyler. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford, UK: Oxford University Press.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113(523):1228–42.

Watts, Duncan J. 2014. "Common Sense and Sociological Explanations." *American Journal of Sociology* 120(2):313–51.

Westreich, Daniel, and Sander Greenland. 2013. "The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients." *American Journal of Epidemiology* 177(4):292–98.

Wildeman, Christopher, Jason Schnittker, and Kristin Turney. 2012. "Despair by Association? The Mental Health of Mothers with Children by Recently Incarcerated Fathers." *American Sociological Review* 77(2):216–43.

Wodtke, Geoffrey T., David J. Harding, and Felix 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):713–36.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R*, 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.

Xie, Yu. 2013. "Population Heterogeneity and Causal Inference." *Proceedings of the National Academy of Sciences* 110(16):6262–68.

Young, Cristobal, and Katherine Holsteen. 2017. "Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis." *Sociological Methods and Research* 46(1):3–40.

Ian Lundberg is a PhD candidate in sociology and social policy at Princeton University. His research develops and applies statistical and machine learning techniques to understand inequality, poverty, and mobility in America.

Rebecca Johnson is an Assistant Professor in Quantitative Social Science at Dartmouth College and an affiliate of Sociology. Her research studies how social service bureaucracies use a mix of data and discretion to allocate scarce resources, focusing on K-12 schools and rental housing.

Brandon M. Stewart is an Assistant Professor of Sociology and Arthur H. Scribner Bicentennial Preceptor at Princeton University and affiliate of the Office of Population Research. His research develops new quantitative statistical methods for applications across computational social science.