The Scandinavian
Journal of Economics

Scand. J. of Economics 124(3), 603–645, 2022
DOI: 10.1111/sjoe.12505

Labor by design: contributions of David Card, Joshua Angrist, and Guido Imbens*

Peter Hull

Brown University, Providence, RI 02912, USA
peter_hull@brown.edu

Michal Kolesár

Princeton University, Princeton, NJ 08544, USA
mkolesar@princeton.edu

Christopher Walters

UC Berkeley, Berkeley, CA 94720, USA
crwalters@econ.berkeley.edu

Abstract

The 2021 Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel was awarded to David Card "for his empirical contributions to labor economics" and to Joshua Angrist and Guido Imbens "for their methodological contributions to the analysis of causal relationships". We survey these contributions of the three laureates, and discuss how their empirical and methodological insights transformed the modern practice of applied microeconomics. By emphasizing research design and formalizing the causal content of different econometric procedures, the laureates shed new light on key questions in labor economics and advanced a robust toolkit for empirical analyses across many fields.

1. Introduction

The 2021 Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel was awarded to David Card "for his empirical contributions to labor economics" and to Joshua Angrist and Guido Imbens "for their methodological contributions to the analysis of causal relationships". To an observer unfamiliar with modern microeconomic research, the connection between these two prize rationales may seem unclear. What does empirical labor economics, which studies the functioning of labor markets and institutions, have to do with methods for inferring causal relationships? Within economics, however, it is now taken for granted that robust empirical

*We are grateful to Alberto Abadie, Joshua Angrist, Orley Ashenfelter, David Card, Guido Imbens, and Pat Kline for excellent suggestions and comments.

study – both inside and outside labor – often rests on a compelling approach for establishing causality. This widespread view reflects the profound impact of the three laureates on the field.

At its core, much of microeconomic theory concerns causal relationships. A good's own-price supply or demand elasticity represents the causal effect of a price increase on its quantity produced by firms or desired by consumers. Human capital theory (Becker, 1964) explains the causal effect of schooling on workers' earning potential. The efficient design of labor market interventions, such as a new training program or a minimum wage increase, hinges on their causal effects on wages and employment. Yet for much of the twentieth century, explicit causal language in microeconomic research was relatively rare. As late as the 1980s, only around 15 percent of National Bureau of Economic Research (NBER) working papers contained the terms "causality" or "causal" (Currie et al., 2020). However, this share increased sharply in the mid-1990s and early 2000s, and has continued to grow. Now, nearly 50 percent of NBER papers feature causal inquiry (Imbens, 2021). What explains this change?

While it is challenging to convincingly answer such a (causal) question from a single time series, we can hypothesize one potential explanation from the timing of this change. Much of David Card's path-breaking empirical work was conducted in the late 1980s and early 1990s, at the Industrial Relations Section of Princeton University. This work brought fresh causal evidence on core questions in labor economics: the employment effects of the minimum wage, the impacts of immigration on labor market outcomes of natives, and the effect of educational investments on labor market outcomes. The answers that Card's research uncovered were surprising and compelling, challenging conventional wisdom, and triggering both debate and many follow-up analyses. Analyzing the impact of a large influx of Cuban migrants into Miami following the Mariel boatlift, for example, Card (1990) found little effect on the employment rates and wages of native workers. Studying the effects of an increase in the New Jersey minimum wage, Card and Krueger (1994) found no evidence for a decline in low-wage employment. Both findings were at odds with simple textbook models of labor markets, and fueled much subsequent empirical and theoretical research on potential explanations – such as the substitutability of foreign and native workers and the market power of low-wage firms. In this way, Card's work illustrated how surprising empirical facts – viewed as specific causal effects – could move strong theoretical prior beliefs and drive decades of new research.

But the impact of David Card's work on empirical practice stems not only from what conclusions he reached but from *how* he reached them. Along with Joshua Angrist, Orley Ashenfelter, Alan Krueger, other colleagues and students at the Princeton Industrial Relations Section, and

other prominent labor economists at the time, Card pioneered a fresh approach to empirical analysis in the late 1980s and early 1990s – one centered on the identification of causal effects. The foundation of this approach is a careful consideration of a study's underlying *research design*: an understanding of where the variation in an economic "treatment", such as high minimum wages, came from, and an empirical approach that leverages this understanding to construct an appropriate comparison group. Desirable treatment variation often comes from so-called "natural experiments" – unanticipated shocks or as-good-as-random shifts in the exposure to treatment, or in the factors determining treatment. This design-based approach can make transparent the key assumptions that drive an empirical study's conclusion, and often guides their empirical validation or falsification.

The design-based approach is compelling in part because of the close connections it draws to true experimentation, such as in a randomized controlled trial (RCT).[1] True randomization is often infeasible for studying important economic questions, such as the effect of immigration or effects of large-scale minimum wage changes on local labor markets. Most economic treatments are not just determined by chance, as in an RCT, but also by individual or institutional choices that are far from random. Earlier solutions to the threat of *selection bias* in such settings focused on models of choice, drawing on a rich body of microeconomic theory. By instead focusing on the quasi-randomness in certain natural experiments, Card and fellow pioneers of the design-based approach showed how such modeling restrictions could be relaxed or even eschewed with by-chance variation.

The design-based approach was made more convincing and rigorous by new econometric insights, recognized in the second half of the 2021 Nobel Prize. In the celebrated local average treatment effect (LATE) theorem of Imbens and Angrist (1994), the laureates showed how a natural (or true) experiment generating randomness in a variable influencing treatment could be leveraged to estimate the causal effects of the treatment – with minimal restrictions on other factors influencing the treatment choice. For example, a randomly drawn lottery number that determines the eligibility of an individual for military draft service can be used to estimate the effects of such service on later-life earnings (Angrist, 1990). However, some individuals might volunteer in the military regardless of draft eligibility,

[1] Card attributes his use of the term "research design" to his exposure to the *New England Journal of Medicine*, which Alan Krueger subscribed to at Princeton, and which often used the term in reference to randomized trials; see "Equitable growth in conversation: an interview with David Card and Alan Krueger", as interviewed by Ben Zipperer, in April 2016, https://eml.berkeley.edu/~card/interviews/interviewwithCardandKrueger.pdf.

while others might find ways of avoiding military service when drafted. Economists typically analyze such settings with instrumental variable (IV) techniques, using draft eligibility as an "instrument" for the military service "treatment". Imbens and Angrist showed that such IV analyses generally recover a LATE: the average earnings effect of military service among *compliers*, who are induced to service as a result of the draft lottery. In contrast, a hypothetical RCT that randomly assigns people to serve in the military would recover the overall ATE in the entire population (including volunteers and those who avoid the draft regardless of their lottery number), not just among compliers.

This central insight of the LATE theorem had a profound effect on how economists interpret evidence produced by natural experiments, and how the field synthesizes evidence accumulated across different studies. Underlying the theorem is a potential outcomes framework, which relaxed the model-based (and often parametric) restrictions from earlier IV analyses. The framework highlighted the potential for treatment effect *heterogeneity*, both across populations and across different natural experiments in the same population. Such heterogeneity explains how the results from one research design may differ from another, despite both yielding valid causal effects. The focus on treatment effect heterogeneity and the associated notions of internal validity versus external validity (i.e., generalizability) has motivated a vast and growing body of literature – including later work by the laureates – showing how structural models of individual behavior and statistical extrapolations can synthesize causal evidence across different research designs and settings.

In this paper, we argue that the face of modern empirical economics was shaped in large part by the empirical and methodological contributions of the 2021 Nobel laureates. Card showed how careful attention to research design can bring new, compelling, and sometimes unexpected evidence to core questions and theories in labor economics. Angrist and Imbens showed the precise strengths and limitations of the design-based approach, allowing new empirical work to be better contextualized and integrated across studies. Together, the laureates strengthened the scientific foundation of economics by showing how robust theory and empirics can interact to advance our understanding of key economic questions where true experimentation is infeasible.

To put the contributions of Card, Angrist, and Imbens in context, we first outline the key empirical challenge the laureates' work focused on – selection bias – and the state of contemporary empirical practice when this work began. We discuss how the design-based approach and search for natural experiments helped to address several critiques of existing empirical approaches. We then turn to the empirical contributions of the laureates, focusing on David Card's work on immigration and minimum wage laws

as well as the three laureates' work on the earnings effects of education and labor market experiences. Next, we detail the methodological piece of the prize, centered around the Angrist and Imbens LATE theorem. We discuss how their result and the general potential outcome framework formalized key strengths and limitations of the design-based IV approach and led to further insights for other methods. We conclude by briefly summarizing other advances, by the laureates and others, that grew out of this prize-winning work.

2. Setting the stage

2.1. The selection challenge

Many important questions in economics hinge on the reliable measurement of causal effects. To decide whether to expand a subsidized post-schooling training program, a policymaker must weigh the effects of the program on trainee labor market outcomes against the costs of the program and the likely effects of alternative programs, such as job-search assistance. When considering health-care reform, a social planner must take into account the effects of health insurance on individual health and welfare as well as the social cost. Should the national minimum wage be raised to $15 an hour? The answer depends, in part, on the likely causal effect of such an increase on low-wage unemployment.

Unfortunately, the answers to such causal questions tend not to be directly revealed by economic data. A researcher cannot simply compare individuals who are "treated" (those who participate in a training program, who have health insurance, or who are subject to a higher local minimum wage law) with individuals who are "untreated" to estimate the treatment's effects, as these two populations are likely very different. Individuals can choose whether to participate in a training program, and this choice might be driven by a wide range of characteristics and circumstances that are relevant for the observed outcomes. Trainees tend to have lower levels of education and past earnings than non-trainees, for example, and differences in both schooling and work history can signal underlying human capital differences. Thus, even if the causal effect of a training program is positive, a researcher might find that trainee earnings after completing the program are lower than non-trainee earnings.

This problem of selection bias has long been recognized in economics. A conceptually simple solution is to remove the element of treatment choice with a randomized experiment. This "gold standard" for evaluating medical treatments (and some social programs) purges bias by ensuring the individuals randomized into the treatment and control groups are, on average, identical prior to the experiment. Unfortunately, such

randomization can be difficult or infeasible for many important economic questions. It is hard to imagine researchers convincing government officials to randomly raise the minimum wage in some areas but not others, or to experiment with other high-stakes economic programs.

In response to this fundamental identification challenge, economists have developed a variety of econometric methods which can purge selection bias in non-experimental settings by incorporating additional data and assumptions. The simplest of these approaches is to adjust for observable differences between the groups, often with linear regression. Another approach is to leverage longitudinal data, for example by comparing the earnings of individuals before and after their participation in a training program. By further contrasting the treatment group's earning change with an analogous change in the untreated group, one arrives at what Ashenfelter and Card (1985) termed a "difference-in-differences" analysis. In both cases, regression-adjusting or time-differencing can address selection bias by making the treated and untreated group more comparable. A different and clever strategy, pioneered by Heckman (1974, 1976, 1979), leverages microeconomic theory to model an individual's decision to self-select into treatment, and derives a statistical bias correction from the selection equation. This model-based approach grew out of a long tradition of simultaneous equation modeling in economics, and it quickly caught on in the early 1980s (for a review, see Blundell, 2001).[2]

2.2. Empirical concerns

Despite decades of empirical work leveraging regression adjustment, longitudinal data, and selection models to answer important questions in labor economics, by the mid-1980s there was growing concern that the collective evidence produced in this work was weak. In a review of studies of the union wage gap, for example, Lewis (1986a,b) documented a range of estimates so large as to be of little use. He further showed that estimates constructed from elaborate selection models appeared even less reliable than simpler regression estimates, noting that "a substantial fraction of [selection method] estimates are [. . .] preposterously large or outlandishly negative" (Lewis, 1986a, p. 1144). In a staff study of training programs for the congressional joint economic committee, Goldstein (1972, p. 14) called for improving the evaluation of training programs, because "the robust expenditures ($179.4 million from fiscal 1962 through 1972) [. . .] are a disturbing contrast to the anemic set of conclusive and reliable findings".

[2]James Heckman was awarded the Nobel Memorial Prize in 2000, alongside Daniel McFadden, for "his development of theory and methods for analyzing selective samples".

This assessment is echoed by Ashenfelter (1978), who documented a possible reason for the unreliability: trainees' earnings tend to fall prior to joining the program, both in absolute terms and relative to a comparison group of non-trainees. This form of self-selection into treatment, dubbed "Ashenfelter's dip", suggests that part of the observed earnings increase following training might reflect mean-reversion (i.e., a return to a permanent earnings path that was temporarily disrupted). Simple longitudinal methods are likely to ascribe such increases to the effect of the treatment, making the program appear more effective than it actually was.

Broadly, these critiques suggested that empirical studies using observational data and existing econometric methods rarely solve the selection challenge in labor economics, and thus might not be a satisfactory substitute for randomized experiments. A direct assessment of the extent to which such studies can reproduce experimental evidence was given in a landmark study by LaLonde (1986). LaLonde first calculated the effect on trainee earnings of the National Supported Work Demonstration, an employment program that randomized participation in a field experiment. He then compared the estimates to those produced by a variety of non-experimental methods based on a modified data set, where the experimental control group was replaced by different comparison groups drawn from the Panel Study of Income Dynamics (PSID) and a matched Current Population Survey (CPS)/Social Security Administration (SSA) file. In spite of using state-of-the-art econometric methods, LaLonde failed to replicate the experimental results without the experimental control group. Different methods and comparison groups produced a wide range of estimates, and standard specification tests were unhelpful in determining which observational estimates were closest to the experimental "ground truth".

Such findings suggested that more and better data, though clearly necessary, were not sufficient for credible causal inference. This conclusion exposed cracks in the foundation of an argument made by Stafford (1986), who documented a stark rise in the share of empirical labor economics papers in 1965–1983 using individual-level data (such as the PSID and CPS) instead of aggregate data on census tracts, states, or countries. Stafford argued that the granularity of these "microdata" protected labor economics from the critiques that had been leveled at empirical economics as a whole, by Leamer (1983) and others.[3] However, the core of Leamer's critique was that many empirical findings were sensitive to small changes in the analytic assumptions – exactly the concern raised by the above studies. Granularity of microdata alone, as it turns out, does not make a study robust.

[3] Other prominent critiques are found in Hendry (1980), Sims (1980), Black (1982), and Leontief (1982).

Leamer worried that such a lack of robustness encouraged specification searches, in which researchers tinkered with the analytic method they used until they found a desired result. The proposed remedy was sensitivity analysis, in which researchers show how their results change with the exact specification or functional form (or, in a Bayesian analysis, by varying the prior distribution). Robustness checks like these are now widely used in economic studies. But while sensitivity analyses can reveal the limitations of an observational analysis, they rarely suggest solutions on their own. Many carefully executed papers at the time, such as the longitudinal studies of training programs by Ashenfelter (1978) and Ashenfelter and Card (1985), were already upfront about the fragility of their results.

LaLonde's analysis also suggested one path forward: putting less emphasis on model-based solutions to self-selection into treatment, which did not clearly dominate simpler regression-based approaches in his study, and more emphasis on the researcher's choice of the comparison group. While "the difficulty of obtaining a reliable comparison group" (Ashenfelter, 1975) had long been noted in labor economics, LaLonde's analysis made clear that sensitivity to this choice could be as – or more – important than the particular econometric approach. Furthermore, LaLonde showed that it can be hard to determine from the data alone which comparison group or method is most likely to address selection bias. Addressing the prevailing concerns in applied microeconomic research would seem to require help from elsewhere.

2.3. Labor economics by design

A primary contribution of the 2021 laureates was to push the field towards approaching causal questions in a fundamentally different way: with an emphasis on *research design* as a means to address the sensitivity concerns raised by Ashenfelter (1978), Lewis (1986a,b), LaLonde (1986), and others. While exact definitions of research design vary (and sometimes overlap with other terms, such as "empirical strategy" or "identification strategy"), it broadly refers to a researcher's understanding of the process determining how units in a study are assigned to different "treatments", or the process determining their outcomes in the absence of a treatment, which can, in turn, be used to construct a sensible non-experimental "control group" (Meyer, 1995). From this perspective, a convincing study makes core selection bias concerns explicit through a clear discussion of the research design, which in turn dictates the appropriate econometric methods. The researcher provides direct or indirect evidence supporting the key assumptions underlying the methods – often loosely called the "identifying assumptions". The laureates' empirical work exemplifies this push towards research

design in the early 1990s, as we discuss in Section 3. Their later methodological work formalized the design-based approach, as we discuss in Section 4.

An emphasis on research design brings new perspective to the argument of Stafford (1986), on the virtues of microdata in empirical labor economics. First, while more detailed data can allow researchers to probe new sources of variation and to consider different identifying assumptions, granular data alone do not make for a convincing study. To leverage and validate a particular research design one needs data on the appropriate variables, which need not be contained in standard datasets. Sometimes such data must be collected by the researcher, as in the ground-breaking study by Card and Krueger (1994) we discuss below. In other cases, a design calls for linking different administrative data to more standard research extracts, as in the study by Angrist (1990), which we also discuss below. New data and linkages could, in turn, enable new lines of research. The rise of such researcher-collected and administratively generated data in applied microeconomics, over the more standard microdata research extracts emphasized in Stafford (1986), clearly coincides with the increase in design-based research (Angrist and Pischke, 2009; Currie et al., 2020).

A design-based perspective also brings insights to the sensitivity critique of Leamer (1983). Identifying assumptions, which involve restrictions on the treatment assignment process or comparability of outcomes for treatment and control groups, are emphasized and distinguished from other choices in the estimation procedure. The "parallel trends" assumption underlying the difference-in-differences analysis of Card and Krueger (1994), or the key IV assumptions of random assignment, exclusion, monotonicity, and relevance in the framework of Imbens and Angrist (1994), are presented and scrutinized. Other parts of estimation, such as exactly how the researcher controls for covariates or weights different subpopulations, are less central when they are not informed by the research design. The effects of such choices can be assessed by more routine sensitivity analysis, while the plausibility of identifying assumptions might require institutional knowledge to assess. This hierarchy of assumptions reflects an emphasis on design, with transparent treatment–control comparisons over potentially complex models and statistical procedures. A state-of-the-art non-parametric IV estimator based on an instrument with unclear assignment might be less convincing than a simple comparison of trends in a well-executed difference-in-differences analysis. As Rubin (2008) puts it, "[for] causal inference, design trumps analysis".

The design-based approach advanced by the laureates also raises new issues for empirical practice. While a clear and plausible research design can lead to *internally* valid estimates of causal effects, which are free from selection bias in a given study population, the *external* validity

(i.e., generalizability) of such estimates to other populations and contexts is far from guaranteed. Similarly, the search for treatments with clear assignment processes or plausible control groups might limit the scope of microeconomic research. Some questions are more easily cast in a design-based approach, while others might seem fully out of reach by concerning treatments that are not easily viewed as manipulable by any design. We return to these and other issues in Section 4.

2.4. Natural experiments

Where do convincing research designs come from? The most obvious source is an RCT, with true experimental assignment. Experimentation might be a natural way to answer some questions in economics, such as the effectiveness of job training programs, as argued in Ashenfelter and Card (1985) and Ashenfelter (1987). Indeed, the past 30 years have seen a steep rise of field experiments through microeconomics (Levitt and List, 2008). Prominent examples range from randomly offering housing vouchers to allow low-income households to move to better neighborhoods through the Moving to Opportunity Program (Kling et al., 2007; Chetty et al., 2016); to randomly offering Medicaid insurance coverage to low-income adults in the Oregon Health Insurance Experiment (Finkelstein et al., 2012; Baicker et al., 2013); to entrepreneurial researchers setting up their own charity to randomize how a charity solicits donations in a study of the determinants of charitable giving (DellaVigna et al., 2012). The field of development economics, in particular, has seen a stunning transformation in the share of experimental papers, which include evaluating the effect of educational policies such as deworming children (Miguel and Kremer, 2004) or changing teaching incentives (Glewwe et al., 2010), studying the effects of microcredit provision (Banerjee et al., 2015), or the rates of return to fertilizer (Duflo et al., 2008, 2011).[4] In a randomized controlled trial, the research design is clear and often under the researcher's control. Validating the identifying assumption of random assignment is also straightforward: if units are assigned to different treatment randomly, observable pre-treatment characteristics should be unrelated to treatment status. Such balance can be checked by testing if the distribution of covariates is the same across the treatment arms.

Much of the work by the 2021 laureates, however, considers economic questions where direct experimentation is rare or infeasible. Here, a convincing research design might come from a natural or quasi-experiment,

[4]Three of the scholars leading this transformation – Abhijit Banerjee, Esther Duflo, and Michael Kremer – were awarded the Nobel Memorial Prize in 2019, "for their experimental approach to alleviating global poverty".

in which some institutional quirk or force of nature generates variation that is plausibly as-good-as-randomly assigned or which otherwise suggests appropriate treatment and control groups.[5] Such variation can take the form of large but unforeseen shocks to regions or markets. Freeman (1989) was an early proponent of basing empirical analyses around such large shocks, illustrating the value of this approach in an analysis of the federal minimum wage imposition on Puerto Rico in the 1970s (Castillo-Freeman and Freeman, 1992), and in an analysis of a post-Sputnik boom in the demand for physicists in the US (Freeman, 1975). Other quasi-experimental analyses leverage more narrow variation, often across individuals in the same region or market, where the research design arises from idiosyncrasies in the rules used to administer some economic variable. An example of this approach is the regression discontinuity (RD) design of Thistlethwaite and Campbell (1960), where a threshold assignment rule (such as a minimum test score for students to avoid taking remedial classes) can be leveraged to study the effect of assignment among individuals just above and just below the threshold. Campbell (1969) was an early proponent of using such discontinuities and other quasi-experimental designs in psychology.[6] Variation in the timing of policy shocks across regions, such as US states, can also provide a basis for a compelling research design: Gruber (1994) exploited such variation in an influential study of the incidence of mandated maternity benefits.

Labor economists were early adopters of natural experiments as the foundation of a research design. This work included studies by Rosenzweig and Wolpin (1980), Meyer et al. (1995), and a group of researchers at the Industrial Relations Section (part of the Princeton University Economics Department). Gary Solon's work (Solon, 1985) showed how one can leverage changes in laws as a source of variation. Orley Ashenfelter and Alan Krueger famously combined a random quirk of nature with innovative data collection to study the returns to education in Ashenfelter and Krueger (1994). The researchers traveled to the annual Twins Days Festival (in Twinsburg, Ohio) to gather data on pairs of identical twins. Independently asking both twins about each other's schooling levels allowed them to account for measurement error in self-reported years of schooling – a prominent concern in existing returns-to-schooling studies. By looking at

[5] DiNardo (2010) differentiates natural experiments as "serendipitous randomized trails", where a variable of interest is as-good-as-randomly assigned across units, and quasi-experiments, which rely on parallel trends assumptions or other restrictions on the comparability of unobservables across treatment and control groups. However, like "research design", exact definitions of these terms vary; see, for example, Titiunik (2021) for alternative definitions and discussion.

[6] The review by Meyer (1995) helped to make economists aware of this quasi-experimental tradition in psychology.

how differences in twins' education levels predict differences in twins' earnings, Ashenfelter and Krueger were further able to address concerns of ability bias in existing estimates. Here, the identifying assumption follows from the natural experiment of twinning, which generates two individuals with identical genetics and thus (arguably) comparable earnings potential. Leveraging the more narrow variation in education within twin pairs, Ashenfelter and Krueger found much greater labor market returns than in previous cross-sectional studies that adjusted for observable demographics and family characteristics.

Two other prominent members of the Industrial Relations Section are two of the 2021 laureates: David Card and Joshua Angrist. We next discuss how their work made both empirical contributions to key questions in labor economics and methodological advances in using design-based approaches to answer them.

3. Empirical landmarks

Empirical work by the laureates is distinguished by its clarity and compelling analysis of several important topics. We organize a partial review of this work into four broad topics: the effects of immigration on local labor markets, the effects of minimum wages on low-wage employment, the labor market returns to schooling, and other determinants of labor earnings.

3.1. Effects of immigration on local labor markets

Immigration raises several important and hotly debated questions in economics and public policy. A key policy concern in many countries is that relaxed immigration laws can induce large inflows of migrants, sometimes with low levels of education and labor market experience, to compete with local workers and potentially reduce wages and employment prospects. Theoretical predictions about the effects of such low-skilled labor supply shocks to local labor markets are ambiguous, since they depend on many factors – including the substitutability of immigrant and native labor, existing labor market institutions, and the potential response in labor and product demand. Early empirical studies of this question typically used cross-sectional variation in immigrant populations to estimate production function parameters that speak to these mechanisms (e.g., Grossman, 1982). As noted by Borjas (1987), however, the self-selection of immigrants to different labor markets might bias such estimates.

Two early studies by Card (Card, 1990; Altonji and Card, 1991) show how an increased focus on research design and large-scale natural

experiments can address such selection concerns. In Card (1990), the so-called 1980 Mariel boatlift was used to study the local labor market effects of low-skill immigration in Miami. This large and arguably unanticipated labor market shock grew out of rising political unrest in Cuba, and led to an unprecedented influx of immigrants in the Miami labor market. The boatlift was announced on 20 April 1980, with Fidel Castro allowing anyone wishing to emigrate from Cuba to leave through the port of Mariel. Between April and October 1980, nearly 125,000 people left Cuba via a flotilla of private vessels. Roughly half settled in Miami, where several processing camps had been established.

Three aspects of the Mariel episode made for a compelling natural experiment. First, the source of the immigration shock was clear and external: the political turmoil that gave rise to the boatlift was plausibly unrelated to other factors affecting labor markets in the United States. Second, the scale of the shock was enormous: the boatlift led to a 7 percent increase in the size of Miami's total labor force over the span of a few months. The Mariel immigrants tended to be less educated and worked in lower-skill occupations than the overall Miami population, implying an even larger increase in labor supply at the lower end of the skill distribution. Third, large survey data sets contained the variables needed to study this large and external shock: baseline labor market conditions immediately prior to the boatlift could be measured in the 1980 census, while the CPS included reasonably large samples for measuring effects in subsequent years. Importantly, unlike most other ethnic groups, Cubans are separately identified in the CPS – allowing Card to study outcomes for both Cubans and other Hispanic immigrants.

Despite these advantages, the appropriate way to leverage the Mariel boatlift experiment to study immigration effects is not immediately clear. Miami is only one city, and neither the timing nor the location of the immigration shock were truly random. Moreover, the onset of the 1982 recession in the aftermath of the Mariel boatlift highlighted the possibility that the effects of the immigration shock could be confounded by changes in other national or regional labor market trends.

To allay these concerns, Card constructed a comparison group that might plausibly account for other time-varying factors and isolate the impact of immigration. Specifically, Card compared Miami's labor market trajectory to those of Atlanta, Los Angeles, Houston, and Tampa-St Petersburg – a group of cities that featured roughly similar demographics and exhibited similar trends to Miami prior to the boatlift. The idea of forming a control group to adjust for time-varying confounders in a non-experimental setting grew out of Card's earlier work on longitudinal earnings models and training programs (Ashenfelter and Card, 1985; Abowd and Card, 1989).

The analysis of Card (1990) revealed that wages and unemployment moved similarly in Miami and the comparison cities before and after the boatlift, including for lower-skilled groups of workers. This suggests that the large influx of Mariel immigrants had limited effects on native outcomes, a surprising finding. Methodologically, the study of Card (1990) provided a clear example of how to combine a natural experiment with a carefully constructed control group to produce compelling empirical findings.[7] This paper presaged the growth of difference-in-differences studies, which have since become one of the most common empirical strategies in applied microeconomics.[8]

The findings in Card (1990) were striking, and not without critique.[9] One clear concern was generalizability: the Mariel experiment was a large shock, but its effects were concentrated in one arguably unique labor market. Indeed, Card (1990) notes that Miami's long history of receiving Cuban immigrants might complicate the interpretation of the findings. For a more comprehensive view of local labor market effects, Altonji and Card (1991) famously devised an IV strategy that translated the logic of the Card (1990) research design to a national level. Just as Cuban immigrants tended to locate to cities such as Miami, where there were large groups of previous Cuban immigrants, immigrants from other countries tend to settle in regions with existing immigrant enclaves. Altonji and Card used previous immigrant settlement patterns to instrument for immigration to different metropolitan areas, finding large negative effects of immigrant inflows on native wages but no effect on employment. Later, Card (2001) refined this IV strategy with a "shift–share" instrument that predicted inflows by city and occupational groups. This shift–share approach is now widely used to study the effects of immigration and other treatments combining large external shocks (e.g., immigrant inflows) and heterogeneous local exposure (e.g., immigrant enclaves).[10]

The design-based approach to immigration study has gained immense popularity in the years following Card (1990) and Altonji and Card (1991), and has generated several strands of literature seeking to explain why

[7]The synthetic control method (Abadie and Gardeazabal, 2003; Abadie et al., 2010), which combines multiple control groups to construct a single synthetic control mimicking the treatment group, can be seen as a further refinement of this idea.

[8]Currie et al. (2020) find that nearly a quarter of NBER working papers in 2020 employed difference-in-differences, constituting around 60 percent of all NBER working papers using an experimental or quasi-experimental approach.

[9]For a recent discussion, see Borjas (2017) and Peri and Yasenov (2018).

[10]Shift–share, or Bartik, instruments can be traced back to Freeman (1980), Bartik (1991), and Blanchard and Katz (1992). A recent methodological literature, including Goldsmith-Pinkham et al. (2020), Borusyak et al. (2022), and Adão et al. (2019), formalizes how such instruments can leverage quasi-experimental variation.

immigration has limited effects on the labor market outcomes of native workers in some – but not all – settings (for a recent review, see Dustmann et al., 2016). Key sources of heterogeneity appear to include the distribution of native skill (particularly communication skills; e.g., Peri and Sparber, 2011) and the ease of technological adjustment (e.g., Dustmann and Glitz, 2016). While there is still ongoing debate over the magnitude of wage and employment effects from immigration overall, such studies of heterogeneity and mechanisms are no doubt helped by a clearer understanding of research design.

3.2. Effects of minimum wages on low-wage employment

A similarly important (and fiercely debated) question in economics and public policy is the effects of federal or local wage floors on low-wage employment. The textbook model of a perfectly competitive labor market predicts that an increase in the minimum wage results in movement along a downward-sloping market labor demand curve, creating unemployment and reducing worked hours among the employed. By the early 1990s, the conventional view among labor economists was that these predictions accurately describe the causal impacts of changes in minimum wage laws. It had long been recognized that increasing the minimum wage can theoretically boost employment by flattening the supply curve facing an employer with market power (Robinson, 1933), but this scenario was generally regarded as specific to situations with a single monopsonistic employer and irrelevant to the functioning of typical low-wage labor markets. Empirical evidence from the 1970s and 1980s, largely based on time-series or cross-sectional variation in minimum wages across states, was broadly consistent with the competitive view (for reviews, see Brown et al., 1982; Card and Krueger, 1995b).

Card and Krueger (1994) famously revisited the effects of the minimum wage using a natural experiment derived from an increase in New Jersey's state minimum wage. Their strategy built on earlier work by Card (1992a), who analyzed the effects of a minimum wage increase in California using a comparison set of unaffected states, and by Katz and Krueger (1992), who studied an increase in the federal minimum wage by comparing establishments paying higher versus lower wages prior to the change (see also Card, 1992b). In 1990, the New Jersey legislature passed a law that would increase the state's minimum wage from $4.25 to $5.05 per hour as of 1 April 1992. Anticipating this change, Card and Krueger (1994) surveyed a set of fast food establishments on both sides of the New Jersey/Pennsylvania border immediately prior to the change (February–March 1992) and again a few months afterward (November–December 1992). Combining elements

of the Card (1992a) and Katz and Krueger (1992) strategies, Card and Krueger (1994) compared changes in outcomes in New Jersey to those in Pennsylvania, as well as changes in outcomes for restaurants with higher versus lower baseline wages within New Jersey.

Like the Mariel boatlift analysis, the New Jersey/Pennsylvania study exhibits several hallmarks of modern studies of natural experiments. The "action" in the variable being studied (the minimum wage) originated from a specific and interpretable source (the New Jersey law change) rather than uncontrolled state-level variation of unclear origin. The size of this shock was large: New Jersey's minimum wage increased by roughly 20 percent, and most fast food restaurants in New Jersey paid below $5.05 before the change. Careful attention was paid to constructing and validating a control group that could plausibly capture the counterfactual path of outcomes in the absence of treatment for affected units. The controls here consisted both of restaurants across the border in Pennsylvania and of high-wage New Jersey restaurants less exposed to the reform. Finally, Card and Krueger (1994) assembled detailed microdata to measure outcomes for the treatment and control groups. In this case, rather than relying on existing surveys, they fielded their own custom survey tailored to the question at hand.

The baseline survey of Card and Krueger in February/March 1992 showed roughly similar wage distributions in New Jersey and Pennsylvania, with average starting wages just over $4.60 and about one-third of restaurants in each state paying exactly the baseline minimum wage of $4.25. By the endline survey in late 1992, about 90 percent of New Jersey restaurants reported paying exactly the new minimum wage of $5.05, while the wage distribution in Pennsylvania appeared roughly unchanged. Despite this large differential change in wages, the survey showed no evidence of a negative employment effect on New Jersey restaurants. In fact, average full-time employees (FTEs) at New Jersey stores increased slightly, while average FTEs in Pennsylvania fell, resulting in a modest positive difference-in-differences estimate. An "exposure design" comparing high- and low-wage employers within New Jersey likewise showed a small relative increase in employment at low-wage restaurants. In contrast to the textbook perfectly competitive model and the earlier time-series evidence, the empirical results of Card and Krueger suggested that increasing the minimum wage did not reduce employment – and, if anything, might have increased it.

The empirical findings of Card and Krueger (1994) upended conventional wisdom on the effects of minimum wages, generating backlash in some quarters of labor economics. Despite the well-known theoretical result that minimum wages could increase employment in settings with employer market power, adherents of the competitive view of labor markets

derided the Card and Krueger (1994) findings as unscientific (see, e.g., Buchanan, 1996). While negative, such reactions highlight the value of natural experiments and careful research design. In contrast to empirical work from earlier eras, in which Leamer (1983) argued that "hardly anyone takes anyone else's data analysis seriously", the results from a compelling natural experiment call out for explanation and further study even among skeptics of the substantive conclusions. As it turns out, the conclusions of several recent studies are broadly consistent with the Card and Krueger (1994) finding of limited effects of the minimum wage on employment (Dube et al., 2010; Giuliano, 2013; Cengiz et al., 2019; Harasztosi and Lindner, 2019; Dustmann et al., 2022). Partially motivated by these findings, an increasing body of work has investigated competitive structure and monopsony power in labor markets; see Card et al. (2018) and Manning (2021) for two recent reviews. Renewed interest in employer monopsony power – following Card and Krueger (1994), the extended treatment in Card and Krueger (1995a), and the analysis of Manning (2003) – has since fueled a large literature on firm wage-setting; for example, see recent work by Azar et al. (2020), Kroft et al. (2020), Lamadon et al. (2022), and Berger et al. (2022).

3.3. Effects of schooling and experience on earnings

Studies in economics and related fields have long considered the effects of education and labor market experience on subsequent earnings and employment. While the theoretical impact of increased human capital is unambiguous (e.g., Becker, 1964), the empirical evidence for such causal effects was limited throughout most of the 20th century. A famous report by Coleman (1966) showed in cross-sectional regressions that the fraction of variance in student achievement attributable to educational inputs was small relative to the contribution of family background. Surveying the large body of empirical literature following the Coleman report, Hanushek (1986) concluded that there was virtually no relationship between educational inputs and subsequent outcomes. Of course, selection bias looms large for such analyses as the deployment of educational resources to students and schools is far from random.

Two influential studies by David Card and Alan Krueger (Card and Krueger, 1992a,b) investigated the effects of school quality on labor market outcomes by isolating a clever source of variation: the movement of students across different US regions. Card and Krueger (1992a) estimated returns to schooling separately by cohort and state of birth, controlling for cohort-specific effects of state of birth and state of residence in the 1980 US census. This strategy compares relationships between earnings

and schooling for individuals in the same birth cohort working in the same state but educated in different states, leveraging cross-state moves to measure differences in cohort-specific school quality across states. Card and Krueger related these estimated returns to measures of school quality for each state and birth cohort, showing that school quality improvements such as reduced pupil/teacher ratios can increase the return to education. To study the role of school quality in the evolution of the racial wage gap, Card and Krueger (1992b) estimated separate returns to schooling by race, state of birth, state of residence, and birth cohort. Between cohorts born in the 1920s and the 1940s, they documented a striking relative increase in the return to education for Southern-born black men compared with both non-Southern-born black men and Southern-born white men within regions of residence. The timing of this differential change in returns coincided with a relative increase in measures of school quality for Southern-born black men, suggesting an important role for school quality in reducing the racial wage gap over time. Though these studies did not take advantage of a sharp policy change, they effectively used individuals moving between locations as a collection of natural experiments – removing permanent differences between locations of birth to flexibly account for unobservables. Recent work in several areas builds on this idea of using "movers" to mitigate selection bias, including studies of firm effects (Abowd et al., 1999; Card et al., 2013, 2016), neighborhood quality (Chetty and Hendren, 2018), and variation in regional health-care utilization (Finkelstein et al., 2016).[11]

These Card and Krueger studies suggest a non-zero "return to schooling": the theoretical parameter governing causal effects of increased education on labor market earnings. Perhaps the most famous estimates of the returns to schooling from this period come from Angrist and Krueger (1991), who used a creative IV strategy – based on an institutional quirk of the US education system – to address the clear self-selection issue of earlier regression-based analyses.[12] Children in the US traditionally start first grade in the calendar year in which they turn six, but historically most state compulsory schooling laws allowed students to drop out on their sixteenth birthday. The combination of these two rules implies that compulsory schooling laws were more stringent for students born early in the calendar year. Consider, for example, two children born in 1930 with one born in January and the other born in December. These children would likely be in the same schooling cohort, starting first grade together in the

[11]Earlier examples of papers using such designs to study industry wage differentials include Murphy and Topel (1987), Krueger and Summers (1988), and Gibbons and Katz (1992).

[12]Another influential contribution is Card (1995), who used the distance to nearby colleges in an individual's birthplace as an instrument for educational attainment.

fall of 1936. The individual born in January would be among the oldest of his or her classmates, reaching the compulsory school age of 16 in January 1946, in the middle of tenth grade. The child born in December, in contrast, would be among the youngest in the class and be compelled to stay in school until the middle of eleventh grade. If both students drop out as soon as the law allows, the December child will thus attain nearly a full year of additional schooling than the child born in January.

This argument suggests a novel instrument for years of schooling: a child's birthday. The interaction between age-at-entry and compulsory schooling rules suggests that birthdays might affect educational attainment. Moreover, it seems plausible that birthdays are as good as randomly assigned, and have no effects on earnings through channels other than completed schooling.[13] Angrist and Krueger (1991) operationalized this idea using instruments based on season (quarter) of birth, the measure of birthdays available in public-use decennial census data.[14]

The Angrist and Krueger analysis revealed a clear relationship between birthdays and education among men born in the 1920s through to the 1940s. On average, children born in the second to fourth quarters of the year stay in school one-tenth of a year longer than those born in the first quarter. Angrist and Krueger presented a battery of falsification exercises suggesting that this pattern is due to their proposed compulsory schooling mechanism. For example, using point-in-time school enrollment for teenagers in the 1960 and 1970 censuses, they demonstrated that the gap in enrollment between first- and later-quarter births emerges at age 16 only in states where the compulsory school-leaving age is 16 rather than 17 or 18. This can be seen as an early example of the placebo and robustness checks that are now commonly used to probe identifying assumptions in design-based studies.[15] In addition to attaining 0.1 fewer years of schooling, individuals born in the first quarter of the year also earn about 1 percent less than those born later. IV estimates formed as the ratio of these two differences (as discussed

[13]Buckles and Hungerman (2013) note that maternal characteristics vary with birthday in recent birth cohorts, suggesting that birthdays might not be fully independent of family background.

[14]With more detailed data on date of birth, this strategy can be sharpened into an RD design leveraging the shift in school entry dates for children born immediately before and after the turn of the calendar year. An example of this approach appears in Clark and Royer (2013).

[15]This exercise strengthens the case for a causal interpretation of the quarter-of-birth first stage. In traditional simultaneous equations models, a causal interpretation of the first stage is unnecessary – the first stage is a linear projection and any bias stems from a relationship between the instrument and unobservables in the outcome equation. In the design-based view, however, it is seen as unlikely that an instrument is as-good-as-randomly assigned unless both the first stage and reduced form are free of selection bias. This idea is made explicit in the framework of Imbens and Angrist (1994), which features a causal model of the first stage as detailed in Section 4.

further in Section 4.1) imply that a year of schooling boosts earnings by roughly 10 percent.

Methodologically, the Angrist and Krueger (1991) analysis differed from several of the above studies of immigration and the minimum wage in two important ways. First, while the Card (1990) and Card and Krueger (1994) studies looked at large and sudden shocks to aggregate labor markets, Angrist and Krueger (1991) leveraged narrow quarter-of-birth variation across individuals within markets. Second, while the 1980 Mariel boatlift and 1992 New Jersey minimum wage change were paired with natural comparison groups, unlike with dates of birth it is difficult to imagine these deliberate policy changes as occurring by chance. The narrow and plausibly as-good-as-randomly assigned IV variation in Angrist and Krueger (1991) thus marked a key methodological shift in the use of natural experiments in economics while – as we discuss below – highlighting new econometric questions.

Substantively, the studies by Card, Angrist, and Krueger helped change the consensus on the effects of educational investment, from the rather pessimistic conclusion of the Coleman (1966) report to the modern consensus that school resources generally matter (see Jackson, 2020 for a recent review). As with the immigration and minimum wage literatures, recent work focuses on the heterogeneity of educational input effects across settings and individuals. IV-based analyses of charter school effects and variation in school quality within urban districts (Angrist et al., 2010, 2017; Abdulkadiroğlu et al., 2011, 2016) is one area where Angrist has continued to study such heterogeneity.[16]

3.4. Other determinants of earnings

Three other studies by the laureates (Angrist, 1990; Imbens et al., 2001; Card and Hyslop, 2005) are worth highlighting, for the creative use of naturally occurring randomization to answer important questions on the determinants of labor market earnings. In Angrist (1990), the random assignment of draft lottery numbers was used to study the effects of Vietnam-era military service on earnings. Paralleling the issues in the LaLonde (1986) analysis of training program effects, military veterans differ from non-veterans on many dimensions, and earlier efforts to address this selection with the available econometric tools yielded unstable and inconclusive estimates. Angrist (1990) leveraged public lotteries that assigned random sequences numbers (RSNs) to dates of birth for men born

[16]Other notable examples include Angrist and Lavy (1999), which uses RD to study class size effects, and the RCT of Angrist et al. (2002), which studied private school vouchers.

between 1950 and 1955. For the 1950–1952 birth cohorts, men whose RSNs fell below a cut-off were conscripted into military service.[17] Randomly assigned draft lottery numbers are clearly independent of earnings potential and plausibly affect outcomes only through military service, making the draft an attractive natural experiment for studying service effects on labor market outcomes.[18]

Angrist (1990) obtained a custom version of the Social Security Administration's Continuous Work History Sample (CWHS), augmented with dates of birth, in order to link earnings to draft RSNs. Veteran status was only partially determined by RSNs, however: some men enlisted regardless of lottery number, while many with low lottery numbers did not serve due to deferrals or performance on pre-induction mental and physical screening tests. As a result, the difference in military service rates for eligible and ineligible men was only around 15 percentage points. This non-compliance calls for an IV, with the RSN serving as instrument for veteran status. Angrist (1990) divided the difference in mean earnings by draft eligibility in the CWHS by the difference in service rates in the Survey of Income and Program Participation (SIPP) to construct IV estimates of the causal impact of military service. The results showed that military service led to a 15 percent earnings penalty for the Vietnam draft cohorts.

The study of Imbens et al. (2001) uses randomization from a lottery to shed light on a different question: what are the wage effects of non-labor income? The effect of unearned income on economic behavior is a foundational question in labor and public economics but is difficult to measure because non-labor income is likely to be correlated with many unobserved determinants of labor supply and other outcomes. To address this challenge, Imbens et al. (2001) conducted a special survey of Massachusetts lottery players. The sampling frame for the survey took advantage of the fact that the state maintains historical records of lottery winners, including some individuals who won millions of dollars and some who won small amounts. Winners of small prizes provide a natural control group for bigger winners. To lend support to the key identifying assumption that the magnitude of the prize is as good as randomly assigned, Imbens et al. (2001) conducted balance checks that showed no correlation between the prize magnitude and individual characteristics (e.g., prior earnings) once the winners of the largest prizes are excluded. Their analysis

[17]As Angrist (1990) notes, RSNs also generated an increase in military service for those born in 1953 even though this cohort was never drafted, as men with low lottery numbers pre-emptively enlisted to improve their terms of service in anticipation of the possibility of conscription.

[18]Earlier work by Hearst et al. (1986) used the Vietnam-era draft lottery to study the effects of draft eligibility on mortality.

revealed modest negative effects of unearned income on labor earnings, with somewhat larger impacts for older workers. Recent studies of the impacts of unearned income have followed the Imbens et al. (2001) lottery-based approach (see, e.g., Cesarini et al., 2017).

Like these two studies, Card and Hyslop (2005) leveraged natural randomization to study the wage effects of a large-scale program: the Self Sufficiency Project (SSP), a Canadian program that made earnings subsidies available for a random pool of long-term welfare recipients over three years. Such effects speak to a large literature in labor and public economics considering possible employment disincentives from mean-tested welfare programs, such as the Earned Income Tax Credit (EITC). Unlike the EITC and earnings subsidies in other countries, the SSP was only available for full-time work; furthermore, participants had to establish eligibility by working full-time within the first year of program participation. Card and Hyslop (2005) showed how this program structure created distinct incentives to find a full-time job in the first year and to continue working once eligibility was established. To tease apart these channels, they developed and estimated a dynamic model with the experimental variation in program participation. Their estimates showed that the combination of "establishment" and "entitlement" incentives generated a striking pattern in the experimental effects, which peaked in the second year following random assignment before fading. Notably, there were no long-run effects on either wages or welfare participation, suggesting that temporary wage subsidies might not induce program dependency. Beyond these substantive findings, the Card and Hyslop (2005) approach to estimate structural models by exploiting natural experiments helped push the frontier of the design-based approach, as we discuss more below.

3.5. Taking stock

These and other early studies of natural experiments produced new and compelling evidence on several classic questions in labor economics. At the same time, they often raised new questions about how such evidence is best interpreted and synthesized into a broader body of scientific knowledge. The increased emphasis on the forces determining the assignment of certain economic "treatments" highlighted that design-based studies often leverage highly specific sources of variation. What does the Card and Krueger (1994) result for fast-food workers in New Jersey teach us about the impact of minimum wage more broadly? Is the lack of labor market effects from a large immigration shock in Card (1990) specific to the 1980s Miami labor market? These questions of interpretation loom especially large in the IV analysis of Angrist and Krueger (1991), where the identifying variation in individual dates of birth led to relatively small differences in completed

education. Comparing the schooling of individuals born in the first and fourth quarters of a year suggests that at most 10 percent were on the margin of dropping out as soon as they are legally allowed. To what extent can this narrow source of variation inform the returns to schooling in the general population?

The interpretation of design-based IV estimates of the returns to schooling was carefully considered in an influential review by Card (1999). He noted that such estimates – including in Angrist and Krueger (1991) – typically exceed corresponding ordinary least-squares (OLS) estimates, often by 30 percent or more. This pattern would seem to present a puzzle: standard selection bias reasoning suggests that OLS estimates should be biased *upward*, not downward, as students with higher unobserved ability are likely to select more schooling. While measurement error in self-reported years of education could explain some of the discrepancy, as it would tend to attenuate the OLS estimates, it is unlikely to explain the often large gap between IV and OLS estimates. To close this gap, Card (1999) offered another explanation: the subpopulations shifted into treatment by the variation in the quarter of birth or other IV strategies might have higher returns to schooling than the overall population. This idea of heterogeneous causal effects driving the interpretation of IV estimates was formalized in the ground-breaking analysis of Imbens and Angrist (1994).

4. A deeper understanding of causality

Interpreting estimates of conceptually similar causal effects across different natural experiments and designs requires a flexible econometric framework. Specifically, it requires a way to think about how the specific source of identifying variation might affect the interpretation of the quantity being estimated. Consider, for example, the Vietnam draft study of Angrist (1990), where draft eligibility (i.e., an RSN below the conscription cut-off) was used to instrument for military service. Most individuals who served in Vietnam were volunteers who would have served no matter their RSN number. Presumably, the draft-based identification strategy of Angrist (1990) cannot speak to the effect of serving in the military for these volunteers. But formally, what effects can it capture?

A standard way of motivating the use of IV methods in such applications is selection (or "omitted variables") bias: individuals who do and do not serve in the military differ in many observed and unobserved ways, some of which might affect their adult earnings. One way to address this concern is to model the selection process; Heckman (1974, 1976, 1979), Heckman and Robb (1985), Chamberlain (1986), and others showed how

IV methods could identify such selection models through a combination of exclusion and functional form restrictions.[19] Selection models can also be used to structure heterogeneity across different instruments and samples – at least when it is clear what observable and unobservable characteristics are relevant and in what way. But such modeling might be hard to do in a flexible manner. More importantly, a model-based approach to bias and heterogeneity might obscure the advantage of a randomly assigned instrument, as in Angrist (1990).

A key innovation in Imbens and Angrist (1994) is to approach the selection bias problem and IV solution from a different direction. Instead of deriving model restrictions sufficient to fully correct selection, and align the IV analysis with a hypothetical randomized experiment, Imbens and Angrist asked what minimal assumptions make a simple linear IV estimand causally interpretable. Their answer helped to separate the conceptual roles of "chance" (i.e., as-good-as-random instrument assignment) and "choice" (assumptions on the selection process) in design-based analyses, clarified how different quasi-experimental studies of conceptually similar economic quantities could be synthesized, and highlighted more general strengths and weaknesses of the design-based approach.

4.1. Potential outcomes framework and LATEs

To keep individual heterogeneity in the response to treatment unrestricted, Imbens and Angrist (1994) cast the IV estimation problem in a potential outcomes framework. This framework dates back to Neyman (1923), who first proposed a version of it for analyzing randomized experiments, and to Rubin (1974, 1978, 1990), who later generalized it for observational studies. The core logic of the potential outcomes framework is also found in the early econometric literature, including work on the IV approach as a method to solve simultaneous causality. Wright (1928), Working (1927), Tinbergen (1930), and Haavelmo (1943) all distinguished between potential economic variables determined by structural relationships and the observed variables determined in market equilibria.[20] The distinction between the observed outcomes of simultaneous equations models (such as equilibrium prices

[19]Other work showed how average effects can be bounded without such restrictions (see, e.g., Robins, 1989; Manski, 1990; Balke and Pearl, 1997).

[20]This early literature also can be seen as laying the foundation of later graphical formalizations of causality and related methods, particularly the path analysis method of Wright (1928). The do-calculus of Pearl (1995, 2000); Pearl and Mackenzie (2018) has evolved in parallel with the potential outcomes framework in recent years, though the latter generally remains more popular in applied economics. See Heckman and Pinto (2014), Pearl (2015), and Imbens (2020) for recent discussions.

and quantities) and the "potential" outcomes that might be realized under certain counterfactuals is especially clear in Haavelmo (1943), who focused on the challenge of interpreting observed data on income and consumption in terms of parameters governing marginal propensities to consume and invest.[21] The emphasis on potential outcomes was revived in the early 1990s by Heckman (1990), Manski (1990), and others, along with Imbens and Angrist (1994); these papers showed the value of the clarity that explicit potential outcome notation delivers.[22]

To sketch the potential outcomes framework, consider the causal effect of some binary treatment D_i (say, enlisting in the army) on some subsequent outcome Y_i (say, adult earnings). We imagine two potential outcomes associated with the treatment, $Y_i(1)$ and $Y_i(0)$, representing the earnings of individual i if they did and did not to enlist in the army. Only one of these potential outcomes is observed, depending on the value of D_i; the other is the individual's *counterfactual* outcome, associated with the unrealized treatment state. Formally, the observed outcome can be written as

$$Y_i = (1 - D_i)Y_i(0) + D_iY_i(1) = Y_i(0) + D_i(Y_i(1) - Y_i(0)), \tag{1}$$

where the quantity $Y_i(1) - Y_i(0)$ represents the effect of D_i on Y_i for individual i.

When D_i is randomly assigned (i.e., we were to randomly enlist some individuals in the military but not others), it becomes independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$. This ensures that the ATE is identified by the difference in average outcomes among treated and untreated individuals:

$$\begin{aligned} E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] &= E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0] \\ &= E[Y_i(1) - Y_i(0)], \end{aligned}$$

where we use equation (1) in the first equality and the random assignment assumption in the second equality. The potential outcome notation makes the magic of a randomized experiment transparent: because we never observe both potential outcomes for each individual, we can never learn their individual treatment effect, $Y_i(1) - Y_i(0)$. However, by virtue of random

[21]Trygve Haavelmo received the Nobel Memorial Prize in 1989, "for his clarification of the probability theory foundations of econometrics and his analyses of simultaneous economic structures". Jan Tinbergen was awarded the first Nobel Memorial Prize in 1969, along with Ragnar Frisch, "for having developed and applied dynamic models for the analysis of economic processes".

[22]As Imbens (2014) notes, much of the post-war econometric literature used a notation only involving realized or observed outcomes; see also Hendry and Morgan (1992) and Imbens (1997) for discussions of this history.

assignment, we can still learn the value of the treatment effect on average, in the population of interest.

To adapt the potential outcomes framework to an IV setting, Imbens and Angrist (1994) defined two sets of potential outcomes: one set for the treatment D_i and one for the outcome Y_i.[23] Specifically, let $D_i(z)$ be the potential treatment status when $Z_i = z$. In the Angrist (1990) example, for instance, $D_i(0)$ is the military status of the individual if they were draft-ineligible, while $D_i(1)$ is the potential military status if they were draft-eligible. Because both the treatment D_i and the instrument Z_i can be manipulated, Imbens and Angrist defined potential outcomes $Y_i(d, z)$ over potential values d of the treatment D_i, and potential values z of the instrument Z_i. These correspond to earnings under each combination of serving in the military and draft eligibility.[24] They then considered four substantive assumptions:

random assignment, $(Y_i(0,0), Y_i(1,1), Y_i(1,0), Y_i(0,1), D_i(0), D_i(1)) \perp\!\!\!\perp Z_i$;

exclusion, $\Pr(Y_i(d,0) = Y_i(d,1)) = 1$ for each $d \in \{0,1\}$;

monotonicity, $\Pr(D_i(1) \geq D_i(0)) = 1$; and

relevance, $\Pr(D_i(1) > D_i(0)) > 0$.

The first assumption requires the instrument to be as-good-as-randomly assigned with respect to the potential outcomes and potential treatment choices. This assumption holds automatically for instruments such as the draft-eligibility, or arguably the quarter-of-birth instrument in Angrist and Krueger (1991). Under random assignment, we can estimate the average effect of the instrument on the treatment, $E[D_i(1) - D_i(0)]$, following the logic of randomized experiments above. By the same logic, we can also estimate the "intent-to-treat effect": the average effect on the outcome of switching the instrument from zero to one. In Angrist (1990), this represents the effect on earnings of being draft-eligible.

The second assumption, or "exclusion restriction", requires any effects of the instrument Z_i on the outcome Y_i to arise from changes in the treatment D_i; varying the instrument while holding the actual treatment fixed has no effect on the outcome. This condition allows us to define potential outcomes $Y_i(d)$ indexed by treatment status d alone, as in the case of a randomly assigned treatment. The first two assumptions together

[23] Footnote 2 in Imbens and Angrist (1994) attributes the adoption of potential outcome notation for D_i to Gary Chamberlain, who – along with Donald Rubin – were faculty members at Harvard when Imbens and Angrist started there as assistant professors.

[24] Imbens and Angrist (1994) derived the LATE theorem with a multivalued Z_i, but we focus on the case with binary Z_i here for simplicity.

capture the sense in which the instrument is "exogenous" in conventional IV analysis. The potential outcomes framework makes it clear that there are actually two separate assumptions underlying this condition. Even if the instrument is randomly assigned, it might fail the exclusion restriction if, for instance, draft-eligible individuals temporarily leave the country in order to avoid the draft and this dodge has an effect on later-life earnings.

The third monotonicity assumption, most original to Imbens and Angrist (1994), requires the instrument to affect the treatment status only in one direction. Without loss of generality, we assume that switching from $Z_i = 0$ to $Z_i = 1$ either increases D_i (i.e., $D_i(1) > D_i(0)$) or has no effect. In other words, being draft-eligible weakly encourages everyone to serve in the military: no individuals would serve in the military if they were draft-ineligible, but refuse to serve if they were draft-eligible – a mild assumption in this case. To help interpret this assumption, Angrist et al. (1996) define four types of individuals, indexed by their potential treatments. First, there are always-takers, who volunteer to serve regardless of their eligibility status: $D_i(1) = D_i(0) = 1$. Second, there are never-takers, who avoid the draft: $D_i(1) = D_i(0) = 0$. Third, there are compliers, who serve only if they are draft-eligible: $D_i(1) = 1$, $D_i(0) = 0$. Finally, there could be defiers, who only serve if they are draft-*ineligible* – the monotonicity assumption can be seen to rule out the presence of such unusual behavior.

The final condition is a relevance assumption, which says that the fraction of compliers in the population is positive. In other words, there are people whose treatment status can be manipulated by changing the instrument. It is not difficult to come up with an instrument that satisfies the first three assumptions – flipping a coin for each individual would do – but finding an instrument that jointly satisfies all four assumption requires some ingenuity. Statistically, the relevance assumption ensures that the correlation between the treatment and the instrument is non-zero: $Cov(Z_i, D_i) \neq 0$. This ensures that we do not divide by zero in the definition of the IV estimand, $\beta^{IV} = Cov(Z_i, Y_i)/Cov(Z_i, D_i)$.[25]

Under these assumptions, Imbens and Angrist (1994) showed that the IV estimand β^{IV} identifies a LATE,

$$\begin{aligned}\beta^{IV} &= \frac{E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]}{E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0]} \\ &= E[Y_i(1) - Y_i(0) \mid D_i(1) > D_i(0)],\end{aligned}$$

[25] Here, β^{IV} is the population analog of the Wald (1940) estimator for a bivariate regression with mismeasured regressors (see Angrist and Pischke, 2009 for a discussion).

where the first equality follows from the definition of β_{IV} and the fact that Z_i is binary.[26] The LATE $E[Y_i(1) - Y_i(0) \mid D_i(1) > D_i(0)]$ is a "local" treatment effect, because it corresponds to the ATE for compliers only. In the Angrist (1990) context, it is the average effect of military service on adult earnings among individuals whose draft lottery numbers compelled them to serve.

The Imbens and Angrist analysis delivers three key insights. First, it clarifies precisely how the source of variation in the treatment induced by the instrument affects the interpretation of the estimand β^{IV}. Second, it helps separate statistical assumptions (random assignment) from substantive economic restrictions (exclusion and monotonicity). Finally, it emphasizes how causal interpretation of IV requires that the treatment and the instrument need to be "manipulable". We discuss each insight in turn.

4.1.1. Internal and external validity. The LATE result showed precisely how the quasi-experimental variation in an instrument affects the IV estimand when no structural restrictions are placed on the treatment effects. The estimand identifies an average effect for the compliers. If the IV leverages a "narrow" source of variation – as in Angrist (1990) and Angrist and Krueger (1991) – then the group of compliers might be a small subset of the overall population. In Angrist and Krueger (1991), the compliers are those who are on the margin of dropping out of school, but are induced to stay on for an additional year because of their exact quarter of birth. These individuals comprise at most 10 percent of the overall population. Intuitively, we can never learn from data alone about the treatment effect for never-takers (or always-takers), as we never see them treated (or untreated) in the data. While the identity of compliers is never directly given by data (because we never observe both $D_i(1)$ and $D_i(0)$ for the same individual i), Abadie (2003) showed how a wide range of functions of their predetermined characteristics and potential outcomes could be estimated.

Relative to an experimental ideal, where we could learn the ATE for the overall population, the LATE result might appear underwhelming. But the group of compliers is often of policy interest. For example, the Angrist and Krueger (1991) compliers might help to inform policies that affect

[26]For the second equality, note that by random assignment $E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0] = E[D_i(1) - D_i(0)]$ and $E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0] = E[Y(D_i(1)) - Y_i(D_i(0))]$, following similar steps as in the above ATE identification proof. By monotonicity, $E[D_i(1) - D_i(0)] = \Pr(D_i(1) > D_i(0))$ and $E[Y(D_i(1)) - Y_i(D_i(0))] = E[(Y_i(1) - Y_i(0))(D_i(1) - D_i(0))] = E[Y_i(1) - Y_i(0) \mid D_i(1) > D_i(0)] \times \Pr(D_i(1) > D_i(0))$, completing the proof.

the minimum school-leaving age. It is true that the LATE is less relevant for predicting effects of other policies, such as the effect of abolishing college tuition. After all, people attending college – or those considering attendance – are not those who are affected by minimum schooling laws that the variation in quarter of birth leverages. Here the value of the LATE result lies in knowing that for considering the effects of such a policy we need to combine the Angrist and Krueger (1991) estimates with an economic model that would allow us to extrapolate the treatment effect estimates to this population, or else look for a more informative natural experiment.

More generally, the LATE result sharpens the distinction between internal validity of the study (when can we interpret the IV estimate as the ATE for compliers?) and its external validity (what are the lessons that carry over to other settings?).[27] A concern raised in Heckman and Urzúa (2010) and Deaton (2010) is that the increased use of natural experiments puts too much emphasis on internal over external validity and that too many studies stop at reporting the IV estimate, which might not answer a question of economic interest. As argued in Imbens (2010), the value of the LATE framework lies in separating the assumptions needed to identify the treatment effect for compliers in the current population from any additional assumptions needed to generalize the internally valid estimate to other populations. This allows for more transparency when researchers complement the quasi-experimental variation in the data with a structural model. For example, it allowed Card and Hyslop (2005) to combine experimental variation in an earnings subsidy with a structural model to identify the impact of the subsidy on welfare entry and exit rates. The Imbens et al. (2001) study, also discussed in Section 3.4, likewise combines experimental variation with a life-cycle model of labor supply. Recent work by Brinch et al. (2017), Mogstad et al. (2019), and Kline and Walters (2019) clarifies connections between the LATE framework and model-based identifying restrictions – and how the former can be used to relax the latter.

4.1.2. Statistical versus substantive restrictions. The second insight of the LATE result lies in separating the substantive restrictions in an IV analysis that always need to be justified by economic reasoning – the exclusion restriction and the monotonicity assumption – from the random assignment assumption, which can hold automatically if the variation in the instrument is as good as random. This clarifies what exactly randomization

[27]Campbell (1957) gives an early formalization of the difference between internal and external validity in the social sciences.

delivers: it allows us to identify the intent-to-treat effect. But additional assumptions are needed to go from this effect to the treatment effect for compliers.

The distinction between the exclusion restriction and the random assignment assumption allows for a more nuanced analysis of potential violations of instrument "exogeneity". Statistical balance checks can be used to verify that the randomization "worked". In contrast, while statistical tests of the exclusion restriction and monotonicity exist (see, e.g., Kitagawa, 2015), most credible applications of IV rely on institutional or theoretical arguments to justify them. Such arguments are often used to develop indirect application-specific diagnostic checks for these assumptions. This targeted probing of the design validity would not be possible if Imbens and Angrist did not make the assumptions clear in the first place.

Making the key assumptions clear also allows for more targeted sensitivity analysis. For example, Angrist et al. (1996) show that, under violations of the monotonicity condition, β^{IV} averages the treatment effect for compliers with the treatment effect for defiers – but the weight on the defiers is negative. On the one hand, this implies that β^{IV} could be negative even if treatment effects are positive for all individuals. On the other hand, the result also implies that the presence of defiers is of lesser concern if their proportion is small, because the weight placed on them is proportional to the size of the defier group.[28]

Another way of interpreting the Imbens and Angrist (1994) assumptions is to think of them as an exploration of model misspecification. Early formalizations of IV methods focused on linear structural models for the outcome and supplemented it with an "exogeneity" assumption that the residual in this equation is uncorrelated with the instrument. The LATE result shows what happens when we drop the parametric restrictions.

4.1.3. The role of manipulation. The potential outcomes framework generally highlights the need for manipulation in causal analyses. To interpret the potential outcomes $Y_i(d, z)$ and potential treatments $D_i(z)$, one needs to be able to manipulate, at least in principle and at least for some subpopulation, the treatment and the instrument. If the treatment is an innate attribute of a unit that cannot be manipulated, this notation makes it clear that we cannot speak of causal effects; in line with the dictum of Rubin (1975) – echoed in Holland (1986) – "no causation without

[28]Similarly, certain violations of the exclusion restrictions might have little effect on the interpretation of the results, as explored, for example, in Kolesár et al. (2015). See also Imbens and Rubin (1997) for a discussion of sensitivity analyses when the exclusion restriction and monotonicity assumptions are violated in a Bayesian framework.

manipulation".[29] For instance, as discussed by Greiner and Rubin (2011), the notation makes it clear that while it is difficult to talk about the causal effect of gender or race, we can talk about causal effects of being perceived as having a certain gender or race. Such perception effects have been studied in evaluating the effects of blind auditions (Goldin and Rouse, 2000), or in countless "audit studies" that manipulate otherwise identical résumés – by, say, changing the name on the résumé (see, e.g., Bertrand and Mullainathan, 2004 for an early example).

4.2. Extensions and connections

While our exposition focuses on the simplest setting with a binary treatment and a binary instrument, the framework extends readily to cases with multi-valued or multi-dimensional instruments (such as indicators for quarter of birth). Angrist and Imbens (1995) consider settings with multi-valued treatment (such as years of education), demonstrating that in this case IV recovers a generalization of LATE known as the average causal response (ACR). Abadie et al. (2002) generalize the set-up to cover estimation of quantile treatment effects.

In another important contribution, Angrist et al. (2000) adapted the LATE framework to cover estimation of demand or supply elasticities in a simultaneous equation system of supply and demand. This is a classic problem that originally motivated the use of instruments by Philip Wright and his son Sewall Wright, Tinbergen, Haavelmo, and other early pioneers of IV methods. This agenda was extended substantially by the members of Cowles Commission, who showed how exclusion and covariance restrictions can be used to identify two-equation supply and demand models as well as more complicated simultaneous equations systems (Christ, 1994). To explain the identification challenge, suppose both the log of the demand curve $Q_i^d(P)$ and the log of the supply curve $Q_i^s(P)$ in market i are linear in log of the price P:

$$\ln Q_i^d(P) = \alpha^d + \beta^d \ln P + \varepsilon_i^d, \tag{2}$$

$$\ln Q_i^s(P) = \alpha^s + \beta^s \ln P + \varepsilon_i^s. \tag{3}$$

Here, $\beta^d < 0$ and $\beta^s > 0$ are demand and supply elasticities, respectively (we assume that these are constant across markets i), and $(\varepsilon_i^d, \varepsilon_i^s)$ are unobserved demand and supply shocks. Because the equilibrium price equates supply and demand, both the observed equilibrium price P_i and

[29] As the recent literature on shift–share and related instruments shows, different views on which components of a treatment or instrument are manipulable can lead to vastly different identifying assumptions, estimation concerns, and inferential procedures (Goldsmith-Pinkham et al., 2020; Borusyak et al., 2022; Adão et al., 2019; Borusyak and Hull, 2022).

the equilibrium quantity Q_i depend on the supply and demand shocks.[30] As a result, a simple regression of observed log quantity on observed log price will recover neither the demand nor the supply elasticity, but a hard-to-interpret mixture of the two. The IV solution to this simultaneity challenge, as first considered by the Wrights, Tinbergen, and others, is to measure some component of the supply shock that does not affect demand. As a result of this "exclusion restriction", one can show that using such a supply shock component as an instrument for log price in a regression of log quantity on log price regression recovers the demand elasticity.[31] If we instead use a component of the demand shock that does not affect supply, we recover the supply elasticity.

But what if we relax the assumption that elasticities are constant across markets, and that the unobserved shocks are additive? Angrist et al. (2000) consider a non-parametric set-up that does not impose any functional form restrictions on the supply and demand curves, $Q_i^d(P)$ and $Q_i^s(P)$. In this unrestricted model, they show that an IV regression using an instrument that shifts the supply curve while not affecting demand identifies a weighted average of market-specific demand elasticities. If the demand elasticity varies with price, the estimand also averages over different prices in the same market. Like the LATE result in the context of estimating causal effects of a binary treatment, this result clarifies the role of internal and external validity of the IV estimates, and the role of functional form restrictions imposed in the classic linear model (2) and (3).

The LATE framework has also been central to understanding RD designs: quasi-experimental settings where treatment eligibility is determined by whether a particular variable (called a running variable) crosses a threshold. For example, to estimate the effects of class size on student test scores, Angrist and Lavy (1999) exploit the fact that class sizes in Israel follow the rule of Maimonides (a twelfth-century rabbinic scholar): a school should not have class sizes bigger than 40. Here the running variable is the class size, and 40 represents the threshold. If a student cohort in a particular school comprises fewer than 40 students, they will all be in one large classroom. But if there are 41 students, they become

[30]Specifically, setting supply equal to demand and solving for price yields

$$\ln Q_i = \frac{\beta^d \alpha^s - \beta^s \alpha^d}{\beta^d - \beta^s} + \frac{\beta^s}{\beta^s - \beta^d}\varepsilon_i^d - \frac{\beta^d}{\beta^s - \beta^d}\varepsilon_i^s$$

and

$$\ln P_i = \frac{\alpha^s - \alpha^d}{\beta^d - \beta^s} + \frac{1}{\beta^s - \beta^d}\varepsilon_i^d - \frac{1}{\beta^s - \beta^d}\varepsilon_i^s.$$

[31]The use of the "exclusion" term in this context can be traced back at least as far as Koopmans (1949).

eligible for a small classroom treatment: the school is allowed to open two classes with an average size of 20.5. If schools follow the Maimonides' rule exactly, we can estimate the effect of the small classroom treatment on test scores by comparing schools with enrolment just below and just above 40 students. More precisely, such a sharp RD design estimates the average causal effect for schools with enrolment at the threshold – those who are at the margin of becoming eligible for the small classroom treatment.

But what if compliance with the Maimonides' rule is imperfect? That is, what if some schools opt for small classrooms even if their cohort size falls below the threshold, and others do not open two classrooms even if their cohort size falls above it? In such "fuzzy" RD designs, the treatment probability would still increase as we cross the threshold, but it does not jump all the way from zero to one as in a "sharp" RD. Hahn et al. (2001) adapted the LATE framework to show that the class size running variable is used as an instrument for the small classroom treatment (again restricting the analysis to schools with enrolment close to the eligibility threshold), we estimate a LATE: the average effect among schools at the threshold of eligibility who comply with the treatment assignment rule. The conditions for this result very much mirror the LATE assumptions in Section 4.1.

This influential result shows that RD designs and IV designs are close cousins, and helped bring about an explosion of RD studies in recent years. Methodological work by the laureates also played an important role in boosting the popularity of RD. Their contributions range from developing a procedure for selecting the estimation window, formalizing what "close to the eligibility threshold" means in practice (Imbens and Kalyanaraman, 2012) to developing a framework for extrapolating the treatment effects to those away from the cut-off (Angrist and Rokkanen, 2015), and to adapting the LATE framework to a closely related regression kink design: where a continuous treatment variable is a piecewise linear function of a running variable (Card et al., 2015), again with possibly imperfect compliance.

The basic approach of Imbens and Angrist – use the potential outcome framework, keep treatment effect heterogeneity unrestricted, and separate the role of any random variation provided by the natural experiment from other substantive restrictions that are needed – has been fruitfully applied by other researchers in many other contexts besides RD. Angrist (1998) used the framework to interpret certain fixed-effect regressions. The modern literature on differences-in-differences methods is still exploring the subtle conceptual issues that this approach highlights (see, e.g., de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2021; Goldsmith-Pinkham et al., 2022, among many others).

Finally, in a series of influential papers, Heckman and Vytlacil (2005, 2007a,b) develop a broader framework of marginal treatment effect (MTE):

the effects for individuals at particular values of an unobserved preference for participation in treatment. Building on work by Björklund and Moffitt (1987), Heckman and Vytlacil (2005) show how the LATE result fits within this framework. Vytlacil (2002) shows that the MTE framework is formally equivalent to the LATE model of Imbens and Angrist (1994), with a binary or multivalued instrument. In particular, the key monotonicity assumption is equivalent to additive separability between instruments and unobservables in a latent index model of treatment choice. In the latent index set-up, LATE can be seen as an average of MTE over a specific range of unobserved preferences, highlighting that effects for compliers for a particular instrument might differ from effects for individuals affected by alternative hypothetical policy changes.

5. Conclusion

The 2021 Nobel laureates helped shape modern applied research in labor economics and beyond. A focus on clear research designs exploiting natural experiments led David Card to several empirical conclusions – on immigration effects, firm monopsony power, and educational quality – which challenged conventional wisdom and fueled large bodies of follow-up literatures. Joshua Angrist and Guido Imbens showed that IV estimators retain internal validity even when restrictive models for the outcome are relaxed. Their LATE theorem, rooted in a flexible potential outcomes framework, is underpinned by clear and interpretable assumptions, and has similarly led to an explosion of subsequent applied econometric research.

More broadly, this methodological and empirical focus on a clear research design, and on understanding what core assumptions underlie its internal validity, helps portability. Follow-up studies can be conducted in other contexts, probing external validity and replicability. The idea that the research design needs to be tied to the institutional features or other forces driving treatment assignment helps to limit specification searches. The laureates' work also showed how, for more complicated questions, simple research designs can be complemented by careful modeling.

Our brief review focuses on these core contributions of the laureates and necessarily omits many other important contributions. As examples, we have not discussed Card's empirical studies of firm wage-setting (e.g. Card et al., 2013, 2016, 2018), Angrist's methodological work on leveraging the randomness in centralized assignment mechanisms (e.g. Abdulkadiroğlu et al., 2017, 2022; Angrist et al., 2021), or Imbens' econometric contributions to the estimation of treatment effects under conditional random assignment (e.g. Imbens, 2000; Hirano et al., 2003; Abadie and Imbens, 2006) or to generalizing the difference-in-differences

framework (e.g. Athey and Imbens, 2006). The laureates also made many contributions that address several technical implementation issues that come up in design-based studies, such as how to deal with many weak instruments (Angrist and Krueger, 1995; Angrist et al., 1999). Not least among these other contributions is the laureates' generosity and dedication when helping their colleagues or advising their students, a trait that we have had the privilege to benefit from first-hand.

References

Abadie, A. (2003), Semiparametric instrumental variable estimation of treatment response models, *Journal of Econometrics 113*, 231–263.

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

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

Abadie, A., Angrist, J. D., and Imbens, G. W. (2002), Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings, *Econometrica 70*, 91–117.

Abadie, A., Diamond, A., and 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*, 493–505.

Abdulkadiroğlu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., and Pathak, P. A. (2011), Accountability and flexibility in public schools: evidence from Boston's charters and pilots, *Quarterly Journal of Economics 126*, 699–748.

Abdulkadiroğlu, A., Angrist, J. D., Hull, P. D., and Pathak, P. A. (2016), Charters without lotteries: testing takeovers in New Orleans and Boston, *American Economic Review 106* (7), 1878–1920.

Abdulkadiroğlu, A., Angrist, J. D., Narita, Y., and Pathak, P. A. (2017), Research design meets market design: using centralized assignment for impact evaluation, *Econometrica 85*, 1373–1432.

Abdulkadiroğlu, A., Angrist, J. D., Narita, Y., and Pathak, P. A. (2022), Breaking ties: regression discontinuity design meets market design, *Econometrica 90*, 117–151.

Abowd, J. M. and Card, D. (1989), On the covariance structure of earnings and hours changes, *Econometica 57*, 411–445.

Abowd, J. M., Kramarz, F., and Margolis, D. N. (1999), High wage workers and high wage firms, *Econometrica 67*, 251–333.

Adão, R., Kolesár, M., and Morales, E. (2019), Shift–share designs: theory and inference, *Quarterly Journal of Economics 134*, 1949–2010.

Altonji, J. G. and Card, D. (1991), The effects of immigration on the labor market outcomes of less-skilled natives, in J. M. Abowd and R. B. Freeman (eds), *Immigration, Trade, and the Labor Market*, University of Chicago Press, Chicago, IL, 201–234.

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

Angrist, J. D. (1998), Estimating the labor market impact of voluntary military service using social security data on military applicants, *Econometrica 66*, 249–288.

Angrist, J. D. and Imbens, G. W. (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.

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

Angrist, J. D. and Krueger, A. B. (1995), Split-sample instrumental variables estimates of the return to schooling, *Journal of Business & Economic Statistics 13*, 225–235.
Angrist, J. D. and Lavy, V. (1999), Using Maimonides' rule to estimate the effect of class size on scholastic achievement, *Quarterly Journal of Economics 114*, 533–575.
Angrist, J. D. and Pischke, J.-S. (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.
Angrist, J. D. and Rokkanen, M. (2015), Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff, *Journal of the American Statistical Association 110*, 1331–1344.
Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), Identification of causal effects using instrumental variables, *Journal of the American Statistical Association 91*, 444–455.
Angrist, J. D., Imbens, G. W., and Krueger, A. B. (1999), Jackknife instrumental variables estimation, *Journal of Applied Econometrics 14*, 57–67.
Angrist, J. D., Graddy, K., and Imbens, G. W. (2000), The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish, *Review of Economic Studies 67*, 499–527.
Angrist, J. D., Bettinger, E., Bloom, E., King, E., and Kremer, M. (2002) Vouchers for private schooling in Columbia: evidence from a randomized natural experiment, *American Economic Review 92* (5), 1535–1558.
Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., and Walters, C. R. (2010), Inputs and impacts in charter schools: KIPP Lynn, *American Economic Review 100* (2), 239–243.
Angrist, J., Hull, P., Pathak, P. A., and Walters, C. R. (2017), Interpreting tests of school VAM validity, *Quarterly Journal of Economics 132*, 871–919.
Angrist, J. D., Hull, P., Pathak, P. A., and Walters, C. R. (2021), Credible school value-added with undersubscribed school lotteries, *Review of Economics and Statistics*, forthcoming (https://doi.org/10.1162/rest_a_01149).
Ashenfelter, O. C. (1975), The effect of manpower training on earnings: preliminary results, in J. L. Stern and B. D. Dennis (eds), *Proceedings of the Twenty-Seventh Annual Winter Meeting*, Industrial Relations Research Association, Madison, WI, 252–260.
Ashenfelter, O. C. (1978), Estimating the effect of training programs on earnings, *Review of Economics and Statistics 60*, 47–57.
Ashenfelter, O. C. (1987), The case for evaluating training programs with randomized trials, *Economics of Education Review 6*, 333–338.
Ashenfelter, O. C. and Card, D. (1985), Using the longitudinal structure of earnings to estimate the effect of training programs, *Review of Economics and Statistics 67*, 648–660.
Ashenfelter, O. C. and Krueger, A. B. (1994), Estimates of the economic return to schooling from a new sample of twins, *American Economic Review 84* (5), 1157–1173.
Athey, S. and Imbens, G. W. (2006), Identification and inference in nonlinear difference-in-differences models, *Econometrica 74*, 431–497.
Azar, J., Marinescu, I., and Steinbaum, M. (2022), Labor market concentration, *Journal of Human Resources 57*, S167–S199.
Baicker, K., Taubman, S. L., Allen, H. L., Bernstein, M., Gruber, J. H., Newhouse, J. P., Schneider, E. C., Wright, B. J., Zaslavsky, A. M., Finkelstein, A. N., and the Oregon Health Study Group (2013), The Oregon health insurance experiment: effects of Medicaid on clinical outcomes, *New England Journal of Medicine 368*, 1713–1722.
Balke, A. and Pearl, J. (1997), Bounds on treatment effects from studies with incomplete compliance, *Journal of the American Statistical Association 92*, 1171–1176.
Banerjee, A., Duflo, E., Glennerster, R., and Kinnan, C. (2015), The miracle of microfinance? Evidence from a randomized evaluation, *American Economic Journal: Applied Economics 7*, 22–53.

Bartik, T. J. (1991), *Who Benefits from State and Local Economic Development Policies?* W. E. Upjohn Institute for Employment Research, Kalamazoo, MI.
Becker, G. (1964), *Human Capital*, Columbia University Press, New York, NY.
Berger, D., Herkenhoff, K., and Mongey, S. (2022), Labor market power, *American Economic Review 112*, 1147–1193.
Bertrand, M. and Mullainathan, S. (2004), Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination, *American Economic Review 94* (4), 991–1013.
Björklund, A. and Moffitt, R. (1987), The estimation of wage gains and welfare gains in self-selection models, *Review of Economics and Statistics 69*, 42–49.
Black, F. (1982), The trouble with econometric models, *Financial Analysts Journal 38*, 29–37.
Blanchard, O. J. and Katz, L. F. (1992), Regional evolutions, *Brookings Papers on Economic Activity 1992*, 1–75.
Blundell, R. (2001), James Heckman's contributions to economics and econometrics, *Scandinavian Journal of Economics 103*, 191–204.
Borjas, G. J. (1987), Self-selection and the earnings of immigrants, *American Economic Review 77* (4), 531–553.
Borjas, G. J. (2017), The wage impact of the marielitos: a reappraisal, *Industrial and Labor Relations Review 70*, 1077–1110.
Borusyak, K. and Hull, P. (2022), Non-random exposure to exogenous shocks, National Bureau of Economic Research (NBER) Working Paper 27845.
Borusyak, K., Hull, P., and Jaravel, X. (2022), Quasi-experimental shift–share research designs, *Review of Economic Studies 89*, 181–213.
Brinch, C. N., Mogstad, M., and Wiswall, M. (2017), Beyond LATE with a discrete instrument, *Journal of Political Economy 125*, 985–1039.
Brown, C., Gilroy, C., and Kohen, A. (1982), The effect of the minimum wage on employment and unemployment, *Journal of Economic Literature 20*, 487–528.
Buchanan, J. (1996), Commentary on the minimum wage, *Wall Street Journal*, 25 April.
Buckles, K. S. and Hungerman, D. M. (2013), Season of birth and later outcomes: old questions, new answers, *Review of Economics and Statistics 95*, 711–724.
Campbell, D. T. (1957), Factors relevant to the validity of experiments in social settings, *Psychological Bulletin 54*, 297–312.
Campbell, D. T. (1969), Reforms as experiments, *American Psychologist 24*, 409–429.
Card, D. (1990), The impact of the Mariel boatlift on the Miami labor market, *Industrial and Labor Relations Review 43*, 245–257.
Card, D. (1992a), Do minimum wages reduce employment? A case study of California, 1987–89, *Industrial and Labor Relations Review 46*, 38–54.
Card, D. (1992b), Using regional variation in wages to measure the effects of the federal minimum wage, *Industrial Labor Relations Review 46*, 22–37.
Card, D. (1995), Using geographic variation in college proximity to estimate the return to schooling, in L. N. Christofides, E. K. Grant, and R. Swidinsky (eds), *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, University of Toronto Press, Toronto, 201–222.
Card, D. (1999), Chapter 30: The causal effect of education on earnings, in O. C. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, Vol. 3A, Elsevier, Amsterdam, 1801–1863.
Card, D. (2001), Immigrant inflows, native outflows, and the local market impacts of higher immigration, *Journal of Labor Economics 19*, 22–64.
Card, D. and Hyslop, D. R. (2005), Estimating the effects of a time-limited earnings subsidy for welfare-leavers, *Econometrica 73*, 1723–1770.

Card, D. and Krueger, A. B. (1992a), Does school quality matter? Returns to education and the characteristics of public schools in the United States, *Journal of Political Economy 100*, 1–40.
Card, D. and Krueger, A. B. (1992b), School quality and black-white relative earnings: a direct assessment, *Quarterly Journal of Economics 107*, 151–200.
Card, D. and Krueger, A. B. (1994), Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania, *American Economic Review 84* (5), 772–793.
Card, D. and Krueger, A. B. (1995a), *Myth and Measurement: The New Economics of the Minimum Wage*, Princeton University Press, Princeton, NJ.
Card, D. and Krueger, A. B. (1995b), Time-series minimum-wage studies: a meta-analysis, *American Economic Review 85* (2), 238–243.
Card, D., Heining, J., and Kline, P. (2013), Workplace heterogeneity and the rise of West German wage inequality, *Quarterly Journal of Economics 128*, 967–1015.
Card, D., Lee, D. S., Pei, Z., and Weber, A. (2015), Inference on causal effects in a generalized regression kink design, *Econometrica 83*, 2453–2483.
Card, D., Cardoso, A. R., and Kline, P. (2016), Bargaining, sorting, and the gender wage gap: quantifying the impact of firms on the relative pay of women, *Quarterly Journal of Economics 131*, 633–686.
Card, D., Cardoso, A. R., Heining, J., and Kline, P. (2018), Firms and labor market inequality: evidence and some theory, *Journal of Labor Economics 36*, S13–S69.
Castillo-Freeman, A. J. and Freeman, R. B. (1992), When the minimum wage really bites: the effect of the US-level minimum on Puerto Rico, in G. J. Borjas and R. B. Freeman (eds), *Immigration and the Work Force: Economic Consequences for the United States and Source Areas*, University of Chicago Press, Chicago, IL, 177–211.
Cengiz, D., Dube, A., Lindner, A., and Zipperer, B. (2019), The effect of minimum wages on low-wage jobs, *Quarterly Journal of Economics 134*, 1405–1454.
Cesarini, D., Lindqvist, E., Notowidigdo, M. J., and Östling, R. (2017), The effect of wealth on individual and household labor supply: evidence from Swedish lotteries, *American Economic Review 107* (12), 3917–46.
Chamberlain, G. (1986), Asymptotic efficiency in semi-parametric models with censoring, *Journal of Econometrics 32*, 189–2018.
Chetty, R. and Hendren, N. (2018), The impacts of neighborhoods on intergenerational mobility I: childhood exposure effects, *Quarterly Journal of Economics 133*, 1107–1162.
Chetty, R., Hendren, N., and Katz, L. F. (2016), The effects of exposure to better neighborhoods on children: new evidence from the moving to opportunity experiment, *American Economic Review 106* (4), 855–902.
Christ, C. F. (1994), The Cowles Commission's contributions to econometrics at Chicago, 1939–1955, *Journal of Economic Literature 32*, 30–59.
Clark, D. and Royer, H. (2013), The effect of education on adult mortality and health: evidence from Britain, *American Economic Review 103* (6), 2087–2120.
Coleman, J. S. (1966), *Equality of Educational Opportunity*, Government Printing Office, Washington, DC.
Currie, J., Kleven, H., and Zwiers, E. (2020), Technology and big data are changing economics: mining text to track methods, *AEA Papers and Proceedings 110*, 42–48.
de Chaisemartin, C. and D'Haultfœuille, X. (2020), Two-way fixed effects estimators with heterogeneous treatment effects, *American Economic Review 110*, 2964–2996.
Deaton, A. (2010), Instruments, randomization, and learning about development, *Journal of Economic Literature 48*, 424–455.
DellaVigna, S., List, J. A., and Malmendier, U. (2012), Testing for altruism and social pressure in charitable giving, *Quarterly Journal of Economics 127*, 1–56.

DiNardo, J. (2010), Natural experiments and quasi-natural experiments, in S. N. Durlauf and L. E. Blume (eds), *Microeconometrics*, *The New Palgrave Economics Collection*, Palgrave Macmillan, London, 139–153.

Dube, A., Lester, T. W., and Reich, M. (2010), Minimum wage effects across state borders: estimates using contiguous counties, *Review of Economics and Statistics 92*, 945–964.

Duflo, E., Kremer, M., and Robinson, J. (2008), How high are rates of return to fertilizer? Evidence from field experiments in Kenya, *American Economic Review 98* (2), 482–488.

Duflo, E., Kremer, M., and Robinson, J. (2011), Nudging farmers to use fertilizer: theory and experimental evidence from Kenya, *American Economic Review 101* (6), 2350–2390.

Dustmann, C. and Glitz, A. (2016), How do industries and firms respond to changes in local labor supply? *Journal of Labor Economics 33*, 711–750.

Dustmann, C., Schönberg, U., and Stuhler, J. (2016), The impact of immigration: why do studies reach such different results? *Journal of Economic Perspectives 30* (4), 31–56.

Dustmann, C., Lindner, A., Schönberg, U., Umkehrer, M., and vom Berge, P. (2022), Reallocation effects of the minimum wage, *Quarterly Journal of Economics 137*, 267–328.

Finkelstein, A. N., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., and the Oregon Health Study Group (2012), The Oregon health insurance experiment: evidence from the first year, *Quarterly Journal of Economics 127*, 1057–1106.

Finkelstein, A. N., Gentzkow, M., and Williams, H. (2016), Sources of geographic variation in health care: evidence from patient migration, *Quarterly Journal of Economics 131*, 1681–1726.

Freeman, R. B. (1975), Supply and salary adjustments to the changing science manpower market: physics, 1948–1973, *American Economic Review 65* (1), 27–39.

Freeman, R. B. (1980), An empirical analysis of the fixed coefficient "manpower requirements" model, 1960–1970, *Journal of Human Resources 15*, 176–199.

Freeman, R. B. (1989), *Labor Markets in Action: Essays in Empirical Economics*, Woodhead Faulkner, Sawston, Cambridge.

Gibbons, R. and Katz, L. F. (1992), Does unmeasured ability explain inter-industry wage differentials? *Review of Economic Studies 59*, 515–535.

Giuliano, L. (2013), Minimum wage effects on employment, substitution, and the teenage labor supply: evidence from personnel data, *Journal of Labor Economics 31*, 155–194.

Glewwe, P., Ilias, N., and Kremer, M. (2010), Teacher incentives, *American Economic Journal: Applied Economics 2*, 205–227.

Goldin, C. and Rouse, C. (2000), Orchestrating impartiality: the impact of "blind" auditions on female musicians, *American Economic Review 90* (4), 715–741.

Goldsmith-Pinkham, P., Sorkin, I., and Swift, H. (2020), Bartik instruments: what, when, why, and how, *American Economic Review 110*, 2586–2624.

Goldsmith-Pinkham, P., Hull, P., and Kolesár, M. (2022), Contamination bias in linear regressions, National Bureau of Economic Research (NBER) Working Paper No. 30108.

Goldstein, J. H. (1972), The effectiveness of manpower training programs: a review of research on the impact on the poor, Technical Report 3, Joint Economic Committee, Washington, DC.

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

Grossman, J. B. (1982), The substitutability of natives and immigrants in production, *Review of Economics and Statistics 64*, 596–603.

Gruber, J. (1994), The incidence of mandated maternity benefits, *American Economic Review 84* (3), 622–641.

Haavelmo, T. (1943), The statistical implications of a system of simultaneous equations, *Econometrica 11*, 1–12.

Hahn, J., Todd, P. E., and van der Klaauw, W. (2001), Identification and estimation of treatment effects with a regression-discontinuity design, *Econometrica 69*, 201–209.
Hanushek, E. A. (1986), The economics of schooling: production and efficiency in public schools, *Journal of Economic Literature 49*, 1141–1177.
Harasztosi, P. and Lindner, A. (2019), Who pays for the minimum wage? *American Economic Review 109*, 2693–2727.
Hearst, N., Newman, T. B., and Hulley, S. B. (1986), Delayed effects of the military draft on mortality, *New England Journal of Medicine 314*, 620–624.
Heckman, J. J. (1974), Shadow prices, market wages, and labor supply, *Econometrica 42*, 679–694.
Heckman, J. J. (1976), The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, *Annals of Economic and Social Measurement 4*, 475–492.
Heckman, J. J. (1979), Sample selection bias as a specification error, *Econometrica 47*, 153–161.
Heckman, J. J. (1990), Varieties of selection bias, *American Economic Review 80* (2), 313–318.
Heckman, J. J. and Pinto, R. (2014), Causal analysis after Haavelmo, *Econometric Theory 31*, 115–151.
Heckman, J. J. and Robb, J., Richard (1985), Alternative methods for evaluating the impact of interventions, in J. J. Heckman and B. H. Singer (eds), *Longitudinal Analysis of Labor Market Data*, Nos 1–2 in Econometric Society Monographs, Cambridge University Press, Cambridge, 145–245.
Heckman, J. J. and Urzúa, S. (2010), Comparing IV with structural models: what simple IV can and cannot identify, *Journal of Econometrics 156*, 27–37.
Heckman, J. J. and Vytlacil, E. J. (2005), Structural equations, treatment effects, and econometric policy evaluation, *Econometrica 73*, 669–738.
Heckman, J. J. and Vytlacil, E. J. (2007a), Chapter 70: Econometric evaluation of social programs, Part I: causal models, structural models, and econometric policy evaluation, in J. J. Heckman and E. E. Leamer (eds), *Handbook of Econometrics*, Vol. 6B, Elsevier, Amsterdam, 4779–4874.
Heckman, J. J. and Vytlacil, E. J. (2007b), Chapter 71: Econometric evaluation of social programs, Part II: using the marginal treatment effect to organize alternative economic estimators to evaluate social programs, and to forecast their effects in new environments, in J. J. Heckman and E. E. Leamer (eds), *Handbook of Econometrics*, Vol. 6B, Elsevier, Amsterdam, 4875–5143.
Hendry, D. F. (1980), Econometrics—alchemy or science? *Economica 47*, 387–406.
Hendry, D. F. and Morgan, M. S. (1992), *The Foundations of Econometric Analysis*, Cambridge University Press, Cambridge.
Hirano, K., Imbens, G. W., and Ridder, G. (2003), Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica 71*, 1161–1189.
Holland, P. W. (1986), Statistics and causal inference, *Journal of the American Statistical Association 81*, 945–960.
Imbens, G. W. (1997), Book review of "The Foundations of Econometric Analysis" by David F. Hendry and Mary S. Morgan, *Journal of Applied Econometrics 12*, 91–94.
Imbens, G. W. (2000), The role of the propensity score in estimating dose-response functions, *Biometrika 87*, 706–710.
Imbens, G. W. (2010), Better LATE than nothing: some comments on Deaton (2009), and Heckman and Urzúa (2009), *Journal of Economic Literature 48*, 399–423.
Imbens, G. W. (2014), Instrumental variables: an econometrician's perspective, *Statistical Science 29*, 323–358.
Imbens, G. W. (2020), Potential outcome and direct acyclic graph approaches to causality: relevance for empirical practice in economics, *Journal of Economic Literature 58*, 1129–1179.
Imbens, G. W. (2021), Prize lecture, https://www.nobelprize.org/prizes/economic-sciences/2021/imbens/lecture/.

Imbens, G. W. and Angrist, J. D. (1994), Identification and estimation of local average treatment effects, *Econometrica 62*, 467–475.
Imbens, G. W. and Kalyanaraman, K. (2012), Optimal bandwidth choice for the regression discontinuity estimator, *Review of Economic Studies 79*, 933–959.
Imbens, G. W. and Rubin, D. B. (1997), Bayesian inference for causal effects in randomized experiments with noncompliance, *Annals of Statistics 25*, 305–327.
Imbens, G. W., Rubin, D. B., and Sacerdote, B. I. (2001), Estimating the effect of unearned income on labor earnings, savings, and consumption: evidence from a survey of lottery players, *American Economic Review 91* (4), 778–794.
Jackson, C. K. (2020), Does school spending matter? The new literature on an old question, in L. Tach, R. Dunifon, and D. L. Miller (eds), *Confronting Inequality: How Policies and Practices Shape Children's Opportunities*, American Psychological Association, Washington, DC, 165–186.
Katz, L. F. and Krueger, A. B. (1992), The effect of the minimum wage on the fast-food industry, *Industrial and Labor Relations Review 46*, 6–21.
Kitagawa, T. (2015), A test for instrument validity, *Econometrica 83*, 2043–2063.
Kline, P. and Walters, C. R. (2019), On Heckits, LATE, and numerical equivalence, *Econometrica 87*, 677–696.
Kling, J. R., Liebman, J. R., and Katz, L. F. (2007), Experimental analysis of neighborhood effects, *Econometrica 75*, 83–119.
Kolesár, M., Chetty, R., Friedman, J., Glaeser, E., and Imbens, G. W. (2015), Identification and inference with many invalid instruments, *Journal of Business & Economic Statistics 33*, 474–484.
Koopmans, T. C. (1949), Identification problems in economic model construction, *Econometrica 17*, 125–144.
Kroft, K., Luo, Y., Mogstad, M., and Setzler, B. (2020), Imperfect competition and rents in labor and product markets: the case of the construction industry, National Bureau of Economic Research (NBER) Working Paper 27325.
Krueger, A. B. and Summers, L. H. (1988), Efficiency wages and the inter-industry wage structure, *Econometrica 56*, 259–293.
LaLonde, R. J. (1986), Evaluating the econometric evaluations of training programs with experimental data, *American Economic Review 76* (4), 604–620.
Lamadon, T., Mogstad, M., and Setzler, B. (2022), Imperfect competition, compensating differentials, and rent sharing in the US labor market, *American Economic Review 112*, 169–212.
Leamer, E. E. (1983), Let's take the con out of econometrics, *American Economic Review 73* (1), 31–43.
Leontief, W. (1982), Academic economics (Letter to the Editors), *Science 217*, 104–107.
Levitt, S. D. and List, J. A. (2008), Field experiments in economics: the past, the present and the future, *European Economic Review 53*, 1–18.
Lewis, H. G. (1986a), Chapter 20: Union relative wage effects, in O. C. Ashenfelter and R. Layard (eds), *Handbook of Labor Economics*, Vol. 2, Elsevier, New York, NY, 1139–1181.
Lewis, H. G. (1986b), *Union Relative Wage Effects: A Survey*, University of Chicago Press, Chicago, IL.
Manning, A. (2003), *Monopsony in Motion: Imperfect Competition in Labor Markets*, Princeton University Press, Princeton, NJ.
Manning, A. (2021), Monopsony in labor markets: a review, *Industrial and Labor Relations Review 74*, 3–26.
Manski, C. (1990), Nonparametric bounds on treatment effects, *American Economic Review 80* (2), 319–323.

Meyer, B. D. (1995), Natural and quasi-experiments in economics, *Journal of Business & Economic Statistics 13*, 151.
Meyer, B., Viscusi, W. K., and Durbin, D. (1995), Workers' compensation and injury duration: evidence from a natural experiment, *American Economic Review 85* (3), 322–340.
Miguel, E. and Kremer, M. (2004), Worms: identifying impacts on education and health in the presence of treatment externalities, *Econometrica 72*, 159–217.
Mogstad, M., Santos, A., and Torgovitsky, A. (2019), Using instrumental variables for inference about policy relevant treatment parameters, *Econometrica 86*, 1589–1619.
Murphy, K. M. and Topel, R. H. (1987), Unemployment, risk, and earnings: testing for equalizing wage differences in the labor market, in K. Lang and J. S. Leonard, *Unemployment and the Structure of Labor Markets*, Blackwell, London, 103–140.
Neyman, J. (1923), On the application of probability theory to agricultural experiments. Essay on principles. Section 9, *Roczniki Nauk Roiniczych Tom X*, 1–51 (in Polish). Translation in 1990, *Statistical Science 5*, 465–480.
Pearl, J. (1995), Causal diagrams for empirical research, *Biometrika 82*, 669–688.
Pearl, J. (2000), *Causality*, Cambridge University Press, Cambridge.
Pearl, J. (2015), Trygve Haavelmo and the emergence of causal calculus, *Econometric Theory 31*, 152–179.
Pearl, J. and Mackenzie, D. (2018), *The Book of Why: The New Science of Cause and Effect*, Basic Books, New York, NY.
Peri, G. and Sparber, C. (2011), Assessing inherent model bias: an application to native displacement in response to immigration, *Journal of Urban Economics 69*, 82–91.
Peri, G. and Yasenov, V. (2018), The labor market effects of a refugee wave: synthetic control method meets the Mariel boatlift, *Journal of Human Resources 54*, 267–309.
Robins, J. M. (1989), The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies, in L. Sechrest, H. Freeman, and A. Mulley (eds), *Health Service Research Methodology: A Focus on AIDS*, National Center for Health Services Research and Health Care Technology Assessment, Public Health Service, US Department of Health and Human Services, Washington, DC, 113–159.
Robinson, J. (1933), *The Economics of Imperfect Competition*, St. Martin's Press, New York, NY.
Rosenzweig, M. R. and Wolpin, K. I. (1980), Testing the quantity–quality fertility model: the use of twins as a natural experiment, *Econometrica 48*, 227.
Rubin, D. B. (1974), Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology 66*, 688–701.
Rubin, D. B. (1975), Bayesian inference for causality: the importance of randomization, in E. D. Goldfield (ed.), *Proceedings of the Section on Government Statistics and Section on Social Statistics*, American Statistical Association, Washington, DC, 233–239.
Rubin, D. B. (1978), Bayesian inference for causal effects: the role of randomization, *Annals of Statistics 6*, 34–58.
Rubin, D. B. (1990), Formal modes of statistical inference for causal effects, *Journal of Statistical Planning and Inference 25*, 279–292.
Rubin, D. B. (2008), For objective causal inference, design trumps analysis, *Annals of Applied Statistics 2*, 808–840.
Sims, C. A. (1980), Macroeconomics and reality, *Econometrica 48*, 1–48.
Solon, G. (1985), Work incentive effects of taxing unemployment benefits, *Econometrica 53*, 295.
Stafford, F. (1986), Chapter 7: Forestalling the demise of empirical economics: the role of microdata in labor economics research, in O. C. Ashenfelter and R. Layard (eds), *Handbook of Labor Economics*, Vol. 1, Elsevier, New York, NY, 387–423.

Sun, L. and Abraham, S. (2021), Estimating dynamic treatment effects in event studies with heterogeneous treatment effects, *Journal of Econometrics 225*, 175–199.
Thistlethwaite, D. L. and Campbell, D. T. (1960), Regression-discontinuity analysis: an alternative to the ex post facto experiment, *Journal of Educational Psychology 51*, 309–317.
Tinbergen, J. (1930), *estimmung und deutung von angebotskurven: ein beispiel*, *Zeitschrift für Nationalökonomie 1*, 669–679.
Titiunik, R. (2021), Chapter 6: Natural experiments, in J. Druckman and D. P. Green (eds), *Advances in Experimental Political Science*, 1st edn, Cambridge University Press, Cambridge, 103–129.
Vytlacil, E. J. (2002), Independence, monotonicity, and latent index models: an equivalence result, *Econometrica 70*, 331–341.
Wald, A. (1940), The fitting of straight lines if both variables are subject to error, *Annals of Mathematical Statistics 11*, 284–300.
Working, E. J. (1927), What do statistical 'demand curves' show? *Quarterly Journal of Economics 41*, 279–292.
Wright, P. (1928), *The Tariff on Animal and Vegetable Oils*, Macmillan, New York, NY.