Causality in Econometrics: Choice vs. Chance[1]

Prize Lecture 8 December 2021 by
Guido W. Imbens
Graduate School of Business, and Department of Economics, Stanford, CA, USA.

This essay describes the evolution and recent convergence of two methodological approaches to causal inference. The first one, in statistics, started with the analysis and design of randomized experiments. The second, in econometrics, focused on settings with economic agents making optimal choices. I argue that the local average treatment effects framework facilitated the recent convergence by making key assumptions transparent and intelligible to scholars in many fields. Looking ahead, I discuss recent developments in causal inference that combine the same transparency and relevance.

1. INTRODUCTION

Knowledge of causal effects is of great importance for decision makers in government, firms, as well as individuals in their private lives. Inferring the values of these effects from observed data is often a major challenge

---

1. This is a revised version of my recorded Prize lecture posted on December 8, 2021. I am grateful for comments from Alberto Abadie, Joshua Angrist, Mohammad Akbarpour, Dmitry Arkhangelsky, Susan Athey, Kevin Bryan, Ambarish Chattopadhyay, Matthew Gentzkow, Chad Jones, Eva Lestant, Alexia Olaizola, Thomas Richardson, Jesse Shapiro and Amar Venugopal. I also want to gratefully acknowledge the discussions on the topics discussed here over many with Alberto Abadie, Joshua Angrist, Susan Athey, Gary Chamberlain, Tony Lancaster, Whitney Newey, and Donald Rubin, as well as my students and collaborators.

when causal mechanisms are not fully understood. These challenges have motivated methodological research in multiple disciplines. This research got a major boost in the 1920s and 1930s, thanks to advances in the design and analysis of randomized experiments in statistics and, separately, methodological work on observational studies in econometrics. More recently, in the late 1980s and early 1990s, there was a sharp increase in empirical and methodological research in economics, as well as other disciplines, with an explicit focus on estimating causal effects. A convergence of the statistical and econometric traditions has been a catalyst for this increase.[2] More than thirty years later, causality is a thriving area of study. Researchers from many disciplines, including economics, statistics, political science, psychology, epidemiology, computer science and other fields, bring new questions and different methodological perspectives to the discussion. Applications range widely from biomedical to social science, with interest coming from academic, government, and private sector organizations.

In this lecture I discuss some of the themes of this field. Per the charge of the committee awarding the prize, this article focuses primarily on my contributions to the study of causality, but I shall place them in the context of the broader interdisciplinary literature.[3] I start by discussing briefly some of the history of methods for causal inference in statistics and econometrics. I then discuss the credibility crisis in the 1980s that provided some of the motivation for the work that was recognized in the prize. After that I discuss some of my contributions to the causal inference literature. In that part of the paper, I will also add some background and color to the specific research I describe, discussing the origins and questions that motivated my collaborators and myself, as well as pivotal moments in my intellectual journey. I see this prize as a recognition of the importance of this general interdisciplinary enterprise and hope it further invigorates the field.[4]

2. CAUSALITY FROM THE 1920S TO THE 1980S

Although there were earlier empirical studies focused on estimation of causal effects, the research on statistical methods for causality and causal inference started in the first half of the twentieth century. In the 1930s, two distinct literatures emerged in new disciplines that both focused on the developing new methodologies for estimating causal effects: one in statistics and one in econometrics.

---

2. See Currie, Kleven, and Zwiers (2020) for a documentation of these trends in economics.
3. See Hull, Kolesár, and Walters (2022) for additional context and references.
4. Evidence of the interdisciplinary nature of this research area is the fact that other prestigious prizes were awarded in 2022 explicitly for research in causal inference, including the BBVA Award to the computer scientist Judea Pearl and the Rousseeuw Prize for Statistics to the epidemiologists Jamie Robins, Thomas Richardson, Andrea Rotnitzky, Miguel Hernán, and Eric Tchetchgen Tchetchgen.

Curiously, in both disciplines, the explicit use of the term "causality" was relatively rare and often discouraged. In statistics, the dictum "correlation is not causality" kept most researchers from using the term outside of randomized controlled trials. Early on in economics, the term was used more widely (e.g., Tinbergen (1941)). That did not last, and the use of the term in either empirical or methodological work became increasingly rare. In the 1950s, a number of foundational studies proposed formal definitions of causality (*e.g.*, Wold (1954), Simon (1955) and references therein). Herman Wold took the position that "The concept of causality is indispensable and fundamental to all science" (abstract, Wold (1954)). He then tried to define it in the context of a relationship $Y = f(X) + \varepsilon$ and wrote that the "relationship is then defined as causal if it is theoretically permissible to regard the variables as involved in a fictive controlled experiment with [X] for cause variables and [Y] for effect variable" (p. 166, Wold (1954)). This somewhat vague definition did not catch on and Herbert Simon, the 1978 Laureate in economics, questioned "whether we wish to retain the word 'cause' in the vocabulary of science" (p. 54 in Hood *et al.* (1953)). Even though Simon argued in favor of doing so, the term causality remained out of favor in the economics literature. In his 2001 Prize lecture, Daniel McFadden argues that "detection of true causal structures is beyond the reach of statistics" and recommends that "For these reasons, it is best to avoid the language of causality" (p. 369, McFadden (2001)), despite analyzing what are clearly causal questions regarding the demand for public transportation under different scenarios.[5] It was not until the 1990s that the term started to gain currency in both empirical and methodological work in micro-economics, as documented in Currie et al. (2020), and also in statistics and other disciplines.

3. CAUSALITY IN THE STATISTICS LITERATURE: THE PRIMACY OF CHANCE

In the statistics literature, the initial focus was on inference for causal effects when the assignment to treatment was based entirely on chance. Both Fisher (1937) and Neyman (1923/1990) developed new methodologies for analyzing Randomized Controlled Trials (RCTs) where the assignment to treatment is completely random and known to the researcher. They focused on different questions. Fisher was primarily interested in testing sharp null hypotheses about causal effects.[6] The

5. During this time a different notion of causality, now typically referred to as Granger-Sims causality, based on presence or absence of predictive relationships, was proposed in the time-series literature, Granger (1969),

6. Sharp null hypotheses are hypotheses where there are no nuisance parameters, and the full distribution of all random variables is known under the null.

leading case considered by Fisher was the null hypothesis that there was no (causal) effect of the treatment whatsoever, against the alternative hypothesis that, for at least some units, there was some effect. Using a sharp null hypothesis allowed Fisher to infer the exact finite sample distribution for any test statistic (that is, any function of the data), under the randomization distribution (the distribution of the statistic induced by the randomization), given the null hypothesis. The most common example of such a test statistic is the difference in average outcomes by treatment status. Based on the randomization distribution, Fisher showed how to calculate the p-value: that is, the probability of observing a value for the statistic more extreme than, or as extreme as, the actual value of the statistic.

The use of p-values calculated in this way continues to be widely used in the analysis of RCTs. However, in recent years there has been increasing concern that such calculations do not answer the primary questions of interest for decision makers, which typically depend on magnitudes of the causal effects and not just their presence (*e.g.*, Wasserstein and Lazar (2016), Imbens (2021)). Even Fisher retreated somewhat from viewing them as central in his later work (Basu, 2011; Rubin, 1980).

Neyman, in line with these modern concerns about the relevance of p-value calculations, viewed the testing of sharp null hypotheses as "only of academic interest," leading to a falling-out with Fisher that was never resolved (Reid, 1998). Instead, Neyman focused on estimating average causal effects and constructing confidence intervals for them. Like Fisher, Neyman studied the properties of such procedures under the randomization distribution. The work by Fisher and Neyman continues to be the basis of the analysis of randomized experiments, not only in agricultural and biomedical settings, but also in online experimentation (Gupta et al. (2019)).

Subsequently, the causal literature in statistics has studied the design and analysis of experiments in more complex settings such as stratified, paired, clustered, and cross-over experiments (Imbens and Rubin (2015), Wu and Hamada (2011)). Currently, much of the interesting research in this area focuses on more complex experimental designs including adaptive (Dimakopoulou et al. (2017)) and multi-stage designs (Bajari, Burdick, Imbens, Masoero, McQueen, Richardson, and Rosen (2021)).

For much of the 20th century, formal discussion of causal inference in statistics was limited to testing and estimating causal effects in randomized experiments. It was primarily Rubin's work that made estimating causal effects in non-experimental (observational) studies a topic of independent research interest. One of Rubin's greatest contributions to this literature was to put the notion of *potential outcomes* center stage. In Rubin (1974), he casts the causal inference problem as one where we wish

to compare two (or more) potential outcomes, $Y_i(E)$ and $Y_i(C)$, defined for the same physical unit i. $Y_i(C)$ is the unit's outcome given exposure to the control treatment, and $Y_i(E)$ is the outcome for the same unit, at the same time, given exposure to the experimental treatment. Let the actual exposure for this unit be denoted by $W_i \in \{C,E\}$. Then the realized (and potentially observed) outcome is

$$Y_i = Y_i(W_i) = \begin{cases} Y_i(C) & \text{if } W_i = C, \\ Y_i(E) & \text{if } W_i = E. \end{cases}$$

The causal effect is the difference $Y_i(E) - Y_i(C)$ (or some other comparison of $Y_i(E)$ and $Y_i(C)$). A benchmark estimand is the average causal effect

$$\tau = \frac{1}{N}\sum_{i=1}^{N} (Y_i(E) - Y_i(C)),$$

where the number of units in the population is N. The causal inference problem now becomes a missing data problem: for the same unit we cannot observe the two potential outcomes, one given exposure to the experimental treatment and one given exposure to the control treatment. Holland (1986) referred to this as "the fundamental problem of causal inference," (p. 947, Holland (1986)), and it provides important links to the missing data literature (e.g., Little and Rubin (2019)).

When I first came across this potential outcome framework, it made a powerful impression on me. It forces the researcher to consider what manipulation could have revealed the *ex post* missing potential outcome. In some cases that is not easy. For example, when studying important societal questions regarding race or gender, it is often unclear what manipulation would allow one to causally interpret any statements about differences in economic outcomes by race or gender. As a result, it is not always clear how to think about causality in such settings. See the discussion between Holland (1986) and Granger (1986), and also in a medical setting, Amutah et al. (2021). A number of studies show that partial progress can be made by manipulating the perception, rather than the actual value, of race, gender, or other characteristics (e.g. Bertrand and Mullainathan (2004), Goldin and Rouse (2000)).

More generally, the potential outcome notation clarified to me the distinction between causal variables or treatments, which are arguments in the potential outcome functions; and covariates, attributes, or pre-treatment variables, which may be correlated with the potential outcomes, but which are not themselves causal and thus not arguments in the potential outcome function. In the econometric textbooks at that time, this distinction was not made. Instead a division was made between endogenous and

exogenous variables with no widely agreed-upon definitions. In that division, causal variables could be exogenous or endogenous, depending on the assignment mechanism.

Rubin's second contribution was the articulation of a key set of assumptions that moved away from completely random assignment. In collaboration with Paul Rosenbaum, Rubin focused on the case where assignment to treatment is not completely random, but conditional on some observed confounders, it can be viewed as random (Rosenbaum and Rubin (1983b)). This key assumption is referred to in various literatures and in various forms as unconfoundedness, ignorable treatment assignment, exogeneity, or selection on observables, and is closely related to the 'missing at random' assumption in the missing data literature (Little and Rubin (2019)). If the observed confounders for unit i are denoted by X_i, unconfoundedness corresponds to conditional independence (conditional on the observed confounders) of the treatment and the potential outcomes:

$$W_i \perp\!\!\!\perp (Y_i(C), Y_i(E)) \mid X_i. \quad (1)$$

Here the distinction between covariates/pre-treatment variables Xi and causes W_i is important. Unconfoundedness is an assumption on the assignment mechanism that determines W_i. It does not place any restrictions on the distribution of the covariates. In the earlier econometric literature, exogeneity-type assumptions would group together the causes and covariates, conflating the substantive assumptions on the assignment with the construction condition on the covariates.

Unconfoundedness, in combination with overlap in the covariate distributions[7] and sometimes with additional functional form assumptions, justifies a wide variety of statistical adjustment methods including (linear) regression, matching, inverse propensity score weighting, and doubly robust methods. See Imbens (2004) for a survey. The particular setting in (1) has become the workhorse model for causal inference and it has spawned vast empirical and methodological literatures. It has also been the leading example used to motivate developments in the semi-parametric literature. The methodological literature continues to study challenging cases involving high dimensional covariates using modern machine learning methods (Chernozhukov et al. (2017), Athey *et al.* (2018), Shi *et al.* (2019)).

7. Overlap means that for all values of the covariates, there are units in the treatment and control groups, or, more formally, that pr(Wi = E|Xi = x) ∈ (0, 1) for all values of x.

4. CAUSALITY IN THE ECONOMETRICS LITERATURE: THE PRIMACY OF CHOICE

The founders of the Econometric Society were also interested in estimating causal effects. However, they took a very different approach to this challenge. Whereas statisticians emphasized the chance aspect of the assignment and took randomized experiments as the starting point, early econometricians focused on settings where the values of the causes were determined by deliberate choices made by economic agents maximizing their utility under constraints. One canonical setting, for example in Tinbergen (1930) (see the translation in Hendry and Morgan (1997)), is that of a market where we observe prices and quantities. To understand the relationship between quantities and prices in a market, economists do not simply look at the correlation between quantities and prices, but rather start with a theoretical economic framework. In his editorial for the first issue of *Econometrica*, Frisch articulated this in a way that continues to resonate, perhaps now more than ever:

> "Statistical information is currently accumulating at an unprecedented rate. But no amount of statistical information, however complete and exact, can by itself explain economic phenomena. If we are not to get lost in the overwhelming, bewildering mass of statistical data that are now becoming available, we need the guidance and help of a powerful theoretical framework. Without this no significant interpretation and coordination of our observations will be possible." (p. 2, Frisch (1933))

In the case of markets, a theoretical framework of the type Frisch refers to comprises a model of supply, demand, and market equilibrium. For example, an early study, Tinbergen (1930), focused on the market for potato flour.[8] Here I use the example from my paper with Joshua Angrist and Kathryn Graddy (Angrist, Graddy, and Imbens (2000)) where we analyze the market for fish (specifically whiting) using data from the Fulton Fish Market collected by Graddy for her PhD thesis at Princeton University. The framework starts with an aggregate demand function, aggregated over all potential buyers coming to the fish market on a given day. For each individual buyer, the demand function comes from that buyer deciding how much fish they would be willing to buy as a function of the price, given their preferences and budget constraint. The aggregate demand function describes how much the buyers collectively are willing to buy at any given price. This setup very closely mirrors Rubin's potential

8. At the time this was an important commodity in The Netherlands, as illustrated by the theme of one of Van Gogh's most famous paintings, "the potato-eaters."

outcome framework (albeit with a continuously valued treatment), and the close connection may partly explain why the potential outcome framework found such a receptive audience in econometrics in the 1990s for studying general causal questions. However, to estimate the demand function, we cannot simply adjust for observed confounders and compare quantities at days with high and low prices, as an approach based on the unconfoundedness assumption would suggest. Prices are not set randomly, not even after conditioning on observed market characteristics. A more plausible framework adds a supply function that describes the quantity that sellers are willing to sell at any given price. A simple model for the assignment mechanism is, then, that prices are determined by market equilibrium: that is, by the intersection of the day-specific demand and supply functions. These ideas are all clearly articulated in Tinbergen (1930) and familiar from introductory econometrics textbooks. Tinbergen proceeds to use instrumental variables to estimate the slope of the (linear) demand function. Despite the lack of data (only eight observations!), the entire approach feels very modern.

Similarly, Haavelmo (1943) appears modern in the explicit causal interpretation of the estimands in his models in terms of a hypothetical experiment:

> "Assume that if the group of all consumers in society were repeatedly furnished with the total income, or purchasing power, r per year, they would, on the average, or 'normally,' spend a total amoun $\overline{u}$ for consumption per year, equal to
>
> $$\overline{u} = \alpha r + \beta,$$
>
> where α and β are constants." (Haavelmo (1943), p. 3)

Interestingly, Haavelmo's notion of the average or normal amount spent for different values of total income is very similar to what Neyman (Splawa-Neyman et al. (1990)) calls the "true yield" in an agricultural setting, although there is no reference to Neyman's work on the analysis of experiments in Haavelmo's work. In his Prize lecture (Haavelmo (1992)), Haavelmo acknowledged the influence of Neyman on his work, but this influence appears primarily in the probabilistic approach (e.g., Haavelmo (1944)), rather than in the area of causality.[9]

Over the years, this literature became the foundation of empirical work in economics. Much of the methodological work generalized the specific

9. Haavelmo met Neyman in 1936 in the UK, as well as during his visit to the United States in 1939, see Bjerkholt (2007).

settings considered by Tinbergen and Haavelmo to accommodate more complex theoretical models.

Over time, some of the clarity in the Tinbergen and Haavelmo work was lost. In the 1960s and 70s, theoretical econometricians had developed more general methods for simultaneous equations models with arbitrary numbers of endogenous and exogenous variables, for example in a common generic form:

$$YB + Z\Gamma = U,$$

where the Y represents endogenous variables and Z the pre-determined variables, either exogenous or lagged dependent variables, and the U represents unobserved error terms, with some restrictions on the unknown parameters B and Γ. In this generalization, the potential outcome notation that was clear in the Tinbergen and Haavelmo work was dropped, along with the explicit assignment mechanism, and omitted variable bias and true simultaneity were lumped together under the generic rubric of "endogeneity" that defied clear and unambiguous definitions. As Tinbergen wrote in his 1969 Prize lecture,

> "Sometimes indeed some of our followers overdo model building." (Tinbergen (1981), p. 18)

While the technical progress in this literature was impressive, it came at a substantial cost. Statisticians lost track of what this literature had to offer them. Phil Dawid complained, "I despair of ever understanding the logic of simultaneous equations well enough to tackle them," (p. 24, Dawid (1984)) in a comment on Pratt and Schlaifer (1984), and David Cox wrote that specification of causal models should satisfy conditions that "precludes the use of y_2 as an explanatory variable for y1 if at the same time y1 is an explanatory variable for y_2," (p. 294, Cox (1992)), ruling out a common specification of simultaneous equations models.

Empirical work using these highly technical methods also faced increasing skepticism within economics, leading to a widening gulf between econometricians and researchers doing empirical work. Hendry (2000) questioned the credentials of econometrics in a paper with the title "Econometrics: Alchemy or Science?" Edward Leamer, in a paper with the famous title "Let's Take the Con Out of Econometrics,"[10] bemoaned that the credibility of empirical work was at a low:

10. Recently, giving a guest lecture in my first-year econometrics class, Leamer suggested that an even better title could have been "Who put the trics into Econometrics."

> "This is a sad and decidedly unscientific state of affairs we find ourselves in. Hardly anyone takes data analyses seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analyses seriously. (p. 37, Leamer (1983))"

Leamer made a clear distinction between observational studies and studies using experimental data, with the latter not subject to his main concerns ("There is therefore a sharp difference between inference from randomized experiments and inference from natural experiments", p. 33, Leamer (1983)). He saw observational studies as the mainstay of empirical economics, with randomized experiments being rare. To illustrate his concerns, he compared various least squares regression estimates of the deterrence effect of the death penalty, using state/year murder rates as the outcome in a linear model setting, and concluded that by choosing different specifications for the regression models one could justify positive and negative effect estimates. To improve the credibility of empirical work, or at least to be clear about its limits, Leamer suggested making sensitivity analyses a more routine part of empirical work. Although there were earlier influential examples of such sensitivity analyses (Cornfield et al. (1959)), they were not then, nor are they now, a routine part of empirical work.[11]

Around the same time LaLonde (1986), with the less controversial title "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," made a point similar in spirit to the Leamer paper, but in a narrower setting. In a paper based on his PhD thesis supervised by Orley Ashenfelter, LaLonde took data from an experimental evaluation of a job training program, the National Supported Work (NSW) program. The program was effective, with a precisely estimated and substantial effect on subsequent annual earnings of around $850. LaLonde then introduced an extremely important validation exercise. He put aside the control group from the randomized experiment and tried to replicate the experimental results (specifically the $850, but also average effects for subsamples) using various non-experimental comparison groups constructed from public use surveys. To estimate the effects with these comparison groups he used a variety of state-of-the-art econometric methods relying on different *identification strategies.*[12] One can interpret the

11. Much interesting work has been done in this direction (Rosenbaum and Rubin (1983a), Imbens (2003), Andrews et al. (2017)). It continues to be an active area, with additional insights from the work on partial identification initiated by Manski (Manski (1990), Manski et al. (1992)). Nevertheless, it has not become as routine as Leamer might have hoped, or as it should be.
12. Strategies that describe how one can infer the causal effects of interest from observational studies. See for general discussion Angrist and Krueger (1999).

LaLonde exercise as the causal equivalent of the out-of-sample validation that is very common in the modern machine learning literature for comparing the performance of predictive estimation methods. The challenge in validating methods for estimating causal effects in observational studies, as opposed to validating predictive methods, is that this cannot be done without additional information or assumptions. Traditionally many econometric studies evaluated the performance of new methods *assuming* the identifying assumptions held. But of course, the real question is whether the maintained assumptions are at least approximately correct. To assess that question, one needs additional information to estimate the ground truth. In LaLonde's case, this comes in the form of a randomized experiment.

LaLonde concluded that the state-of-the-art methods did not deliver on their promise. In the abstract, he writes that:

> "This comparison [of experimental and observational methods] shows that many of the econometric procedures do not replicate the experimentally determined results, and it suggests that researchers should be aware of the potential for specification errors in other non-experimental evaluations." (p. 604, LaLonde (1986))

This is followed by the recommendation in the conclusion that:

> "... policymakers should be aware that the available nonexperimental evaluations of employment and training programs may contain large and unknown biases resulting from specification errors." (p. 617, LaLonde (1986))

Although subsequent research has questioned part of the conclusion by bringing in additional flexible methods based on unconfoundedness assumptions (see Dehejia and Wahba (1999)), LaLonde's paper was influential and led policymakers in the United States Congress to insist on the inclusion of experimental evaluation components in many labor market programs. It also played a role in motivating the work by Abhijit Banerjee, Esther Duflo and Michael Kremer that made experimentation a regular tool in the empirical economist's toolkit, leading to their Prize in 2019 (Banerjee (2020), Duflo (2020), Kremer (2020)).

In the long run, these two papers, Leamer (1983)) and LaLonde (1986), had a dramatic impact on empirical and methodological work in general, and both featured prominently in all three Prize lectures in 2021. Even though the two specific recommendations (sensitivity analyses as suggested by Leamer, and randomized experiments as suggested by LaLonde), had some immediate effect, the long-term impact went far

beyond that in clarifying the standards by which empirical work as well as econometric methods should be measured.

This long-term impact took two directions. First, the Princeton labor group led by Orley Ashenfelter took a route that was not suggested, explicitly or implicitly, in the Leamer and LaLonde papers. In what Joshua Angrist and Steve Pischke later called the "credibility revolution," the Princeton group focused on applications that were believable, and that not just the authors but also other researchers and policymakers would, in Leamer's words, "take seriously." But importantly, these studies were not based on explicit randomization in controlled experiments. Instead, they relied on *natural experiments*, exploiting idiosyncratic variation induced in the causes of interest. As my ten-year-old daughter said in an interview with the Stanford media team on October 11th, "you do experiments without actually doing your own experiments." Well-known examples of such natural experiments include Angrist (1990), Angrist and Krueger (1991), Card (1990), Ashenfelter and Krueger (1994), Meyer, Viscusi, and Durbin (1995), Card and Krueger (1994), Imbens, Rubin, and Sacerdote (2001).

Second, the Leamer and LaLonde papers inspired new methodological work with the goal of ensuring that empirical research would be taken seriously, and that it could be demonstrated to be credible. It is the econometric methods part of this credibility revolution to which my work with Joshua Angrist contributed.

6. CAUSAL INFERENCE AND THE CREDIBILITY REVOLUTION

Most of my career has been focused on developing econometric methods that enable empirical researchers to obtain credible estimates of causal effects to inform decision makers, in the spirit of Gary Chamberlain's view of econometrics as applied decision science (Chamberlain (2000)). The key themes of this research are transparency around the critical assumptions, understanding the limits of what data can credibly tell us, and making the research accessible to the broader social science community. Many of these methods take what David Card in his 2021 Prize lecture (Card (2022)) calls a *design-based* perspective, with an emphasis on understanding the assignment mechanism. These methods have helped bring together the statistics and econometrics traditions, partly through collaborations with Donald Rubin (*e.g.*, Imbens and Rubin (2015)). As a result, they have inspired research in the statistics literature on traditionally econometric topics like instrumental variables, and research in econometrics that build on traditional statistics topics such as matching methods and experimental design.

In this section I discuss three sets of papers that illustrate these themes. In the first, I discuss my work on local average treatment effects

with Joshua Angrist (Imbens and Angrist (1994)), later extended in (Angrist, Imbens, and Rubin (1996)) and summarized in Imbens (2014). This work built on the potential outcome framework developed in the statistics literature by Donald Rubin (Rubin (1974)) and combined it with traditional econometric ideas involving instrumental variables. In the second subsection I discuss my work with Joshua Angrist and Kathryn Graddy (Angrist, Graddy, and Imbens (2000)), which extended these ideas to the classic supply and demand models studied in Tinbergen (1930), and also Angrist and Imbens (1995) which extended the local average treatment effect ideas to the multi-valued treatment setting. The third paper is my most applied paper, Imbens, Rubin, and Sacerdote (2001), in which we estimated the effect of unearned income on labor earnings. This paper motivated additional methodological research on the problems related to estimating causal effects in settings with multi-valued and continuous treatments, extending the potential outcome literature from the binary treatment case and freeing up functional forms.

6.1. Local Average Treatment Effects

One of the early conversations I had with Joshua Angrist after I joined the Harvard economics department in 1990 was about his PhD thesis, published as Angrist (1990). In this paper Angrist is interested in the causal effect of serving in the military (denoted by W_i) on earnings (denoted by Y_i).13 The concern is that a simple comparison of earnings between veterans who served in the military, and non-veterans who did not serve, is not credible. Like the comparisons of murder rates in states with and without the death penalty in Leamer (1983), it would be difficult to convince readers that there is no omitted variable bias in a comparison of veterans and non-veterans, even if one controls for various observed characteristics of the individuals in an approach based on the unconfoundedness assumption. Angrist follows a different approach, in what became one of the canonical examples of a natural experiment. During the Vietnam War there was a compulsory draft, where draft priority within a birth year cohort was determined by a lottery tied to an individual's date of birth. Simplifying the procedure somewhat: think of the lottery as randomly assigning men born in a particular year to two groups, those who were draft-eligible and those who not, denoted by $Z_i \in \{0, 1\}$. Just as in a randomized experiment, these two groups are comparable *ex ante.* Of course, the causal effect of being draft-eligible is not itself a very interesting object. Instead, Angrist focused on the effect of *actually serving* in the military on earnings, using the indicator for draft

13. To simplify the discussion I will here use the notation Wi $\in \{0, 1\}$ rather than $W_i \in \{C,E\}$ as in Section 3.

eligibility as an econometric instrument. It was clear that the instrument did change the probability of serving in the military substantially, and so the instrument is correlated with the endogenous variable, and thus satisfies what Staiger and Stock (1997) called the *relevancy condition*. It also seemed plausible that it satisfied the *exclusion restriction,* that the only effect of being draft-eligible was through actually serving in the military, although I return to that assumption later. Even if we believe those two assumptions, it is not clear that they are sufficient to credibly estimate the effect of military service on earnings. On the one hand, from a classical textbook econometrics' perspective, this approach seems perfectly fine. In such a textbook version, and in line with Angrist (1990), one might write the economic model as

$$Y_i = \alpha + \tau \times W_i + \varepsilon_i,$$

with the instrument Z_i (a binary indicator for draft-eligibility) uncorrelated with ε_i so that the standard instrumental variables estimator

$$\hat{\tau} = \frac{\widehat{Cov}(Y_i, Z_i)}{\widehat{Cov}(W_i, Z_i)},$$

is a consistent estimator for τ. Two influential papers gave us pause regarding this argument: Heckman (1990) and Manski (1990). Both went beyond the textbook instrumental variables set up with constant effects and used something akin to the potential outcome framework that Rubin was advocating for in the statistics literature.

Let $(Y_i(0), Y_i(1))$, denote the two potential outcomes for earnings as a function of veteran status. This set up naturally allowed for general heterogeneity in the causal effect. Allowing for general heterogeneity was critical because an assumption that the causal effect was identical for all individuals was not credible. As Heckman wrote in his 2001 Prize lecture on microeconometrics, "Accounting for heterogeneity and diversity and its implications for economics and econometrics is ... a main theme" (p. 675, Heckman (2001)). However, Heckman and Manski both showed, in different ways, that in this setting it was not possible to identify the average effect of military service,

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)],$$

without additional assumptions. Heckman showed that identification of this average effect, which appeared to be a natural benchmark, required values z for the instrument, such that the probability of being a veteran, conditional on the instrument, $\text{pr}(W_i = 1|Z_i = z)$ was arbitrarily close to

zero and one. This clearly did not hold for the draft lottery application, where many men who were draft-eligible did not serve, and many who were not draft eligible did. At the same time, Manski derived large sample sharp bounds on the average effect of veteran status for this binary instrument case. These bounds would collapse into a single point only under the same condition as used by Heckman: that the probability of being a veteran, conditional on the instrument, took on values arbitrarily close to zero and one.

In my early discussions with Angrist, many on Saturday mornings in the local laundromat,[14] our initial, narrow goal was to reconcile the apparent credibility of the draft lottery example with the negative identification results in Heckman (1990) and Manski (1990). To do so, a key step was to go beyond the latent index models that were popular in the discrete choice literature at the time. In that approach, the decision to serve in the military would be modelled through a latent index crossing a threshold:

$$W_i = \begin{cases} 1 & \text{if } \gamma + \pi \times Z_i + \nu_i > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Although we initially worked within that traditional latent index framework, our then colleague at Harvard, Gary Chamberlain, suggested that it would improve transparency to remove what he called "the somewhat mysterious variable νi," and to use a potential outcome notation not just for the outcomes, but also for the decision to serve in the military. Here, the pair of potential treatment values,

$$W_i(0), W_i(1),$$

denotes whether a particular individual would serve if draft-eligible (the potential outcome $W_i(1)$), and whether they would serve if not draft-eligible (the potential outcome $W_i(0)$). This notation greatly clarified our argument and made clear that there are, in principle, four different types of individuals:[15]

		$W_i(0)$ (veteran status if not draft-eligible)	
		0	1
$W_i(1)$ (veteran status if draft-eligible)	0	never-taker	defier
	1	complier	always-taker

Table 1: Compliance Types

14. As junior faculty we were both living in Harvard faculty housing units that shared a laundromat.
15. A similar classification appears in Permutt and Hebel (1989).

Why is it useful to think about these four latent types? There are three advantages to this setup. First, the latent types clarify why we could not hope to estimate the overall average effect of military service, the result that Heckman and Manski had proved earlier. Never-takers by definition do not serve, so we cannot use these data to learn about what would happen if they did. Thus, it immediately implies that we cannot estimate the average effect for the never-takers, and as a result we cannot estimate the overall average effect. Similarly, we only see always-takers in the veteran state and cannot estimate what their earnings would be like if they did not serve.

Second, the setup allows us to assess more deeply the plausibility of the exclusion restriction: the assumption that the instrument, draft eligibility, does not have a direct effect on earnings, but only an indirect effect through military service. Third, the setup showed that we have a randomized experiment for compliers. Thus, we could directly estimate the average effect for compliers if only we could observe who these compliers were.

Let us examine the last two benefits in more detail. To assess the plausibility of the exclusion restriction that there is no direct effect of draft-eligibility on outcomes, consider separately the never-takers and always-takers. These are very different subpopulations, and the force of the exclusion restriction is very different for them. Consider an always-taker, who had already made up his mind, prior to the draft lottery, to volunteer for the armed services. There is no reason why he would even be interested in knowing his draft lottery number. The value of the lottery number does not affect his options or his choices at that point. For such an individual, the exclusion restriction appears highly plausible. Things are less clear for never-takers. Suppose a never-taker does not serve in the military, even if drafted, because of a medical exemption. In that case it is plausible that draft-eligibility is irrelevant and does not have a substantive effect on subsequent labor market outcomes. On the other hand, a never-taker could be someone who is in principle fit to serve but has strong preferences not to do so. If such a person has a favorable draft lottery number, they do not have to serve. However, if they have an unfavorable draft lottery number, they must take specific actions to stay out of the military. There is anecdotal evidence that people did this by moving to Canada, staying in school longer to get educational deferments, or even injuring themselves. If there are many such individuals, and if such actions affect the outcomes of interest, the exclusion restriction could be violated in a meaningful way. By defining the four compliance types, it is easier to assess the plausibility of the critical assumptions.

It was clear that, if the type of an individual was observed, we could simply partition the sample by type. Conditional on these latent types, unconfoundedness holds, and so we can give the comparison of outcomes for veterans and non-veterans of the same type a causal interpretation. However, even if we observed the type of each individual directly, we would not be able to estimate the treatment effect for never-takers and compliers, because never-takers (always-takers) are not observed in the treatment (control) group (though their observed outcomes could be used to estimate bounds). Nevertheless, if we were to observe the latent type, we could directly estimate the average effect of the treatment for compliers and defiers. The problem with this approach is that we cannot infer the type of an individual with certainty from the data on draft-eligibility, veteran status, and earnings.

To make progress, Angrist and I added one more assumption, the monotonicity condition. We assumed there were no defiers who do the opposite of their assignment: that is, there are no people who serve if they do not get drafted, but do not serve if they do get drafted. In that application monotonicity appears to be fairly plausible, and the fraction of defiers would appear to be small, if there are any such individuals. In other applications, monotonicity may be a controversial assumption. In particular monotonicity is less plausible in what are now called judge leniency designs, (*e.g.*, Kling (2006), Aizer and Doyle Jr (2015)).

In that class of applications, the instrument is the identity of a judge (or other screening agent) who changes the likelihood of sentencing (or other treatment, see also Example 2 in Imbens and Angrist (1994)). It is plausible that one of the judges is, on average, more lenient than another judge, making the instrument relevant. However, it is less plausible that monotonicity holds because that would require anyone whom the lenient judge sentences to also be sentenced by the stricter judge. It may well be that, despite being more lenient overall, the lenient judge cares more about some offenses than the strict judge.

Given the monotonicity assumption, we still cannot infer the type for all individuals, but we can infer it for some. As illustrated in Table 2, we can infer that someone who does not serve in the military despite being draft-eligible must be a never-taker. Similarly, individuals who serve despite not being draft-eligible must be always-takers. We cannot infer for sure whether someone who does not serve and who was not draft-eligible is a never-taker or a complier, and similarly for someone who serves and who is draft-eligible we cannot tell whether they are compliers or always-takers.

		Z_i (Draft-eligible) 0	1
W_i (Veteran)	0	complier/never-taker	never-taker
	1	always-taker	complier/always-taker

Table 2: Compliance Type by Treatment and Instrument given Monotonicity

The final step in this project involves disentangling the mixtures probabilistically. We know that the subsample that was not draft-eligible and that did not serve (individuals with $Z_i = 0$ and $W_i = 0$) is a mixture of compliers and never-takers. To infer the distribution of outcomes for compliers in this subsample, we use the fact that we can estimate the distribution of outcomes for never-takers by looking at a different subsample, namely the subsample of individuals who did not serve despite being draft-eligible, and that therefore consists of never-takers. This allows us to disentangle the mixture and estimate the distribution of non-veteran outcomes for compliers. Similarly, we can estimate the veteran outcomes for compliers using the other two subsamples.

Combining the estimates of the two distributions of veteran and non-veteran outcomes for compliers allowed us to estimate the average effect of veteran status for compliers, or what Angrist and I called the *Local Average Treatment Effect*. The Local Average Treatment Effect or "LATE" is a somewhat unusual estimand. Questions were raised initially regarding its relevance for policy and therefore its usefulness for empirical research.[16] The main concern at the time was that it appeared opportunistic. The standard approach to identification was to first state what the target estimand was, and then to articulate the identification strategy through assumptions that would allow one to identify that estimand.[17] Angrist and I turned this strategy around and introduced an alternative way to study identification questions. Rather than start with an estimand and ask if and how we could identify that, we started by asking what we could identify under reasonable assumptions. This led to the discovery that a popular estimation method, instrumental variables, did estimate a meaningful object, one with a clear causal interpretation. We then characterized what exactly that interpretation was. Because these assumptions were substantially weaker than those required for the identification of a benchmark estimand, such as the overall average treatment effect, this enabled researchers to interpret

16. See the discussion in Deaton (2010) and Imbens (2010).
17. If the assumptions that guaranteed point-identification were too strong, Manski (1990) argued for reporting bounds on the estimand.

their results under more plausible assumptions than they used previously.

Another, related, concern is that the LATE estimates the average effect for a subpopulation that cannot directly be identified. We do not know whether any given individual is a complier or not. Moreover, this subpopulation is indexed by the instrument, and so the LATE is not necessarily invariant to the assignment mechanism. However, we may be able to make some inferences about this subpopulation. If we observe characteristics, say age, for all individuals, we can estimate the average age for compliers, and thus learn something about the characteristics of this subpopulation, using the results in Abadie (2003).

A third concern is the policy relevance of the LATE estimand. The relevance varies by application. In the draft lottery example, one could argue that the average effect for the compliers, who were actually affected by the draft, is more interesting than, say, the average effect for never-takers or always-takers, or even than the overall average effect. Analogously, in the Angrist and Krueger (1991) study of the returns to education, the complier subpopulation is that of people who might be induced to stay in school a little longer by changing compulsory schooling laws. Again, that is likely to be a subpopulation of great interest to policymakers working on high school attendance. On the other hand, this LATE is clearly less informative about, say, policies that affect college attendance.

These concerns notwithstanding, the LATE is an important component in the decision process for policymakers, as it provides credible estimates of causal effects under transparent assumptions that are substantially weaker than those employed previously. Most importantly, it liberated part of the econometric causal literature from functional form and homogeneity assumptions. It also improved the understanding of popular estimation methods that allowed for more transparent communication with other disciplines.

6.2. Multi-valued Endogenous Variables

The LATE paper focused primarily on the simplest possible instrumental variables setting, a single binary endogenous variable, a single binary instrument, and no exogenous variables. That led to a paper that, in its published version, was very short and clear,[18] no doubt a part of its subsequent popularity. As a result, we did not consider many generalizations. This includes the case with covariates which did not appear to raise substantial conceptual issues. In the LATE paper, we did briefly discuss the case with multi-valued discrete instruments. At that time, we were not

18. This was not all our doing, but partly the result of the handling editor pushing us to shorten the paper substantially from its first submitted version, unquestionably improving it in the process.

particularly concerned with the extension to multiple distinct instruments; finding good instruments is rare enough that finding multiple distinct instruments is uncommon in empirical work. One special case where multiple instruments are common arises when the instruments are generated by interacting a single basic instrument with indicators for subpopulations. The leading example is Angrist and Krueger (1991), where indicators for quarter of birth were interacted with indicators for year and state of birth. That generated interesting theoretical work later, e.g., Bekker (1994), Staiger and Stock (1997). See Chamberlain and Imbens (2004) for a hierarchical random effects approach to such settings.

The most challenging case not considered in the LATE paper concerned multi-valued or multiple endogenous variables. We made some progress towards understanding this case in two subsequent papers. First, in Angrist and Imbens (1995), we focused on the setting analyzed in Angrist and Krueger (1991) which studied the effect of years of education on earnings using compulsory schooling laws as instruments (through the way these laws differentially affect people born in different parts of the year and in different states). Second, we teamed up with Kathryn Graddy (in Angrist, Graddy, and Imbens, 2000) to study the classical simultaneous equations setup in the form of a supply and demand setting, using the data Kathryn had collected for her PhD thesis at Princeton (Graddy, 1995) to illustrate our theoretical results.

The first part of the LATE setup with the potential outcomes does not change if the treatment takes on more than two values. The combination of the exclusion restriction and the assumed exogeneity of the instrument implies that we can think of a set of potential outcomes $Y_i(w)$ that are all independent of the instrument:

$$Z_i \perp\!\!\!\perp Y_i(w), \quad (2)$$

possibly after conditioning on some exogenous covariates. The monotonicity assumption is different, however. In Angrist and Imbens (1995) we extended the monotonicity condition from the binary case (maintaining the binary instrument setting) to the multi-valued treatment case by assuming

$$W_i(1) \geq W_i(0), \forall\, i, \quad (3)$$

(or $W_i(1) \leq W_i(0)$ for all units). In the compulsory schooling case, this makes sense: it is satisfied if tightening compulsory schooling laws increases the amount of schooling received or leave it unchanged, but

cannot decrease it.[19] However, there is an important difference with the binary treatment case. In the multi-valued treatment case, the monotonicity condition in (3) implies a stochastic dominance relation for the conditional distribution of treatments, given the two values for the instrument, leading to a number of testable inequality restrictions.[20] In contrast, in the binary treatment case there are no restrictions implied by this assumption.[21] Under these two assumptions (2) and (3), the same estimand as in the binary treatment case (the ratio of the covariance of the outcome and the instrument and the covariance of the treatment and the instrument), has an interpretation as a weighted average of average causal effects of unit increases in the treatment, weighted by the share of compliers for that level of the treatment:

$$\frac{Cov(Y_i,Z_i)}{Cov(W_i,Z_i)} = \sum_{w=1}^{J} \lambda_w \mathbb{E}[Y_i(w) - Y_i(w-1) | W_i(1) \geq w > W_i(0)], \quad (4)$$

With

$$\lambda_w = \frac{\text{pr}(W_i(1) \geq w > W_i(0))}{\sum_{j=1}^{J} \text{pr}(W_i(1) \geq j > W_i(0))},$$

This is obviously a more complex estimand than in the binary case, and it highlights the challenges when a single binary instrument is used to infer causal effects in a more complex environment with multi-valued treatments.[22]

In the second paper, Angrist, Graddy, and Imbens (2000), we focused on the classic problem in econometrics: disentangling supply and demand functions with data on quantities and prices over time. This problem goes back to Tinbergen (1930). In his study, Tinbergen was interested in estimating the demand for potato flour using data on prices and quantities of potato flour traded between the Netherlands and France (see Hendry and Morgan (1997) for additional discussion). Kathryn Graddy had collected data on prices and quantities from the Fulton Fish Market in New York.[23] In our joint paper we use just a subset of these data. See Graddy (1995)

19. By tightening the compulsory schooling laws, I mean changing them in a way that the schooling required from someone who wishes to leave as early as is legally allowed is weakly increased.
20. These restrictions are different from the ones on the outcome distributions implied by the instrumental variables set up in the binary treatment / binary instrument case that are discussed in detail in Kitagawa (2015).
21. There are some restrictions on the outcome distributions, see Kitagawa (2015).
22. Recent work has explored alternative extensions of the monotonicity condition to the multi-valued treatment setting in the case with multiple instruments, see Mogstad, Torgovitsky, and Walters (2021).
23. Collecting data in this case sounds much simpler than it was. Graddy directly collected information on prices and quantities of whiting for all transactions by one particular dealer throughout the early morning that the market was open.

for more details. For our purpose, quantities were aggregated daily, and a single price was calculated for each day. Simply looking at the correlation between the logarithm of quantities and the logarithm of prices obviously does not tell us very much. Instead, we imposed a standard supply and demand model. We assumed buyers would come to the market on day t with an aggregate demand function $D_t(p)$, describing the quantity they would be willing to purchase for any given price p. This part is similar to the structure Tinbergen (1930) used and naturally matches the Rubin potential outcome framework that by the time of our fish paper had become popular in the causal inference literature. Similarly, we postulated the existence of a supply function for each market, $S_t(p)$. We then assumed the price observed in market t was the equilibrium price that cleared the market:

$$P_t^E \; solves \; D_t(p) = S_t(p),$$

and the quantity observed in market t is the quantity corresponding to demand and supply at that equilibrium price, $Q_t^E = D_t(P_t^E) = S_t(P_t^E)$. None of this was new, and arguably is underlying the textbook simultaneous equations model. However, being explicit about the potential outcomes and the assignment mechanism (in this case the market clearing equilibrium condition) allowed us to make this accessible to statisticians.[24]

Now suppose we have an instrument that affects the supply function, but not the demand function. In Angrist, Graddy, and Imbens, 2000, following Graddy (1995), these are weather conditions at sea in the preceding days, specifically wave height and wind speed. Assume that these satisfy the instrument conditions: that they are both exogenous with respect to the unobserved components in the demand function, and do not directly affect demand, but do affect supply, all of which are plausible assumptions in this setting. In the case with a linear demand function, the textbook results on simultaneous equations imply that the instrumental variables estimator then delivers the slope of the demand function, the price elasticity of demand. More generally, we showed in Angrist, Graddy, and Imbens, 2000 that when demand functions are nonlinear and may differ between markets beyond an additive shift, an instrumental variables strategy identifies a weighted average of conditional demand elasticities similar to that in (4). The conditioning is on markets where the supply function is affected by the instruments, and the averaging is proportional to the effect of the instruments on the equilibrium price in each market.

24. When I presented this at a conference in Belgium, David Cox, who had earlier dismissed the logic simultaneous equations (Cox (1992)), commented that this was the first time he had understood what economists really meant by simultaneous equations models.

6.3. Unconfoundedness and Continuous Treatments

The third paper, or set of papers, I want to discuss originated in a graduate class Donald Rubin and I taught at Harvard in 1995. This was probably the first-ever graduate class devoted entirely to causal inference, although the registrar's office changed that to *casual inference*, leading to a much bigger turnout the first day than what we eventually ended up with. This class symbolized the convergence of the statistical and econometric traditions in causal inference, with students attending from both departments and getting exposure to early work from both literatures.

From the outset, we asked which thought experiment or conceptual manipulation would allow you to observe the counterfactual you need for a causal effect. At some point, the discussion in class focused on the causal effect of child poverty on the labor market and other life outcomes. We discussed the fact that many estimates in the literature essentially compared children growing up in middle-class families with children growing up in poor families. Obviously, the treatment of interest is *not* moving the children from poor to middle class families. A more relevant treatment is making the poor families better off by giving them more financial resources. This eventually led to a suggestion by Bruce Sacerdote (at the time a student in this class) to use the lottery as a natural experiment. The lottery can be thought of as randomly assigning yearly payments of substantial sums of money to individuals randomly selected from those buying lottery tickets. Of course, that is not quite the experiment we would have liked to have done, but it is close. The payments do not continue forever; they stop after twenty years. The population is not the general population, but only those buying lottery tickets. Comparisons with the Current Population Survey suggested that this latter concern was not a major one. In the paper that came out of this discussion,[25] Imbens, Rubin, and Sacerdote (2001), we ended up focusing on a slightly different outcome. Instead of studying the effect of unearned income on child poverty, we studied the effect on labor market outcomes.[26] Specifically the focus was on the propensity to earn out of unearned income, per dollar of unearned income. This measures how much, on average, people reduce their labor earnings for every dollar they win in the lottery. It is the subject of an extensive literature using observational data (Pencavel, 1986), with estimates ranging from −0.3 to 0.

Although collecting the data, with the help of the Massachusetts State Lottery Commission, was perhaps the biggest challenge in this project, it turned out there was still a need for careful statistical analysis despite the

25. Another influential paper that originated in that class was Dehejia and Wahba (1999).
26. The reason for changing the focus away from child poverty was that the data we collected had too few families with young children.

explicit randomization in the lottery. Comparing earnings in the year prior to winning the lottery for winners and losers (that is, individuals who won small one-time prizes) showed a difference in average earnings of −$3.5K, with a standard error of about $1.4K (winners earning less than losers). There were two main suspects for this difference. One was that the lottery does not randomize over individuals, it randomizes over tickets, and different individuals buy different numbers of tickets. On average, the winners reported having bought 4.6 tickets per week, compared to 2.2 for losers. The number of tickets bought may well be related to income. This did not turn out to be a big concern, with a regression of tickets bought on earnings leading to a coefficient of −0.78 with a standard error of 0.93. Second, the survey response rate was less than 50%. Non-response was clearly not random. We could in fact infer directly that the non-response was not random, because we knew the size of the prize won for all individuals, irrespective of whether they responded to the survey. Estimating a logistic regression of the indicator for response with the logarithm of the yearly prize as the regressor leads to a t-statistic of −3.5, showing that large winners were significantly less likely to respond to the survey. Given the clear evidence that in the sample the size of the prize was not completely random, we focused on analyses that adjust for the rich set of observable pre-winning variables we had collected, including six years of pre-winning annual earnings. This led to adjusted estimates for the marginal propensity to earn per dollar of unearned income, averaged over the six years post-winning, of −0.051, with a standard error of 0.014.

Is this credible as a causal estimate? In Leamer's words, should anyone take these numbers seriously? The answer is yes, and there are two reasons why. First, in the spirit of Leamer (1983) we investigated the sensitivity to a range of specifications and found the results to be robust. Second, and this is the recent type of analysis that is an integral part of the credibility revolution, we could look at *placebo analyses*. In the lottery example, we looked at the causal effect of prizes on earnings prior to winning. Of course, the causal effects on these (pseudo) outcomes are zero because people cannot anticipate winning the lottery. But given that we see a difference in average earnings between winners and losers prior to the lottery (−$3.5K, or 20%), the question is whether the statistical methods are effective in removing that difference. With the lottery data, we did find that the adjustment methods we used (excluding earnings in the year prior to winning) did remove essentially all the differences between winners and losers in the year prior to winning the lottery (see Imbens (2015)). Where the adjusted average difference between winners and losers postlottery was about −$5.00K (standard error 1.36K), in the year prior to winning the lottery the raw difference was -$3.50K, but the

adjusted difference was -$0.10K (standard error 0.95K), lending support to the claim that the statistical adjustment was effective in removing pre-lottery differences between winners and losers, and thus that the estimates for the post-lottery differences are credible causal estimates.

In the lottery paper, we used relatively simple adjustment methods based on least squares regression to estimate the propensity to earn out of unearned income, partly motivated by the relative similarity of the winners and losers samples. However, this led me to wonder whether the propensity score methodology developed in Rosenbaum and Rubin (1983b) could be extended to the setting with multi-valued treatments to estimate the dose-response function $\mathbb{E}[Y_i(w)]$, that is, the average of the potential outcomes as a function of the treatment. Recall that in the binary treatment case, the unconfoundedness assumption is formulated as

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i,$$

implying here that we can compare people with high and low lottery prizes within subpopulations with the same values for all covariates. This can be challenging if this assumption relies on the researcher having many covariates. The Rosenbaum-Rubin propensity score result states that one can eliminate all biases associated with the covariates by adjusting just for a scalar function of the covariates, the propensity score $e(x) = \mathrm{p}r(W_i = 1|X_i = \mathrm{x})$, irrespective of the number of covariates. In other words, unconfoundedness implies that

$$W_i \perp\!\!\!\perp Y_i(0),\ Y_i(1) \mid e(X_i),$$

where we condition only on the propensity score instead of the full covariate vector Xi. One can immediately extend the Rosenbaum-Rubin result to the multi-valued case where $W_i \in \{0, 1, \ldots, J\}$ by assuming the multiple treatment version of the unconfoundedness assumption:

$$W_i \perp\!\!\!\perp Y_i(0), Y_i(1), \ldots, Y_i(\mathrm{J}) \mid X_i. \quad (5)$$

This in turn implies that

$$W_i \perp\!\!\!\perp Y_i(0), Y_i(1), \ldots, Y_i(\mathrm{J}) \mid \mathrm{p}r(Y_i = 1|X_i), \ldots, \mathrm{p}r(W_i = J|X_i). \quad (6)$$

However, this *strong unconfoundedness* result requires conditioning on J propensity scores, rather than a scalar one as in the binary case. Especially in settings where the cardinality of the treatment is substantial, this result does not reduce the dimension of the problem meaningfully. This

puzzled me at the time. It seemed counterintuitive that increasing the number of treatment values would fundamentally change the nature of the problem and make it much harder statistically. In 1998, Alan Krueger invited me to give a seminar at Princeton, and during that visit I spent a day sitting in the Industrial Relations Section thinking about this problem. Perhaps inspired by the location where so many contributions to the credibility revolution had been conceived, I realized that one can simplify the problem of estimating average treatment effects to one that requires adjusting only for a scalar score even when $J \geq 2$. A key step is the insight that, for estimation of the average treatment effects or the dose-response function, we do not need independence of the treatment and the full set of J potential outcomes. Instead, it suffices to have a weaker form, what I labeled *weak unconfoundedness* in Imbens (2000), for one treatment level and the corresponding potential outcome at a time:

$$\mathbf{1}_{W_i=w} \perp\!\!\!\perp Y_i(w) \mid X_i, \forall w.$$

Although formally weaker, this assumption is not substantively weaker than (5). However, weak unconfoundedness has the advantage that it implies a propensity score type result where, just as in the original Rosenbaum-Rubin result, we only need to adjust outcomes for a scalar covariate:

$$\mathbf{1}_{W_i=w} \perp\!\!\!\perp Y_i(w) \mid r(w, X_i), \forall w.$$

where $r(w, x) = \text{pr}(W_i = w|X_i)$ is the *generalized propensity score.* The key is that the conditioning variable, $r(w, X_i)$ is different for different levels of the treatment. How can we use this result given knowledge of this generalized propensity score to estimate the dose-response function $\mathbb{E}[Y_i(w)]$? Imbens (2000) proposed a two-stage procedure. First, estimate the conditional mean function

$$\beta(w, r) = \mathbb{E}[Y_i|W_i = w, r(W_i, X_i) = r],$$

with two scalar arguments, the realized treatment wi and the probability of receiving the treatment actually received, r(Wi,Xi). This function is not causal in the sense that β(w, r) – β(w', r) does not have an interpretation as an average causal effect. However, it can be used to estimate average potential outcomes through the equality

$$\mathbb{E}[Y_i(w)] = \mathbb{E}[\beta(w, r(w, X_i)],$$

where we evaluate $r(w, X_i)$ at w, not at the realized treatment value Wi. An alternative to this two-stage procedure is to reweight the observations, exploiting the equality

$$\mathbb{E}[Y_i(w)] = \mathbb{E}\left[\frac{Y_i \mathbf{1}_{W_i=w}}{r(W_i, X_i)}\right].$$

Both estimation approaches readily extend to the continuous treatment case. Hirano and Imbens (2004) illustrate these approaches to estimating causal effects in using the lottery data where the continuous treatment is the lottery prize. They consider parametric models for the generalized propensity score $r(w, x)$ (the conditional distribution of the treatment (the lottery prize) given the covariates). In addition they use flexible regression models for the conditional mean function $\beta(w, r)$ involving higher order moments and find that the results are robust to changes in the specification.

7. LOOKING AHEAD

After these discoveries in the 1990s to the early 2000s, causal inference continues to be an exciting area. Researchers in a number of fields are developing new methods for credibly learning about causal effects in observational settings. They also continue to propose new designs for experiments that go beyond the simple treatment/control group experiments or A/B tests for which Fisher and Neyman laid the groundwork. Here I want to highlight three areas where exciting progress is being made, and where important challenges remain. All are characterized by the strong connections between empirical and methodological research that motivated the collaboration between Joshua Angrist and myself in the early 1990s.

7.1. Synthetic Control Methods and Difference-In-Differences

One particularly interesting strand of research is that on *synthetic control* methods, initiated by Abadie and Gardeazabal (2003) and Abadie et al. (2010), and the closely related recent work on *difference-in-differences* (see Roth et al. (2022) for a recent survey). In one of the canonical examples, Abadie et al. (2015), the authors study the causal effect of German re-unification on West German Gross Domestic Product using data on GDP for West Germany and other countries both before and after the German re-unification. Traditionally, researchers might have employed regression methods, or matching methods where West Germany would be compared to another country, or a simple average of other countries. It is not clear

that such approaches would be satisfactory. Even taking out country-specific averages in a difference-in-differences strategy is unlikely to be credible, given the substantial heterogeneity that is likely to be present in differences between countries, both in trends and in levels. Abadie and coauthors proposed constructing a *synthetic* version of West Germany as a convex combination of the other countries. In practice this leads to a much better match for West Germany than any single other country. The basic synthetic control method as well as various extensions (Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021), Ben-Michael, Feller, and Rothstein (2021)), have in a short time found many applications in a wide range of fields, including studies of the effects of Brexit and the effects of country- or state-level COVID-19 policies.

7.2. Interactions, Spillovers and Peer Effects

A second exciting area with much ongoing work has focused on settings with substantial interactions between units. This is important in settings with infectious diseases in epidemiology, but also in labor markets and other marketplaces. For example, the effect of a training program on a particular individual may well depend on the fraction of treated individuals in their labor market. Crépon et al. (2013) illustrate this in an experimental evaluation consisting of multiple experiments in separate labor markets. Concerns about interactions are generally important in economic settings where individuals interact strategically in marketplaces. Understanding how to address these interactions in randomized experiments and in observational studies is at the core of a rapidly developing literature. One strand of this literature has focused on the design of experiments in settings with interactions (Bajari et al. (2021)) and the analysis of such experiments (Carrell, Sacerdote, and West (2013)). A second strand has focused on estimating peer effects in observational studies (Manski (1993)).

7.3. Causality and Computer Science

Third, there is a fast-growing literature in computer science on causality. Taking some of its cues from the path analyses developed by Wright (1920) (see Tinbergen (1940) for another early example of a graphical model) the work by Judea Pearl and others (Pearl (1995, 2000), Bareinboim and Pearl (2016), Richardson and Robins (2013), Peters et al. (2017)) uses *Directed Acyclic Graphs* (DAGs) and *Structural Causal Models* (SCMs) to study identification issues for causal questions. These methods have yet not caught on in econometrics as much as in other disciplines. The DAG approach has, potentially, two distinct benefits to offer to research-

ers. The first is primarily pedagogical, by formulating the critical assumptions in a form that may capture naturally the way some researchers think of causal relationships. DAGs can be a powerful way of illustrating the key assumptions underlying causal models, the same way that path analyses and arrow schemes in (Wright, 1928, 1934, Tinbergen, 1940) did earlier. A second potential benefit is the mathematical tools developed in the recent DAG literature. For example, the do-calculus developed by Pearl (2000) can be used by researchers to answer causal questions in a novel way. This second benefit is particularly relevant for questions in complex models with a large number of distinct components. One concern is that, in such settings, credible causal inference is particularly challenging irrespective of the methods. That concern in the economics literature led to the 1980s credibility crisis in econometrics. See Imbens (2020) for additional discussion of the relation of the graphical causal literature with the potential outcome literature.

A different part of causal inference literature where computer science ideas have made an impact has focused on improved methods for estimating causal effects in more traditional settings using modern machine learning methods. Here the use of supervised learning methods such as deep neural nets and random forests have proven helpful in improving methods for estimating average treatment effects (*e.g.*, Chernozhukov et al. (2017)), and for estimating heterogeneous treatment effects and treatment policies (*e.g.*, Wager and Athey (2018), Athey and Wager (2021)). In addition, generative adversarial networks and reinforcement learning methods are making inroads into econometrics (see Athey *et al.* (2021) and Chen (2022)).

8. CONCLUSION

The research on causality and causal inference recognized by the 2021 Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel continues to rapidly evolve and influence many traditional academic disciplines. Its impact has been felt in many substantive areas where randomized experiments are difficult to implement. It is my hope that the award will bring more young researchers into this area and lead them to develop and apply methods that will enable policymakers to make more informed decisions, or in Tinbergen's words from his 1969 Prize lecture, to find "ways of influencing actual development in some desired direction" (p. 17 Tinbergen (1981)).

REFERENCES

Abadie, Alberto (2003): "Semiparametric instrumental variable estimation of treatment response models," *Journal of econometrics,* **113**, 231–263. [21]

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010): "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American statistical Association,* **105**, 493–505. [31]

— (2015): "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 495–510. [32]

Abadie, Alberto and Javier Gardeazabal (2003): "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review,* **93**, 113–132. [31]

Aizer, Anna and Joseph J Doyle Jr (2015): "Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges," *The Quarterly Journal of Economics*, **130**, 759–803. [19]

Amutah, Christina, Kaliya Greenidge, Adjoa Mante, Michelle Munyikwa, Sanjna L Surya, Eve Higginbotham, David S Jones, Risa Lavizzo-Mourey, Dorothy Roberts, Jennifer Tsai, et al(2021): "Misrepresenting race – the role of medical schools in propagating physician bias," *New England Journal of Medicine,* **384**, 872–878. [6]

Andrews, Isaiah, Matthew Gentzkow, and Jesse M Shapiro (2017): "Measuring the sensitivity of parameter estimates to estimation moments," *The Quarterly Journal of Economics,* **132**, 1553–1592. [11]

Angrist, Joshua D (1990): "Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records," *American Economic Review,* 313–336. [13, 14, 15]

Angrist, Joshua D, Kathryn Graddy, and Guido W Imbens (2000): "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish," *The Review of Economic Studies,* **67**, 499–527. [8, 14, 23, 24, 25, 26]

Angrist, Joshua D and Guido W Imbens (1995): "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American Statistical Association,* **90**, 431– 442. [14, 23]

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin (1996): "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association,* **91**, 444–455. [14]

Angrist, Joshua D and Alan Krueger (1991): "Does Compulsory Schooling Affect Schooling and Earnings," *Quarterly Journal of Economics,* **CVI**, 979–1014. [13, 22, 23]

Angrist, Joshua D and Alan B Krueger (1999): "Empirical strategies in labor economics," in *Handbook of labor economics*, Elsevier, vol. 3, 1277–1366. [12]

Arkhangelsky, Dmitry, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager (2021): "Synthetic difference-in-differences," *American Economic Review*, **111**, 4088–4118. [32]

Ashenfelter, Orley and Alan Krueger (1994): "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review,* **84**, 1157–1173. [13]

Athey, Susan, Guido W Imbens, Jonas Metzger, and Evan Munro (2021): "Using wasserstein generative adversarial networks for the design of monte carlo simulations," *Journal of Econometrics*. [34]

Athey, Susan, Guido W Imbens, and Stefan Wager (2018): "Approximate residual balancing: debiased inference of average treatment effects in high dimensions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* **80**, 597–623. [7]

Athey, Susan and Stefan Wager (2021): "Policy learning with observational data," *Econometrica*, **89**, 133– 161. [34]

Bajari, Patrick, Brian Burdick, Guido W Imbens, Lorenzo Masoero, James Mcqueen, Thomas Richardson, and Ido M Rosen (2021): "Multiple Randomization Designs," arXiv preprint arXiv:2112.13495. [5, 32]

Banerjee, Abhijit Vinayak (2020): "Field experiments and the practice of economics," *American Economic Review,* **110**, 1937–51. [13]

Bareinboim, Elias and Judea Pearl (2016): "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Sciences,* **113**, 7345–7352. [33]

Basu, Debabrata (2011): "Randomization analysis of experimental data: the Fisher randomization test," *Selected Works of Debabrata Basu,* 305–325. [4]

Bekker, Paul A (1994): "Alternative approximations to the distributions of instrumental variable estimators," *Econometrica: Journal of the Econometric Society,* 657–681. [23]

Ben-Michael, Eli, Avi Feller,and Jesse Rothstein (2021): "The augmented synthetic control method," *Journal of the American Statistical Association,* **116**, 1789–1803. [32]

Bertrand, Marianne and Sendhil Mullainathan (2004): "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American Economic Review*, **94**, 991– 1013. [6]

Bjerkholt, Olav (2007): "Writing "the probability approach" with nowhere to go: Haavelmo in the United States, 1939–1944," *Econometric Theory,* **23**, 775–837. [9]

Card, David (1990): "The Impact of the Mariel Boatlift on The Miami Labor Market," *Industrial and Labor Relation,* **43**, 245–257. [13]

— (2022): "Design-Based Research in Empirical Microeconomics," *American Economic Review,* **112**, 1773– 81. [14]

Card, David And Alan Krueger (1994): "Minimum Wages and Employment: Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, **84**, 772–793. [13]

Carrell, Scott E, Bruce I Sacerdote, and James E West (2013): "From natural variation to optimal policy? The importance of endogenous peer group formation," *Econometrica*, **81**, 855–882. [32]

Chamberlain, Gary (2000): "Econometrics and decision theory," *Journal of Econometrics*, **95**, 255–283. [14]

Chamberlain, Gary And Guido Imbens (2004): "Random Effects Estimators with many Instrumental Variables," *Econometrica*, **72**, 295–306. [23]

Chen, Jiafeng (2022): "Synthetic Control As Online Linear Regression," arXiv preprint arXiv:2202.08426. [34]

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey (2017): "Double/debiased/Neyman machine learning of treatment effects," *American Economic Review*, **107**, 261–65. [7, 33]

Cornfield, Jerome, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, And Ernst L Wynder (1959): "Smoking and lung cancer: recent evidence and a discussion of some questions," *Journal of the National Cancer Institute,* 22, 173–203. [11]

Cox, David R (1992): "Causality: some statistical aspects," *Journal of the Royal Statistical Society: Series A (Statistics in Society),* **155**, 291–301. [10, 25]

Cr.Pon, Bruno, Esther Duflo, M. Gurgand, R. Rathelot, and P. Zamora (2013): "Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment," *Quarterly Journal of Economics*, **128**, 531–580. [32]

Currie, Janet, Henrik Kleven, And Esm.E Zwiers (2020): "Technology and big data are changing economics: Mining text to track methods," in *AEA Papers and Proceedings*, vol. **110**, 42–48. [2, 3]

Dawid, Ap(1984): "Comment: Causal inference from messy data," *Journal of the American Statistical Association*, **79**, 22–24. [10]

Deaton, Angus (2010): "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, **48**, 424–455. [21]

Dehejia, Rajeev H And Sadek Wahba (1999): "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs," *Journal of the American Statistical Association,* **94**, 1053–1062. [12, 27]

Dimakopoulou, Maria, Zhengyuan Zhou, Susan Athey, and Guido Imbens (2017): "Estimation considerations in contextual bandits," arXiv preprint arXiv:1711.07077. [5]

Duflo, Esther (2020): "Field experiments and the practice of policy," *American Economic Review*, **110**, 1952– 73. [13]

Fisher, Ronald Aylmer (1937): *The design of experiments*, Oliver and Boyd; Edinburgh; London. [4]

Frisch, Ragnar (1933): "Editorial," *Econometrica*, 1–4. [8]

Goldin, Claudia And Cecilia Rouse (2000): "Orchestrating impartiality: The impact of 'blind' auditions on female musicians," *American Economic Review*, **90**, 715–741. [6]

Graddy, Kathryn (1995): "Testing for imperfect competition at the Fulton fish market," *The RAND Journal of Economics,* 75–92. [23, 25]

Granger, Cwj (1986): "Comment on Statistics and causal inference by P. Holland," *Journal of the American Statistical Association*, **81**, 967–968. [6]

Granger, Clive Wj (1969): "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, 424–438. [3]

Gupta, Somit, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al (2019): "Top challenges from the first practical online controlled experiments summit," ACM SIGKDD *Explorations Newsletter*, **21**, 20–35. [5]

Haavelmo, Trygve (1943): "The statistical implications of a system of simultaneous equations," *Econometrica: Journal of the Econometric Society*, 1–12. [9]

— (1944): "The probability approach in econometrics," *Econometrica: Journal of the Econometric Society,* iii–115. [9]

— (1992): "Econometrics and the welfare state," *Economic Sciences, 1981-1990: The Sveriges Riksbank (Bank of Sweden) Prize in Economic Sciences in Memory of Alfred Nobel,* **2**, 261. [9]

Heckman, James (1990): "Varieties of selection bias," *American Economic Review*, **80**, 313–318. [16, 17]

Heckman, James J (2001): "Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture," J*ournal of Political Economy*, **109**, 673–748. [16]

Hendry, David F (2000): *Econometrics: alchemy or science?: Essays in econometric methodology,* Oxford University Press on Demand. [10]

Hendry, David F and Mary S Morgan (1997): *The foundations of econometric analysis,* Cambridge University Press. [8, 25]
Hirano, Keisuke and Guido W Imbens (2004): "The propensity score with continuous treatments," *Applied Bayesian modeling and causal inference from incomplete-data perspectives,* 226164, 73–84. [31]
30 Holland, Paul W (1986): "Statistics and causal inference," *Journal of the American Statistical Association,* **81**, 945–960. [6]
Hood, William C, Tjalling Charles Koopmans, et al. (1953): *Studies in econometric method*, John Wiley. [3]
Hull, Peter, Michal Koles.R, And Christopher Walters (2022): "Labour by Design: Contributions of David Card, Joshua Angrist, and Guido Imbens," arXiv preprint arXiv:2203.16405. [2]
Imbens, Guido (2000): "The Role of the Propensity Score in Estimating Dose–Response Functions," *Biometrika*, **87**, 706–710. [30]
— (2004): "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics,* 1–29. [7]
— (2010): "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature,* 399–423. [21]
— (2014): "Instrumental Variables: An Econometrician's Perspective," *Statistical Science,* 323–358. [14]
Imbens, Guido W (2003): "Sensitivity to exogeneity assumptions in program evaluation," *The American Economic Review, Papers and Proceedings,* **93**, 126–132. [11]
— (2015): "Matching methods in practice: Three examples," *Journal of Human Resources,* **50**, 373–419. [28]
— (2020): "Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics," *Journal of Economic Literature*, **58**, 1129–79. [33]
— (2021): "Statistical significance, p-values, and the reporting of uncertainty," *Journal of Economic Perspectives*, **35**, 157–74. [4]
Imbens, Guido W and Joshua D Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, **61**, 467–476. [14, 20]
Imbens, Guido W and Donald B Rubin (2015): *Causal Inference in Statistics, Social, and Biomedical Sciences,* Cambridge University Press. [5, 14]
Imbens, Guido W, Donald B Rubin, and Bruce I Sacerdote (2001): "Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players," *American Economic Review,* 778–794. [13, 14, 27]
Kitagawa, Toru (2015): "A test for instrument validity," *Econometrica*, **83**, 2043–2063. [24]
Kling, Jeffrey R (2006): "Incarceration length, employment, and earnings," *American Economic Review*, **96**, 863–876. [19]
Kremer, Michael (2020): "Experimentation, innovation, and economics," *American Economic Review*, **110**, 1974–94. [13]
Lalonde, Robert J (1986): "Evaluating the econometric evaluations of training programs with experimental data," *American Economic Review,* 604–620. [11, 12, 13]
Leamer, Edward E (1983): "Let's take the con out of econometrics," *American Economic Review,* **73**, 31–43. [11, 13, 15, 28]
Little, Roderick Ja and Donald B Rubin (2019): *Statistical analysis with missing data,* vol. **793**, Wiley. [6, 7]

Manski, Charles (1993): "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies,* **60**, 531–542. [33]
Manski, Charles F (1990): "Nonparametric bounds on treatment effects," *American Economic Review*, **80**, 319–323. [11, 16, 17, 21]
Manski, Charles F, Gary D Sandefur, Sara Mclanahan, and Daniel Powers (1992): "Alternative estimates of the effect of family structure during adolescence on high school graduation," *Journal of the American Statistical Association,* **87**, 25–37. [11]
Mcfadden, Daniel (2001): "Economic choices," *American Economic Review,* **91**, 351–378. [3]
Meyer, Bruce D, W Kip Viscusi, and David L Durbin (1995): "Workers' compensation and injury duration: evidence from a natural experiment," A*merican Economic Review*, 322–340. [13]
Mogstad, Magne, Alexander Torgovitsky, and Christopher R Walters (2021): "The causal interpretation of two-stage least squares with multiple instrumental variables," *American Economic Review*, **111**, 3663–98. [24]
Neyman, Jerzey (1923/1990): "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Statistical Science,* **5**, 465–472. [4]
Pearl, Judea (1995): "Causal diagrams for empirical research," *Biometrika*, **82**, 669–688. [33]
— (2000): *Causality: Models, Reasoning, and Inference*, New York, NY, USA: Cambridge University Press. [33]
Pencavel, John (1986): "Labor supply of men: a survey," *Handbook of labor economics*, **1**, 3–102. [27]
Permutt, Thomas and J Richard Hebel (1989): "Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight," *Biometrics*, 619–622. [17]
Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017): *Elements of causal inference: Foundations and learning algorithms,* MIT press. [33]
Pratt, John W and Robert Schlaifer (1984): "On the nature and discovery of structure," *Journal of the American Statistical Association*, **79**, 9–21. [10]
Reid, Constance (1998): *Neyman*, Springer Science & Business Media. [4]
Richardson, Thomas S and James M Robins (2013): "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality," *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper,* 128, 2013. [33]
Rosenbaum, Paul R and Donald B Rubin (1983a): "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome," *Journal of the Royal Statistical Society. Series B (Methodological),* 212–218. [11]
— (1983b): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, **70**, 41–55. [6, 29]
Roth, Jonathan, Pedro Hc Sant'anna, Alyssa Bilinski, and John Poe (2022): "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature," arXiv preprint arXiv:2201.01194. [31]
Rubin, Donald B (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology,* 66, 688. [5, 14]
— (1980): "Randomization analysis of experimental data: The Fisher randomization test comment," *Journal of the American Statistical Association*, **75**, 591–593. [4]

Shi, Claudia, David Blei, and Victor Veitch (2019): "Adapting neural networks for the estimation of treatment effects," *Advances in neural information processing systems*, 32. [7]

Simon, Herbert A (1955): "Causality and econometrics: comment," *Econometrica: Journal of the Econometric Society,* 193–195. [3]

Sims, Christopher A (1972): "Money, income, and causality," *American Economic Review,* **62**, 540–552. [3]

Splawa-Neyman, Jerzy, Dorota M Dabrowska, and TP Speed (1990): "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." *Statistical Science*, 465–472. [9]

Staiger, Douglas and James H Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica,* **65**, 557–586. [15, 23]

Tinbergen, Jan (1930): "Determination and interpretation of supply curves: an example," *Zeitschrift für Nationalokonomie,* **1**, 669–679. [8, 9, 14, 24, 25]

— (1940): "Econometric business cycle research," *Review of Economic Studies,* **7**, 73–90. [33]

— (1941): "Econometrie," . [3]

— (1981): "The use of models: experience and prospects," *American Economic Review*, **71**, 17–22. [10, 34]

Wager, Stefan and Susan Athey (2018): "Estimation and inference of heterogeneous treatment effects using random forests," J*ournal of the American Statistical Association,* **113**, 1228–1242. [34]

Wasserstein, Ronald L And Nicole A Lazar (2016): "The ASA statement on p-values: context, process, and purpose". [4]

Wold, Herman (1954): "Causality and econometrics," *Econometrica: Journal of the Econometric Society,* 162–177. [3]

Wright, Philip G (1928): *Tariff on animal and vegetable oils,* Macmillan Company, New York. [33]

Wright, Sewall (1920): "The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs," *Proceedings of the National Academy of Sciences*, **6**, 320–332. [33]

— (1934): "The method of path coefficients," *Annals of Mathematical Statistics,* **5**, 161–215. [33]

Wu, Cf Jeff And Michael S Hamada (2011): E*xperiments: planning, analysis, and optimization,* vol. **552**, John Wiley & Sons. [5]