Chapter 6
Treatment Effects

Robert P. Lieli

Abstract This chapter outlines the emergence of the treatment effects framework in econometrics and its relationship to the older structural equation tradition. It starts from a simple question: what gives a regression coefficient a causal interpretation? The structural equation approach answered this question by embedding regressions in larger models of economic behavior. Identification then relied on exclusion restrictions, functional form assumptions, and behavioral assumptions that were difficult to defend. The potential outcome framework offered a different and more transparent approach. It made causal analysis start with an explicit comparison between potential outcomes, a clearly defined estimand, and an identification argument. The chapter follows this shift from the statistical roots of potential outcomes through their migration into applied econometrics, also covering modern developments such as machine learning aided causal inference, heterogeneous treatment effects, panel data models, external validity, and so on. The central claim is not that treatment effects replaced structural econometrics. Rather, treatment effects reshaped the language of causal inference by making the connection between assumptions, empirical comparisons, and causal parameters explicit, while structural econometrics remains essential for questions involving mechanisms, equilibrium and counterfactual policy regimes outside the support of observed data.

6.1 Introduction

What gives a linear regression model a causal interpretation? This is a surprisingly hard question that econometrics textbooks had struggled to answer transparently before the potential outcome framework and the associated treatment effect vocabulary became part of the standard discourse in the field (Angrist & Pischke, 2017). The conceptual difficulties are rooted in the fact that the linear regression model is

Robert P. Lieli ✉
Central European University, Vienna, Austria e-mail: LieliR@ceu.edu

traditionally posited as a relationship between actually observed values of variables without an explicit reference to an intervention or counterfactuals. Given a dependent variable Y and independent variables $X_1, \ldots, X_k$, it is always possible to write Y as a linear combination of $X_1, \ldots, X_k$ plus an additive error term or, more precisely, a residual ϵ that is uncorrelated with all the X variables. This representation is called the linear projection of Y on $X_1, \ldots, X_k$ and it exists (and is unique) as long as the second moment matrix of the X variables is nonsingular. Technically, it is the coefficients in this linear projection that OLS consistently estimates under general conditions (e.g., Hamilton, 1994, Ch. 4.1).

The somewhat tautological but honest answer then is that regression coefficients have a causal interpretation if one *chooses* to interpret them causally. The first step toward this interpretation is to regard the X variables not just as covariates but determinants of Y, and ϵ not just as the residual from a projection, but as a combination of unobserved determinants omitted from the regression (as well as pure random variation). By 'determinants' we mean variables whose external manipulation, if only via a thought experiment, is expected to bring about a change in the value of Y while holding the other determinants constant.

If the error term is given such a 'structural' interpretation, then the coefficients in the model no longer necessarily coincide with the linear projection coefficients because the omitted variables could be correlated with the included regressors $X_1, \ldots, X_k$. Instead, one can then regard the coefficient on, say, X_1 as representing the partial effect of the variable X_1 on Y holding $X_2, \ldots, X_k$ and any other relevant factors, buried in ϵ, constant. The problem is that the latter is impossible to do in the data, since, by definition, there are no observations on the variables that are part of ϵ. In a large data set one could, in principle, compare data points that are different in terms X_1 and similar in terms of $X_2, \ldots, X_k$, but these data points may also have systematically different values of ϵ associated with them. However, if the researcher can argue that there is no systematic relationship between the X variables and ϵ (e.g., because nothing important is omitted from the regression model) then ϵ represents only 'pure' noise, and not being able to hold it constant does not matter. In other words, the coefficients in the model can again be identified as linear projection coefficients and hence they are consistently estimated by OLS.

While the story given so far matches the classical arguments used to causally interpret regressions, it is somewhat loose and ambitious at the same time. It is loose because regression notation does not distinguish between hypothetical and actual values of a variable, leaving any underlying thought experiments implicit. It is ambitious because it suggests that in a properly specified regression model the list of determinants of Y is in a sense complete, and *all* coefficients have a causal interpretation.

The potential outcome framework, the foundation of treatment effect models, provides a more focused and clearer development. One starts by outlining a thought experiment that involves an explicit treatment or intervention, targeted at a population of interest, whose effect one wishes to estimate. Assuming for simplicity that the treatment is binary, the experiment envisions two hypothetical outcomes: the outcome of a randomly chosen population unit if they were exogenously exposed to the

treatment and the outcome of the same unit if they were exogenously excluded from the treatment. The first quantity is the treated potential outcome $Y(1)$ and the second is the untreated potential outcome $Y(0)$. Loosely speaking, these potential outcomes represent two parallel universes that are initially identical, but in one the unit receives the treatment and in the other it does not. The two universes are then left to evolve on their own and outcomes are recorded at a later stage. For any given unit, the difference between $Y(1)$ and $Y(0)$ represents the individual treatment effect, which can in general vary across units.

The fundamental problem of causal inference is that for any given unit only one of the potential outcomes is observed—that which corresponds to the unit's actual treatment status. Denoting the treatment status by the binary variable D, the actually observed outcome Y is given by $Y(1)$ if $D = 1$ and $Y(0)$ if $D = 0$. To simplify the exposition, let us assume that all individual treatment effects are the same, given by the constant δ. Therefore, we can write the observed outcome as $Y = Y(0) + \delta D = \alpha + \delta D + \epsilon$, where $\alpha = E[Y(0)]$ and $\epsilon = Y(0) - E[Y(0)]$. Thus, we have arrived at a simple linear regression model in which the coefficients and the error term have a clear causal foundation.

The remaining question is how to use the observed data to construct a consistent estimator of the coefficient δ, i.e., how to identify the treatment effect of interest. A simple OLS regression of Y on D identifies the expected difference in outcomes between the treated and untreated groups, but this quantity does not generally coincide with the treatment effect because the treated and untreated groups may have had systematically different outcomes even in the absence of the treatment (D and ϵ are generally correlated). Multiple linear regression can be viewed as a tool that adjusts for pre-existing observable differences between the two groups. The precise condition under which a set of control variables is sufficient for identifying the treatment effect (i.e., it eliminates all omitted variable or selection bias) is again stated in terms of potential outcomes. Specifically, we seek variables $X_1, \ldots, X_k$ so that treatment participation D is mean-independent of $Y(0)$ conditional on these X's. If, in addition, the conditional mean of $Y(0)$ given the X's is linear, then it is easy to show that in the linear projection of Y on D and $X_1, \ldots, X_k$, the slope coefficient on D is precisely δ. Hence, OLS consistently estimates this quantity. By contrast, the slope coefficients on the X's do not generally have a causal interpretation; the model is focused on one particular causal question.

The foregoing semi-formal discussion demonstrates how the conceptual framework and language of treatment effects have changed how we think about the representation and estimation of causal effects in econometrics. The rest of the chapter is a historical overview of how this literature, originating from statistics, permeated econometrics, displacing some older traditions. This is an incomplete and subjective account; there are many surveys that document the paradigm shift in detail. For example, Imbens and Wooldridge (2009) is an influential survey paper from the time when the treatment effect approach had solidified its place in econometric practice. Angrist and Pischke (2009) remains an important text that popularized econometrics in the treatment effects 'style'. Heckman and Pinto (2024) is a more recent review that contrasts the potential outcome framework with a more model based account of causality, and also

discusses directed acyclical graphs and Pearl's do calculus—topics which are omitted from this chapter.

The chapter outline is as follows. Section 6.2 introduces simultaneous equation models—the classical framework that was meant to capture causal relationships in econometrics. The section includes a discussion of the reasons why this paradigm has fallen largely out of favor in the 21st century. Section 6.3 offers a brief review of the origins of the potential outcome framework in the statistics literature, while Section 6.4 explains why the framework was later adopted in econometrics as part of the 'credibility revolution'. Section 6.5 is a high-level evaluation of the treatment effects approach. Section 6.6 reviews some of the recent lines of research in this large and expanding literature. Section 6.7 provides further discussion of the relationship between treatment effects and today's structural econometrics. Section 6.8 concludes. The introduction was the most technical part of the chapter; the rest of it does not contain any mathematical notation.

6.2 The Structural Equation Paradigm (1940s-1980s)

6.2.1 Origins and Achievements

In his foundational paper on structural/simultaneous equation models (SEMs), Haavelmo (1944) outlined a comprehensive framework for empirical work in economics. In this framework one employs economic theory to set up a system of linear equations that describes the (hypothesized) relationships between the observed variables and also incorporates latent causal factors through probabilistic error terms. Each equation of the system can be read as encoding a thought experiment, involving external manipulations of the values of the constituent variables, lending an appropriate causal interpretation to the associated coefficients. The aim is to provide a complete description of how the joint distribution of the observed variables is generated and what restrictions it obeys. Perhaps the simplest example of such a system describes the observed price and quantity in a given market through a structural supply and demand equation. In other classical applications much larger systems were constructed with the aim of describing entire economic sectors or the macroeconomy (Klein, 1950; Klein & Goldberger, 1955; Duesenberry, Fromm, Klein & Kuh, 1965; Christ, 1994).

The theory of SEMs was developed further by the research program of the Cowles Commission (see Christ, 1994 for an overview). Koopmans (1949) studied the important question of how to construct a consistent estimate of the model coefficients. To this end, the model is converted to its reduced form, which corresponds directly to some moments of the joint distribution of the observed variables, and hence its coefficients are consistently estimable by construction. The identification problem then is to determine whether the parameters of the structural equations could be recovered from the reduced form. For example, in a regression of quantity on price, the slope coefficient does not recover either the demand slope or the supply slope; it

is a mixture of the two, since quantity and price are endogenous variables determined jointly in the system. The reduced form expresses the quantity and price as a function of demand and supply shifters that are considered exogenous, i.e., they are determined outside the model. The reduced form is directly estimable by OLS, and the question is whether it can be solved for the two slopes of interest.

In a linear SEM identification can be achieved through rank and order conditions. The latter corresponds to exclusion restrictions, i.e., claims that some of the exogenous variables do not show up in all equations of the system. For example, for the demand curve to be identified, the system must contain exogenous variables that shift supply but not demand. The rank condition asks whether these excluded exogenous variables generate enough independent variation to distinguish a given structural equation from the rest of the system.[1] In the SEM literature identification quickly became a technical question about whether the restrictions imposed by economic theory were strong enough to allow one to move from the observable covariance structure back to the underlying structural equations.

This framework was an impressive achievement. It brought endogeneity and simultaneity to the forefront of econometrics, distinguishing it from statistics in general. In addition, it gave rise to an estimation theory tailored to linear systems with jointly determined variables. Instrumental variables, limited-information methods, and full-information methods such as two-stage or three-stage least squares were all part of this machinery. On the practical side, large macroeconomic SEMs developed after the Second World War were employed in forecasting and counterfactual policy analysis. Structural equation econometrics was not a minor technical detour in the history of the field; for several decades it was the main language in which econometricians articulated the possibility of learning causal economic relationships from nonexperimental data.

6.2.2 What did not Age Well

The great ambitions of structural equation modeling also made it fragile in practice. The framework was meant to capture the structure of the multiple causal mechanisms that gave rise to equilibrium outcomes, requiring strong assumptions about functional form, omitted variables, exclusion restrictions, and the stability of parameters across environments. These assumptions were not arbitrary; they were typically motivated by economic theory. Nevertheless, in practice researchers focused on fulfilling the formal conditions for identification, especially the necessary exclusion restrictions, and the justifications for this often strained credibility. Thus, identification was precise as a mathematical property of the model, but often less transparent as an argument about the source of identifying variation in the data.

The concept of identification is perhaps the most important contrast with the later treatment effects literature. In the classical SEM framework, the identification

[1] Technically, this corresponds to a certain matrix having full rank, hence the name.

problem was whether structural parameters could be recovered from the reduced form. Rank and order conditions gave a formal answer to this question. From the perspective of modern causal inference, the more basic question is what comparison is being made with the aid of a statistical model and why this comparison isolates the causal effect of interest. A system may be formally identified because some variables are excluded from some equations, but the causal interpretation of the resulting coefficients still depends on whether these exclusions are believable as restrictions on behavior or on the economic environment. In this sense, the older framework tended to make identification look mechanical, while the treatment effects literature made the credibility of the identifying comparison more explicit.

The limitations were especially visible in applied microeconomics. The classical linear simultaneous equations model was built around systems of equations with constant parameters, while many empirical questions in labor, education, health, and development economics concern treatments whose effects may differ across individuals. Although it may be tempting to interpret the estimated coefficients as some type of average treatment effect, the availability of this interpretation was not formally justified, and is in fact a nontrivial question. Modeling selection into treatment, especially on the basis of unobserved characteristics, also strained the linear simultaneous-equations framework. Once treatment choice depends on latent variables, the model is no longer a simple linear system with constant causal coefficients. Selection models such as Heckman (1979) addressed this problem, but only by introducing nonlinear and distributional structure that made the causal interpretation more model-dependent. By contrast, the treatment effects framework made it natural to tie causal interpretation to a specific research design or identifying comparison, and then to ask what causal object—for example, an average treatment effect, an effect on the treated, or a local average treatment effect—that design could identify. This made it more natural to separate the conceptual problem of causal interpretation from the statistical problem of estimation (Imbens & Wooldridge, 2009).

In macroeconomics the difficulty took a different form. Large macroeconometric models were designed for forecasting and policy analysis, but this required the estimated equations to remain stable when policy changed. Lucas (1976) famously argued that such stability could not be taken for granted. If agents change their optimal behavior in response to a new policy regime, then equations estimated under the old regime may no longer describe behavior under the new one. The Lucas critique therefore challenged the claim that conventional macroeconometric equations were structural in the sense required for counterfactual policy evaluation.

Sims (1980) developed a related but distinct critique. His objection was not simply that large SEMs were too theoretical, but that they often imposed many exclusion restrictions whose credibility was difficult to assess. These restrictions carried much identifying power, yet they were frequently defended by convention or by a loose appeal to theory. Sims's advocacy of vector autoregressions (VARs) was partly a response to this problem: VARs offered a less restrictive way to summarize dynamic relationships among macroeconomic variables before adding stronger structural interpretations. This did not eliminate the burden of causal identification, and Sims

himself continued to regard policy analysis as a legitimate econometric objective (Sims, 1982). It did, however, reduce reliance on large-scale SEMs in empirical work.

The decline of the classical structural equation paradigm did not mean that econometricians abandoned causal modeling. Rather, the standards for a credible causal claim changed. In macroeconomics, dissatisfaction with large simultaneous-equations models led toward VARs, structural VARs, and later DSGE models. In applied microeconomics, it encouraged a turn toward the statistical literature on causal inference, natural experiments, instrumental variables interpreted through potential outcomes, difference-in-differences, regression discontinuity designs, and other design-based approaches. The common thread was a shift away from asking only whether a system was formally solvable and toward asking whether the comparison or policy counterfactual used for identification was convincing.

6.3 The Statistical Roots of the Potential Outcomes Framework (1920s-1980s)

The potential outcome framework did not originate in econometrics. Its earliest formal statement is usually traced to Neyman's (1923) work on randomized agricultural experiments, available in English as Neyman (1990). This setting was narrower than the later econometric literature in that the treatment was assigned by an experimenter, and the main problem was repeated-sampling inference for average treatment effects. But the core idea of causal inference was already present: each unit could be associated with more than one possible outcome, one under each treatment condition, even though only the outcome corresponding to the realized assignment could be observed. Randomization provided the basis for learning about average causal effects without specifying a complete model of the outcome-generating process, while allowing individual-level treatment effects to be completely heterogeneous. Around the same time, Fisher (1935) further advanced the practice of causal inference through randomization and the comparison of outcomes under alternative assignments. This statistical language and viewpoint were rather different from the econometric language of simultaneous equations. Its point of departure was the design of appropriate comparisons rather than a system of behavioral equations.

A crucial step forward in establishing the potential outcome framework was Rubin's formulation of causal effects in both randomized and nonrandomized studies (Rubin, 1974, 1978). He generalized the experimental logic by treating causal inference as a problem of missing data given that, for any unit, only one potential outcome is observed and the others are missing. It is the treatment assignment mechanism that determines whether the missing data problem can be treated as *ignorable* for the purpose of estimating causal effects. In a randomized experiment ignorability of the assignment mechanism is achieved by design; in an observational study it must be justified as an assumption. What needs to be argued is that the data set contains a sufficiently rich set of observed covariates such that, conditional on these covariates, the potential outcomes are independent (or, for most purposes, mean-independent)

of the treatment status. Intuitively, this means that if we take groups of units that are similar in terms of these characteristics, then within such a group treatment assignment can again be treated as if it were decided by, say, a coin flip. This condition is now known by many other names in the literature such as unconfoundedness, conditional independence, selection-on-observables, etc. It is the basis of important causal inference methods such as regression adjustment or matching.

The technical result by Rosenbaum and Rubin (1983) on the propensity score function made the potential outcome framework, combined with the unconfoundedness assumption, an even more effective setting for causal inference with observational data. Under unconfoundedness, adjustment for the probability of treatment conditional on the observed covariates, called the propensity score, is sufficient to balance the distribution of the covariates between the treated and untreated units. The importance of this result was not merely that it supplied another estimator. It showed how the assignment mechanism could be modeled and checked separately from the outcome equation, and how a high-dimensional adjustment problem could be reduced to a scalar balancing score.[2] This made the potential outcome framework especially attractive in fields such as biostatistics and epidemiology, where randomized experiments were an ideal benchmark but observational data were common.

The framework remained largely statistical for several decades, shaped by experimental design, clinical trials, biostatistics, and observational-study methodology rather than by mainstream econometrics. Holland's influential essay helped codify this way of thinking and popularized the phrase "the fundamental problem of causal inference" (Holland, 1986). Its later appeal to econometricians came from the same features that had made it useful in statistics: it separated causal assumptions from functional-form assumptions, connected individual-level counterfactuals to aggregate treatment effects, and made the assignment mechanism or identifying comparison central to the interpretation of empirical estimates.

6.4 The Migration of Potential Outcomes and Treatment Effects into Econometrics (1990s–2000s)

The migration of potential outcomes into econometrics was not simply a case of statistical theory being imported into a static field. It happened in interaction with a marked shift in applied microeconometric research designs that made the source of identifying variation a central concern. Angrist's work on the Vietnam-era draft lottery, Angrist and Krueger's use of quarter of birth as an instrument for schooling, Card's study of the Mariel boatlift, and Card and Krueger's minimum-wage study all exemplified this style (Angrist, 1990; Angrist & Krueger, 1991; Card, 1990; Card & Krueger, 1994). These papers did not fully rely on the potential outcome notation, but they emphasized quasi-experimental variation, institutional detail, and

[2] More precisely, the curse of dimensionality was relegated to another domain; namely, the modeling and estimation of the propensity score. This is now typically handled by machine learning methods (see Section 6.6 for additional discussion).

transparent comparisons. The empirical question was typically framed around a particular historical event, policy rule, or institutional discontinuity that generated something close to an experiment. In short, applied microeconomics was already moving toward a style in which the credibility of a comparison mattered more than the completeness of a structural system.

This empirical movement created a demand for a theoretical language that could express more precisely what such designs identified. The potential outcome framework met this demand well. It made it possible to describe the counterfactual comparison implicit in a natural experiment, to state the assumptions under which this comparison had a causal interpretation, and to distinguish among different treatment effect parameters. This was especially useful because the new empirical style often made the source of identifying variation narrow or local. A draft lottery, a compulsory schooling rule, an immigration shock, or a state-level policy change may be credible precisely because it is specific and reasonably exogenous, but that specificity also raises the question of what type of causal effect is being estimated. The question was no longer simply whether a coefficient could be interpreted as a structural parameter. It was also what treatment effect, and for which (sub)population, was identified by a particular source of variation.

Instrumental variables (IVs) provided a formal bridge between the older econometric toolkit and the new treatment effects language. In the traditional view, a valid instrument solved an endogeneity problem by inducing exogenous variation in an explanatory variable. The potential outcome interpretation, developed most influentially by Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996), showed that this description was incomplete when treatment effects were heterogeneous. Under assumptions such as independence, exclusion, a nonzero first stage, and monotonicity, IV identifies the average treatment effect for the complier subpopulation, that is, for units whose treatment status is changed by the instrument. This parameter, called the local average treatment effect (LATE), was not merely a new label for an old estimator. It changed the interpretation of IV by making it clear that the design determined the estimand.

The point is particularly visible in the paper by Angrist, Graddy and Imbens (2000) on the Fulton fish market, which nicely illustrates the transition between the older simultaneous equations tradition and the treatment effects interpretation of IV. The setting is classical: price and quantity are jointly determined by supply and demand, and weather conditions at sea provide instruments that shift supply but not demand. In the older SEM language, the objective is to identify *the* (linear) demand curve. The authors instead ask what a linear IV estimator actually estimates when demand may be nonlinear and heterogeneous across days. The answer is that IV estimates a weighted average derivative along a segment of the demand curve, with weights determined by the variation in supply induced by the instrument. Thus, the treatment effects perspective turns the question from whether the demand equation is identified in a linear system to what local causal response is recovered by a particular instrument.

Manski's work on treatment effect bounds and partial identification provided another important, and in some ways more radical, counterpoint to the older structural

equation tradition (Manski, 1990, 1995). The guiding idea is to combine the data with (very) weak, explicitly stated assumptions and ask what can be learned about the average treatment effect. The resulting bounds can be interpreted as a form of minimal common ground. They describe what follows from the data and from assumptions weak enough so that any researcher should share these conclusions regardless of their preferred modeling approach. Sometimes this common ground is informative; sometimes it is not. This is not a failure of the analysis, but part of the message. Stronger conclusions require stronger assumptions, and the gap between what is observed and what is claimed should be made visible rather than obscured by convention. If a researcher is wary of making sufficiently strong assumptions, the analysis may need to proceed under partial identification rather than point identification.

A parallel literature associated with Heckman also developed treatment effect concepts through selection models, missing data, and program evaluation, but with a stronger commitment to structural modeling and explicit behavioral assumptions (Heckman, 1979; Heckman, LaLonde & Smith, 1999). The marginal treatment effect is a useful example. It represents treatment effects as varying with the unobserved resistance to treatment, and uses this object to organize average treatment effects, effects on the treated, local average treatment effects, and policy-relevant treatment effects within a common framework (Heckman, Urzua & Vytlacil, 2006). This tradition is important because it shows that the treatment effects vocabulary does not belong exclusively to the design-based literature. There is, however, a contrast. The Heckman tradition tended to ask what could be learned about policy-relevant parameters by modeling selection and outcome processes, while the design-based tradition emphasized what could be learned from particular sources of quasi-experimental variation.

By the 2000s the empirical work in applied microeconomics and the theoretical work in econometrics became mutually reinforcing. Natural experiments created demand for a causal vocabulary and theoretical insights; potential outcomes supplied both. The result was the methodological style later described as the "credibility revolution" in empirical economics (Angrist & Pischke, 2010). Its hallmark is not simply to prefer simple regressions over complicated models, but to place the identifying argument at the center of the empirical analysis. In much of applied microeconometrics, the central question became less whether a coefficient could be interpreted as structural in a fully specified system and more whether the empirical design supplied a convincing comparison and what treatment effect parameter that comparison identified.

6.5 Taking Stock: Why Potential Outcomes Dominate Causal Inference Today

The potential outcome framework came to dominate causal inference because it offered a clear path that connects empirical designs to causal claims. Its central

question is not whether an entire economic system has been specified correctly, but what comparison identifies what causal object under what assumptions. This made the framework especially well suited to the design-based style of applied microeconometrics that developed in the 1990s and 2000s. It also made the evaluation of empirical work more transparent by making it routine for authors to explain why the proposed comparison was credible, what parameter it identified, and how far that parameter could reasonably be generalized. Nevertheless, this shift in focus was not entirely costless. In addition to summarizing the advantages of the treatment effects approach, the 'balance sheet' below also presents some criticisms and limitations.

Advantages

The advantages are related to the discipling of empirical research. The framework does not by itself make causal inference easy, but it forces several questions to be asked in a fixed order: what is the treatment, what is the causal object, what comparison identifies it, and what assumptions make that comparison credible?

- *Transparent identifying assumptions.* Assumptions such as unconfoundedness, exclusion, monotonicity, parallel trends, and continuity at a cutoff may be strong, but they are tied to particular designs and estimands. This makes them easier to state and debate than the restrictions embedded in a full simultaneous equations system.
- *Separation of identification and estimation.* The framework imposes a logical order on causal inference. First, one defines the treatment and the causal effect of interest. Second, one devises an identification strategy, i.e., proposes a comparison that captures some causal parameter such as the average treatment effect, the average treatment effect for the treated, a local average treatment effect, or some other object. Third, a suitable estimator is employed to implement the strategy. Thus, regression becomes a tool for making appropriate comparisons rather than the source of the causal interpretation itself.
- *Treatment effect heterogeneity.* Heterogeneity is built into the framework from the beginning. This makes it natural to ask for which part of the population the effect is being estimated and why different designs may identify different parameters (Heckman et al., 2006).
- *Compatibility with machine learning.* Many causal identification strategies involve estimating predictive relationships (nuisance functions) such as propensity scores, conditional means, or treatment rules. Machine learning is especially suitable for this purpose when these functions are high-dimensional. At the same time, causal identification remains anchored in the design. This division of labor is central to double/debiased machine learning and related methods (Chernozhukov et al., 2018; Athey & Imbens, 2019; see Section 6.6 for more detail).

These advantages explain why the treatment effects approach became so useful as a language for empirical work. Nonetheless, the framework cannot not supply a

complete account of every causal or policy question economists care about. The same focus that makes the framework powerful also defines its limits.

Limitations and Critiques

- *Weaker attention to mechanisms.* Treatment effect estimates often say what happened under a particular intervention, but not why it happened. This is a limitation when the goal is behavioral explanation, welfare analysis, or mechanism-based policy design.
- *Equilibrium and policy counterfactuals.* Standard treatment effects applications are usually partial equilibrium in spirit. Traditional design-based estimates do not account for general equilibrium effects, and may be less informative about policies that change prices, expectations, market structure, or equilibrium behavior. These questions often require structural modeling.
- *External validity and transportability.* In an effort to ensure internal validity, credible designs are often local or setting-specific. This feature can make it difficult to transport estimates to other populations, policy margins, or institutional settings (Deaton & Cartwright, 2018).
- *Overreliance on standard designs.* Garg and Fetzer (2026) document how design-based causal analysis came to dominate publications in economics over time. This trend also raises some concerns. First, economists may eschew large and important questions in favor of niche topics where identification is easier to argue. Second, just because the analysis is presented in a now-standard design framework (such as difference-in-differences, two-way fixed effects, and so on), it does not mean that the causal claim is automatically well-justified. Design quality matters and it is heterogeneous across studies.

The balance sheet is therefore mixed but, at least in the view of the present author, the advantages rightfully justify the paradigm shift in modern empirical work. Potential outcomes became dominant because they improved the clarity and credibility of many empirical claims. Their success, however, should not be understood as a full replacement of structural econometrics. Rather, the framework clarified one central part of causal inquiry: what can be learned from a particular comparison. Other questions—about mechanisms, equilibrium, welfare, and policy regimes not yet observed—continue to require additional economic structure. The lasting contribution of the potential outcome framework is thus not that it made causal inference model-free, but that it made the connection between assumptions, comparisons, and causal parameters much harder to leave implicit.

6.6 New Directions in Treatment Effects (2010s-2020s)

6.6.1 Machine Learning and Double/Debiased Machine Learning

A first recent direction is the use of machine learning methods[3] in the estimation of standard treatment effect parameters in settings with many covariates, interactions, and possible functional forms. The potential outcome framework is especially helpful here because it clearly separates the identification of the causal effect from the prediction tasks needed to implement the identifying strategy. Even when the target is a single low-dimensional parameter such as an average treatment effect, an average effect on the treated, or a treatment coefficient in a partially linear model, credible adjustment may require estimating high-dimensional nuisance functions. These include propensity scores, conditional mean functions, first stages, or other auxiliary objects. Machine learning methods are well suited to these prediction problems because they can account for many covariates in a flexible way, without requiring the researcher to specify all relevant interactions and nonlinearities in advance. Belloni, Chernozhukov and Hansen (2014) provided an early and influential discussion of how high-dimensional methods could be used for inference on structural and treatment effects.

The double/debiased machine learning (DML) literature developed this logic further by combining flexible nuisance estimation with moment conditions that are designed for valid inference on the target parameter (Chernozhukov et al., 2018). The first key ingredient is orthogonality. Orthogonal moments are constructed so that small errors in the first-stage nuisance estimates have only a second-order effect on the estimator of the causal parameter. This matters because machine learning methods are typically chosen for predictive performance and may converge at slower rates than classical parametric (or even nonparametric) estimators. Orthogonality prevents these first-stage imperfections from contaminating the estimation of the target parameter too severely.

The second key ingredient is cross-fitting. In analogy with cross-validation, the procedure is based on partitioning the sample into several parts. To estimate the target parameter, the nuisance functions are estimated on one part of the sample and evaluated on the other parts. The roles of the subsamples are then interchanged and the resulting fold-specific estimates are combined to produce a final estimate. The separation between nuisance function estimation and evaluation reduces overfitting bias and makes it possible to avoid the restrictive empirical-process conditions that would otherwise be needed to achieve theoretical guarantees when flexible prediction methods are employed. In this sense, DML is a particularly clear expression of the modern division of labor: causal identification comes from the research design or identifying assumptions, while machine learning provides flexible control for the covariates and adjustment functions needed to implement the design. This perspective is also emphasized in broader reviews of machine learning methods for treatment effect estimation (Athey & Imbens, 2019; Knaus, 2022; Lieli, Hsu & Reguly, 2022).

[3] Chapter 12 of this volume reviews several machine learning methods and their use in econometrics.

6.6.2 Heterogeneous Treatment Effects

A closely related but distinct direction treats treatment effect heterogeneity itself as the object of interest. The conditional average treatment effect (CATE) function asks how the average effect of a treatment varies with observed characteristics or a pre-specified subset of them. This object is natural in applications where policy makers do not only want to know whether a program works on average, but for whom it works, and possibly why. Work on CATE estimation developed nonparametric, semiparametric, dimension-reduction, and debiased machine learning methods for estimating such functions and conducting inference on them (Abrevaya, Hsu & Lieli, 2015; Knaus, Lechner & Strittmatter, 2021; Semenova & Chernozhukov, 2021; Fan, Hsu, Lieli & Zhang, 2022).

A different strand of the literature is more explicitly discovery-oriented. In the work just mentioned, the dimension along which heterogeneity is analyzed is typically chosen in advance by the researcher, or the object is a CATE function indexed by a pre-specified set of covariates. Tree-based methods instead use the data to search for subgroups in which treatment effects differ. Causal trees and causal forests adapt recursive partitioning and random forests to the estimation of heterogeneous effects, while preserving the distinction between discovering subgroups and estimating effects within them (Athey & Imbens, 2016; Wager & Athey, 2018). Their appeal is that they can reveal which covariates, or combinations of covariates, are most relevant for treatment effect heterogeneity, rather than requiring the researcher to specify those dimensions ex ante.

The broader lesson is that the potential outcome framework naturally accommodates treatment effect heterogeneity at the conceptual level, while recent methods make it easier to study it empirically. Of course, traditional regression models can also display heterogeneity by including interaction terms between the treatment and observed covariates. The difference is that such interactions impose a particular functional form and typically need to be specified in advance; a naive model search over many possible interaction terms exposes the researcher to the risk of overfitting. Machine learning methods are more flexible in this respect; they allow for richer functional forms, facilitate discovery, and help summarize heterogeneity in richer covariate spaces. This does not remove the need for identifying assumptions, but it changes the question from whether there is a single treatment effect to what distribution or function of treatment effects can be learned from the design.

6.6.3 Panel Data and Event-Study Methods

The new difference-in-differences and two-way fixed effects literature is one of the clearest and most recent examples of the potential outcome framework disciplining the interpretation of familiar regression estimators. In the simplest two-group, two-period case, the logic of the difference-in-differences design is rather transparent. One compares the change in outcomes for a treated group to the change in outcomes for

an untreated group, and under a parallel trends assumption this comparison identifies an average treatment effect for the treated group. The potential outcome framework is useful even in this simple setting because it makes clear what the unobserved counterfactual is: the path that the treated group would have followed in the absence of the treatment. This counterfactual path is then identified with the actually observed path of the untreated group. The two-way fixed effects (TWFE) regression *appeared* to generalize this logic naturally to settings with many groups, many periods, and policies adopted at different dates.

These generalizations are, however, not innocuous. With staggered adoption, some units are treated earlier than others, and a TWFE regression uses many implicit comparisons across groups and time periods. Some of these comparisons are the natural ones: newly treated units are compared to units not yet treated or never treated. Others are less natural: already-treated units may serve as controls for later-treated units, even though their outcomes may already embody treatment effects. If treatment effects are constant, these complications are less important. But if treatment effects vary across cohorts or over event time, and especially if treatment timing is related to the magnitude or dynamic path of group-time treatment effects, the TWFE coefficient need not correspond to a simple average of the effects researchers usually have in mind.

The recent literature clarified this problem by using the potential outcome framework to define the underlying causal objects explicitly. One can define group-time average treatment effects, for example, as the effect for units first treated in period g and observed in period t. The TWFE coefficient can then be written as a weighted average of such underlying causal effects, with weights determined by the residualized treatment variable rather than by a policy-relevant aggregation chosen by the researcher. These weights may be unintuitive and, in some cases, negative. Thus, the coefficient may place negative weight on some individual or group-time treatment effects, so that it can be difficult to interpret even when all the underlying effects have the same sign (Goodman-Bacon, 2021; de Chaisemartin & D'Haultfoeuille, 2020). The point is very much in the spirit of the broader treatment effects literature—regression output must be tied back to a clearly defined estimand before it can be given a causal interpretation.

Similar issues arise in event-study specifications, where researchers estimate regressions in 'event time' in order to study dynamics and assess pre-trends.[4] In staggered adoption designs with heterogeneous effects, a coefficient on a particular lead or lag is not automatically the average effect at that event time. It may be contaminated by treatment effects from other periods or other cohorts. As Sun and Abraham (2021) show, apparent pre-trends may arise mechanically from treatment effect heterogeneity rather than from genuine anticipation effects or violations of parallel trends. This result is important because event-study graphs had become a standard visual language for design-based causal inference. The newer literature shows that even this visual language requires a careful definition of the underlying cohort-specific and event-time estimands.

[4] Event time means measuring time relative to a pre-specified event that may actually happen on different calendar dates for different individual units.

The constructive response has been to define the causal objects first and then build estimators that target them directly. Callaway and Sant'Anna (2021), for example, define group-time average treatment effects and propose aggregation schemes that make explicit which effects are being averaged. Borusyak, Jaravel and Spiess (2024) develop an imputation approach that estimates untreated potential outcomes and then forms treatment effect estimates from the gap between observed and imputed outcomes. de Chaisemartin and D'Haultfoeuille (2020, 2023) develop related heterogeneity-robust alternatives and provide a broader synthesis of the new difference-in-differences literature. The common lesson is that panel regressions, just as cross-sectional regressions, do not speak for themselves. Their causal interpretation depends on the comparisons that are implicit in the estimation method used.

6.6.4 Policy-Focused Innovations

Other developments in treatment effects are motivated by policy questions that do not fit neatly into the simple binary treatment, many units framework. Some policy interventions happen at the aggregate level and are historically unique such as a regional conflict, a tobacco-control program, a national policy change, or a country joining an international alliance. Such events are often documented and analyzed as case studies. Traditional case studies can be rich in institutional detail, but they often leave counterfactual comparisons implicit or ambiguous. Synthetic control methods retain the case study focus while making the construction of the counterfactual explicit and systematic (Abadie & Gardeazabal, 2003; Abadie, Diamond & Hainmueller, 2010, 2015).

In potential outcome terms, the problem is to approximate the treated unit's untreated potential outcome after the intervention; that is, to show how the treated unit would have evolved in the absence of treatment. The synthetic control method does this by constructing a weighted average of untreated units that resembles the treated unit before treatment, usually in terms of pre-treatment outcomes and selected covariates. The post-treatment path of this weighted average is then used as the comparison path for the treated unit. This makes the counterfactual visible instead of being hidden inside a regression coefficient.

At the same time, synthetic control methods rely on a form of *external validity* or stability assumption. A good pre-treatment fit is taken as evidence that the same weighted combination of untreated units would have continued to approximate the treated unit's untreated potential outcome after the intervention. Thus, the method makes an external validity assumption as part of its design. This assumption may be rather transparent but it is still an inherently untestable assumption about how relationships learned in one part of the data carry over to another.

A related but distinct policy-focused literature asks more directly when treatment effects can be transported or extrapolated across settings. Policy makers often want to know not only what happened in one experiment, program, or natural experiment, but also what would happen under a similar intervention in another population, location,

or time period. This problem is especially salient in IV settings. The LATE framework gives an internally credible interpretation of an IV estimate as the average effect for compliers but the identity and the size of the complier group depends on the instrument. Moving from this local effect to an effect for another population, another instrument, or a broader policy target requires additional assumptions about treatment effect heterogeneity and selection into treatment (Angrist, 2004; Heckman & Vytlacil, 2005; Angrist & Fernández-Val, 2013).

More generally, work on external validity and counterfactual treatment effects formalizes the information and procedures needed to extrapolate from an internally credible estimate to a new target setting (Hotz, Imbens & Mortimer, 2005; Athey & Imbens, 2017; Deaton & Cartwright, 2018; Hsu, Lai & Lieli, 2022). This line of work does not abandon the potential outcome framework, but it also reveals one of its limitations. Potential outcomes are very good for defining causal effects and clarifying the assumptions under which a comparison identifies them. They are less naturally suited, by themselves, to saying why an effect should remain stable across environments. One response is to add explicit transportability assumptions; another is to return, at least partly, to structural modeling, where behavioral mechanisms are used to support counterfactual claims outside the original empirical setting. The common theme is that policy relevance requires more than internal validity. It requires an account of the new population or policy environment to which the estimated effect is meant to apply.

6.6.5 Spillovers, Multiple Treatments and Interactions

The basic potential outcome notation ($Y(1)$ and $Y(0)$) suppresses several complications. It assumes, first, that a given unit's outcome depends only on the unit's own treatment status and not on the treatment status of other units. It also assumes that the treatment is well defined, in the sense that there are no relevant hidden versions of treatment. These conditions are commonly summarized by the 'stable unit treatment value assumption' (SUTVA). In its simplest form, SUTVA is what makes it possible to write the causal problem as a comparison between two potential outcomes for each unit.

One important violation of SUTVA arises through spillovers or interference. Vaccination is the classic example: an individual's health outcome may depend not only on whether the individual is vaccinated, but also on whether others in the relevant community are vaccinated. Similar issues arise in schools, neighborhoods, social networks, labor markets, and peer-effect settings. The potential outcome framework does not break down in such cases, but the notation has to be expanded. A unit's potential outcome may need to be indexed by its own treatment and by the treatment status of others, or by an exposure mapping that summarizes the relevant treatments received by others. This makes it possible to distinguish direct effects, spillover or indirect effects, total effects, and overall effects (Hudgens & Halloran, 2008; Aronow & Samii, 2017; Vazquez-Bare, 2022).

A related complication is that the treatment itself may not be binary. Sometimes treatment is multivalued or continuous, as in the case of treatment intensity, dosage, program type, or policy packages. In other applications there are several binary treatments. Voter mobilization experiments may combine different modes of contact, while educational interventions may combine tutoring and financial incentives. In these settings the simple pair $Y(1)$ and $Y(0)$ must be replaced by potential outcomes indexed by treatment levels, treatment combinations, or exposure states. The generalized propensity score and multiple-treatment literatures extend the selection-on-observables logic to such cases (Imbens, 2000; Lechner, 2001; Hirano & Imbens, 2004), while recent IV applications make clear how quickly the relevant counterfactual comparisons and compliance types multiply (Blackwell, 2017; Kormos, Lieli & Huber, 2025).

Interactions and endogenous takeup add another layer of nuance. If there are two treatments, the effect of treatment A may depend on whether treatment B is also received. With noncompliance, the realized treatment vector may differ from the assigned treatment vector, and with multiple instruments the set of compliance types becomes much richer than in the binary LATE framework. IV estimands may then combine direct effects, interaction effects, spillover effects, and treatment effect heterogeneity. Blackwell's analysis of voter mobilization experiments and Vazquez-Bare's work on IV with spillovers illustrate how the potential outcome framework can be used to unpack these estimands (Blackwell, 2017; Vazquez-Bare, 2022). Kormos, Lieli, and Huber show that with interacting treatments and endogenous takeup, it can be difficult to separate treatment interaction from treatment effect heterogeneity without additional restrictions (Kormos et al., 2025; Bhuller & Sigstad, 2024).

Mediation analysis raises a closely related issue. Here the object is not only whether a treatment affects an outcome, but through which intermediate variable or causal channel the effect operates. In the simplest language, one wants to decompose the total treatment effect into an indirect effect operating through a mediator and a direct effect that does not operate through that mediator. This again requires expanding the potential outcome notation: potential outcomes must be indexed not only by treatment status, but also by the value the mediator would take under different treatment states. As in the multiple-treatment and spillover settings, the framework clarifies the estimands but also reveals the strength of the identifying assumptions required. Huber's work on causal mechanisms and mediation analysis is a useful example of how direct and indirect effects can be formulated and identified within the treatment effects framework (Huber, 2014).

The common lesson is that the potential outcome framework remains useful precisely because it makes these complications visible. It forces the researcher to define the relevant counterfactuals explicitly: own treatment, others' treatment, treatment A alone, treatment B alone, both treatments, neither treatment, and so on. The cost is that the number of potential outcomes and estimands grows quickly. Once the treatment and exposure space becomes richer, simple regression coefficients and simple IV ratios no longer have transparent causal interpretations without additional assumptions. The framework therefore clarifies the problem, but it also makes clear why the problem is hard.

6.7 The Place for Structural Econometrics Today

Structural econometrics did not disappear; its role changed and became more specialized. In much of applied microeconomics, the treatment effects framework became the natural language for defining causal objects and for thinking about empirical research designs. But there remain many questions that go beyond what can be answered by comparisons based on research design between treated and untreated units. Policy makers often want to evaluate a policy that has not yet been tried, a different version of a policy, or a policy applied in a new environment. They may also care about welfare, equilibrium responses, dynamic incentives, or the distribution of gains and losses. These questions often require an explicit model of how agents respond to a changed policy environment (Heckman & Vytlacil, 2005; Wolpin, 2013).

Industrial organization (IO) is perhaps the clearest example. The modern structural IO literature often begins from detailed models of demand, supply, and equilibrium pricing. The purpose is not only to estimate the effect of a past price change or regulation, but to ask what would happen under a merger, a tax, a subsidy, a new product, or a different competitive regime. The classic and influential work of Berry, Levinsohn and Pakes (1995) on automobile demand and Nevo (2001) on the ready-to-eat cereal industry illustrates this logic. The estimand is a counterfactual market outcome, often including prices, quantities, profits, consumer surplus, and total welfare. A treatment effect comparison can be very informative about a particular historical intervention, but it does not by itself describe the equilibrium that would arise under a large-scale policy that changes the incentives of economic actors.

The same point appears in structural microeconometrics more broadly. Many applications in labor, education, development, and public economics involve forward-looking decisions, expectations, dynamic constraints, or endogenous program participation. Rust (1987)'s model of bus engine replacement became a canonical example because it showed how a dynamic decision problem could be taken seriously in estimation. The later dynamic discrete choice literature extended this logic to many settings in which agents choose today while anticipating future states and future policies (Aguirregabiria & Mira, 2010). In policy evaluation, Todd and Wolpin (2006) used experimental evidence to validate a dynamic behavioral model that could then be used for additional counterfactual policy exercises. This is one productive way in which structural methods and treatment effect evidence can complement each other: credible causal evidence can discipline the model, while the model can be used to derive counterfactuals beyond the original experiment.

Macroeconometrics followed its own path. Macroeconomic questions often concern aggregate shocks, policy rules, expectations, and general equilibrium responses. For this reason, macro did not adopt much of the treatment effects language or the natural experiment approach of applied microeconomics. The critique of large macroeconomic SEMs led partly to VAR methods (Sims, 1980), but it also left room for estimated or calibrated structural macro models, including DSGE models, that try to connect shocks and frictions to aggregate dynamics and policy counterfactuals (Smets & Wouters, 2007; Fernández-Villaverde & Rubio-Ramírez, 2007). The lesson is that the two traditions now occupy different, partly overlapping spaces. Treatment effects

methods are especially powerful when the central problem is credible identification of a well-defined causal contrast. Structural econometrics remains important when the central problem is mechanism, equilibrium, welfare, dynamics, or counterfactual policy evaluation outside the support of the original empirical design.

6.8 Conclusion

The history reviewed in this chapter is not simply the history of a new notation. The potential outcome framework changed econometric practice because it changed how researchers were expected to justify the causal interpretations of their estimates. A regression coefficient, an IV coefficient, a difference-in-differences coefficient, or a matching estimate could no longer be treated as self-explanatory. The researcher had to specify the causal object of interest, the comparison that would identify it, and the assumptions under which that comparison was credible. This was a major shift in emphasis. It did not eliminate modeling, but it made the empirical comparison itself the center of the causal argument.

This shift helped explain the success of the treatment effects approach in applied microeconometrics. It fit naturally with randomized experiments, quasi-experiments, instrumental variables, regression discontinuity designs, and difference-in-differences designs. It also provided a common language for evaluating these designs. The key question became not whether a whole economic system had been specified correctly, but what a particular source of variation identified. This more modest ambition was also a source of strength. It made causal claims easier to state, easier to criticize, and often more credible.

At the same time, the chapter has also emphasized that this gain in clarity came with tradeoffs. The treatment effects approach is at its strongest when the object is a well-defined causal contrast in a well-specified population or design. It does not directly provide answers to questions about mechanisms, equilibrium adjustments, welfare, dynamic behavior, and the effects of policies that have not yet been observed. New literatures on machine learning, heterogeneous treatment effects, modern difference-in-differences, synthetic controls, external validity, spillovers, and multiple treatments extend the reach of the framework, but they also reveal the same basic tension. Once the empirical question moves away from a simple treatment-control contrast, the relevant potential outcomes, estimands, and identifying assumptions multiply quickly.

This is why the relationship between treatment effects and structural econometrics is best understood as a division of labor rather than replacement. The treatment effects framework disciplined causal inference by making assumptions and estimands explicit. Structural econometrics remains important where economic theory is needed to model counterfactuals beyond the original design. In contemporary econometrics, the two traditions often coexist. Credible research designs can discipline structural models, and structural models can extend causal evidence to questions about mechanisms, welfare, dynamics, and policy environments not directly observed in the data. The lasting contribution of the treatment effects framework is therefore not that it made

econometrics model-free, but that it raised the standards for justifying the causal content of empirical work.

References

Abadie, A., Diamond, A. & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505.

Abadie, A., Diamond, A. & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*(2), 495–510.

Abadie, A. & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American Economic Review*, *93*(1), 113–132.

Abrevaya, J., Hsu, Y.-C. & Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, *33*(4), 485–505.

Aguirregabiria, V. & Mira, P. (2010). Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, *156*(1), 38–67.

Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. *The American Economic Review*, *80*(3), 313–336.

Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *The Economic Journal*, *114*(494), C52–C83.

Angrist, J. D. & Fernández-Val, I. (2013). Extrapolate-ing: External validity and overidentification in the late framework. *Advances in Economics and Econometrics*, *3*, 401–434.

Angrist, J. D., Graddy, K. & Imbens, G. W. (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*(3), 499–527.

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

Angrist, J. D. & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014.

Angrist, J. D. & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Angrist, J. D. & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30.

Angrist, J. D. & Pischke, J.-S. (2017). Undergraduate econometrics instruction: Through our classes, darkly. *Journal of Economic Perspectives*, *31*(2), 125–144.

Aronow, P. M. & Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of*

Applied Statistics, *11*(4), 1912–1947.

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

Athey, S. & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Athey, S. & Imbens, G. W. (2019). Machine learning methods that economists should know about. In A. Agrawal, J. Gans & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda* (pp. 507–552). Chicago: University of Chicago Press.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29–50.

Berry, S., Levinsohn, J. & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, *63*(4), 841–890.

Bhuller, M. & Sigstad, H. (2024). 2sls with multiple treatments. *Journal of Econometrics*, *242*(1), 105785.

Blackwell, M. (2017). Instrumental variable methods for conditional effects and causal interaction in voter mobilization experiments. *Journal of the American Statistical Association*, *112*(518), 590–599.

Borusyak, K., Jaravel, X. & Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *The Review of Economic Studies*, *91*(6), 3253–3285.

Callaway, B. & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230.

Card, D. (1990). The impact of the mariel boatlift on the miami labor market. *Industrial and Labor Relations Review*, *43*(2), 245–257.

Card, D. & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, *84*(4), 772–793.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Christ, C. F. (1994). The cowles commission's contributions to econometrics at chicago, 1939-1955. *Journal of Economic Literature*, *32*(1), 30–59.

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

de Chaisemartin, C. & D'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964–2996.

de Chaisemartin, C. & D'Haultfoeuille, X. (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal*, *26*(3), C1–C30.

Duesenberry, J. S., Fromm, G., Klein, L. R. & Kuh, E. (Eds.). (1965). *The brookings quarterly econometric model of the united states*. Chicago: Rand McNally.

Fan, Q., Hsu, Y.-C., Lieli, R. P. & Zhang, Y. (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business &*

Economic Statistics, *40*(1), 313–327.

Fernández-Villaverde, J. & Rubio-Ramírez, J. F. (2007). Estimating macroeconomic models: A likelihood approach. *Review of Economic Studies*, *74*(4), 1059–1087.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Garg, P. & Fetzer, T. (2026). *Causal claims in economics.* Retrieved from https://arxiv.org/abs/2501.06873

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, iii–115.

Hamilton, J. D. (1994). *Time series analysis*. Princeton: Princeton University Press.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–161.

Heckman, J. J., LaLonde, R. J. & Smith, J. A. (1999). The economics and econometrics of active labor market programs. In O. C. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3A, pp. 1865–2097). Amsterdam: Elsevier.

Heckman, J. J. & Pinto, R. (2024). Econometric causality: The central role of thought experiments. *Journal of Econometrics*, *243*(1), 105719.

Heckman, J. J., Urzua, S. & Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, *88*(3), 389–432.

Heckman, J. J. & Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, *73*(3), 669–738.

Hirano, K. & Imbens, G. W. (2004). The propensity score with continuous treatments. In A. Gelman & X.-L. Meng (Eds.), *Applied bayesian modeling and causal inference from incomplete-data perspectives* (pp. 73–84). Chichester: Wiley.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Hotz, V. J., Imbens, G. W. & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, *125*(1–2), 241–270.

Hsu, Y.-C., Lai, T.-C. & Lieli, R. P. (2022). Counterfactual treatment effects: Estimation and inference. *Journal of Business & Economic Statistics*, *40*(1), 240–255.

Huber, M. (2014). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, *29*(6), 920–943.

Hudgens, M. G. & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, *103*(482), 832–842.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710.

Imbens, G. W. & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475.

Imbens, G. W. & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, *47*(1), 5–86.

Klein, L. R. (1950). *Economic fluctuations in the united states, 1921–1941* (No. 11). New York: John Wiley & Sons.

Klein, L. R. & Goldberger, A. S. (1955). *An econometric model of the united states, 1929–1952*. Amsterdam: North-Holland.

Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, *25*(3), 602–627.

Knaus, M. C., Lechner, M. & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, *24*(1), 134–161.

Koopmans, T. C. (1949). Identification problems in economic model construction. *Econometrica, Journal of the Econometric Society*, 125–144.

Kormos, M., Lieli, R. P. & Huber, M. (2025). Interacting treatments with endogenous takeup. *Journal of Applied Econometrics*.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. *Econometric Evaluation of Labour Market Policies*, 43–58.

Lieli, R. P., Hsu, Y.-C. & Reguly, Á. (2022). The use of machine learning in treatment effect estimation. In F. Chan & L. Mátyás (Eds.), *Econometrics with machine learning* (pp. 79–109). Springer.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In K. Brunner & A. H. Meltzer (Eds.), *The phillips curve and labor markets* (Vol. 1, pp. 19–46). North-Holland.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, *80*(2), 319–323.

Manski, C. F. (1995). *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, *69*(2), 307–342.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, *5*(4), 465–472. (Translated and edited by D. M. Dabrowska and T. P. Speed)

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*(1), 34–58.

Rust, J. (1987). Optimal replacement of GMC bus engines: An empirical model of harold zurcher. *Econometrica*, *55*(5), 999–1033.

Semenova, V. & Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, *24*(2), 264–289.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, *48*(1), 1–48.

Sims, C. A. (1982). Policy analysis with econometric models. *Brookings Papers on Economic Activity*, *1982*(1), 107–164.

Smets, F. & Wouters, R. (2007). Shocks and frictions in US business cycles: A bayesian DSGE approach. *American Economic Review*, *97*(3), 586–606.

Sun, L. & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175–199.

Todd, P. E. & Wolpin, K. I. (2006). Assessing the impact of a school subsidy program in mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *American Economic Review*, *96*(5), 1384–1417.

Vazquez-Bare, G. (2022). Causal spillover effects using instrumental variables. *Journal of the American Statistical Association*, *117*(538), 667–678.

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

Wolpin, K. I. (2013). *The limits of inference without theory*. Cambridge, MA: MIT Press.