KUNGL. VETENSKAPS-AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

11 OCTOBER 2021

Scientific Background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021

ANSWERING CAUSAL QUESTIONS USING OBSERVATIONAL DATA

The Committee for the Prize in Economic Sciences in Memory of Alfred Nobel

BOX 50005 (LILLA FRESCATIVÄGEN 4 A), SE-104 05 STOCKHOLM, SWEDEN
TEL +46 8 673 95 00, KVA@KVA.SE ⋆ WWW.KVA.SE

ANSWERING CAUSAL QUESTIONS USING OBSERVATIONAL DATA

Most applied science is concerned with uncovering causal relationships. In many fields, randomized controlled trials (RCTs) are considered the gold standard for achieving this. The systematic use of RCTs to study causal relationships — assessing the efficacy of a medical treatment for example — has resulted in tremendous welfare gains in society. However, due to financial, ethical, or practical constraints, many important questions — particularly in the social sciences — cannot be studied using a controlled randomized experiment. For example, what is the impact of school closures on student learning and the spread of the COVID-19 virus? What is the impact of low-skilled immigration on employment and wages? How do institutions affect economic development? How does the imposition of a minimum wage affect employment? In answering these types of questions, researchers must rely on observational data, i.e., data generated without controlled experimental variation. But with observational data, a fundamental identification problem arises: the underlying cause of any correlation remains unclear. If we observe that minimum wages and unemployment correlate, is this because a minimum wage causes unemployment? Or because unemployment and lower wage growth at the bottom of the wage distribution leads to the introduction of a minimum wage? Or because of a myriad of other factors that affect both unemployment and the decision to introduce a minimum wage? Moreover, in many settings, randomized variation by itself is not sufficient for identification of an average treatment effect.

This year's Prize in Economic Sciences rewards three scholars: David Card of the University of California, Berkeley, Joshua Angrist of Massachusetts Institute of Technology, and Guido Imbens of Stanford University. The Laureates' contributions are separate but complementary. Starting with a series of paper from the early 1990s, David Card began to analyze a number of core questions in labor economics using "natural experiments", i.e., a study design in which the units of analysis are exposed to as good as random variation caused by nature, institutions, or policy changes. These initial studies — on the minimum wage, on the impact of immigration, and on education policy — challenged conventional wisdom, and were also the starting point of an iterative process of replications, new empirical studies, and theoretical work, with Card remaining a core contributor. Thanks to this work, we have gained a much deeper understanding of how labor markets operate.

In the mid-1990s, Joshua Angrist and Guido Imbens made fundamental contributions to the challenge of estimating an average treatment effect. In particular, they analyzed the realistic scenario when individuals are affected differently by the treatment and choose whether to comply with the assignment generated by the natural experiment. Angrist and Imbens showed that even in this general setting it is possible to estimate a well-defined treatment effect — the local average treatment effect (LATE) — under a set of minimal (and in many cases empirically plausible) conditions. In deriving their key results, they merged the instrumental variables (IV) framework, common in economics, with the potential-outcomes framework for causal inference, common in statistics. Within this framework, they clarified the core identifying assumptions in a causal design and provided a transparent way of investigating the sensitivity to violations of these assumptions.

The combined contribution of the Laureates, however, is larger than the sum of the individual parts. Card's studies from the early 1990s showcased the power of exploiting natural experiments to uncover causal effects in important domains. This early work thus played a crucial role in shifting the focus in empirical research using observational data towards relying on quasi-experimental variation to establish causal effects. The framework developed by Angrist and Imbens, in turn, significantly altered how researchers approach empirical questions using data generated from either natural experiments or randomized experiments with incomplete compliance to the assigned treatment. At the core, the LATE interpretation clarifies what can and

cannot be learned from such experiments. Taken together, therefore, the Laureates' contributions have played a central role in establishing the so-called design-based approach in economics. This approach – aimed at emulating a randomized experiment to answer a causal question using observational data – has transformed applied work and improved researchers' ability to answer causal questions of great importance for economic and social policy using observational data.

The design-based approach: background

At least until the 1980s, the traditional approach to causal inference in economics relied on structural equation models — that is, on the specification of systems of equations capturing behavioral relationships; see Wright (1928) and 1989 Laureate Trygve Haavelmo (1943, 1944). A key concern with the structural equation approach, however, is that in order to establish a causal relationship, the proposed structure has to be correctly specified.

By the early 1980s, the difficulties associated with correctly specifying a structural model for causal inference became clear.[1] Ashenfelter (1978) pointed to the difficulties of evaluating job training programs and Lalonde (1986) showed that experimental estimates, obtained from a randomized evaluation of a job-training program, were systematically different from the estimates obtained by applying standard econometric methods to observational data from the same program. These results provided an important impetus to the Laureates' innovations that were to come.

The design-based approach: new credible evidence based on natural experiments

In response to the findings in, e.g., Lalonde (1986), labor economists in the late 1980s turned to exploiting data generated by natural experiments. Under certain circumstances, the variation induced by (changes in) nature, policy, or institutions, implies that we can separate those affected by the treatment from those that were not, as if assignment to treatment was random. Examples of early, and highly influential contributions include Angrist (1990), Card (1990), Angrist and Krueger (1991), Card and Krueger (1992a, 1994). While economists were not the first to advocate a quasi-experimental approach for causal inference, the systematic use of quasi-experimental variation to address questions of great significance to economic and social policy quickly changed the practice in applied microeconomic research and beyond.[2] This change was not primarily about using new empirical methods, but rather about how to approach a causal question. The natural experiment agenda required researchers to understand the process determining which units receive which treatments. The new approach thus required an understanding of the source of identifying information, i.e., it required institutional knowledge about the natural experiment.

The use of arguably exogenous variation to estimate causal effects provided new, credible, and highly policy-relevant evidence. These initial findings stimulated new research, and research results are thus cumulating. Card's work on minimum wages provides a good example of this iterative process.

According to the textbook competitive model of the labor market, an increase in the minimum wage yields a substantial reduction in employment — a result that was broadly supported by the available evidence in the late 1980s. Card and Krueger (1994) challenged this conclusion. Using a natural experiment, with data from two neighboring states in the US — one in which the minimum wage was increased — they showed that the increased minimum wage did not have a negative employment impact. This finding stimulated reanalysis using data for the US as well as other countries; the overall conclusion from this extensive literature is that the negative employment impact is limited. Card and Krueger's results also generated empirical and theoretical

[1] At the time there was a general concern with the inability to credibly establish causal claims. See, for example, Hendry (1980), Sims (1980), and Leamer (1983).

[2] Already in the 1960s, Donald Campbell discussed using institutional reforms and institutional rules in education to estimate causal effects (see Section 1).

work trying to explain the lack of negative employment effects. Several empirically supported explanations have been offered. One explanation is that the cost per efficiency unit of labor does not rise one for one with minimum wages; for instance, evidence shows that worker productivity improves following the introduction of a higher minimum wage. Another explanation is that local service providers, which are most affected by the minimum wage, are able to shift the cost increase onto consumers in the form of higher prices without much loss in product demand. A third explanation is that firms have monopsony power in the labor market; with monopsony power, the effect of the minimum wage is ambiguous because of countervailing effects on labor demand and labor supply.[3] The minimum wage literature thus prompted a revival of research on the consequences of monopsony in the labor market and the contribution of firm wage setting to (changes in) wage inequality. While it would be inappropriate to conclude that increases in the minimum wage never have negative employment effects, we have a much deeper understanding of why this may or may not be the case than we did 30 years ago.

The design-based approach: the LATE framework

The findings from the initial wave of studies exploiting natural experiments raised conceptually important issues. Quasi-experimental estimates of the earnings return to schooling, for example, suggested that causal returns were higher than simple estimates using ordinary least squares (OLS). This was surprising, since researchers expected the OLS estimator to suffer from upward "ability bias".[4] These findings pointed to the importance of considering heterogeneous effects.

A second issue relates to compliance. With the exception of certain types of controlled clinical trials, compliance with the quasi-experimental or experimental assignment is almost always incomplete. If you prolong compulsory schooling by a year, for instance, educational attainment will increase by less than a year, on average, since some individuals already would have pursued more than compulsory schooling prior to the policy change. Similarly, individuals may deviate from the treatment protocol in a randomized controlled trial. That is, complete compliance is the exception rather than the rule.

How can one estimate a treatment effect when the responses vary in the population and there is incomplete compliance? In medical science, the norm was, and still is, to estimate and report the intention-to-treat estimate: i.e., to analyze outcomes with individuals assigned to their initial treatment status, even if they deviated from that during the course of the trial. In many contexts, the intention-to-treat estimate is a parameter of interest; in others, it is the only causal parameter that can be estimated without imposing additional assumptions. The intention-to-treat estimate, however, does not estimate the treatment effect for individuals following the treatment protocol, which almost always is a key parameter. Importantly, and as discussed, for example, by Hernán and Robins (2017), intention-to-treat analyses may make unsafe interventions appear safe, or an effective intervention appear to be ineffective.

In the late 1980s, researchers began to investigate under what conditions a treatment effect can be estimated, when effects are heterogeneous and compliance is incomplete. The early contributors to this literature, including Chamberlain (1986), Robins (1989), Heckman (1990), and Manski (1990), focused either on specifying conditions under which an average treatment effect can be estimated or on bounding an average treatment effect. These contributions showed that the causal effect for those who took part in the intervention could only be estimated under special

[3] Monopsony refers to a situation where employers have market power in the labor market (Robinson, 1933). Employers may use that market power to set the wage below the level generated in a competitive labor market. In a monopsonistic labor market, a marginal increase in the minimum wage can raise employment because of a positive labor supply response.

[4] The concern is that more "able" individuals have higher earnings as well as more schooling. Since all dimensions of ability cannot be observed, the OLS estimator may be positively biased (see Section 1).

circumstances;[5] moreover, in practical applications, the bounds around the average treatment effect were often too wide to be informative.

In seminal work, Imbens and Angrist (1994) pushed this literature forward in a crucial way (see also Angrist and Imbens, 1995, and Angrist, Imbens, and Rubin, 1996). Specifically, Angrist and Imbens reformulated the problem by asking what causal treatment effect can be estimated from a randomized or quasi-experimental study without placing restrictions on the behavior of study objects when, as is reasonable, the responses are heterogeneous in the population. Using a minimal set of assumptions, they showed that an average causal effect can be estimated among those who complied with the assignment generated by the experiment or the quasi-experiment, and that this effect is identified by instrumental variables. Angrist and Imbens referred to this causal effect as the local average treatment effect — sometimes it is also referred to as the complier average causal effect. Angrist and Imbens thus made clear exactly what one can say about treatment effects in settings where the source of the heterogeneity is not known (and cannot be modeled and estimated).

In establishing their key results, Angrist and Imbens made a broader contribution. By casting their analysis in terms of potential outcomes, they merged the IV framework, invented in economics, with the potential-outcomes framework for causal inference, developed in statistics. This, in turn, yielded a general framework that has improved researchers' ability to establish causal effects and interpret their results; in particular, it makes the nature of the identifying assumptions transparent and allows researchers to assess the sensitivity of an empirical design to deviations from these assumptions. These advantages have turned the framework into the dominant one for both quasi-experimental and experimental work in economics and beyond. Moreover, the basic framework provided by Angrist and Imbens has been used to test the assumptions required to identify LATE and to investigate the conditions under which a causal effect can be identified when using other methods for causal inference, including the regression discontinuity design (Hahn, Todd, and van der Klaauw, 2001) and the differences-in-differences design (e.g., de Chaisemartin and D'Haultfoeuille, 2020).[6]

Outline

The remainder of this overview is organized as follows. Section 1 places natural experiments in context by discussing different approaches to causal inference in economics. Section 2 focuses on Card's research contributions to three areas: the minimum wage, the impact of immigration, and education policy. The discussion covers his initial seminal work, the subsequent work that these papers stimulated, and what we have learned from almost 30 years of research in these three areas. The early and later literature involved several researchers, and Card has been a key contributor throughout. Another central researcher was Alan Krueger, who passed away in 2019; among other things, he coauthored a number of the early key studies with Card. Angrist contributed to the early literature as well, in part with a paper coauthored with Krueger that we discuss in Section 1.

Section 3 turns to Angrist's and Imbens' methodological contributions to the design-based approach. It illustrates how casting instrumental variables in the potential-outcomes framework clarifies the assumptions underlying such analyses in the presence of heterogeneous treatment effects and it discusses the interpretation of the treatment effects. Section 4 gives a brief account of the scientific discussion that has followed the advent of the design-based approach. Section 5 offers concluding remarks.

[5] For example, if individuals are unaware of how the treatment affects them, or do not act on such treatment heterogeneity, it is possible to estimate the average causal effect among the treated.
[6] These methods are briefly described in Section 1.

1. SETTING THE STAGE: CAUSAL INFERENCE IN ECONOMICS

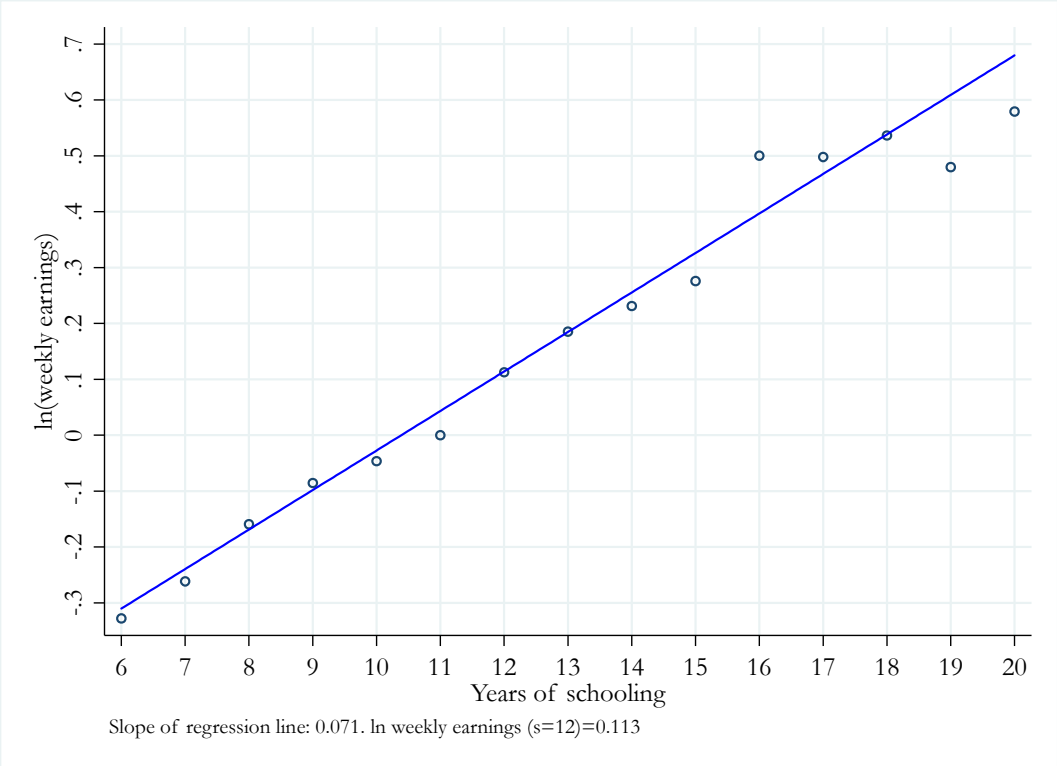
What are the benefits of completing high school? This question is surely in the minds of many parents and children. One component of the answer is how a high school degree impacts a child's future earnings. Since this "rate of return to schooling" is both a highly policy-relevant parameter and helps in understanding inequality in the labor market, much research in economics has probed the relationship between educational attainment and future labor market outcomes. How does one proceed to answer a causal question such as the one above? We will introduce the contributions of this year's Laureates and place them in context of earlier work by focusing on this application.

Before we begin, let us note that we need to isolate the effect of only one factor — completing high school or not — and compare two outcomes that cannot occur at the same time. Only one path will eventually be chosen, but we need to compare this path to a "counterfactual" path to determine the effect on earnings. Virtually all decisions have this feature — not only for parents and children, but also for firms and other organizations. Suppose a government considers a policy change, such as the introduction of a minimum wage. Before making this decision, it would ideally like to know what would happen to the economy in two alternative scenarios, with and without the minimum wage. However, it is only possible to observe one of these scenarios, *ex post*.

Let us return to the question of the impact of education on earnings. A natural start is to look at data. Figure 1 shows the cross-sectional relationship between the logarithm of earnings, y, and schooling, s, for a number of individuals observed in the US Census. Earnings and schooling are clearly positively related. On average, an additional year of schooling is associated with about 7 percent higher earnings. If we focus on the margin of high school completion, i.e., compare individuals with 12 and 11 years of schooling, the estimated return is higher — it exceeds 11 percent. Extensive work by Jacob Mincer (e.g., Mincer, 1958), and others, using similar kinds of data has shown that the earnings-schooling relationship is a strong empirical regularity that exists in all conceivable contexts.

However, we cannot directly interpret the relationship in Figure 1 causally. The main reason for skepticism is that earnings outcomes may be caused by other factors that also affect schooling choices; after all, the data cover a large range of individuals who differ in a number of ways. For example, an individual may simply be inherently very "able" at many activities, including obtaining good grades in school as well as generating high earnings. Indeed, it is possible that more schooling doesn't cause higher earnings at all, and that the graph simply illustrates that able individuals have high earnings as well as more schooling. We could address this problem if it were possible to control for all relevant individual characteristics. Many individual characteristics are indeed observable, but many relevant factors — "ability," "motivation," "willingness to work hard" — are arguably not. Thus, we cannot simply sort the question out by constructing the same graph for all possible subgroups of people: we do not have the data. Unobserved heterogeneity is more generally a major challenge in empirical analysis based on observational data. So, what to do?

Figure 1: The cross-sectional relationship between earnings and schooling



Notes: The figure is based on the data used by Angrist and Krueger (1991). The data set comes from the 1980 US Census. It covers men born 1930-1939. Log earnings for individuals with 11 years of schooling are normalized to 0.

The 1989 Laureate Haavelmo suggested one way forward, by combining economic theory and empirical methodology. His applications were different, but translated to our context a theory would offer a causal mechanism in the form of two functions $f(\cdot)$ and $g(\cdot)$ that relate earnings (y) and schooling (s) to one another: $y = f(s, X)$ and $s = g(y, X)$. The $f(\cdot)$ mapping represents a theory of how earnings are generated; the $g(\cdot)$ mapping describes how schooling choices are made; and X is a (long) vector, which contains the characteristics of the individuals, some of which may not be observable, and purely random shocks.

The obvious issue here is that a hard-to-observe element of X (say, ability) may affect both y and s. Haavelmo's proposed way forward was that one could imagine that some element of X, called z, appeared in $g(\cdot)$ but not in $f(\cdot)$: z affects s but does not directly cause y. In econometric terminology, z can thus be used to construct an "instrument" for s; the method of instrumental variables was originally proposed by Wright (1928). Haavelmo's approach has been followed by many others. It has had significant success in economics and yielded many insights. The Prizes in 2000 to the labor economist James Heckman and in 2011 to the macroeconomists Thomas Sargent and Christopher Sims rewarded related contributions, relying on various degrees of theoretical assumptions in their respective subfields. Different subfields face specific challenges; in macroeconomics, it is arguably particularly difficult to find fully convincing instruments, and hence theoretical models have historically played a larger role in this field.

Econometric analysis using instrumental variables remains at the core of the subject, and this is also true for the research rewarded by this year's prize. The challenge, however, is a practical one: Where do valid instruments come from? How can we find our z? In our application, the instrument should affect schooling but have no direct effect on earnings. Almost inevitably, when we try to imagine any candidate z, this exclusion restriction becomes a concern.

A key aspect of the research rewarded this year is that it has demonstrated new ways of defining valid instruments, ways that involve much less input in the form of theoretical

assumptions. The path taken by the researchers has been to make use of so-called natural experiments, which in the schooling example can be seen as ways of varying s independently of y: the natural experiment gives us the equivalent of the z. Natural experiments appear frequently in observational data and they surprisingly often offer instances that are "as-if" controlled experiments. In particular, the Laureates, and the researchers following their original contributions, have demonstrated the abundance of natural experiments in contexts of significant relevance for economics, in understanding human behavior, and for policymaking. The idea of using natural experiments had been proposed in the social sciences before (see below), but their major importance in economics was yet to be uncovered.

To understand how a natural experiment works, let us first return to the core question of causality in controlled experiments. To formally define a causal effect, the notion of potential outcomes — introduced by Neyman (1923) and further refined and extended by Rubin (1974, 1977) — is useful. Suppose we are interested in the causal effect of completing high school on earnings. For each individual, there are two potential outcomes — one pertaining to the earnings outcome if the individual completes high school, $Y_i(1)$, and the other if the individual does not complete high-school, $Y_i(0)$. The causal effect for a single individual i is defined as the difference between these two potential outcomes, i.e., $\beta_i = Y_i(1) - Y_i(0)$. Since one cannot complete and not complete high school at the same time, one of the potential outcomes — the counterfactual — is always missing. Thus, we cannot estimate the causal effect for an individual, without making strong assumptions. Yet, with a suitable empirical design and data on a large set of individuals, we can sometimes estimate an average causal effect of completing high school on earnings.

Suppose we have access to observational data for a large number of individuals on their earnings, Y_i, and whether they have completed high school, $D_i = 1$, or not, $D_i = 0$. The observed average difference in earnings across the two groups, is an estimate of $\Delta = E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0]$. It is unlikely that Δ equals the causal effect of interest, however. To see this formally, subtract and add $E[Y_i(0)|D_i = 1]$, a counterfactual term, which yields:

$$\underbrace{\Delta}_{\text{Difference in means}} = \underbrace{E\left[\left(Y_i(1) - Y_i(0)\right)\middle|D_i = 1\right]}_{\text{causal effect on high school completers}} + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{\text{selection bias}}$$

The first term in this expression, $E\left[\left(Y_i(1) - Y_i(0)\right)\middle|D_i = 1\right]$, is the causal effect of interest. In our high school example, it provides the answer to the question: What is the effect on wage earnings of completing high school, among those that did so? The second term, $E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]$, represents selection bias. It measures how different the earnings would have been in the two populations — completers and non-completers — had they not completed high school. For reasons described above, high school completers would likely earn more than non-completers, even without high-school, implying that the selection effect is positive in this example. In such a case, the comparison of means, Δ, provides an upward biased estimate of the causal effect of high school education on earnings.

1.1 The controlled randomized experiment

In the medical sciences, double-blind randomized experiments are often used for determining the effects of a treatment. For example, a drug and a placebo may be randomly given to patients and the health effects then compared between those receiving the drug and those given a placebo.[7] This method is of tremendous value, though it still does not produce counterfactuals at the

[7] With double blinding, neither the study object (e.g., a patient) nor the implementer of the treatment is aware of which group the study object is assigned to. If participants in the experiment know which treatment was given to the subjects, their behavior may be affected, which may bias the estimate of the treatment effect from the experiment.

individual level; rather, because of random assignment (and double-blind trial protocols) patients in the group receiving the drug and the placebo are expected to be alike (and behave similarly) in the absence of the intervention. If individuals comply completely with random assignment, this feature allows inference to the average causal effect of the drug on a health outcome for the population in the experiment.

In an ideal experiment, where D is randomly assigned, potential outcomes are independent of treatment status. In expectation, observed outcomes differ across individuals assigned to treatment and control groups only through their exposure to the treatment. This result depends crucially on perfect compliance with random assignment, i.e., that all individuals assigned to the treatment group take up treatment, while no individuals in the control group do so. Thus, under certain conditions, a randomized experiment, provides an unbiased estimate of the average treatment effect in the study population.[8] The ideal experiment, therefore, has a strong claim for internal validity, i.e., it answers the question at hand for the population and context studied in the experiment. Further trials are required to determine whether the results can be generalized to other populations and other settings, i.e., whether the estimate is externally valid.[9]

The potential to uncover causal relationships using RCTs has revolutionized many scientific disciplines. For example, RCTs have been the norm in medical research for decades; more recently, field experiments have become an increasingly used tool, for example, in development economics. Randomized experiments can be used to answer a broad range of causal questions, as illustrated, for example, by the work of the 2019 Laureates Abhijit Banerjee, Esther Duflo, and Michael Kremer. But the approach requires that the researcher can assign the treatment of interest to the research subjects; for many important questions in economics this is not possible — neither practically nor ethically. In our schooling example, an RCT would single out some students and not allow them to finish high school, whereas others would be allowed to — or even forced to — finish.

The notion of RCTs as a useful thought experiment goes back a long time. In the context of the research mentioned above, Haavelmo (1944) in fact argued that any quantitative theory should include a description of the *notional experiment* that the researcher would like to run in order to isolate a particular theoretical mechanism. The identification of such a mechanism may also be assisted by "the stream of experiments that Nature is steadily turning out of her own enormous laboratory" (Haavelmo, 1944, p. 14). We turn to these natural experiments next.

1.2. Natural experiments

A natural experiment is an event or a situation, that is not under the control of the subjects under study, which generates variation in the variable of interest that is as if it had been randomly assigned. The underlying variation (i.e., the natural experiment) can come from policy changes, administrative rules, naturally occurring random variation (e.g., birth dates, weather shocks), or from unforeseen events (e.g., immigration flows). Natural experiments provide a powerful complement to controlled experimentation; in applied microeconomics, the use of natural experiments for causal inference has exploded over the past 30 years.

The idea to exploit natural experiments for causal inference, however, goes further back in time (as illustrated by the quote from Haavelmo above). Donald Campbell was an early proponent of a quasi-experimental approach (e.g., Campbell, 1969) and also developed one of the main empirical methods — the regression discontinuity (RD) design (Thistlewaite and Campbell, 1960) — to estimate causal effects from natural experiments. This design is built on the existence of threshold rules where, e.g., individuals on one side of the threshold get the treatment and

[8] Since potential outcomes are independent of treatment status, the causal effect among the treated is in expectation equal to the causal effect in the study population

[9] The distinction between internal and external validity was introduced by Campbell (1957).

individuals on the other side do not. If the (observed and unobserved) characteristics of the individuals around the threshold are very similar to one another, the RD design can be used to estimate the treatment effect. The next subsection gives several examples of such threshold rules that lead to RD designs.

One part of this year's Prize in Economic Sciences rewards David Card for improving our understanding of how the labor market operates. In a number of key contributions from the early 1990s, he brought new evidence on core topics in labor economics using natural experiments. Card's influence, however, goes beyond the substantive results discussed in Section 2. His work, along with the work of Joshua Angrist and Alan Krueger, among others, helped shape the approach to empirical research based on natural experiments. This new approach — the design-based approach — uses the notional experiment as a guiding framework.

Nature's stream of experiments

Let us now briefly describe a handful among the thousands of examples of natural experiments that have been used in the literature in economics and beyond. The objective here is to illustrate how researchers can address important questions using plausibly exogenous variation, generated in naturally occurring as-if experiments.

How does fertility affect parental labor supply? Answering this question requires an empirical strategy that can address reverse causality, i.e., the fact that fertility likely depends on labor market prospects. Angrist and Evans (1998) thus used parental preferences for mixed gender composition of their children. Hence, families where the first two children are of the same gender are more likely to have a third child. Moreover, child gender is random, implying that the gender composition of the children can serve as an instrument. Using this natural experiment, Angrist and Evans find that the negative effects on female labor supply are slightly larger than the corresponding ordinary least squares estimates.

In healthcare, it is not uncommon that threshold rules are used to determine when a medical intervention is administered. For example, newborns below 1,500 grams are typically classified as low birth weight and they receive additional medical treatment. To investigate to what extent these medical interventions save and improve the lives of low-birth-weight children, Almond, Doyle Jr., Kowalski, and Williams (2010) compared outcomes among children just above and below the threshold at 1,500 grams. They found that those just below 1,500 grams have lower one-year mortality rates, even though the correlation between mortality and birth weight is negative in the data.

Related to our example of the returns to education is the question of the returns to specific educational programs or certain types of schools. When a program or a school is oversubscribed, there is typically an admittance threshold where students above the threshold are admitted and those below are not. Since students just above and below the admittance margin can be expected to be very similar to one another, one can credibly compare future outcomes for those who were marginally admitted with the outcomes for those who were not. Several studies have used variation coming from such admittance thresholds (see Kirkebøen, Leuven, and Mogstad, 2016, and Pop-Eleches and Urquiola, 2013, for example).

A long-standing question in political science is whether the incumbent is more likely to win the next election, just because he or she is in power. Answering this question is difficult, since we need to distinguish the effect of incumbency per se from the qualities that led the incumbent win the last election. However, in close elections the qualities of the politician who barely won and the politician who barely lost should be approximately the same in the eyes of the voters, and therefore one can use close elections to identify the effects of incumbency. Using this strategy, Lee (2008) concluded that the incumbency advantage is substantial.

A key policy question is whether extending unemployment insurance (UI) receipt contributes to increasing unemployment. This question has been addressed using several natural experiments.

Lalive (2008), for example, used variation coming from a massive extension of UI benefits — from 7 months to 4 years — in Austria during the late 1980s. The extension applied to workers above age 50 who lived in certain regions. Lalive compared individuals on either side of the age threshold, as well as individuals close to the geographic border of the treatment regions, and concluded that unemployment duration increased by around 15 weeks. Subsequently Lalive, Landais, and Zweimuller (2015) used the same reform to examine whether the effects of the UI extension spilled over to other workers not covered by the reform. Using a difference-in-differences (DiD) design, comparing the evolution of labor market outcomes among ineligible workers in treated and untreated regions, they concluded that labor market prospects for non-treated workers improved.[10]

Sometimes governments wish to stabilize the economy using temporary fiscal stimulus packages. To examine whether a tax rebate affected household consumption expenditures in the US, Johnson, Parker, and Souleles (2006) used the fact that the exact week when households received the rebate was effectively randomly assigned (it depended on the second-to-last digit in the taxpayer's social security number). They could thus credibly compare weekly consumption expenditure among households receiving the rebate in different weeks during a month. They find that households spent around 30 percent of the tax rebate on non-durable goods.

In all of the examples above, policy-related or "natural" happenstances generated experiments that were not planned, as such. However, there are also cases where there was indeed a lottery, although entirely without a scientific aim. An interesting Swedish example is the use of a government-issued savings vehicle — bonds where part of the interest was randomized. Savers could thus win large (or small) amounts in actual lotteries, and a large fraction of the Swedish population invested in these bonds. Researchers have used data on individuals investing in these bonds to examine how additional wealth affects a number of important variables such as hours worked, health, and psychological well-being (see Cesarini, Lindqvist, Östling, and Wallace, 2016, and Cesarini, Lindqvist, Notowidigdo, and Östling, 2017, for example).

Revisiting the schooling example

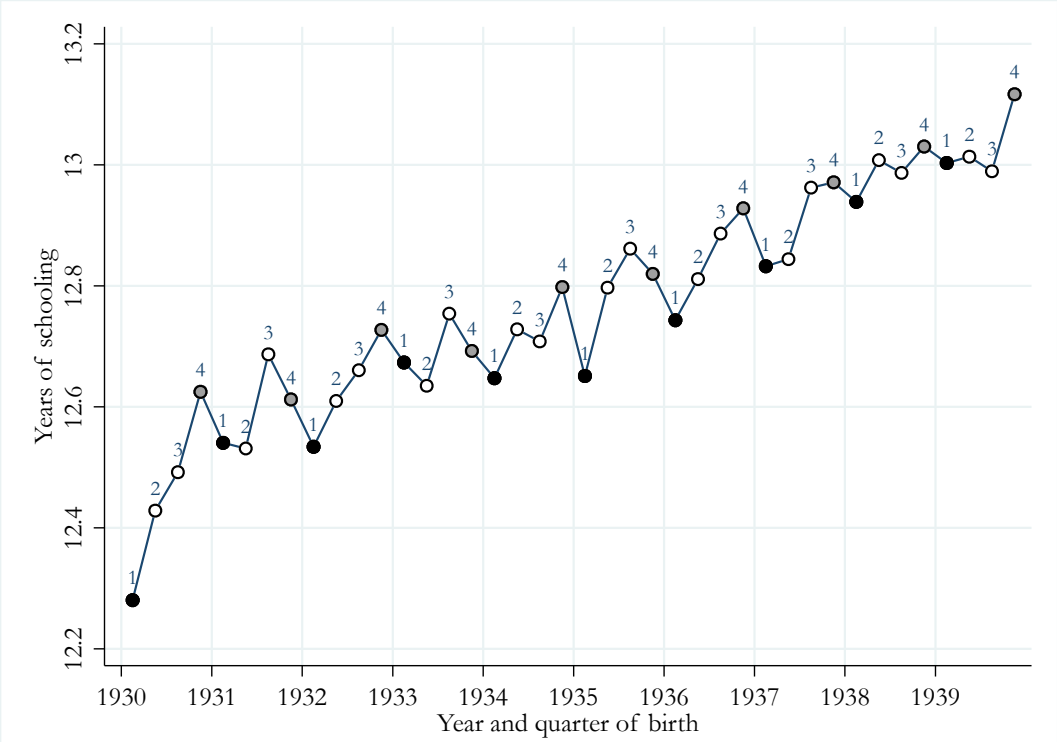
Several natural experiments have been used in order to improve our understanding of how schooling affects earnings. Here we focus on the landmark study by Angrist and Krueger (1991), which we also discuss in Sections 2 and 3. They noted that an individual's birth date could have an effect on high school completion, and educational attainment more generally. The reason is that US legislation features a compulsory schooling-leaving age: pupils are allowed to leave school when they reach a certain age. However, all students born in a given state and year start school at the same time. This means that students born early in a year would reach an age where it was legal to drop out of high school earlier than others, and some of the students do. Thus, Angrist and Krueger argued that the birth date of a student could serve as an instrument, z: something that affects schooling but, in itself, had no direct impact on earnings. In the data they observed birth quarters rather than birth dates, but it was still plausible that birth quarter appeared in $g(\cdot)$ but not in $f(\cdot)$. This was not a controlled experiment that assigned children to high school completion or non-completion, but it had those features.

Figure 2 illustrates the relationship between years of schooling and quarter of birth using the same data as Angrist and Krueger (1991) used. The black circles show average years of schooling for those born in the first quarter, while the gray circles show average years of schooling for individuals born in the fourth quarter. Those born in the first quarter consistently have attained fewer years of schooling than those born in the fourth quarter within a birth year. On average, the difference between the two groups is 0.15 years of education. Figure 3 presents analogous evidence

[10] The DiD design is probably the most commonly used research design for estimating causal effects. It is based on comparing the change in outcomes (before and after the intervention) among subjects affected by the treatment, to the corresponding change among subjects who are not. The DiD design is usually attributed to John Snow (1855).

for log earnings. On average, the earnings of those born in the fourth quarter exceed the earnings of those born in the first quarter by 1.4 percent.

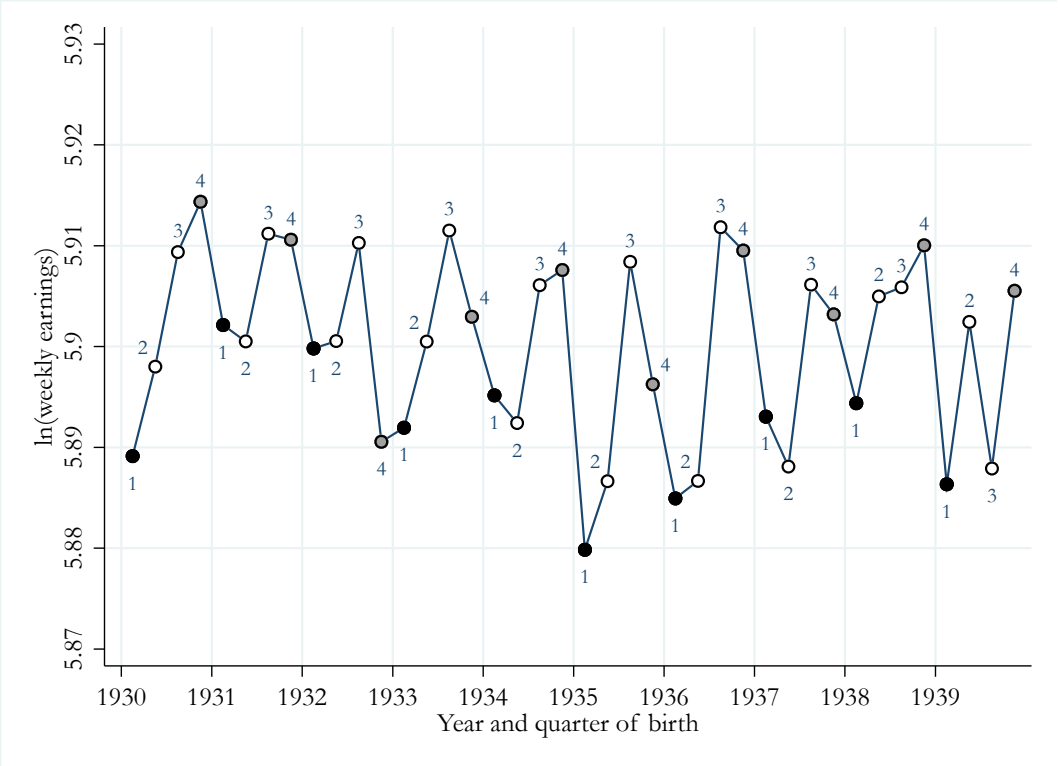
Figure 2: Years of schooling by quarter of birth



Notes: The figure is based on the data used by Angrist and Krueger (1991). The data set comes from the 1980 US Census and covers men born 1930–1939. Black circles show mean years of schooling among men born in the first quarter; gray circles pertain to men born in the fourth quarter. The difference between the two groups is 0.1514 years.

The ratio between the difference in earnings to the difference in schooling is an instrumental variables estimate of the rate of return to schooling. In this particular case, the IV estimate equals 0.089. In other words, the causal return to an additional year of schooling is roughly 9 percent. Thus, perhaps surprisingly, Angrist and Krueger found a causal return to schooling that is slightly higher than indicated by the regression line in Figure 1. As we shall see, however, this is not the end of the story. The quasi-experimental variation mainly affected those with a high probability of dropping out of school as soon as possible. It may well be that the returns to schooling in this part of the population are not representative of the overall population. In the language of controlled experiments, those who were unaffected by the natural experiment are "non-compliers", and their returns to schooling are potentially different than among the "compliers", because of heterogeneous treatment effects.

Figure 3: Log weekly earnings by quarter of birth



Notes: The figure is based on the data used by Angrist and Krueger (1991). The data set comes from the 1980 US Census and covers men born 1930–1939. Black circles show mean log earnings among men born in the first quarter; gray circles pertain to men born in the fourth quarter. The difference between the two groups is 0.0135.

1.3 Causal effects in a world with imperfect compliance and individual heterogeneity

The discussion above makes clear that the individual causal effect $\beta_i = Y_i(1) - Y_i(0)$ likely varies across individuals and across contexts. This raises a number of new issues, independently of whether the source of variation in the data arises from an RCT or from a natural experiment. Thus, even in medical sciences, where compliance is seldom complete, these issues arise.

Imperfect compliance with treatment assignment makes it more difficult to identify the average effect of the treatment, in particular when causal effects vary in the population under study. At the heart of the problem is that when only the assignment, but not the actual treatment received, can be controlled, the selection problem resurfaces. Since selected subsets of the population decide to comply with the assignment, the difference in means across the two treatment groups no longer provides an unbiased estimate of the average treatment effect. The difference in means, however, still captures a causal effect — namely that of the assignment. This effect is often labelled the intention-to-treat (ITT) effect. An ITT analysis thus provides an unbiased estimate of the effect of the treatment assignment in the study population, but not the causal effect of the treatment itself.

In Section 3 we discuss the core contributions of Joshua Angrist and Guido Imbens. In contrast to the earlier literature they asked the fundamental question: What can and cannot be learned from a randomized or natural experiment without placing additional restrictions on the behavior of study objects, when, as is reasonable, the study population is heterogeneous and compliance is imperfect? To answer this question, they introduced a framework that connected instrumental variables to the randomized experiment. Using a minimal set of assumptions, they then showed that it is possible to estimate an average causal effect among those who complied

with the assignment.[11] Their work has proved very valuable not only in economics but also in other sciences, since imperfect compliance is a general problem.

Let us return to the schooling example. As a result of the contributions by Angrist and Imbens, we now know how to properly interpret the findings in Angrist and Krueger (1991). In particular, their IV estimate should be interpreted as the average return to schooling among compliers. Their estimate thus generalizes to the subset of the population where different dates of birth relative to the school entry cutoff affected educational attainment. Without further assumptions, little can be said about those who were unaffected by the natural experiment.

2. UNDERSTANDING LABOR MARKETS

The past 50 years have seen rather substantial changes in inequality. After declining during the 1970s, income inequality has increased dramatically since the early 1980s. Most of the rapid increase in income inequality was driven by a surge in earnings inequality (Atkinson and Piketty, 2010). That surge was observed in most industrialized countries, but it has been much greater in the US and UK than in continental Europe and the Nordic countries.

Earnings inequality is fundamentally shaped by the demand and supply of skills. How changes in demand and supply affect labor market outcomes depends on the design of institutions (e.g., collective bargaining and labor law). Some policies, such as education and immigration policy, directly influence the supply of skills, while other policies, such as minimum wage policies, mainly affect the demand side. Understanding labor market outcomes requires a realistic model of how the labor market operates, as well as information about how different policies affect wages and employment.

To evaluate how different policies (or any other changes) affect labor market outcomes is a major task. As emphasized in the previous section, we need a solid understanding of counterfactual outcomes to estimate the causal impact of a particular policy: What would have happened in the alternative state of the world? Answering that counterfactual question is complicated by the fact that changes usually happen for a reason; policies, for instance, are typically introduced to solve a particular problem.

In a series of papers from the early 1990s, David Card brought rigor and transparency to the analysis of century-old questions: the employment effects of the minimum wage, the labor market impact of immigration, and the effect of educational investments on labor market outcomes. By addressing these questions in novel and, *a priori*, in more credible ways, Card was able to produce new and more reliable answers. The results from the initial studies stimulated reanalysis and theoretical work on how the labor market operates, with Card himself being a key contributor. Thanks to the research initiated by Card, as well as the work that followed, we have gained a much better understanding of the potential of policy to influence labor market outcomes, the impact of immigration on wage and employment differences, and the role of firms in shaping earnings inequality.

The remainder of this section provides brief reviews of the initial seminal work by Card on minimum wages, immigration, and educational policy, respectively. We discuss the subsequent work that these initial studies generated, and conclude with summarizing where we stand today in terms of cumulative knowledge.

While this section focuses on Card's work in a few selected areas, his contributions extend far beyond these. For instance, he also has important work relating to labor market programs and unemployment insurance; Ashenfelter and Card (1985); and Card, Chetty, and Weber (2007) are

[11] Note that "compliers" is a conceptual construct. We cannot tell whether an individual is a complier. As discussed in Section 3, however, the share of compliers in the population can be determined, and their characteristics can be described.

two examples. Another theme in his work is the how unions and wage bargaining affect wage inequality; Card (1996) is an example. His research is not only empirical; in many cases, he combines the empirical work with an interpretative framework or an explicit theory.

2.1. The employment effects of minimum wages

The minimum wage is potentially an important policy tool to reduce poverty among low-wage earners. However, a binding minimum wage increases wage costs and may thus reduce employment; consequently, it is not clear that low-wage workers benefit from the minimum wage in the end. Economists have long studied the employment effects of the minimum wage. The textbook competitive model suggests that employment will decline substantially if the imposition of a higher minimum wage raises wages above their equilibrium levels. Prior to 1990, the evidence on the effects of the minimum wage tended to agree with the textbook model. The evidence was mainly based on time series data, where relative employment among teenage workers (those most likely affected by a minimum wage) was related to the minimum wage "bite", i.e., the minimum wage relative to average (or median) wages (Brown, Gilroy, and Kohen, 1982, surveyed this literature).

However, identifying causal effects of minimum wages is challenging, in particular in a time series setting.[12] Minimum wages are implemented (or changed) for a reason, and the underlying reason may be linked to changes in employment prospects. For example, a downturn of the business cycle manifests itself in a fall in employment and lower wage growth, in particular for low-wage earners. Lower wage growth at the bottom of the distribution may, in turn, create calls for increases in the minimum wage to protect the working poor. If policy-makers act on such demands, we have a scenario where a fall in employment leads to an increase in the minimum wage, rather than the other way around.

Card's early contributions

In 1992, the *Industrial and Labor Relations Review* published the output from a symposium entitled "New Minimum Wage Research". The papers included in the symposium reported results based on variation within states over time to estimate the employment impact of minimum wage. In contrast to work using time-series data, these studies could control flexibly for common time trends. The findings in these papers were partly at odds with previous evidence. By way of background, we discuss three of these papers – Card (1992a), Card (1992b), and Katz and Krueger (1992) – in more detail;[13] thereafter, we turn to the paper by Card and Krueger (1994). The papers we cover in this subsection represent a break with the previous time series evidence, as they clearly spelled out the source of the minimum wage change and the nature of the comparison used to estimate the employment effect.[14]

Card (1992a) compared the evolution of wages and employment in California, which raised its minimum wage in 1988 by 27 percent, to the corresponding evolution in a set of comparison states that did not change their minimum wage policy. Between 1987 and 1989, the wages among teenagers increased 10 percent more in California in relation to the comparison states. Despite this increase, there is no evidence of a decline in employment among teenagers. In fact, the employment-to-population ratio increased 4 percent more in California relative to the comparison states. This DiD estimate appears to be driven by an increase in the labor force participation rate.

---

[12] The early literature also includes work using cross-sectional data, but similar problems of interpretation apply to this kind of studies.

[13] There was also a study by Neumark and Wascher (1992) that presented evidence which was more in line with the previous time series evidence.

[14] They actually revived a case study approach that was common prior to the advent of the time series evidence; Kennan (1995) described the older literature.

Card (1992b) leveraged the fact that national changes in the minimum wage have differential impacts across states, depending on the initial wage distribution in each state. The increase in the US federal minimum wage in 1990 potentially affected more than 50 percent of teenagers in some Southern states, while only 5 percent in some New England states, for example. Card found that wages increased more in the states with a greater fraction of affected teens, yet teen employment-to-population rates did not change.

Katz and Krueger (1992) utilized the fact that establishments will be differentially affected by minimum wage changes depending on the share of the workforce that used to be paid below the new minimum. Surveying fast-food restaurants in Texas before and after the 1990 and 1991 increases in the US federal minimum wage, they reported results in line with Card (1992a) — starting wages increased in more affected establishments, and, if anything, employment in these establishments increased.

A lingering concern with studies based on panel data is that it is not always clear why some states increased minimum wages while others did not. Perhaps states implementing stricter minimum wage policies have been exposed to negative labor market shocks, or vice versa. Ideally, one would like to hold employment prospects constant, and only vary the extent to which minimum wages change. To accomplish this, Card and Krueger (1994) compared two adjacent areas, separated by a state border, that were exposed to different minimum wages policies. The idea is that these two adjacent areas experienced similar economic shocks; one area can thus be used as a counterfactual for the other.[15]

More specifically, Card and Krueger (1994) studied the impact of the New Jersey increase in the state minimum wage, from $4.25 to $5.05 per hour, in April 1992. Following Katz and Krueger (1992), they surveyed 410 fast-food restaurants in New Jersey and eastern Pennsylvania (where the minimum wage remained at $4.25 per hour), before and after the introduction of a higher minimum wage in New Jersey.

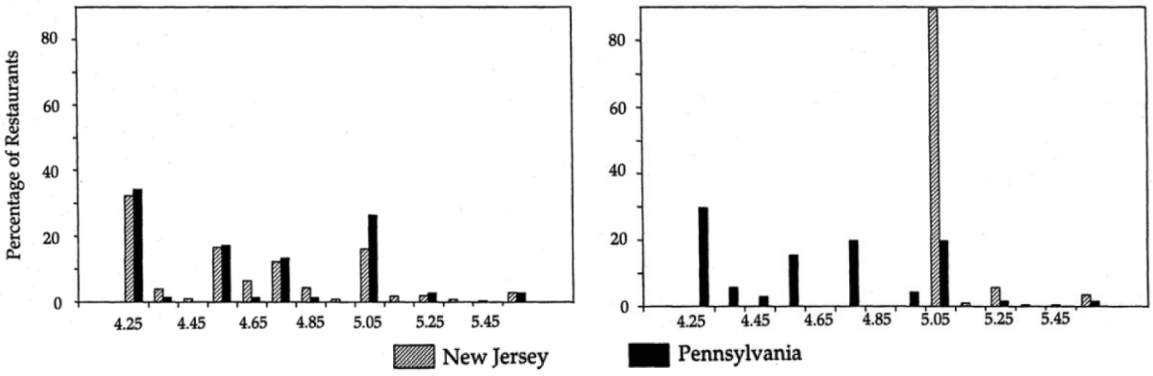
Figure 4 shows the distribution of starting wages in New Jersey before and after the increase in the minimum wage, compared to eastern Pennsylvania. Before the policy change, the distributions look similar. After the policy change, there is a sharp increase in the share of New Jersey restaurants paying the new minimum. On average, starting wages increased 11 percent more in New Jersey than in Pennsylvania. How was employment affected? The left-hand side of Figure 5 shows how employment changed in New Jersey and eastern Pennsylvania, respectively. While employment declined in Pennsylvania, there was a slight increase in New Jersey. There is thus no evidence of a decline in employment as a result of the minimum wage increase.

The right-hand side of Figure 5 compares restaurants that were forced to raise wages a lot (those at $4.25 initially) with restaurants that were broadly unaffected (those at $5.00+ initially). If the employment effects of increases in the minimum wage are negative, we should see employment in low-initial-wage restaurants fall relative to the high-initial-wage restaurants. Again, the data reveal no such a pattern.[16]

[15] In general, Card and Krueger's (1994) research design is akin to a county-level border discontinuity design. In that type of design, researchers compare two contiguous counties straddling a state border, where one of the counties experienced a minimum wage increase while the other did not; see Dube, Lester, and Reich (2010).

[16] The quality of the data in Card and Krueger (1994) was criticized by Neumark and Wascher (2000). Card and Krueger (2000) thus redid the analysis using administrative data from the Bureau of Labor Statistics. They found a positive estimate on an indicator for New Jersey, but it was not statistically significant.

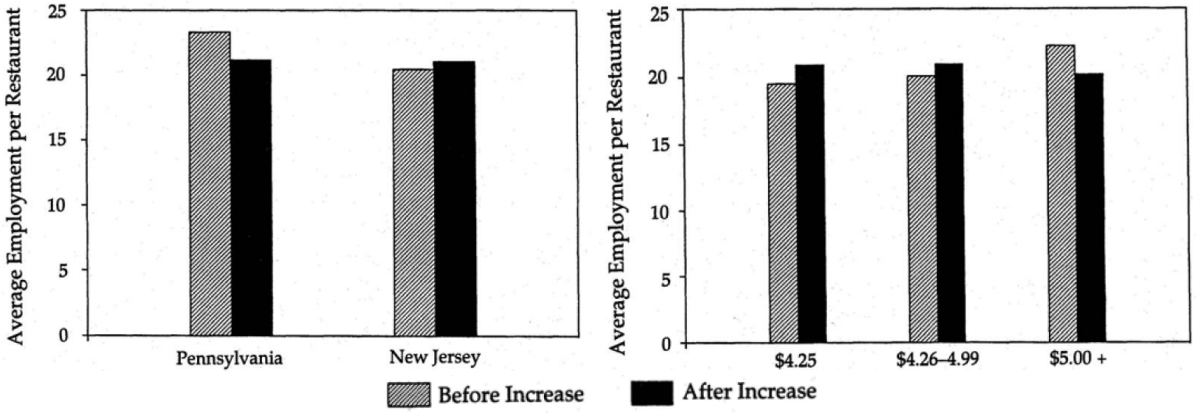
Figure 4: Distribution of starting wage rates, before and after minimum wage increase



Notes: The left-hand side shows the distribution prior to the policy change (Feb-March 1992), the right-hand side shows the distribution after the policy change (Nov-Dec 1992).

Source: Card and Krueger (1995).

Figure 5: Reduced form employment effects



Notes: The left-hand side shows average employment per restaurant before/after the increase for restaurants in New Jersey and Pennsylvania. The right-hand side compares restaurants within New Jersey, by initial wage prior to the increase.

Source: Card and Krueger (1995).

Card and Krueger (1995) summarized and extended the evidence on the effects of the minimum wage using a variety of research methods. Their analyses suggested no detrimental employment impacts of the minimum wage.[17]

Subsequent evidence

The findings summarized in Card and Krueger (1995) were at odds with both previous evidence and the standard textbook model. Unsurprisingly, the findings stimulated reanalysis, for the US as well as other countries.

Cengiz, Dube, Lindner, and Zipperer (2019) examined the impact of 138 minimum wage changes in the US and concluded that the employment prospects of affected workers were unrelated to the minimum wage. Minimum wage changes led to a reduction of employment below the new minimum — implying that the minimum wage changes were binding — but this reduction was compensated by an equal-sized increase in employment just above the new minimum.

[17] The book by Card and Krueger (1995) also contained new analyses covering, for example, the impact of minimum wages on firm values, the robustness of previous time series and panel data evidence, and the international evidence on minimum wages as well as updates and extensions of their 1992 and 1994 papers.

Wolfson and Belman (2019) presented a meta-analysis of 37 minimum wage studies published since 2000. The (precision-weighted average) elasticity of employment with respect to the minimum wage is –0.024; an increase in the minimum wage by 10 percent thus reduces employment by 0.24 percent. The estimate of –0.024 is small but statistically significant; it is an order of magnitude smaller, in absolute value, than the consensus range of [–0.3, –0.1] based on the time series evidence reported by Brown, Gilroy, and Kohen (1982).

Another way of summarizing the evidence is to report the own-wage elasticity of employment. The own-wage elasticity takes the wage impact of the minimum wage increase into account; hence, this elasticity is easier to compare across studies based on different groups or countries. Dube (2019) summarized the evidence on the own-wage elasticity from 36 US studies. The median estimate across the 36 studies is –0.17. A (minimum wage induced) wage increase by 10 percent thus reduces employment by 1.7 percent. The subsequent evidence for the US thus suggests small employment effects.

It is not clear that the US evidence generalizes to other countries. The federal US minimum wage is low compared to the national minimum wages observed elsewhere.[18] This was true in the beginning of the 1990s (see Card and Krueger, 1995), and it is still true today (see Manning, 2021). Whether the minimum wage is raised from a baseline of 30 percent of median wages or from 70 percent of median wages most likely makes a difference. Nevertheless, the international evidence, summarized by Manning (2021) and Dube (2019), suggests relatively small employment effects, despite the fact that minimum wages are higher in other countries than in the US. The median of the estimated own-wage elasticities across 48 estimates drawn from various countries is –0.16 (see Dube, 2019).

Most of the evidence concerns the short-run impact of the minimum wage. Because of capital-labor substitution, however, the longer-run employment effects may be more negative than the short-run effects. Harasztosi and Lindner (2019) studied the impact of a large minimum wage increase in Hungary. They were able to track the employment effects 1–4 years after the minimum wage increase. The 4-year impact is only slightly larger in absolute size than the 1-year impact.[19]

Explanations

How should we understand the limited evidence of negative impacts on employment? Below we discuss some of the possible explanations put forward and the evidence supporting them.[20]

Labor costs. Compensation packages have more components than wages. In response to a minimum wage increase, firms may reduce the value of other parts of the package, which implies that total labor costs do not rise one-for-one with the minimum wage. Another reason why there is limited evidence of an employment decline is that firms face frictions in the form of hiring and training costs. An increase in the wage, likely reduces the outflow of workers from the firm, which in turn saves on turnover costs for the firm (see, e.g., Portugal and Cardoso, 2006, and Dube, Lester, and Reich, 2016, for evidence). Thus, total labor costs do not rise to the same extent as the minimum wage increase.

[18] Since most other countries have nationally determined minimum wages, researchers have to rely on comparisons across firms that are more or less affected by the minimum wage (as in Katz and Krueger, 1992), across individuals who are differentially affected, or across regions that are more or less affected (as in Card 1992b).

[19] The 4-year own-wage elasticity is –0.18, which is remarkably close to the –0.17 reported by Dube (2019) for the US. Harasztosi and Lindner (2019) also presented results suggesting capital-labor substitution: firms that are more affected by minimum wage hikes thus increase investments in capital. Therefore, capital-labor substitution contributes to the somewhat larger longer-run effect on employment.

[20] This discussion follows Manning (2021).

Productivity. Minimum wages may also raise productivity, since they make it more valuable for workers to hold on to their jobs after a minimum wage increase. Coviello, DeSerrano, and Persico (2020) showed that productivity among workers at a large US retailer improved, while profits and employment were unaffected, in response to minimum wage hikes. Productivity may also rise in the aggregate because of a reallocation effect. Dustmann et al. (2020) showed that the introduction of the minimum wage in Germany induced low-wage workers to move from small, low-paying, firms to larger, high-paying, firms.

Price responses. Minimum wages primarily "have a bite" in low-wage service sectors provided in local markets, such as the fast-food restaurants analyzed by Katz and Krueger (1992) and Card and Krueger (1994). A change in the minimum wage, whether at the state or national level, will affect all local service providers. And they may all raise their prices without much of a reduction in product demand. When the product demand elasticity is low, minimum wage increases can be shifted onto consumers without much loss in product demand and employment. Several papers document that prices respond to minimum wage changes (e.g., Aaronson, 2001; Renkin, Montialoux, and Siegenthaler, 2020).

Things are different in tradeable sectors of the economy. Here, firms facing a higher minimum wage compete with firms that do not. Such firms cannot raise prices without losing demand, and, thus, we should expect more negative effects of the minimum wage in the tradeable sector. Recent studies lend some support for this hypothesis (Harasztosi and Lindner, 2019, and Cengiz, Dube, Lindner, and Zipperer, 2019). Note, though, that few minimum wage workers work in the tradeable sector.

Imperfect competition in the labor market. In a model with search frictions, employers have some market power (see Burdett and Mortensen, 1998). Employers may use that market power to set wages lower than in a perfectly competitive market. Fewer workers are willing to work at this lower wage than in the competitive equilibrium. In such a monopsonistic setting, the employment impact of a minimum wage increase is *a priori* ambiguous. A marginal increase in the minimum wage can increase employment because of a positive labor supply response.

Broader research impact: Monopsony and firm wage setting

As discussed above, the results regarding the minimum wage are consistent with the view that firms have market power in the labor market. Such market power may come from employers being large relative to the local labor market, or from search frictions of the kind considered by Burdett and Mortensen (1998). An important implication is that firms' wage-setting policies will matter for wage dispersion, and hence for inequality in the labor market. The renewed interest in the monopsony model following the findings from Card and Krueger's papers from the early 1990s has stimulated an extensive research literature on the impact of firms' wage-setting behavior.[21]

Following Abowd, Kramarz, and Margolis (1999), several studies have used matched employer-employee data to decompose wages into a firm component and an individual component as a way of quantifying the importance of firms in explaining wage inequality. The general result is that 10–20 percent of the variance of earnings is attributable to stable firm effects. Card is an important contributor to this literature. Among other things, Card, Heining, and Kline (2013) showed that about a quarter of the increased wage inequality in Germany between 1985 and 2009

[21] The recent literature on the consequences of imperfect competition in labor (and product) markets is vibrant. An incomplete list includes papers by Azar, Marinescu, and Steinbaum (2020), Berger, Herkenhoff, and Mongey (2019), Kroft, Luo, Mogstad, and Setzler (2020), and Lamadon, Mogstad, and Setzler (2020). Azar et al. (2019) asked whether employment effects of minimum wages are more positive in more concentrated labor markets; the answer to this question is yes.

can be attributed to firms. Card, Cardoso, and Kline (2016) found that women are less likely to work at firms paying higher premiums overall, and that women receive only 90 percent of the firm-specific pay premiums earned by men.

Card has also made theoretical contributions to this field by proposing a micro-founded model of monopsony that has become something of a workhorse model in the very recent literature (Card, Cardoso, Heining, and Kline, 2018). The model, which uses insights from industrial organization, is based on workers having idiosyncratic preferences for different jobs.

What have we learned?

Research over the past 30 years has taught us that the employment impact of the minimum wage is not as negative as one would think based on a textbook competitive model of the labor market. This is the typical result in the most recent US studies as well as studies using data for other countries. The longer-run employment effect tends to be slightly larger in absolute value, but the most compelling study suggests it is still small.

Several empirically supported explanations have been offered for the finding that the employment effects are limited. One is that the cost per efficiency unit of labor does not rise one for one with minimum wages. Another is that local service providers, which are most affected by the minimum wage, are able to shift the cost increase onto consumers by raising prices, without much loss in product demand. A third explanation is that firms have monopsony power in the labor market; with monopsony power, employment effects are ambiguous because of countervailing impacts on labor demand and labor supply.

There has been a recent revival of research on the consequences of monopsony in the labor market and on the contributions of firms to (changes in) wage inequality. All in all, we have a much better understanding of how the minimum wage affects the labor market today than we had 30 years ago.

2.2. The labor market impact of immigration

Immigration is a hotly debated policy question in many countries. The concern is that a large inflow of migrants, i.e., a positive labor supply shock, may hamper the labor market opportunities of native workers through declining wage and employment prospects. Such concerns are moderated by a number of factors, however. First, the outcome for resident workers depends on whether their labor services substitute or complement the labor services of new immigrants. Second, we expect changes on the labor supply side to generate changes on the labor demand side. Thus, we expect firms to enter regions with a larger inflow of immigrants and to invest in technologies that are better tailored to the characteristics of immigrant workers. Third, immigrant inflows generate changes in the demand for goods and services, which might spill over to the labor market prospects of resident workers.

How immigration affects native workers residing in a certain area, and low-skilled residents in particular, is thus not clear *a priori*. Answering this question empirically is challenging because it is hard to tell what would have happened in an area had there not been an inflow of migrants. The issue is that migrants are likely to move to growing labor markets, and economic outcomes for natives in growing markets are different from other markets even if there is no immigration.

Early attempts at answering the question used variation in the number of immigrants across metropolitan areas in the US to estimate the parameters of a production function, aggregated to the local level (see Grossman, 1982, and Borjas, 1987). The general result was that the impact on natives was small, while the impact on the immigrants themselves was large. As pointed out by Borjas (1987), however, these estimates might be biased because the location choices of natives as well as immigrants are endogenous to economics prospects, as mentioned above.

Card's early contributions

Two studies by Card (Card, 1990, and Altonji and Card, 1991) reinvigorated the literature on the labor market impact of immigration. Both of these studies dealt (as best possible) with the confounding impact of immigrants moving to booming local labor markets.

Card (1990) utilized a unique event in US history, the so-called Mariel Boatlift. In late April 1980, Fidel Castro declared that Cubans wishing to emigrate to the US could leave from the port of Mariel. Between May and September 1980 around 125,000 individuals left Cuba, and 50 percent of them settled permanently in Miami. In just a few months, the Miami labor force increased by a staggering 7 percent. Card compares the evolution of wages and employment in the Miami labor market before and after the Mariel boatlift, to the evolution of employment in four comparison cities. Given an appropriate choice of comparison cities, the Mariel event provides a quintessential natural experiment.

Despite the massive inflow of new and unskilled migrants to Miami, Card (1990) found no evidence suggesting that the wage rates and unemployment of less skilled non-Cuban workers were affected. Card put forward two explanations for why wages and unemployment did not respond to the immigration event: first, evidence suggests that migration of natives and previous immigrants to Miami was reduced as a result of the Mariel episode; second, because of the history of immigration in the past, the Miami labor market had an industrial structure that was capable of absorbing a large inflow of unskilled migrants.

One can discuss whether the results from the Mariel Boatlift generalize to other events.[22] As Card noted himself, the fact that Miami had a history of receiving immigrants from Cuba likely influenced the results. In general, there is considerable history dependence in migrants' location choices. There is thus good reason that most of the Cuban emigres went to Miami rather than elsewhere in the country. Along these lines, Joseph Altonji and David Card (1991) noted that immigrants tend to settle with previous immigrants (Bartel, 1989), and thus used the previous settlement pattern as an instrument for the immigrant inflow. Examining the longer run (decadal) responses to changes in immigration between 1970 and 1980 across 120 US cities, their instrumental variable estimates suggested a rather substantial negative impact of immigrants on native wages; employment was unaffected, however.

The paper by Altonji and Card (1991) left two main legacies to the economics of immigration literature. First, it set the most commonly used conceptual framework — essentially a model of the demand side where the arrival of immigrants is modeled as a labor supply shock — for analyzing the effects of an increase in immigration. Second, the approach the authors took to estimate the effects of immigration has been applied repeatedly in the literature. This "shift-share" approach was further refined by Card (2001a), where he used the previous settlement pattern of each ethnic group to generate a prediction of the overall immigrant inflow by city and occupation group. In general, such shift-share approaches to generating instruments are very common in applied microeconomic studies and the approach has been applied repeatedly in the economics of immigration.[23]

[22] With 30 years of hindsight, the samples provided by the analysis data set (the Current Population Surveys) seem to have been too small to allow precise conclusions; see the interchange between Borjas (2017) and Peri and Yasenov (2019). The latest reanalysis of the Mariel Boatlift by Peri and Yasenov (2019) reproduced Card's original results, however.

[23] They are usually referred to as Bartik (1991) instruments, although their intellectual legacy can probably be traced back to Freeman (1980). The recent literature on when these shift-share instruments have all the properties of a valid instrument includes Adao, Kolesar, and Morales (2019), Borusyak, Hull, and Javarel (2020), and Goldsmith-Pinkman, Sorkin, and Swift (2020). Jaeger, Ruist, and Stuhler (2018) critically discussed the use of shift-share instruments in the immigration context.

Subsequent evidence

The results in Card (1990) were stark and thus triggered reanalyses. Following Card (1990), several studies examined the effects of large immigration episodes for other countries.[24] These studies generally concluded that the effects for the average native worker were negligible. Dustmann, Schönberg, and Stuhler (2017) analyzed the consequences of a commuting policy that led to a sharp and unexpected inflow of Czech workers to areas along the German-Czech border. In contrast to many earlier studies (particularly those using US data), they were able to follow individuals over time. They found no displacement of native workers who resided in these areas and a small negative impact on native wages. Yet local native employment declined, through diminished inflow of natives into the affected regions, a pattern that is similar to the one observed in Card (1990).

Other studies examine the impact of more normal immigration events. These studies typically compare outcomes across skill groups that are differentially affected by an inflow of immigrants. Hence, they focus on distributional effects of immigration, rather than the general question of how immigration affects wages and employment at large. Evidence reported by Borjas (2003) and Card (2009) points to negative wage effects for the skill groups mostly affected by immigration.[25]

Explanations and broader research impact

The surprising results in Card (1990) also led to a body of literature trying to explain why (or why not) immigration affects the labor market outcomes of native workers. Here we briefly review this research.

Do immigrant inflows lead to native outflows? Native migration responses are interesting in their own right. But they can also be a source of bias for estimates of wage and employment outcomes built on comparing regions that are differentially affected by immigration. If native migration responses are large and primarily found among negatively impacted native workers, the upward bias can be large (Borjas, Freeman, and Katz, 1996). This issue has been addressed in several papers (e.g., Card and Di Nardo, 2000; Card, 2009; and Peri and Sparber, 2011). The conclusion is that native responses are rather limited, and too small to yield substantial bias.

To what extent are different groups substitutable for one another? This is a key question in determining the overall and distributional impact of immigration. Ottaviano and Peri (2012) and Manacorda, Manning, and Wadsworth (2012) used time-series variation to ask whether immigrant and native labor seem to be substitutable within education and experience cells. They concluded that immigrant and native labor are not perfect substitutes. The estimates in Ottaviano and Peri (2012) in fact suggest that the average US native worker would experience a small gain, while previous immigrants would suffer a rather substantial wage loss. Cortes (2008) also found that those most affected by an inflow of low-skilled immigrants were previous low-skilled immigrants; the effect on low-skilled native workers was much smaller (see also Lalonde and Topel, 1991).

Peri and Sparber (2009) developed a framework for explaining why low-skilled native and immigrant labor are imperfect substitutes. The essence of the argument is that natives and

[24] This literature includes papers by Hunt (1992), Friedberg (2001), Angrist and Kugler (2003), and Glitz (2012). Borjas and Monras (2017) revisited some of these studies. Their analysis suggested substantial heterogeneity in the impact on natives; they conclude that natives with similar skills were adversely affected while there were positive effects for complementary native workers.

[25] Dustmann, Schönberg, and Stuhler (2016) surveyed the literature and discussed why the estimates vary substantially across studies. The estimates in Borjas (2003) and Card (2009) are relatively close to one another, however. According to Borjas (2003), an immigration-driven increase in relative supply by 1 percent reduces relative wages by 0.57 percent. According to Card (2009), the corresponding change would yield a reduction of relative wages by 0.42 percent.

immigrants possess different skills, in particular language skills, and therefore, are sorted into different kinds of occupations: native workers into occupations requiring communication skills, and immigrant workers into occupations requiring manual skills. They also presented evidence consistent with the theory: in local labor markets experiencing a larger (predicted) inflow of low-skilled immigrants, more low-skilled natives are found in occupations requiring communication skills. Foged and Peri (2015) used data for Denmark to show that low-educated natives move to more complex and less manual occupations in response to an influx of immigrants. In part due to this adjustment, the wage and employment effects of immigration are zero or positive.

Dustmann, Frattini, and Preston (2013) estimated the impact of immigration across the entire native worker wage distribution. They documented striking differences across the distribution in the UK: native workers in the bottom decile of the wage distribution are negatively affected, while all other deciles are positively affected. As a result, the average UK native gains from immigration.

To what extent is there directed technical change? One potential effect of immigration is adjustment on the labor demand side as a consequence of the changes in the structure of labor supply induced by immigration. Lewis (2011) found evidence of such adjustment; in particular, he showed that there was less investment in automation machinery in areas that saw greater increases in low-skilled immigration. Similarly, Peri (2012) reported evidence that immigration increased total factor productivity and reduced the skill bias of production technologies in the long run. Finally, Dustmann and Glitz (2015) showed that firms start using the skill group that has become more abundant as a result of immigration, without a concomitant wage adjustment. This result is consistent with demand-side adjustments, e.g., investments in technology, in response to a change in supply.

What have we learned?

What have we learned from 30 years of research following Card (1990)? There is still an ongoing discussion about the magnitude of the wage and employment effect for the natives who are most likely substitutes with new immigrants. Nevertheless, the literature has converged on a number of conclusions. First, the labor market prospects of prior immigrants are most detrimentally affected; many native groups, in fact, appear to benefit from new immigrants. Second, native workers seem to avoid the negative consequences by entering occupations requiring communication skills where there is less competition from immigrants. Third, technological investments adjust to immigrant inflows, which also reduces the detrimental impact on the groups most affected by immigration.

2.3. The effects of investments in education

One potential route to improve labor market outcomes, particularly for the disadvantaged, is via education. Investing in school resources, to provide better educational opportunities, plausibly improves school achievement and labor market outcomes in the longer run. Yet the empirical literature on the relationship between achievement and resources prior to 1990 suggested that this association was weak. This result is perhaps best exemplified by the Coleman (1966) report. The report (among other things) used regression to examine the fraction of the variance in achievement that could be accounted for by variation in school inputs. There was little association between school inputs and achievement, however: beyond the importance of family background, schools added little to the explained variance. Hanushek (1986, 2003) surveyed the large literature that followed after the Coleman report, concluding that the relationship between achievement and resources was close to nonexistent.[26]

[26] This conclusion has been challenged. A formal meta-analysis of the studies surveyed by Hanushek suggests that the association between achievement and resources is positive; see Hedges, Laine, and Greenwald (1994).

The previous literature should arguably be regarded as descriptive. Answering questions about the causal impact of school resources requires an empirical strategy that can deal with a variety of confounding variation in student, school, or community characteristics. For example, there is typically compensatory resource allocation such that low-performing pupils are placed in small classes. If such compensatory allocation is substantial, the correlation between student achievement and resources can be negative, even though the true causal impact is positive.

Card's early contributions

Card and Krueger co-authored two papers in the beginning of the 1990s, where they examined the importance of school quality in general for labor market outcomes. Both papers use plausibly exogenous variation in school quality, coming from a period of large investments in school resources (during the 1930s to 1950s), particularly in the American South.[27]

In contrast to much of the previous literature, Card and Krueger analyzed how school quality affects labor market outcomes rather than test scores. This is a significant innovation since school quality may affect many more of the abilities that make individuals succeed in the labor market, than those narrowly measured by test scores. They also broke out of the mold of the existing literature by clearly spelling out their empirical research design.

Card and Krueger (1992a) asked whether school quality affects the earnings return to a year of schooling. The idea is that better schools improve labor market returns for given educational attainment. School quality and labor market returns are measured at the state level. Variation in school resources across states may reflect differences in population and labor market characteristics, however; thus, there is need for a strategy that can separate between these effects. Card and Krueger (1992a) tackled this problem by focusing on interstate movers, i.e., by comparing individuals residing in the same state, but having grown up in different states.[28] Hence, they asked whether the earnings return to schooling is higher for individuals having grown up in states where educational spending is higher than elsewhere. To illustrate the main idea, take a set of individuals (born at the same time) who either grew up in Alabama and Iowa, but reside in California at the time of observation. They have all elected to move to California and are in that sense similar. The Alabama/Iowa comparison then potentially gets at the difference in the causal return to schooling associated with growing up in Alabama relative to Iowa.[29]

In practice, Card and Kreuger (1992a) implemented this "movers design" by estimating the earnings return to schooling separately by state of residence, state of birth, and birth cohort, and then teasing out the return to schooling associated with growing up in a particular state for a given cohort. In a second step, they examined whether the variation in these returns across birth states (b) and cohorts (c) is related to different measures of school quality. Using data for white men born in the 1920s, 1930s, and 1940s, they thus ran the fixed effects regression

---

[27] Today, there is a large experimental literature on the educational production function in low-income countries (KVA, 2019). Randomized experiments on school inputs and policies are rarer in higher-income settings, partly because of institutional constraints and partly because the costs of implementing such trials are high. There are exceptions, however. The most well-known example is the so-called STAR experiment, where students were randomly assigned to small and normal-sized classes; see Finn and Achilles (1990). Data from STAR have subsequently been used in many research papers; see Krueger (1999), Krueger and Whitmore (2001), and Chetty et al. (2011). These studies typically find that being assigned to a smaller class in early grades has a beneficial impact on educational outcomes in both the short run and in the long run. There is also a quasi-experimental literature on class size, e.g., Angrist and Lavy (1999) and Fredriksson, Öckert, and Oosterbeek (2013) that has reached similar conclusions.

[28] The empirical strategy applied by Card and Krueger (1992a) has been followed by others in the literature; Carneiro and Lee (2009) and Chetty and Hendren (2018) are two relatively recent examples.

[29] This "movers design" realistically allows for a correlation between the unobserved determinants of earnings and years of schooling. If we want to interpret the estimate of the returns to schooling by birth cohort and state as causal, the crucial assumption is that this correlation is the same for all individuals moving to a particular state (California) from different states of birth (Alabama or Iowa).

$$\text{Returns}_{bc} = \lambda_b + \lambda_c + \beta(\text{School Quality})_{bc} + \varepsilon_{bc} .$$

Card and Krueger considered three measures of school quality: the pupil-teacher ratio, teacher relative wages, and the length of the school term. They found that individuals having grown up in states that reduced the pupil-teacher ratio, for example, have systematically higher returns to schooling than those growing up in areas with constant or reduced pupil-teacher ratios. Thus, their conclusion is that school quality did matter for labor market outcomes.[30]

There were particularly large reductions in the pupil-teacher ratios in the US South: school quality was very different for black students than for white students within a state in the beginning of the 20th century. The situation then improved for black students, especially between 1920 and 1950 and for the states in the South. Card and Krueger (1992b) asked whether this improvement of relative school quality affected the relative returns to education earned by black and white men. Using information on individuals who had moved to states in the North, they show that returns to education do indeed increase with school quality. Since black men experienced a larger improvement of school quality, changes in school quality contributed to narrowing the black-white earnings gap between 1960 and 1980.

Additional evidence on resources

The results in the two papers by Card and Kruger surprised the research community and contributed to renewed interest in educational production. The results led to a discussion on whether school quality and school resources mattered for school and labor market outcomes. The different perspectives on this issue are summarized in the book *Does Money Matter?* (Burtless, 1996).

Later work, based on quasi-experimental research designs, has found evidence in line with the initial studies of Card and Krueger. The US evidence is based on variation coming from court-mandated school finance reforms; Hoxby (2001) and Card and Payne (2002) are two early studies. Jackson, Johnson, and Persico (2016) documented substantial effects on educational attainment and wages;[31] moreover, the effects were much larger for children from low-income families. Similarly, Rothstein and Schanzenbach (2021) found evidence of increasing returns to schooling, particularly for black students.

There is also evidence from broad compulsory schooling reforms. In practice, these reforms implied that resources were targeted to students from disadvantaged family backgrounds, by improving access to education and raising school quality mainly for them. Evidence from the Nordic countries shows that educational attainment and earnings increased for individuals with low-educated parents (see Meghir and Palme, 2005, Kerr, Pekkarinen, and Uusitalo, 2009, Aakvik, Salvanes, and Vaage, 2010).

Broader research impact: Interpreting the evidence on the returns to schooling

The rate of return to schooling is a key parameter for evaluating educational expansions in general.[32] Here the landmark study is arguably the paper by Angrist and Krueger (1991) described

[30] Heckman, Layne-Ferrar, and Todd (1996) replicated and extended Card and Krueger's (1992a) analysis. Their reanalysis suggested that school quality matters for the returns to schooling, but mainly at the higher levels (some college or more).

[31] They found that a 10-percent increase in per pupil spending throughout elementary and high-school improved educational attainment by 0.31 years and raised wages by 7 percent.

[32] Note, however, that the individual and the social returns to education are typically not the same; for instance, taxes and transfers drive a wedge between the individual and social return to schooling.

in Section 1. Another early contribution is Card (1995a), who used the existence of a nearby college when the individual grew up as an instrument for educational attainment.

Card (1995b, 1999, 2001b) summarized the literature on the returns to schooling. A striking feature of the early quasi-experimental literature is that the IV estimates tend to be larger than the corresponding OLS estimates using data from the same context.

More importantly, perhaps, Card also interpreted the literature through the lens of a Becker (1967) model of optimal schooling choices. According to this model, optimal investment choices equate the return to schooling to the cost of borrowing (discount rates). Children from families who face greater costs of borrowing choose fewer years of schooling, while children who are more able, and hence have a higher return to schooling, choose to invest in more human capital. This model implies that the returns to schooling vary across the schooling distribution. Under the (strong) assumption that ability and discount rates are uncorrelated, the marginal return to schooling is higher for individuals at the bottom of the distribution since their choices are constrained by their high discount rates.[33]

The model discussed by Card thus features heterogeneity in the returns to schooling. Following Imbens and Angrist (1994) and Angrist and Imbens (1995), Card noted that instrumental variables identify the average effect of schooling for groups who were mainly affected by the instrument; moreover, many of the instruments that have been used in the literature tend to influence individuals towards the lower end of the schooling distribution. One example is the quarter-of-birth instrument used by Angrist and Krueger (1991), which mainly influenced those with a high probability of leaving school. Another example is the distance instrument used by Card (1995a). This instrument is probably less relevant for children from well-off families, since they would go to college no matter where they grew up.

When a particular natural experiment mainly affects individuals who are likely to have low levels of schooling, and the main reason for this is that they face higher than average discount rates, IV estimates of the returns to schooling are larger than the average return in the population. In such a scenario, policy initiatives to increase the educational attainment among individuals from disadvantaged backgrounds can have a substantial wage return.

What have we learned?

The overall conclusion from the past 30 years of research is that school resources appear to matter for labor market outcomes in industrialized countries.[34] The recent survey of the US evidence by Jackson (2020) presented strong support for the conclusion that educational resources have an impact on average. Jackson carefully pointed out that this does not imply that all types of spending increases, in all types of contexts, yield improvements in achievement and labor market outcomes. On average, however, school finance reforms tend to reduce student-teacher ratios, increase teacher salaries and prolong school years (Jackson, Johnson, and Persico, 2016). These expenditure components are most likely to improve school quality and the earnings return to schooling (see Card and Krueger, 1992a).

The impact of resources on school achievement tends to be greater for non-advantaged students, suggesting that their schooling choices are constrained to a greater extent than for students from advantaged backgrounds. Whether increases in school spending reduces wage and earnings inequality is a more complicated matter. Nevertheless, the quasi-experimental literature is

[33] More generally, Card concluded that if the variation in optimal schooling choices is mainly driven by variation in discount rates (or credit constraints) rather than ability, the marginal return to schooling is higher towards the lower end of the schooling distribution.

[34] By contrast, the major concern in developing countries does not appear to be educational resources, but how teaching is performed and at what level; see KVA (2019).

consistent with the view that earnings effects of investments in education are higher for individuals from disadvantaged backgrounds.

3. IDENTIFYING CAUSAL TREATMENT EFFECTS USING INSTRUMENTAL VARIABLES

The initial wave of studies exploiting natural experiments raised new and conceptually important issues. For example, as just described, IV estimates of the returns to schooling were typically larger than the corresponding OLS estimates. To rationalize such findings, it is natural to turn to a framework where the return to schooling is heterogeneous in the population; in such a setting, the source of the variation used in estimation matters.

The study by Angrist (1990) also pointed to the importance of paying attention to heterogeneity. Angrist leveraged the Vietnam-era draft lottery to estimate the effect of military service on earnings later in life. In the beginning of the 1970s, draft eligibility was determined by a randomized draw of birth dates.[35] Angrist thus used eligibility for the draft as an instrument for doing service in the Vietnam War. However, most of the individuals who served in Vietnam were volunteers who would have served no matter their lottery number. Thus, the draft lottery only affected individuals who would not have served voluntarily in the military. Consequently, Angrist's estimates are probably not representative of those who volunteered for service in the Vietnam War.

In most situations, responses are likely *heterogeneous*. When treatment effects vary across individuals and people make choices, there is likely to be *incomplete compliance* with the (natural) experiment. The fact that individuals volunteered during the Vietnam war is an example of incomplete compliance because such individuals signed up for service no matter whether they were draft eligible or not. Conversely, some of those who were draft eligible were exempted for health reasons or because they were still in school. In general, incomplete compliance to quasi-experimental and experimental variation is ubiquitous.

The combination of treatment heterogeneity and incomplete compliance poses a problem for causal analysis. Prior to the work of Joshua Angrist and Guido Imbens, researchers explored the conditions allowing identification of the average treatment effect in the overall population or in the treated population (Chamberlain 1986; Heckman 1990); the general finding is that these conditions tend to be stringent. An alternative approach is to bound these effects (Robins 1989; Manski 1990; Balke and Pearl 1997); unfortunately, such bounds are often too wide to be practically informative.

To illustrate that identification of the average treatment effect among the treated requires stringent assumptions, consider a scenario where treatment eligibility is randomized and has no direct impact on the outcome of interest. Because of incomplete compliance, however, treatment status and treatment eligibility are not the same. In general, treatment status may depend on individual treatment effects, and, therefore, additional assumptions have to be invoked.[36]

One assumption that allows identification of the average causal effect for the treated is that individuals do not know how they will be affected by the intervention of interest. In such a situation, the decision to take part in the treatment cannot be based on individual returns to participation. Therefore, it is possible to estimate the average causal effect for the treated with a straightforward application of instrumental variables (e.g., Heckman, 1997). In many real-world

[35] In televised draws, each birth date in draft-eligible cohorts was assigned a number between 1 and 365, and all men with a lottery number below a certain ceiling were eligible for drafting.

[36] For example, Garen (1984) allowed for heterogeneous returns to schooling using a random-coefficients framework. To identify the average return to schooling in the population, he assumed that the random coefficients were normally distributed.

settings, however, individuals are at least partially aware of the consequences of the intervention and then naturally respond to that fact.

Another possibility is that noncompliance is one-sided, i.e., that the probability of participating is zero for individuals who are not eligible for treatment. If this holds true for the entire population under study, then one can estimate the average causal effect among the treated; see Bloom (1984).[37] If there is one-sided noncompliance in an observable subset of the population, then one can estimate the average effect for the treated in that particular group. In many contexts, however, one-sided noncompliance is likely violated.[38]

Imbens and Angrist (1994) approached the difficulties associated with heterogeneity and noncompliance in a different way. Specifically, they took a step back and asked: what can and cannot be learned from a randomized or natural experiment without placing additional restrictions on the behavior of study objects whose behavior one cannot directly observe?

In the remainder of this section, we describe Angrist and Imbens' main results and the framework they used to derive them. This framework integrates instrumental variables analysis into the potential outcome framework for causal inference. The resulting framework makes the nature of the identifying assumptions transparent and allows researchers to assess the sensitivity of their empirical designs to deviations from these assumptions. In the sequel we illustrate this claim. We also discuss implications for intention-to-treat (ITT) analysis, which is the norm for analyzing experiment with incomplete compliance in medicine.

This section focuses on a particular methodological innovation by Angrist and Imbens, but their contributions of course extend beyond the studies covered here. Angrist is an applied labor economist who has worked in many fields, but perhaps most extensively on the economics of education.[39] In addition to the work already mentioned, he has, for example, examined how student performance is affected by: class size (Angrist and Lavy, 1999); attending charter schools (Abdulkadiroğlu et al., 2011); and private school vouchers (Angrist et al., 2002). Imbens is an econometrician whose other work mainly concerns evaluation methods. His work on matching methods (e.g., Abadie and Imbens, 2006; Hirano, Imbens, and Ridder, 2003), the regression discontinuity estimator (e.g., Imbens and Kalyanaraman, 2012), and refinements of the difference-in-differences design (Athey and Imbens, 2006) are examples.

3.1. A framework for identification and interpretation of treatment effects

The methodological framework laid out by Imbens and Angrist (1994) built on the potential outcome framework introduced by Neyman (1923), in the context of randomized experiments, and by Rubin (1974), in the context of observational data. At the core of such a framework is the assignment mechanism, i.e., the process that determines which units receive which treatments, and thus which potential outcomes are realized and can be observed, and which potential outcomes are missing. Unlike Neyman, who assumed random assignment or Rubin, who linked the assignment to the propensity score, Angrist and Imbens linked the assignment to the existence of an instrument.[40] In this way, the new approach merged the IV framework invented in economics

[37] Bloom (1984) showed that the average causal effect among the treated can be estimated by IV, since it equals the ratio of the ITT effect and the compliance rate. This works because the treated population coincides with the complier population when there is one-sided noncompliance. In this scenario, the average causal effect among the treated is thus equal to the local average treatment effect (see below).

[38] A related approach is to assume that some observed covariate has unlimited support. Then the probability of participation tends to zero as the values of the observed covariate tends to infinity, implying that it is possible to identify the average causal effect among the treated using "identification at infinity"; see Chamberlain (1986) and Heckman (1990).

[39] We should also mention the Angrist and Pischke (2008) textbook that has completely changed how econometrics is taught in graduate schools.

[40] The propensity score is the probability of being treated given observed covariates.

with the potential outcomes framework for causal inference developed in statistics.[41] The instrument could be the result of physical randomization, as in RCTs; alternatively, the source of exogenous variation could be a natural experiment. Angrist and Imbens thus provided a general framework applicable to both quasi-experimental and experimental work.

Below we outline the framework developed by Angrist and Imbens.[42] Suppose we are interested in a binary treatment. For concreteness, take the example of the effect of completing high school on earnings from Section 1. Remember that Y_i denotes the earnings of individual i, while $D_i = 1$ for those who completed high school and $D_i = 0$ for those who did not. There are two potential outcomes: $Y_i(1)$ and $Y_i(0)$. The causal effect for a single individual can be defined as the difference between the two potential outcomes, $Y_i(1) - Y_i(0)$. As pointed out in Section 1, one cannot be treated and untreated at the same time: the individual causal effect is not identified.

The target of estimation (the "estimand") is typically an average of the individual causal effects. The most ambitious of these targets is the average treatment effect (ATE) in the entire population under consideration, $ATE = E\big(Y_i(1) - Y_i(0)\big)$. But this average treatment effect is only identified under stringent assumptions when there is heterogeneity and incomplete compliance.

As emphasized in Section 1, we need an instrument to estimate the effect of completing high school on earnings. Let us thus elaborate on the birth date example used by Angrist and Krueger (1991). Recall that date of birth is relevant for high school completion in the US context because schools start once a year, say the year when a child turns six, but you are allowed to leave school at a certain age, say the day when a student turns 18.[43] Now, suppose we compare two individuals born a day apart — one on December 31 and the other on January 1. Both children start school on the same date, say September 1. The December-born child is then 5 years and 8 months old at school start, while the January-born child is 6 years and 8 months old. Since they are allowed to drop out at age 18, only the January-born child could do so prior to high school completion. Random variation in birth dates thus yield differences in educational attainment for institutional reasons that are plausibly unrelated to other determinants of schooling. Therefore, one could potentially use an indicator for being born before the school entry cutoff as an instrument. Denote the instrumental variable by Z, where $Z = 1$ for individuals born July–December, and $Z = 0$ for individuals born January–June. We expect those born July–December to have a higher probability of completing high-school than those born January–June.

Since the treatment indicator (completing high school or not) is endogenous, it is useful to think of it in terms of potential outcomes. With a binary instrument Z_i, we thus have two potential treatments, $D_i(0)$ and $D_i(1)$; without further assumptions, there are now four potential outcomes $Y_i(z, d)$, where $z = 0,1$, $d = 0,1$.

A minimum set of key assumptions

Angrist and Imbens started their analysis by asking: What assumptions should a valid instrument satisfy? First of all, the instrument should be as good as randomly assigned. The instrument should thus be unrelated to all potential outcomes, formally:

Assumption 1 (random assignment): $[\{Y_i(z, d) \; \forall \; d, z\}, D_i(1), D_i(0)] \perp Z_i$.

---

[41] An alternative conceptual model for causal inference in statistics is the structural causal models (SCM); see Pearl (2000). Pearl (2009) discusses the relationship between the SCM and the potential-outcomes framework.

[42] Our presentation of the results in Imbens and Angrist (1994) follows Angrist, Imbens, and Rubin (1996). Throughout the presentation, we assume that the potential outcomes for each individual i is unrelated to the treatment status of all other individuals (Cox, 1958, and Rubin, 1980). This assumption precludes general equilibrium effects, for example.

[43] Currently, 16 US states have a drop-out age of 18, 9 states a drop-out age of 17, and the remainder (the majority) has a drop-out age of 16.

In our running example, this corresponds to assuming that the potential outcomes are independent of the dates on which individuals are born.

If the instrument is randomly assigned, it is possible to consistently estimate the so-called reduced form, i.e., the relationship between the outcome of interest and the instrument, which also corresponds to the estimate of the ITT effect in an RCT. In other words, the regression of Y on Z identifies the causal effect of the instrument on the outcome:

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

where the last step follows from random assignment. In our example, $E\big[Y_i\big(1, D_i(1)\big)\big] - E\big[Y_i\big(0, D_i(0)\big)\big]$ is the causal effect of being born late rather than early in a given year on earnings.

Analogously, the relationship between the treatment of interest and the instrument — the first stage — is given by

$$E[D_i|Z_i = 1] - E[D_i|Z_i = 0] = E[D_i(1)|Z_i = 1] - E[D_i(0)|Z_i = 0] = E[D_i(1)] - E[D_i(0)] \quad (2)$$

Thus, $E[D_i(1)] - E[D_i(0)]$ is the causal effect of being born late on the likelihood of completing high school.

For the instrument to be useful, it must also be relevant, formally expressed as:

Assumption 2 (relevance): $E[D_i(1)] - E[D_i(0)] \neq 0$.

In the birth date example, this requires that there is a stronger response to the possibility of dropping out of high school among those who are born early in the year compared to those born late in the year.

As Z is not equal to D, the identification of the effect of the treatment (high school completion) on the outcome (earnings) requires additional assumptions. Specifically, it requires that the only way the instrument affects the outcome is via the treatment. This is the exclusion restriction, formally:

Assumption 3 (exclusion): $Y_i(1, d) = Y_i(0, d) = Y_i(d), d = 0,1$.

The combination of assumptions 1 and 3 implies that the instrument is "exogenous" to the error term in a conventional linear regression. By framing instrumental variables analysis in the potential-outcomes framework, it becomes clear that there are two distinct assumptions going into the exogeneity assumption, namely randomization and exclusion; Angrist, Imbens, and Rubin (1996) made this distinction particularly clear. This point is conceptually important because these are two separate assumptions that deserve separate attention, and violations of them can have different sources and consequences. Consider our concrete example. Suppose you focus attention on individuals born close to the school entry cutoff, say individuals born in the last week of December and individuals born in the first week of January. Since it is essentially random whether you are born one or two weeks apart, the instrument is randomly assigned. But is it excludable? This is not obvious, as the instrument also generates variation in the school starting age. If the school starting age matters for earnings, the exclusion restriction would be violated.

Under the exclusion restriction, the reduced-form/ITT coefficient simplifies to

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = E[(D_i(1) - D_i(0))(Y_i(1) - Y_i(0))]$$

The right-hand side of this expression can be further decomposed by noting that the difference $(D_i(1) - D_i(0))$ can only attain three distinct values: -1, 0, 1. We can thus write

$$\begin{aligned} &E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] \\ &\quad = \sum_{k=\{-1,0,1\}} k\Pr\big[\big(D_i(1) - D_i(0)\big) = k\big]E\big[\big(Y_i(1) - Y_i(0)\big)\big|\big(D_i(1) - D_i(0)\big) = k\big] \\ &\quad = \Pr\big[\big(D_i(1) - D_i(0)\big) = 1\big]E\big[\big(Y_i(1) - Y_i(0)\big)\big|\big(D_i(1) - D_i(0)\big) = 1\big] \\ &\quad - \Pr\big[\big(D_i(1) - D_i(0)\big) = -1\big]E\big[\big(Y_i(1) - Y_i(0)\big)\big|\big(D_i(1) - D_i(0)\big) = -1\big] \end{aligned}$$

This expression makes it immediately clear that individuals who do not respond to the instrument — the individuals for whom $D_i(1) - D_i(0) = 0$ — do not contribute to identification. Intuitively, we cannot estimate a causal effect for the population that did not change behavior in response to the instrument, since there is no control group for this subset of the population. Although straightforward, this is an immensely important insight.

As it stands, interpreting the reduced form relationship is problematic. In fact, the reduced form effect (which is proportional to the IV estimate under assumption 3) can be negative, even though all the underlying individual causal effects of D on Y are positive. The problem is that the treatment effect for those who shift from nonparticipation to participation when Z is switched from 0 to 1 (the "compliers") can be cancelled out by the treatment effect of those who shift from participation to nonparticipation (the "defiers").

How can we make progress? One approach is to assume that there is no treatment heterogeneity in the population, but this is of course the same as assuming that the problem does not exist. A less restrictive approach is to assume "monotonicity". Monotonicity — introduced to IV analysis by Imbens and Angrist (1994) — is the assumption that all individuals are affected in the same direction, or not at all, by the instrument.[44]

Assumption 4 (Monotonicity): $D_i(1) \geq D_i(0)$ (or vice versa).

Focusing on $D_i(1) \geq D_i(0)$, this implies $\Pr\big[\big(D_i(1) - D_i(0)\big) = -1\big] = 0$ and that the causal effect of the instrument on the treatment equals $E\big(D_i(1) - D_i(0)\big) = \Pr\big[\big(D_i(1) - D_i(0)\big) = 1\big]$.

Identification

The monotonicity assumption, combined with the other three assumptions, implies that the ITT effect equals

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = \Pr\big[\big(D_i(1) - D_i(0)\big) = 1\big]E\big[\big(Y_i(1) - Y_i(0)\big)\big|\big(D_i(1) - D_i(0)\big) = 1\big]$$

The left-hand side is the causal effect of the instrument on the outcome of interest. The first component of the right-hand side is the causal effect of the instrument on the treatment. The ratio of the two is the IV estimand.

Therefore, under assumptions 1-4, the IV estimand

[44] Robins (1989) used assumptions 1, 3, and 4 (as well as additional assumptions) to bound the average treatment effect in the population.

$$\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = E\big[\big(Y_i(1) - Y_i(0)\big)\big|\big(D_i(1) - D_i(0)\big) = 1\big]$$

identifies the average causal effect for the population that changed their treatment status in accordance with the change in the value of the instrument. Angrist and Imbens labeled this causal effect, the local average treatment effect (LATE). Angrist, Imbens, and Rubin (1996) referred to the population of individuals who changed their behavior in accordance with the value of instrument as compliers; therefore, LATE is sometimes referred to as the complier average causal effect (CACE).

In our running example, the complier population corresponds to the subset where different dates of birth relative to the school entry cutoff, affected high school completion. The population that would always complete high school, no matter when they were born, are referred to as "always-takers"; the population that would always drop out are "never-takers," and the population that would drop out because they were born just before the school entry cutoff are "defiers". The monotonicity assumption rules out the existence of defiers.

3.2. Extensions and generalizability

The work of Angrist and Imbens spans more general cases than the special case considered in Section 3.1. We begin this section by reviewing some of these extensions. We then turn to a discussion of what can be done to investigate whether estimates of LATE generalize to a broader population than the compliant subpopulation.

Extensions

Imbens and Angrist (1994) not only considered binary instruments. They in fact showed that when the instruments are multi-valued and discrete, the IV estimand is a weighted average of local average treatment effects. The weights are all positive and sum to one. They also showed that the two-stage least square (2SLS) estimator — the most efficient (one-step) IV estimator — consistently estimates the weighted-average LATE, given that the instruments have a monotonic effect on the probability to be treated. The 2SLS estimator is asymptotically normal, with a variance that can be estimated using a robust estimator.

Angrist and Imbens (1995) made the important generalization to the case where the treatment is multi-valued, say years of schooling. This adds a layer of complication because for every given individual, there are now many causal effects — say going from 11 to 12 years of schooling, 12 to 13, etc. In fact, there are conceptually as many potential outcomes as there are values of the treatment variable.

Suppose we have a single binary instrument at our disposal and that we are interested in estimating the returns to schooling. Angrist and Imbens (1995) showed that the IV estimand is a weighted average of unit causal responses. The unit causal response is the difference in potential outcomes among compliers at a particular point in the schooling distribution. The weights reflect to what extent individuals are affected by the instrument at different points in the schooling distribution; they are all positive and sum to one.

DiNardo and Lee (2011) showed that one can relax the monotonicity assumption to "probabilistic monotonicity" and still identify a well-defined causal treatment effect. Probabilistic monotonicity requires the probability of treatment to increase for all individuals randomized to treatment. The IV estimand is then interpretable as a weighted average of individual causal effects, where the weights reflect the effect of the instrument on the probability of treatment for each (type of) individual. This extension is useful because it implies that IV estimates may generalize beyond the population of compliers.

Finally, Imbens and Rubin (1997a) presented a framework for conducting sensitivity analyses with respect to the underlying assumptions. In particular, they maintained the randomization and relevance assumptions and asked what happens to causal inference when the exclusion restriction and the monotonicity assumption do not hold. To do the analysis, they had to invoke alternative assumptions.[45] Such sensitivity analysis is primarily interesting for the exclusion restriction since this restriction is a concern in many applications.

The generalizability of LATE estimates

The local average treatment effect is what can be identified under the minimum set of assumptions. As such, estimates of LATE are internally valid for an unobserved population complying with the instrument.[46] In the absence of one-sided noncompliance, in which case the populations of treated and compliers are identical, this means that the effect for one group of compliers (defined by one instrument) need not generalize to another complier population (defined by another instrument) in an analysis of the same data. A number of analyses can be done to shed light on the question of generalizability from a specific study.

First, even though the identity of compliers is unobserved, it is straightforward to estimate the fraction of compliers in the data (this is the first-stage coefficient — see equation 2) and the fraction of compliers relative to the number of treated; it is also possible to describe the complier population in terms of observed covariates (Abadie, 2003; Angrist, 2004). Such descriptions shed light on the question of how special compliers are relative to the population one wishes to conduct inference to; if the policy-target is low-income individuals, for example, and low-income individuals are overrepresented among compliers, then LATE should be highly relevant for policy purposes.

Second, Imbens and Rubin (1997b) showed that under assumptions 1-4, the potential outcome distributions among compliers are identified; one can thus provide a sense of effect heterogeneity among compliers.[47] Third, with access to more than one instrument one can test the extent of heterogeneous responses to the instrument (Angrist, Lavy, and Schlosser, 2010). If heterogeneity is not substantial, the estimates may generalize to a broader population.

A fourth option is that of filling in the gaps in the data using auxiliary assumptions. For example, Heckman, Tobias, and Vytlacil (2001, 2003) as well as Angrist (2004) used parametric latent-index models to identify and compare different causal effects, such as the average causal effect in the population, the average causal effect among the treated, and the average causal effect among compliers. Brinch, Mogstad, and Wiswall (2017) imposed structure on the marginal treatment effects to go beyond LATE. Chamberlain (2010) developed a Bayesian semi-parametric procedure for extrapolation that relies on models for variation in outcome distributions as a function of the first stage.

3.3. Examples

The framework provided by Angrist and Imbens showed when and how the effect identified by the IV estimator can be interpreted. By casting the analysis in terms of potential outcomes, they

[45] They thus considered a Bayesian framework involving parametric models for the posterior distributions which allow them to generate unobserved potential outcomes. This assumption allowed them to estimate ITT effects for always-takers and never-takers in addition to compliers, which gives a way to examine the validity of the exclusion restriction (since ITT effects should be zero for always-takers and never-takers if the exclusion restriction is valid). Note that the Bayesian models are consistent with the LATE framework under exclusion and monotonicity.

[46] Heckman and Vytlacil (1999) showed that nonparametric identification of the average treatment effect is possible with continuous instrumental variables that drive the probability of treatment from zero to one. In practice, instruments having this property are extremely rare; most instruments are discrete with finite support.

[47] In addition, one can compare the distribution of the potential outcome with the treatment for compliers and always-takers as well as the potential outcome without the treatment for compliers and never-takers.

also provided a framework for validating and discussing the plausibility of particular assumptions in an empirical design. In this section, we illustrate this by discussing a few concrete studies.

Effects of serving in the Vietnam War

Let us return to the question of how serving in the military affects subsequent earnings outcomes (Angrist, 1990). During the Vietnam War, a large number of American men were affected by the draft. Among men born 1950–1952, around 38 percent were draft eligible. An important question in this context is whether experience from the war has some use on the labor market. And if wartime experience is of limited use, how much compensation should be offered to veterans?

For obvious reasons, a regression of labor market outcomes on actual veteran status is likely to produce a misleading answer regarding the causal effect of veteran status, since the subset of the population doing military service is likely selected on observed and unobserved characteristics. Angrist (1990) used variation coming from the randomized draft lottery to estimate the effects of veteran status on earnings. More precisely, he used eligibility for the draft as an instrument for doing service in the Vietnam War. Angrist, Imbens, and Rubin (1996) used an analogous approach to analyze the impact on mortality.[48]

The probability of serving in the military falls with lottery number. The monotonicity assumption requires that someone who would serve in the military with lottery number k would also serve in the military with lottery number l less than k, which is plausible. Thus, it is reasonable to assume that there are no, or very few, defiers.

The exclusion restriction requires that potential earnings with and without military service be independent of the lottery number. This would be violated if lottery numbers are related to earnings through some variable other than veteran status. Always-takers would serve in the military no matter the lottery number. For such a person, it is reasonable to assume that the draft lottery number has no direct effect on earnings. Never-takers, on the other hand, would not serve in the military no matter their lottery number. Here, the exclusion restriction might be problematic. If a draftee avoided military service by staying in school or moving abroad, such behavior could have a direct impact on earnings which would violate the exclusion restriction (see Angrist, Imbens, and Rubin, 1996, for a more extensive discussion of the exclusion restriction).

Judge leniency designs

Interest is growing in using instruments based on discretionary decisions of judges, caseworkers, etc., which we refer to as "judge leniency" designs. Given that decision-makers differ in leniency, and that individuals are randomly assigned to decision-makers, leniency could be used in an IV design to study the effects of the decision on an outcome of interest.

A classic example is provided by Kling (2006), who studied the impact of incarceration length on subsequent labor outcomes. The problem for such a study is that people who get longer prison sentences might be different from those who get shorter sentences, in ways that matter for future labor earnings. Kling used the fact that judges are randomly assigned to court cases, conditional on the date and location that the case is filed. This is potentially useful because some judges appear to be systematically harsher when sentencing than others.

Another example comes from Dahl, Kostøl and Mogstad (2014) who used Norwegian data to examine how parental participation in disability insurance affected subsequent welfare participation among their children. They used random assignment of judges to disability insurance applicants whose cases were initially denied. A further example is Dobbie, Goldin and Yang (2018)

[48] The first researchers to use the draft lottery to solve the selection problem were Hearst, Newman, and Hulley (1986), who present reduced-form effects of eligibility for the draft on mortality after the Vietnam War. By contrast, Angrist, Imbens, and Rubin (1996) analyzed the impact of serving in the Vietnam War on mortality.

who used the detention tendencies of (quasi-)randomly assigned bail judges to estimate the causal effect of pretrial detention on subsequent outcomes.

The exclusion restriction may be one concern in these types of studies, in particular if there is some interaction between the decision-maker and the individual. The exclusion restriction seems easier to justify when there is no interaction between the decision-maker and the case being decided. The monotonicity assumption may be a bigger problem, however. As Dobbie, Goldin, and Yang (2018) noted, monotonicity "requires that individuals released by a strict judge would also be released by a more lenient judge, and that individuals detained by a lenient judge would also be detained by a stricter judge" (Dobbie, Goldin, and Yang, 2018: p. 222). The issue is that judges are likely to differ not just in leniency, but also in preferences; court cases are multidimensional and some judges may be harsher towards certain types of defendants than towards others.

3.4. The LATE framework outside of economics

The LATE framework developed by Angrist and Imbens is nowadays extensively used in economics but also in the other social sciences.[49] It has also found increasing use in disciplines such as epidemiology and medicine. This section briefly discusses the use of the IV method, i.e., the estimation of LATE, as a complement to ITT analysis, with an emphasis on medical research. For a more comprehensive review of methods to estimate average treatment effects in medical science and epidemiology, including LATE (or CACE, which is the more common term in the medical literature), see e.g., Little and Rubin (2000) and Hernán and Robins (2017).

Examining the causal effect of a medical intervention is typically done through an explanatory trial and a pragmatic trial. An explanatory trial examines the efficacy of the intervention, i.e., its impact under ideal or controlled conditions. A pragmatic trial estimates the intervention's effectiveness, i.e., its impact in a real-world scenario or in normal clinical conditions. In pragmatic trials, incomplete adherence (i.e., incomplete compliance) to the assigned treatment sequences is common.[50] The traditional approach, therefore, has been to focus on ITT analysis (see Shrier et al., 2014, and Dodd, White and Williamson, 2012). As noted above, ITT estimates measure the causal effect of the assignment rather than the treatment.

Two justifications are usually given for focusing on ITT analysis. First, the ITT effect may be most policy-relevant, since one cannot in general force people to take a treatment. Second, under the exclusion restriction, the ITT effect is a diluted treatment effect — it may thus be a conservative strategy for evaluating new treatments. Yet none of these justifications necessarily hold true (see, e.g., Hernán and Hernandez-Diaz (2012) for a discussion).

One situation when the justification for ITT analysis does not necessarily hold true is when comparing a new treatment with an existing treatment, and adherence rates differ across treatments. It is then possible that ITT estimates suggest that one treatment is more effective than the other, even if the treatment effects are the same across two treatments. In short, a treatment may appear more effective simply because subjects adhered to it to a greater extent.

Another situation when ITT analysis may be problematic is when examining possible negative side effects of a treatment. Dilution of the true treatment effect may then make unsafe treatments appear safe. For this reason, the international guidelines for good clinical practice suggest that the analysis of negative side effects should be according to treatment received (ICHGCP, 1999). Yet,

[49] For example, Dunning (2012) discussed the use of natural experiments in the social sciences. Sovey and Green (2011) surveyed studies in political science using instrumental variables.

[50] In this section we follow the terminology used in the medical literature. Treatment adherence implies following the treatment sequence specified in the trial protocol. When the treatment is given just once, nonadherence is the same as noncompliance as defined earlier.

and as discussed above, comparing those who are treated with those that are not generally yields a biased estimate of the treatment effect.

Dodd, White and Williamson (2012) reviewed nonadherence to treatment protocol in 98 RCTs published in the *New England Journal of Medicine* and the *Journal of the American Medical Association* during 2008. They found that, in addition to ITT analyses, 49 of 98 trials included "as-treated" or "per-protocol" analyses designed to address the problems caused by nonadherence. As-treated analysis compares individuals taking the treatment with those who did not. Per-protocol analysis, instead, excludes individuals who did not follow the treatment protocol. Again, both of these procedures provide biased estimates of treatment effects.

When information on treatment participation is available, which often is the case, LATE can be estimated and has a clear interpretation: it captures the effect on the patients complying with the protocol in the trial. For this reason, several scholars argue that LATE ought to be the main causal estimand.[51]

3.5. What have we learned?

Causality is central in all sciences. Experiments are not easily conducted in economics, however, and for most questions of interest we have to rely on observational data. By focusing on the assignment to treatment, Imbens and Angrist (1994) tied Haavelmo's (1944) ideas of identification to quasi-experimental and experimental variation. They showed that instrumental variables can identify a causal treatment effect under minimal, and in many cases empirically plausible assumptions, even when there is heterogeneity and incomplete compliance. The effect identified is the average causal effect among compliers, i.e., the causal effect for the subset of the population that changed behavior because of the value of the instrument.

In deriving their key results, Angrist and Imbens established a general framework that makes the nature of the identifying assumptions more transparent, and allows researchers to assess the sensitivity of their empirical designs to deviations from these assumptions. These advantages have turned the framework into the dominant one for both quasi-experimental and experimental work.

Moreover, the basic framework provided by Angrist and Imbens has been used to investigate the conditions under which a treatment effect can be identified when using existing methods for causal inference, such as the regression discontinuity (RD) design (Hahn, Todd, and van der Klaauw, 2001) and the difference-in-differences (DiD) design (e.g., de Chaisemartin and D'Haultfoeuille, 2020) or new methods, such as the regression kink design (Card, Lee, Pei, and Weber, 2015) and the synthetic control method (Abadie, Diamond, and Hainemuller, 2010).

4. THE DESIGN-BASED APPROACH: SUBSEQUENT DISCUSSION

The design-based approach has transformed applied work in economics, and beyond. As with all "revolutions", there has been a scientific discussion about the virtues and drawbacks of this approach. Here we address some issues that been raised in this discussion.

4.1 Internal validity at the expense of external validity?

One concern is that the emphasis on credible identification has caused the pendulum to change too much. Perhaps the focus is too much on internal validity as opposed to external validity.

In a response to such criticism (by, e.g., Deaton (2010); Heckman and Uruzua, 2010), Imbens (2010) argued that the issue of study design is distinct from the use of theory in estimation and interpretation. Moreover, he made the point that it is useful to separate the assumptions needed

[51] See, for example: McNamee (2009); Shrier et al. (2014); Shrier, Verhagen, and Stovitz (2017); Steele, Shrier, Kaufman, and Platt (2015).

to identify a causal effect in the population studied from the assumptions needed to generalize an internally valid estimate to other populations. A similar point was made by DiNardo and Lee (2011).

In fact, quasi-experimental variation and structural models may usefully complement one another. For example, Card and Hyslop (2005) used experimental variation to aid the identification of a structural model for welfare participation. Design-based estimates can also be used to validate structural models as done by Blundell (2013). Relatedly, Kline and Walters (2019) showed that instrumental variables and selection-correction type of estimates (e.g., Heckman, 1979) of LATE are numerically equivalent. In other words, the choice between these two estimators is unimportant for estimating treatment effects that are identified in the data. Kline and Walters showed how this equivalence can form the basis for validation exercises of structural models that impose parametric assumptions to identify e.g., the average causal effect in the population.

4.2 Inference, specification searches and p-hacking

The past 30 years have seen an improvement in procedures for inference. Following a hallmark study by Moulton (1986), researchers are now aware of the importance of taking grouping structures in the data into account. If individuals belonging to the same group are exposed to the same variation, one should take the correlation across individuals within group into account. This is a particular concern for DiD designs which typically utilize variation across groups (e.g., regions) over time for identification. For important contributions on inference in DiD designs, see Bertrand, Duflo, and Mullainathan (2004), Donald and Lang (2007), and Hansen (2007).

Instrumental variables analyses rely on having strong instruments (Nelson and Startz 1990; Staiger and Stock 1997). If the instruments are weak, estimates may be severely biased and the inference is misleading; Andrews, Stock, and Sun (2019) and Young (2020) discussed improvements of prevailing strategies for inference.

An intricate problem in empirical research is that it is easier to publish statistically significant results than insignificant ones. This creates incentives for researchers to engage in so-called p-hacking. Card and Krueger (1995) conducted a meta-analysis of the prior literature on the minimum wage and concluded that it suffered from publication bias. More recently, Brodeur, Lé, Sangnier, and Zylberberg (2016) suggested that there is publication bias because there is excess mass of estimates having a p-value just below 0.05 than just above 0.05. Brodeur, Cook, and Heyes (2020) focused in particular on the methods associated with the design-based approach. They concluded that p-hacking is more common for DiD and IV methods than for RCTs and RD designs. This pattern is likely due to the fact that RCTs and RD designs are more "scripted" than IV and DiD (see Ioannidis, 2005). RCTs and RD designs, by now, have established codes of conduct that limit degrees of freedom for the researcher. Note that IV analysis in RCT studies with partial compliance exhibit substantially less p-hacking than other types of IV studies. In general, the p-hacking problem is lower when the IV analysis has a strong first stage.

The existence of p-hacking may compromise the credibility of empirical results. Nevertheless, there has been some progress; see Christensen and Miguel (2018). An increased focus on the research design — which is the essence of the design-based approach — and requirements to register studies and pre-analysis plans, limit the scope for specification searches. A clear statement of the identifying assumption (as in the LATE framework) invites replication and robustness analysis since other researchers can focus on the validity of these assumptions. Over time, replication and robustness analysis have become more common in economics. Most journals today require researchers to post their data and programs (if the data are not proprietary).

5. Concluding Remarks

In a series of papers from the early 1990s, David Card and co-authors analyzed a number of core questions in labor economics using natural experiments. These studies brought rigor and transparency to the analysis, and by addressing these questions in novel and in, *a priori*, more credible ways, Card was able to produce new and more reliable answers. These initial papers stimulated reanalyses as well as new theoretical and empirical work aiming at understanding the mechanisms behind the findings. As a result of this iterative process, in which Card has been a key contributor, we have a much better understanding of how the labor market operates today than we did 30 years ago.

Joshua Angrist and Guido Imbens showed that it was possible to estimate a well-defined causal treatment effect under a minimal set of conditions, even if individuals are differently affected by the intervention and even if there is imperfect compliance. They established how this causal effect should be interpreted and showed that it could be estimated by instrumental variables. In deriving their results, they linked instrumental variables to the potential-outcomes framework for causal inference commonly used in statistics. Their framework makes the nature of the identifying assumptions more transparent; it has improved researchers' ability to establish causal effects, assess the sensitivity of their empirical designs, and interpret their results.

Together the work by this year's Laureates laid the ground for the design-based approach, which has drastically changed how empirical research is conducted over the past 30 years. The design-based approach uses mainly quasi-experimental, but also experimental, variation to estimate the causal effect of interest. Quasi-experimental variation can come from the many experiments provided by nature, administrative borders, institutional rules, and policy changes. The design-based approach features a clear statement of the assumptions used to identify the causal effect and validation of these identifying assumptions. This approach has become dominant within economics and has also spread to the other social sciences. As a result, researchers' ability to answer causal questions of substantial importance for economic and social policy has improved tremendously, greatly benefitting society at large.

REFERENCES

Aakvik, A., K. Salvanes, and K. Vaage (2010). "Measuring heterogeneity in the returns to education using an education reform." *European Economic Review*, 54: 483-500.

Aaronson, D. (2001). "Price pass-through and the minimum wage." *The Review of Economics and Statistics*, 83(1): 158-169.

Abadie, A. (2003). "Semiparametric instrumental variable estimation of treatment response models." *Journal of Econometrics*, 113(2): 231-263.

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

Abadie, A. and G.W. Imbens (2006). "Large sample properties of matching estimators for average treatment effects." *Econometrica*, 74(1): 235-267.

Abdulkadiroğlu, A., J.D. Angrist, S.M. Dynarski, T.J. Kane, and P.A. Pathak (2011). "Accountability and flexibility in public schools: Evidence from Boston's charters and pilots." *Quarterly Journal of Economics*, 126(2): 699-748.

Abowd, J., F. Kramarz, and D. Margolis (1999). "High wage workers and high wage firms." *Econometrica*, 67(2): 251-333.

Adao, R., M. Kolesar, and E. Morales (2019). "Shift-share designs: Theory and inference." *Quarterly Journal of Economics*, 134(4): 1949–2010.

Almond, A., J.J. Doyle, Jr., A.E. Kowalski, H. Williams (2010). "Estimating marginal returns to medical care: Evidence from at-risk newborns." *Quarterly Journal of Economics*, 125(2): 591–634.

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

Andrews I., J. Stock J, and L. Sun (2019), "Weak instruments in IV regression: Theory and practice," *Annual Review of Economics*, 11: 727-753.

Angrist, J.D. (1990). "Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records." *American Economic Review*, 80: 313-385.

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

Angrist, J.D., E. Bettinger, E. Bloom, E. King, and M. Kremer (2002). "Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment." *American Economic Review*, 92(5): 1535-1558.

Angrist, J.D. and W.N. Evans (1998). "Children and their parent's labor supply: Evidence from exogenous variation in family size." *American Economic Review*, 88: 450-477.

Angrist, J.D. and G.W. Imbens (1995). "Two-stage least squares estimation of average causal effect in models with variable treatment intensity." *Journal of the American Statistical Association*, 90(430): 431-442.

Angrist, J.D., G.W. Imbens, and D.B. Rubin (1996). "Identification of causal effects using instrumental variables." *Journal of the American Statistical Association*, 91: 444-472.

Angrist, J.D. and A.B. Krueger (1991). "Does compulsory schooling attendance affect schooling and earnings?" *Quarterly Journal of Economics*, 106: 976-1014.

Angrist, J.D. and A.D. Kugler (2003). "Protective or counter-productive? Labour market institutions and the effect of immigration on EU natives." *Economic Journal*, 113(488): F302-F331.

Angrist, J.D. and V. Lavy (1999). "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." *Quarterly Journal of Economics*, 114: 533-575.

Angrist, J.D., V. Lavy, and A. Schlosser (2010). "Multiple experiments for the causal link between the quantity and quality of children." *Journal of Labor Economics*, 28: 773-824.

Angrist, J.D. and J-S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. Princeton.

Ashenfelter, O.A. (1978). "Estimating the effect of training programs on earnings." *Review of Economic and Statistics*, 58: 47-57.

Ashenfelter, O.A. and D. Card (1985). "Using the longitudinal structure of earnings to estimate the effect of training programs." *Review of Economic and Statistics*, 67: 648-660.

Athey, S and G.W. Imbens (2006). "Identification and inference in nonlinear difference-in-differences models." *Econometrica*, 74: 431-97.

Atkinson, A.B. and Thomas Piketty (2010). *Top Incomes: A Global Perspective*. Oxford: Oxford University Press.

Azar J., E. Huet-Vaughn, I. Marinescu, B. Taska, and T. von Wachter (2019). "Minimum wage employment effects." Manuscript.

Azar J., I. Marinescu, and M. Steinbaum (2020). "Labor market concentration." *Journal of Human Resources*, forthcoming.

Balke and J. Pearl (1997). "Bounds on treatment effects from studies with incomplete compliance." *Journal of the American Statistical Association*, 92: 1171-1176.

Bartel, A.P. (1989). "Where do new U.S. immigrants live?" *Journal of Labor Economics*, 7(4): 371-391.

Bartik, T.J. (1991). *Who Benefits from State and Local Economic Development Policies?* W.E Upjohn Institute.

Becker, G. (1967). *Human Capital and the Personal Distribution of Income*. Ann Arbor: University of Michigan Press.

Berger, D., K. Herkenhoff, and S. Mongey (2019). "Labor market power." NBER Working Paper, 25719.

Bertrand, M., E. Duflo, and S. Mullainathan (2004), "How much should we trust differences-in-differences estimates?" *Quarterly Journal of Economics*, 119: 249-275.

Bloom, H.S. (1984). "Accounting for no-shows in experimental evaluation designs." *Evaluation Review*, 8(2): 225-46.

Blundell, R. (2013). "Empirical evidence and tax reform." In Acemoglu, D., M. Arellano, and E. Dekel (eds.) *Advances in Economics and Econometrics: Theory and Applications, Tenth World Congress,* Vol 3, Chapter 14, Cambridge University Press, Econometric Society Monographs.

Borjas, G.J. (1987). "Immigrants, minorities, and labor market competition." *Industrial and Labor Relations Review*, 40(3): 382-392.

Borjas, G.J. (2003). "The labor demand curve is downward-sloping: Reexamining the impact of immigration on the labor market." *Quarterly Journal of Economics*, 118(4): 1335-1374.

Borjas, G.J. (2017). "The wage impact of the *Marielitos*: A reappraisal." *Industrial and Labor Relations Review*, 70(5): 1077–1110.

Borjas, G.J., R.B. Freeman, and L.F. Katz (1996). "Searching for the impact of immigration in the labor market." *American Economic Review*, 86(2): 246-251.

Borjas, G.J. and J. Monras (2017). "The labor market consequences of refugee supply shocks." *Economic Policy*, 32(91): 361–413.

Borusyak, K., P. Hull, and X. Javarel (2020). "Quasi-experimental shift-share research designs." *Review of Economic Studies*, forthcoming.

Brinch, C.N, M. Mogstad, and M. Wiswall (2017). "Beyond LATE with a discrete Instrument." *Journal of Political Economy*, 125: 985-1039.

Brodeur, A., N. Cook, and A. Heyes (2020). "Methods matter: p-hacking and publication bias in causal analysis in economics." *American Economic Review*, 110(11): 3634-3660.

Brodeur, A., M. Lé, M. Sangnier, and Y. Zylberberg (2016). "Star Wars: The empirics strike back." *American Economic Journal: Applied Economics*, 8(1): 1–32.

Brown, C., C. Gilroy, and A. Kohen (1982). "The effect of the minimum wage on employment and unemployment." *Journal of Economic Literature*, 20(2): 487-528.

Burdett, K. and D.T. Mortensen (1998). "Wage differentials, employer size, and unemployment." *International Economic Review*, 39(2): 257-273.

Burtless, G. (1996) (ed.), *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success.* The Brookings Institution.

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 experiment." *The 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 1987–1989." *Industrial and Labor Relations Review*, 46(1): 38–54.

Card, D. (1992b). "Using regional variation in wages to measure the effects of the federal minimum wage." *Industrial and Labor Relations Review*, 46(1): 22-37.

Card, D. (1995a). "Using geographical variation in college proximity to estimate the return to schooling." In Christofides, N.L, E.K. Grant, and R. Swidinsky (eds.) *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp,* University of Toronto Press, Toronto.

Card, D. (1995b). "Earnings, schooling, and ability revisited." *Research in Labor Economics*, 14: 23-48.

Card, D. (1996). "The effect of unions on the structure of wages: A longitudinal analysis." *Econometrica*, 64(4): 957-979.

Card, D. (1999). "The causal effect of education on earnings." In Ashenfelter, O. and D. Card (eds.) *Handbook of Labor Economics,* Vol. 3A*,* Elsevier, Amsterdam.

Card, D. (2001a), "Immigrant inflows, native outflows, and the local labor market impacts of higher immigration." *Journal of Labor Economics,* 19(1): 22-64.

Card, D. (2001b). "Estimating the return to schooling: Progress on some persistent econometric problems." *Econometrica*, 69(5): 1127-1160.

Card, D. (2009). "Immigration and Inequality." *American Economic Review*, 99(2): 1-21.

Card, D., A. Cardoso, and P. Kline (2016). "Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women." *Quarterly Journal of Economics*, 131(2): 633–686.

Card, D., A. Cardoso, J. Heining, and P. Kline (2018). "Firms and labor market inequality: Evidence and some theory." *Journal of Labor Economics*, 36(S1): S13-S69.

Card, D., R. Chetty, and A. Weber (2007). "Cash-on-hand and competing models of intertemporal behavior: New Evidence from the labor market." *Quarterly Journal of Economics*, 122(4): 1511–1560.

Card, D. and J. DiNardo (2000). "Do immigrant inflows lead to native outflows?" *American Economic Review*, 90(2): 360-367.

Card, D., J. Heining, and P. Kline (2013). "Workplace heterogeneity and the rise of West German wage inequality." *Quarterly Journal of Economics*, 128(3): 967-1015.

Card, D. and D. Hyslop (2005). "Estimating the effects of a time-limited earnings subsidy for welfare-leavers." *Econometrica*, 73(6)**:** 1723-1770.

Card, D. and A.B. Krueger (1992a). "Does school quality matter? Returns to education and the characteristics of public schools in the United States." *Journal of Political Economy*, 100(1): 1-40.

Card, D. and A.B. Krueger (1992b), "School quality and black-white relative earnings: A direct assessment." *Quarterly Journal of Economics*, 107(1): 151-200.

Card, D. and A.B. Krueger (1994). "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania." *American Economic Review*, 84: 772-784.

Card, D. and A.B. Krueger (1995), *Myth and Measurement: The New Economics of the Minimum Wage*, Princeton University Press, Princeton.

Card, D., and A.B. Krueger (2000). "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Reply." *American Economic Review*, 90(5): 1397-1420.

Card, D., D. Lee, Z. Pei, and A Weber (2015). "Inference of causal effects in a generalized Regression Kink Design." *Econometrica*, 83(6): 2453-2483.

Card, D. and A. Payne (2002). "School finance reform, the distribution of school spending, and the distribution of student test scores." *Journal of Public Economics*, 83: 49-82.

Carneiro, P. and S. Lee (2009). "Trends in quality-adjusted skill premia in the United States, 1960–2000.", *American Economic Review*, 109(6): 2309–2349.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019). "The effect of minimum wages on low-wage jobs", *Quarterly Journal of Economics*, 134(3): 1405–1454.

Cesarini, D., E. Lindqvist, M.J. Notowidigdo, and R. Östling (2017). "The effect of wealth on individual and household labor supply: Evidence from Swedish lotteries." *American Economic Review*, 107(12): 3917-3946.

Cesarini, D., E. Lindqvist, R. Östling, and B. Wallace (2016). "Wealth, health, and child development: Evidence from administrative data on Swedish lottery players." *Quarterly Journal of Economics*, 131(2): 687–738.

Chamberlain, G. (1986) "Asymptotic efficiency in semiparametric models with censoring." *Journal of Econometrics*, 32: 189–218.

Chamberlain, G. (2010). "Binary response models for panel data: Identification and information." *Econometrica*, 78: 159–168.

Chetty, R., J.N. Friedman, N. Hilger, E. Saez, D.W. Schanzenbach, and D. Yagan (2011). "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *Quarterly Journal of Economics*, 126, 1593-1660.

Chetty, R. and N. Hendren (2018). "The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects." *Quarterly Journal of Economics*, 133(3): 1107–1162.

Christensen, G. and E. Miguel (2018). "Transparency, reproducibility, and the credibility of economics research." *Journal of Economic Literature*, 56(3): 920-980.

Coleman, J.S., et al. (1966). *Equality of Educational Opportunity*. Washington: Government Printing Office.

Cortes, P. (2008). "The effect of low-skilled immigration on U.S. prices: Evidence from CPI Data." *Journal of Political Economy*, 116(3): 381-422.

Coviello, D., E. DeSerranno, and N. Persico (2020). "Minimum wages and individual worker productivity." Manuscript, Northwestern University.

Cox, D. (1958). *Planning of experiments*. New York: John Wiley and Sons.

Dahl, G., A. Kostøl and M. Mogstad (2014). "Family welfare cultures." *Quarterly Journal of Economics*, 129(4): 1711–1752.

Deaton, A. (2010). "Instruments, randomization, and learning about development." *Journal of Economic Literature*, 48: 424-455.

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

DiNardo, J. and D. Lee (2011). "Program evaluation and research designs." In O. Ashenfelter and D. Card (Eds) *Handbook of Labor Economics*, Vol. 4A. Amsterdam: Elsevier

Dobbie, W., J. Goldin, and C.S. Yang (2018). "The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges." *American Economic Review*, 108(2): 201-240.

Dodd S., I. White, and P. Williamson (2012). "Nonadherence to treatment protocol in published randomised controlled trials: a review." *Trials*, 13: 84.

Donald S. and K. Lang (2007). "Inference with difference-in-differences and other panel data." *The Review of Economics and Statistics*, 89(2): 221-233.

Dube, A (2019). *Impacts of Minimum Wages: Review of the International Evidence*. London: HM Treasury.

Dube, A., W. Lester, and M. Reich (2010). "Minimum wage effects across state borders: Estimates using contiguous counties." *Review of Economics and Statistics*, 92(4): 945-964.

Dube, A., W. Lester, and M. Reich (2016). "Minimum wage shocks, employment flows, and labor market frictions." *Journal of Labor Economics*, 34(3): 663-704.

Dunning, T. (2012), *Natural Experiments in the Social Sciences,* Cambridge University Press, Cambridge, UK.

Dustmann, C., T. Frattini, and I. Preston (2013). "The effect of immigration along the distribution of native wages." *Review of Economic Studies*, 80(1): 145–173.

Dustmann, C. and A. Glitz (2015). "How do industries and firms respond to changes in local labor supply?" *Journal of Labor Economics*, 33(3): 711-750.

Dustmann, C., A. Lindner, U. Schönberg, M. Umkehrer, and P. vom Berge (2020). "Reallocation effects of the minimum wage." CREAM DP 07/20, UCL.

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

Dustmann, C., U. Schönberg, and J. Stuhler (2017). "Labor supply shocks, native wages, and the adjustment of local employment." *Quarterly Journal of Economics*, 132(1): 435–448.

Finn, J.D. and C.M. Achilles (1990). "Answers and questions about class size: A statewide experiment." *American Educational Research Journal*, 28: 557-577.

Foged, M. and G. Peri (2016). "Immigrants' effect on native workers: New analysis on longitudinal data." *American Economic Journal: Applied Economics*, 8(2): 1-34.

Fredriksson, P., B. Öckert, and H. Oosterbeck (2013). "Long-term effects of class size." *Quarterly Journal of Economics*, 128: 249-285.

Freeman, R. (1980). "An empirical analysis of the fixed coefficient "manpower requirement" model, 1960-1970." *Journal of Human Resources*, 15(2): 176–199.

Friedberg, R.M. (2001). "The impact of mass migration on the Israeli labor market." *Quarterly Journal of Economics*, 116(4): 1373–1408.

Garen, J. (1984). "The returns to schooling: A selectivity bias approach with a continuous choice variable." *Econometrica*, 52: 199-1218.

Glitz, A. (2012). "The labor market impact of immigration: A quasi-experiment exploiting immigrant location rules in Germany." *Journal of Labor Economics,* 30(1): 175-213.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020). "Bartik Instruments: What, when, why, and how." *American Economic Review,* 110(8): 2586-2624.

Grossman, J.B. (1982). "The substitutability of natives and immigrants in production." *Review of Economics and Statistics*, 54(4): 596-603.

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

Haavelmo, T. (1944). "The probability approach in econometrics." *Econometrica*, 12: 1–115.

Hahn, J., P. Todd, and W. van der Klaauw (2001). "Identification and estimation of treatment effects with a Regression-Discontinuity Design." *Econometrica*, 69: 201-209.

Hansen, C.B. (2007). "Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects." *Journal of Econometrics*, 140: 670-694.

Hanushek, E.A. (1986). "The economics of schooling: Production and efficiency in public schools." *Journal of Economic Literature*, 49(3): 1141-1177.

Hanushek, E.A. (2003). "The failure of input-based schooling policies." *Economic Journal*, 113: 65–98.

Harasztosi, P. and A. Lindner (2019). "Who pays for the minimum wage?" *American Economic Review*, 109(8): 2693-2727.

Hearst, N., T.B. Newman, and S.B. Hulley (1986). "Delayed effects of the military draft on mortality. A randomized natural experiment." *New England Journal of Medicine*, 314: 620-624.

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: Papers and Proceedings*, 80(2): 313–318.

Heckman, J.J. (1997). "Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations." *Journal of Human Resources*, 32: 441-462.

Heckman, J.J., A. Layne-Farrar, and P. Todd (1996). "Human capital pricing equations with an application to estimating the effect of schooling quality on earnings." *Review of Economics and Statistics* 78: 562–610.

Heckman, J.J., J.L. Tobias, and E. Vytlacil (2001). "Four parameters of interest in the evaluation of social programs." *Southern Economic Journal*, 68(2): 210-223.

Heckman, J.J., J.L. Tobias, and E. Vytlacil (2003). "Simple estimators for treatment parameters in a latent-variable framework." *Review of Economics and Statistics,* August 2003, 85(3): 748–755.

Heckman, J.J. and S. Urzua (2010). "Comparing IV with structural models: What simple IV can and cannot identify." *Journal of Econometrics* 156: 27-37.

Heckman J.J. and E. Vytlacil (1999). "Local instrumental variables and latent variable models for identifying and bounding treatment effects.", *PNAS,* 96: 4730–4734.

Hedges, L., R.D. Laine, and R. Greenwald (1994). "Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes." *Educational Researcher*, 23(3): 5–14.

Hendry, D. (1980). "Econometrics – alchemy or science?" *Economica*, 7(188): 387-406.

Hernán M. and S. Hernandez-Diaz (2012). "Beyond the intention-to-treat in comparative effectiveness research." *Clinical Trials*, 9(1): 48-55.

Hernán, M. and J. Robins (2017). "Per-protocol analyses of pragmatic trials." *The New England Journal of Medicine: Statistics in Medicine*, 377(14): 1391-1398.

Hirano, K., G.W. Imbens, and G. Ridder (2003). "Efficient estimation of average treatment effects using the estimated propensity score." *Econometrica*, 71: 1161-1189.

Hoxby, C. (2001). "All school finance equalizations are not created equal." *Quarterly Journal of Economics*, 116: 1189–1231.

Hunt, J. (1992). "The impact of the 1962 repatriates from Algeria on the French labor market." *Industrial and Labor Relations Review*, 45(3): 556-572.

ICHGCP (1999). "International Conference on Harmonization of Good Clinical Practice E9 expert working group: ICH harmonised tripartite guideline. Statistical principles for clinical trials." *Statistics in Medicine*, 18: 1905–1942.

Imbens, G.W. (2010). "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature*, 48(2): 399-423.

Imbens, G.W. and J.D. Angrist (1994). "Identification and estimation of local average treatment effects." *Econometrica*, 61: 467-476.

Imbens, G.W. and D.B. Rubin (1997a). "Bayesian inference for causal effects in randomized experiments with noncompliance." *Annals of Statistics*, 25: 305-377.

Imbens, G.W. and K. Kalyanaraman (2012). "Optimal bandwidth choice for the regression discontinuity estimator." *Review of Economic Studies*, 79(3): 933–959.

Imbens, G.W. and D.B. Rubin (1997b). "Estimating outcome distributions for compliers in instrumental variables models." *Review of Economic Studies*, 64: 555-574.

Ioannidis, J.P. (2005). "Why most published research findings are false." *PLoS Medicine*, 2(8): e124.

Jackson, K. (2020). "Does school spending matter? The new literature on an old question." An Equal Start: Policy and Practice to Promote Equality of Opportunity for Children.

Jackson, K., R. Johnson, and C. Persico (2016). "The effects of school spending on educational and economic outcomes: Evidence from school finance reforms." *Quarterly Journal of Economics*, 131(1): 157-218.

Jaeger, D., J. Ruist, and J. Stuhler (2018). "Shift-share Instruments and the Impact of Immigration." NBER Working Paper 24285.

Johnson, D.S., J.A. Parker, and N.S. Souleles (2006). "Household expenditure and the income tax rebates of 2001." *American Economic Review*, 96(5): 1589-1610.

Katz, L. and A.B. Krueger (1992). "The effect of the minimum wage on the fast-food Industry." *Industrial and Labor Relations Review*, 46(1): 6-21.

Kennan, J. (1995). "The elusive effects of minimum wages." *Journal of Economic Literature*, 33(4): 1950-1965

Kerr S.P., T. Pekkarinen, and R. Uusitalo (2009). "School tracking and intergenerational income mobility: Evidence from the Finnish comprehensive school reform." *Journal of Public Economics*, 93: 965-973.

Kirkebøen, L.J., E. Leuven, and M. Mogstad (2016). "Field of study, earnings, and self-selection." *Quarterly Journal of Economics*, 31(3): 1057–1111.

Kline, P and C. Walters (2019). "On Heckits, LATE, and numerical equivalence." *Econometrica*, 87(2): 677-696.

Kling, J. (2006). "Incarceration length, employment, and earnings." *American Economic Review*, 96(3): 863-876.

Kroft, K., Y. Luo, M. Mogstad, and B. Setzler (2020). "Imperfect competition and rents in labor and product markets: The case of the construction industry". NBER Working Paper 27325.

Krueger, A.B. (1999). "Experimental estimates of education production functions", *Quarterly Journal of Economics*, 114: 497-532.

Krueger, A.B. and D.M. Whitmore (2001). "The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR." *Economic Journal*, 111: 1-28.

KVA (2019). "Understanding development and poverty alleviation." Scientific background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2019.

Lalive, R. (2008). How do extended benefits affect unemployment duration? A regression discontinuity approach. *Journal of Econometrics*, 142(2): 785-806.

Lalive, R., C. Landais, and J. Zweimüller (2015). "Market externalities of large unemployment insurance extension programs." *American Economic Review*, 105(12): 3564-396.

LaLonde, R.J. (1986). "Evaluating the econometric evaluations of training programs using experimental data." *American Economic Review*, 76: 602-620.

Lalonde, R.J. and R. Topel (1991). "Labor market adjustments to increased immigration." In J. Abowd and R.B. Freeman (eds.) *Immigration, Trade, and the Labor Market*. University of Chicago Press.

Lamadon, T., M. Mogstad, and B. Setzler (2020), Imperfect competition, compensating differentials and rent sharing in the U.S. labor market. Manuscript. University of Chicago.

Leamer, E. (1983). "Let's take the con out of econometrics." *American Economic Review*, 73: 31-43.

Lee, D.S. (2008). "Randomized experiments from non-random selection in U.S. house elections." *Journal of Econometrics*, 142(2): 675-697.

Lewis, E. (2011). "Immigration, skill mix, and capital skill complementarity." *Quarterly Journal of Economics*, 126(2): 1029–1069.

Little R. and D. Rubin (2000). "Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches." *Annual Review of Public Health*, 21:121–145.

Manacorda, M., A. Manning, and J. Wadsworth (2012). "The impact of immigration on the structure of wages: Theory and evidence from Britain." *Journal of the European Economic Association*, 10(1), 120-151.

Manning, A. (2021). "The elusive employment effect of the minimum wage." *Journal of Economic Perspectives*, 35(1): 3-26.

Manski, C. (1990). "Nonparametric bounds on treatment effects." *American Economic Review: Papers and Proceedings*, 80(2): 319-323.

McNamee R. (2009). "Intention to treat, per protocol, as treated and instrumental variable estimators given non-compliance and effect heterogeneity." *Statistics in Medicine*, 28(21): 2639-2652.

Meghir, C. and M. Palme (2005). "Educational reform, ability, and family background." *American Economic Review*, 95(1): 414-424.

Mincer, J. (1958). "Investment in human capital and personal income distribution." *Journal of Political Economy*, 66(4): 281–302.

Moulton, B. (1986). "Random group effects and the precision of regression estimates." *Journal of Econometrics*, 32(3): 385-397.

Nelson, C.R. and R. Startz (1990). "The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor One." *Journal of Business*, 63: S125–S140.

Neumark, D. and W. Wascher (1992). "Employment effects of minimum and subminimum wages: Panel data on state laws." *Industrial and Labor Relations Review*, 46(1): 55–81.

Neumark, D and W. Wascher (2000). "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Comment." *American Economic Review*, 90(5): 1362-1396.

Neyman, J. (1923/1990). "On the application of probability theory to agricultural experiments, Essays on principles, section 9." translated in *Statistical Science,* (with discussion) 5, 465-480.

Ottaviano, G. and G. Peri (2012). "Rethinking the effect of immigration on wages." *Journal of the European Economic Association*, 10(1), 152-197.

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference,* Cambridge University Press, New York.

Pearl, J. (2009). "Causal inference in statistics: An overview." *Statistics Surveys*, 3: 96-146.

Peri, G. (2012). "The effect of immigration on productivity: Evidence from U.S. states." *Review of Economics and Statistics*, 94(1): 348-358.

Peri, G. and C. Sparber (2009). "Task specialization, immigration, and wages." *American Economic Journal: Applied Economics*, 1(3): 135-69.

Peri, G. and C. Sparber, (2011). "Assessing inherent model bias: An application to native displacement in response to immigration." *Journal of Urban Economics*, 69(1): 82-91.

Peri, G. and V. Yasenov (2019). "The labor market effects of a refugee wave: Synthetic control method meets the Mariel Boatlift." *Journal of Human Resources*, 54: 267-309.

Pop-Eleches, C. and M. Urquiola (2013). "Going to a better school: Effects and behavioral responses." *American Economic Review*, 103 (4): 1289-1324.

Portugal, P. and A. Cardoso (2006). "Disentangling the minimum wage puzzle: An analysis of worker accessions and separations." *Journal of the European Economic Association* 4(5): 988–1013.

Renkin, T., C. Montialoux, and M. Siegenthaler (2020). "The pass-through of minimum wages into U.S. retail prices: Evidence from supermarket scanner data." *Review of Economics and Statistics*, forthcoming.

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 Sechrest, L. Freeman, H. and Bailey, A. (eds.), *Health Service Research Methodology: A Focus on AIDS* (vs. Public Health Service).

Robinson, J (1933), *The Economics of Imperfect Competition*, MacMillan.

Rothstein, J. and D.W. Schanzenbach (2021). "Does money still matter? Attainment and earnings effects of post-1990 school finance reforms." NBER Working Paper No. 29177.

Rubin, D.B. (1974). "Estimating causal effects of treatments in randomized and non-randomized studies." *Journal of Educational Psychology*, 66: 688-701.

Rubin, D.B. (1977). "Assignment to treatment group on the basis of a covariate." *Journal of Educational Statistics*, 2: 1–26.

Rubin, D.B. (1980). "Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *Journal of the American Statistical Association*, 75(371): 591–93

Shrier I., R. Steele, E. Verhagen, R. Herbert, C. Riddell, J. Kaufman J.S. (2014). "Beyond intention to treat: what is the right question?" *Clinical Trials*, 11(1): 28-37.

Shrier I, E. Verhagen, S. Stovitz (2017). "The intention-to-treat analysis is not always the conservative approach." *American Journal of Medicine*, 130(7): 867-871.

Sims, C.A. (1980). "Macroeconomics and reality." *Econometrica*, 48(1): 1-48.

Snow, J. (1855). *On the Mode of Communication of Cholera*. 2nd ed. John Churchill, London.

Sovey, A. and D. Green (2011). "Instrumental variables estimation in political science: A readers' guide." *American Journal of Political Science*, 55: 188–200.

Staiger, D. and J. Stock (1997). "Instrumental variables regression with weak instruments." *Econometrica*, 65(3): 557–586.

Steele R., I. Shrier, J. Kaufman, R. Platt (2015). "Simple estimation of patient-oriented effects from randomized trials: An open and shut CACE." *American Journal of Epidemiology*, 182(6): 557-566.

Thistlewaite, D.L. and D.T. Campbell (1960). "Regression-discontinuity analysis: An alternative to the ex post facto experiment." *Journal of Educational Psychology*, 51: 309-317.

Vytlacil, E. (2002). "Independence, monotonicity, and latent index models: An equivalence result." *Econometrica*, 70: 331-341.

Wolfson, P. and D. Belman (2019). "15 years of research on U.S. employment and the minimum wage." *Labour* 33(4): 488-506.

Wright, P.G. (1928). *The Tariff on Animal and Vegetable Oils,* Macmillan, New York.

Young, A. (2020). "Consistency without inference: Instrumental variables in practical application." mimeo: London School of Economics and Political Science.