Treatment for natural experiments:

How to improve causal estimates using conceptual definitions and substantive interpretations

Author: Stefan Öberg, Unit for Economic History, School of Business, Economics and Law, University of Gothenburg, PO Box 625, 405 30 Gothenburg, Sweden (email: stefan.oberg@econhist.gu.se).

October 21, 2021

Abstract

The local average treatment effect (LATE) estimator has enabled a wide range of empirical studies that use natural experiments to estimate causal effects from observational data. However, this empirical literature has overlooked a crucial assumption regarding the definition of the treatment—the part of the stable unit treatment value assumption referred to as the consistency assumption in epidemiology. The consequence of ignoring this assumption has been that results have been misinterpreted and over-generalized. I illustrate these problems using examples from seminal studies from the natural experiments literature and present how to improve definitions of the treatment in future studies. Correctly interpreted LATEs are substantially more specific than previously acknowledged but reclaim their internal validity. By providing clear and careful definitions of the treatment and substantive interpretations of the results, future studies will increase their usefulness by presenting the causal effect that has been actually estimated.

Keywords: causal effects, instrumental variables, interpreting causal effects, local average treatment effects, SUTVA, consistency, proxy instruments, surrogate instruments

JEL classification codes: C18, C26, C36, C52, C90

Introduction

Design-based empirical studies that aim to estimate the causal effect of a treatment from observational data by using some source of exogenous variation in a so-called natural experiment are common in economics and are also becoming widespread in other research fields. I will refer to this literature as the “natural experiments literature.” In this paper, I criticize how the empirical natural experiments literature conceptualizes and interprets their (natural) experiment design and estimated effects.

The natural experiments¹ framework has been presented as an improvement upon empirical research because of better and explicitly discussed research designs (e.g., Angrist and Pischke, 2009, 2010). I argue that this literature has been overly focused on finding as-good-as-random assignment mechanisms (compare Dunning, 2008, p. 283; Imbens, 2020, p. 1152; Esterling et al., 2021, p. 26). An experiment consists of (at least) five different aspects: a well-defined study population, a well-defined (and pre-selected) outcome, a well-defined (and pre-selected) treatment, and a clear understanding of the mechanism for assignment to the treatment as well as the mechanism determining reception of the treatment. When designing a natural experiment, it is equally important to have clear and detailed definitions of all the other aspects of the experiment as having an as-good-as-random assignment mechanism. A (natural) experiment is not defined only by its assignment mechanism but, rather, by all of its aspects. Different (natural) experiments can be designed based on the same assignment mechanism.

The natural experiments literature uses the potential outcomes framework to define and estimate causal effects (Angrist et al., 1996; Imbens and Rubin, 2015, chaps. 1–3; Morgan and Winship, 2015, chap. 2; Imbens, 2020). But the natural experiments literature has, from its beginning, overlooked one of the necessary assumptions of

¹ I use the term “natural experiments” as defined in, for example, Haavelmo (1944) and Angrist and Krueger (2001).

the potential outcomes framework to estimate causal effects: the second part of the stable unit value assumption (SUTVA) (Imbens and Rubin, 2015, pp. 9–12, 517, 589).² The second part of SUTVA states that we must define our treatment so that it has the same effect on all treated units (Cox, 1992[1958], pp. 17–19; see also Hernán and Robins, 2020, chaps. 1–3).

This assumption has largely been overlooked in the social sciences, including economics, but is necessary for estimating reliable and useful causal effects (Esterling et al., 2021). The assumption has been discussed and elaborated upon in the epidemiology literature, where it is referred to as the consistency assumption (e.g., Cole and Frangakis, 2009; VanderWeele, 2009; Rehkopf et al., 2016; Hernán and Robins, 2020, pp. 31–36).

We need clear and careful conceptual definitions of the substantive meanings of all the arguments of our estimator, and the second part of SUTVA states this necessity specifically for the treatment. It is useful to have this as a separate assumption because, as we will see, it is not always straightforward to specify a sufficiently clear and careful definition of the treatment.

A clear and careful definition of the treatment is absolutely fundamental to our ability to ensure that our conceptual definitions correspond to the data and to interpret the estimated effect correctly (e.g., Hernán, 2016; Kaufman, 2019). We

² The second part of SUTVA is not mentioned in Imbens and Angrist’s papers introducing the LATE (Imbens and Angrist, 1994; Angrist and Imbens 1995). (Angrist and Imbens [1995] discuss only the first part relating to no interference between units.) Angrist, Imbens, and Rubin (1996) only mention it implicitly: “Our notation has already restricted both [the IV] Z and [the treatment] D to have only two levels; that is, there is no partial compliance.” Because there is no partial compliance with the treatment, it only has two levels. There are, therefore, no versions of the treatment.

also need to have a clear and careful definition of the treatment to be able to evaluate the assumptions of the estimator and specify the causal model (e.g., Hernán and VanderWeele, 2011, p. 373; Imbens, 2020, p. 1157). In this paper, I will show how we can improve the conceptual definitions of the treatment and how this helps us to improve the validity of our (natural) experiments and the interpretation of the effects we estimate.

The conceptual definition of the substantive meaning of the arguments of an estimator cannot be expressed using equations. There is nothing inherent in equations that require us to define precisely what is meant by their arguments. Estimators do not need meaning to produce estimates, but the definition of precise meanings for the arguments is necessary to produce useful and interpretable estimates. The conceptual definitions and substantive interpretations are separate from mathematically deriving the estimator but are equally crucial parts of any empirical analysis as the derivation.³ I argue that we need to be just as careful with our words as with our algebraical arguments because “[e]ven when all the statistical analyses in a study are impeccable, errors of causal reasoning may still lead to the wrong causal conclusion” (Shadish et al., 2002, p. 63).

³ Sander Greenland (2017) expresses this better than I could ever do: “(T)he interpretations of terms in the model do not have to be settled for deductive aspects of modeling to proceed. Logic and mathematics operate with undefined (‘primitive’) concepts and abstract objects defined only from those concepts, which together form the logical structure (syntax) of the model. The interpretation (meaning or semantics) of the model and the deductions from it are left to the user. There are always a huge number of interpretations that would follow the logical structure; more precise definitions narrow down the vast space of logically acceptable interpretations to a far more practical range” (Greenland, 2017, p. 9).

As you will have realized by now, my arguments are related to external validity and construct validity (e.g., Shadish et al., 2002, chaps. 1–3; Esterling et al., 2021). The natural experiments literature has focused almost exclusively on the internal validity of its estimates, but we need all types of validity to provide useful and reliable results. The aim of any empirical study should, therefore, be to produce causal estimates that are valid overall, meaning that they are true: “A causal claim is valid if the claim is true, that is, if the purported cause and purported effect are the actual cause and actual effect” (Esterling et al., 2021, p. 8).

Natural experiments and the LATE estimator

Several different estimators are used to analyze natural experiments. I will discuss the case of using natural experiments to create instrumental variables (IVs). My arguments also apply to other methods that use variation from natural experiments, such as difference-in-difference or regression discontinuity estimators. I will discuss cases with binary instruments and, therefore, also a binary treatment of the experiment because this is the most common in the literature. Again, the argument extends to other situations. I will assume that the IVs are used for a two-stage least squares estimator, with a first stage modeling the share of units treated by assignment and a second stage modeling the outcome on the assignment.

The exogenous variation from the natural experiment should create an assignment mechanism determining assignment to treatment that is “as good as random.” It is exceedingly rare to find natural experiments involving people (as individuals or aggregations) that affect everyone in the same way.⁴ In other words, it is

⁴ The standard interpretation of effects estimated using IVs requires that the assignment mechanism (the instrument) has the same effect on the treatment (the endogenous variable of interest) for everyone (including at all levels of the treatment). When the assignment mechanism has the same effect on the treatment for everyone, the assignment mechanism “assign the variable of interest randomly” (Angrist and Krueger, 2001, p. 72; see also, Angrist and Pischke,

exceedingly rare to find assignment mechanisms that “assign the variable of interest randomly” (Angrist and Krueger, 2001, p. 72; see also, Angrist and Pischke, 2009, pp. 21, 151–152). What we can find are situations that induce an exogenous change in the variable of interest for a specific subpopulation, in other words, the “assignment is as good as random at least for a subpopulation” (Imbens, 2020, p. 1152). What we are then analyzing are natural experiments in which the assignment mechanism affects different units differently (i.e., we have a “heterogenous effect of the instrument on [the treatment] D_i ”, e.g., Angrist and Pischke [2009], p. 152).

We can still estimate useful effects using IVs with heterogenous effects by interpreting them as local average treatment effects (LATEs) (Imbens and Angrist, 1994; Angrist and Imbens, 1995; Angrist et al., 1996).⁵ The LATE is the effect of the treatment for the subpopulation that has their treatment changed by the

2009, pp. 21, 151–152). Unfortunately, proponents of natural experiments have not made it sufficiently clear that this is not the typical situation analyzed when estimating a LATE.

⁵ Imbens and Rubin (1997) proposed that these effects should instead be termed complier average causal effects (CACEs). This name makes the interpretation of the effect clearer. LATEs (or CACEs) are the (average) causal effect for the subpopulation that has their level of treatment affected by the assignment mechanism (i.e., the compliers).

assignment mechanism.⁶ This subpopulation is referred to as the “compliers” because they comply with the intended design of the study.⁷

The LATE has the added benefit of allowing for heterogeneous effects of the treatment on the outcome (i.e., different units in the study population can be affected differently by the treatment). What we are estimating when using the LATE estimator is the (average) effect of the treatment for the subpopulation of units that have their treatment changed by the assignment mechanism. (As we will discuss further below, the treatment must have the same effect on all units within this subpopulation.) The LATE estimator has enabled a large number of empirical analyses of natural experiment situations and is now widely used in many different fields of research (e.g., Gallen and Raymond, 2021, fig. 1).

The LATE estimator is based on the potential outcomes framework (for an introduction, see Imbens and Rubin, 2015, chaps. 1–3, 23–24; Morgan and Winship, 2015, chaps. 2, 9; Öberg, 2021). The potential outcomes framework defines a causal effect as the difference in outcome for a specific unit being treated and not treated.

⁶ Hearst and Newman’s (1988) discussion on the use of the draft lottery as a basis for causal inferences regarding Vietnam veterans is a useful, non-technical description of how the method works and the substantial thinking that is required to evaluate its validity.

⁷ With two-sided non-compliance, the units can be classified as one of four “types”: compliers, never-takers, always-takers, and defiers. See Angrist, Imbens, and Rubin (1996), Imbens and Rubin (2015, chaps. 23–24), or Winship and Morgan (2015, chap. 9) for the definitions of these types. Greenland and Robins (1986) also discuss the definitions of these types using a different terminology. With one-sided non-compliance, there is only one type of non-complier (i.e., always-takers or never-takers). In the three examples provided in this paper, there are, for example, never-takers but no always-takers.

The definition of what it means for a unit to be treated or not is, therefore, a fundamental aspect of defining a causal effect in the potential outcomes framework.⁸ For any specific unit, at least one of the alternatives of either being treated or not treated is a potential, or hypothetical, outcome. Hence, we need to assume that different units are comparable to obtain empirically observable differences between treated and not treated units.

I argue that the empirical natural experiments literature must begin to include clear and careful definitions of the experiment's treatment and the potential outcomes to be able to continue claiming to estimate credible causal effects. As we will return to later, we cannot separate the experiment's treatment from how it is delivered. This makes it equally important to determine the reception mechanism (i.e., how and why the treatment is affected [or not] for units that are assigned to the treatment [or not]).

Some of the issues with the empirical natural experiments literature that I raise here have been discussed, more or less directly, by critics of the literature. The issues I raise are at the root of much of the criticism of natural experiments. Addressing the issues I raise would, therefore, tackle important parts of this criticism. The natural experiments literature has, for example, been criticized for estimating effects that are vague (e.g., Dawid, 2000a; Deaton, 2010). A key reason for the vagueness is that it is sometimes difficult to know “what, exactly, is the treatment” (Dunning, 2008, p. 287; see also, e.g., Heckman, 2005, p. 5). Critics and proponents of natural experiments all agree that a well-defined causal effect requires a clear and careful definition of the treatment (e.g., Rubin, 1978, pp. 39–40; Angrist et al., 1996, p. 446; Robins and Greenland, 2000, p. 434; Heckman, 2005, p. 2; Lewbel, 2019, p. 872).

⁸ Proponents of the potential outcomes framework indeed point out that it is a defining difference of this framework compared to other branches of econometrics that it defines the causal effect as a comparison of potential outcomes rather than “in terms of parameters in some model” (Imbens and Rubin, 2010[2008], p. 229).

Still, there has not been sufficient focus on providing clear and careful definitions of what the treatment of the experiment is.

Another key reason for why the effects are deemed vague is that it is not possible to identify empirically which units are compliers (i.e., for which units we are estimating the effect) (Angrist et al., 1996, p. 449; Deaton, 2010, p. 430; Heckman, 2010, p. 372; Imbens and Rubin, 2015, p. 545; Lewbel, 2019, pp. 869–870). Even proponents of the natural experiment framework lament that it is not possible to identify the compliers (e.g., Angrist and Pischke, 2009, p. 171).

The criticism that the LATE is estimating an effect for an unidentifiable group of units is relevant for how this method has been implemented until now but is not relevant for how it should be implemented. It is a fact that we cannot conclusively identify the compliers empirically. But I argue that we absolutely must have a clear and careful conceptual definition of which units are compliers for the analysis to make sense. This critique (and the current practice in the literature) confuses the empirical and conceptual aspects of the method.

Another critique of the natural experiments literature has been that it is not explicit about “how and why things work” (Deaton, 2010, p. 432). That is to say, there is a lack of explicit explanations for which units have their treatment affected by the assignment mechanism and why. This is, in essence, a critique of the lack of clear and careful definitions of the treatment including the reception mechanism. Heckman (1997, 2005, 2010) and Rosenzweig and Wolpin (2000), among others, have argued for defining structural models as the method to remedy this problem.

Structural models solve one issue by specifying explicit models for the reception mechanism. However, it is just as important to provide clear and careful definitions and interpretations of all aspects of a structural model as it is with natural experiment designs. And whereas, there is nothing in the structural approach that requires that we provide such definitions and interpretations, I argue that the

potential outcomes framework, when applied correctly, requires that we do provide such clear and careful definitions and interpretations of all aspects of our design.

Therefore, I argue that the key to improving the natural experiments literature is to provide clear and careful definitions of the experiment's treatment. This, in turn, leads to clear and careful ex ante conceptual definitions of who is a complier. Clear and careful definitions of what it is we are estimating the effect of can, further, help us improve the underlying theoretical causal models and evaluate the necessary assumptions.

The rest of the paper provides a seemingly basic guide of how we should make clear and careful definitions of the treatment, the compliers, and the underlying causal model. It is "basic" because it draws on the fundamentals of the potential outcomes framework but only "seemingly" basic because these fundamentals have not been implemented in practice in the empirical literature. I will use three well-known examples from the literature to illustrate how things have gone wrong in previous work.⁹

Back to the basi(c)s

Defining the treatment: the treatment of the experiment vs. the treatment of interest

We need to begin with some terminology. The "treatment of interest" is that which we would ideally like to estimate the effect of (i.e., the "effect of interest"). What I will term the effect of interest here is also referred to as the estimand.¹⁰

⁹ The examples are quite old (published in 1990, 1998, and 2010). This is not because the issues I raise are a thing of the past. I chose these examples because they are well-known in the natural experiments literature.

¹⁰ Lundberg, Johnson, and Stewart (2021) suggest that we refer to the effect of interest as "the theoretical estimand" to distinguish it from what is actually

In the examples used below, the effect of interest is “the long-term labor market consequences of military service during the Vietnam era” (Angrist 1990, p. 313), “the causal link running from sibship size to human capital, economic well-being, and family structure” (Angrist, Lavy, and Schlosser 2010, p. 816), and “the causal link running from fertility to the work effort of both men and women” (Angrist and Evans 1998, p. 451). I will show that, when we estimate LATEs, there will (almost) always be a difference between the treatment of interest and the “treatment of the experiment”. This difference will, quite naturally, lead to the effect estimated by the (natural) experiment being different from the effect of interest.

That the effect estimated by the (natural) experiment might not be the effect of interest has sometimes been recognized in comments and in the theoretical literature (e.g., Dunning, 2008, p. 290; Imbens and Rubin, 2010[2008], p. 232; Hernán and Robins, 2020, p. 202). Critics of the LATE have discussed this issue as the LATE not producing the results that are of theoretical interest (e.g., Rosenzweig and Wolpin, 2000; Heckman, 2005; Mogstad and Torgovitsky, 2018). I follow Haavelmo (1944, p. 14) in arguing that the difference between the effect of interest and the effect estimated from the (natural) experiment is an inescapable consequence of estimating causal effects using observational data from natural experiments rather than prospective data from controlled experiments.

It is currently too common in the empirical natural experiments literature to apply wishful thinking when conceptually defining the design of the study and interpreting the estimated effect. Oftentimes, it is merely stated that the estimated effect is an estimate of the causal effect of interest. This approach is neither correct nor helpful. The LATE estimator cannot provide estimates of anything but the effect on the

possible to estimate empirically (which they, in turn, call “the empirical estimand”). Because I limit my discussion to natural experiments, I will refer to the empirical estimand here as the “effect of the experiment”.

compliers from the change some of them have experienced as a result of being assigned to treatment through the assignment mechanism.

Although we might not end up estimating the effect of interest, we need to begin by defining our effect of interest (i.e., our estimand). This is what we then try to identify (compare Lundberg et al., 2021). Identification of an effect is separate from estimating it and is the logical first step (e.g., Heckman, 2005; Imbens and Rubin, 2015, chap. 1). However, identification is dependent on the data available to us,¹¹ on the causal model we have formulated for the effect of interest, and on the potentially useful estimators available to us. Identification of an effect that is empirically possible to estimate is, therefore, in practice, “somewhat circular or recursive” (Lewbel, 2019, p. 840). We will have to update the definition of the treatment of the experiment and, hence, also the estimated effect because of limitations due to the data, model, and estimator available to us.

This insight has been missing from the empirical natural experiments literature. With natural experiments, we cannot control the data-generating process. Thus, we need to update the design of our study, including the definition of the treatment and estimated effect, to match the data that are available to us. In the words of Haavelmo (1944, p. 14): “We try to choose a theory and a design of experiments to go with it, in such a way that the resulting data would be those which we get by passive observation of reality.”¹² This will (in almost all cases) mean that we have to acknowledge that we cannot estimate the effect that we are most interested in.

¹¹ Including how and why the data were generated and collected.

¹² Proponents of natural experiments have repeatedly quoted Haavelmo’s (1944) lament that few economists are explicitly defining the design of their studies, as well as his definition of a natural experiment (Angrist and Pischke, 2009, pp. 7–8; see also, Angrist and Krueger, 2001, pp. 80–81; Angrist and Pischke, 2010, p. 16). Just a few lines further down on the same page, Haavelmo (1944) discusses

This paper argues that there is also a lot to be gained from recognizing the substantive interpretation of the effect estimated by the LATE. The effect that is estimated by LATE is inherently different from the effect of interest and will, most often, be quite specific. I argue that we should acknowledge and utilize the fact that the LATE is different from the effect of interest rather than ignoring it.

Next, I will provide an example of how the empirical literature has failed in defining the treatment and interpreting the estimated effect.

Example 1. IVs based on twin births: Examples of how the treatment has been ill-defined in previous work

Studies of how children are affected by their number of siblings face problems because the number of children in the family is endogenous in the model. Using twin births as an IV for the number of children was introduced by Rosenzweig and Wolpin (1980a, 1980b). The method was later refined to focus on parity-specific twin births in samples of families that had at least a specific number of births (Angrist et al., 2005, 2010; Black et al., 2005).

It is typically claimed that IVs based on (parity-specific) twin births make it possible to estimate the “causal effect of family size” (e.g., Black et al., 2005, p. 670; Tan, 2019, p. 1171) or “the causal effect of fertility” (e.g., Bhalotra and Clarke, 2020, p. 3121) on child outcomes. This is an example of wishful thinking when defining the treatment of the experiment and the interpretation of the estimated effect.

the consequences of analyzing natural rather than controlled experiments. He points out that when we analyze natural experiments, we cannot change what is in the data. Therefore, we have to adjust what it is we think we can learn from the data. Tellingly, this section has, to my knowledge, never been quoted in the natural experiments literature.

Angrist, Lavy, and Schlosser (2010) have one of the best discussions on the effect estimated by twin birth IVs in the literature known to me. Nonetheless, the closest they come to an explicit definition of the treatment of the natural experiment is the statement that the “treatment is defined as a dummy for having another child” (Angrist et al., 2010, p. 788); in other words, the experiment’s treatment is “having another child”. At other places in the paper, they make claims that correspond to a different definition of the treatment. They, for example, also discuss how “the subpopulation of compliers consists of individuals who switch from having fewer than j to at least j children because of the instrument” (Angrist et al., 2010, p. 787). The clause “because of the instrument” acknowledges that some families would have had another child even without the twin birth. This acknowledgment is also reflected in their discussion of their first-stage coefficients, where they explain why these are lower than the coefficients estimated by Angrist and Evans (1998): “Multiple births result in a smaller increase in family size when families would have been large even in the absence of a multiple birth” (Angrist et al., 2010, p. 785). Bringing the desired number of children into the definition moves the interpretation in the direction of unintended children.

As we can see, it is somewhat unclear what Angrist et al. (2010) define the treatment to be. However, their de facto definition becomes clear from another statement in the paper. Angrist et al. (2010, pp. e.g., 776, 788; see also Angrist and Pischke, 2009, pp. 160–161) argue that IVs based on twin births are a special case among IVs because all families that experience a twin birth comply with their assigned treatment (i.e., there is “perfect compliance”). There will be perfect compliance in the twin birth IV case only if the treatment is defined as experiencing a twin birth instead of a singleton birth at the studied parity. In other words, there will only be perfect compliance with twin birth IVs if we (conf)use the assignment mechanism as the treatment. This definition of the treatment is far removed from an exogenous increase of the number of children in the family.

When we use (parity-specific) twin births as an IV for the number of children born in a family, we can estimate the effect from the exogenous increase in the number of children that the (parity-specific) twin birth creates in *some* families. If we, for example, use a twin birth as the second birth as the IV, we can estimate the effect of being “forced” to have a third, unintended child because of the twin birth in families that only intended to have two children.¹³ This is, admittedly, a more cumbersome definition of the estimated effect than the “causal effect of family size.” It has the crucial benefit of corresponding to the causal effect that is actually estimated when using these IVs.

¹³ Using (parity-specific) twin births as an IV for the number of children in a family requires that we rely on a large number of implausible assumptions (Öberg, 2021). For example, we need to assume that parents have an in advance fixed desired number of children (Rosenzweig and Wolpin, 1980a, p. e.g., 232). This is not a realistic assumption, but we still need to rely on it. Otherwise, the parents cannot be “forced” to have another, unintended child being born because of the twin birth. We further need to assume that having two children being born at once is no different from having two children born with some time in between. This is not a plausible assumption either for the children or the parents. Any economies or diseconomies of scale in caring for twins compared to children born with some time in between violates the exogeneity of this IV. Behavioral adjustments as a reaction to a twin birth also threaten the validity. There can also be no positive or negative selection into giving birth to twins (with fertility treatments as a special case). This assumption has also been questioned in a number of studies published recently. IVs based on twin births estimate the effect of having another, unintended birth. This is likely to be different from the effect of choosing to have another (i.e., intended) child and will limit the generalizability of the effects estimated using these IVs. See Öberg (2021) for a detailed discussion of IVs based on twin births.

Defining the treatment: requirements and tools

The causal effects estimated using the LATE estimator are, as mentioned, based on the definition of a causal effect in the potential outcomes framework. Hence, we also need to define the treatment of our natural experiments according to the requirements for a treatment in the potential outcomes framework. These requirements are clearly stated for example in the foundational texts on the potential outcomes framework by Donald Rubin.

“By a treatment we mean a series of well-defined actions that can be applied to a unit of study” (Rubin, 1978, p. 36).¹⁴ Thus, our definition of the treatment needs to specify what the treatment is, and how it is assigned to and received by units in our study population. In the twin birth IV example, the series of actions included in the treatment are having another, unintended child being born because of a twin birth. The definition of the treatment also needs to include any pretreatment events that are common to all units regardless of their assignment. In the twin birth IV example, all families have already had one child and are having (at least) one more.¹⁵

When we have defined the series of actions that constitute the treatment of the experiment, “we cannot attribute cause to one particular action in the series of actions that define a treatment” (Rubin, 1978, p. 40). The treatment of the experiment—that which we are estimating the effect of—is the treatment as

¹⁴ In the words of James Heckman (2005, p. 29): “A treatment generally consists of a package of characteristics and if we vary the package ... we get different treatment effects.”

¹⁵ It is useful to think about the effect as the difference in outcome between units receiving the treatment of the experiment (also called “active treatment”) and the units receiving the “control treatment”. Defining the control treatment can remind us to consider the experiences of the units that are “not treated.”

experienced by the units (see also Hernán and Robins, 2020, pp. 35–36). This might seem self-evident but, as we saw in the case of IVs based on twin births, it means that we need to be much more specific than is typically the case today. In the twin birth IV example, we cannot claim to be estimating the effect of having “another” child, not even having “another” child at the parity specified by the instrument. We cannot even claim to be estimating the effect of having another child born at the specified parity because of a twin birth. We are estimating the effect of having another, unintended child born at the specified parity because of a twin birth (given that we choose to accept all the other necessary assumptions).

When we specify a clear and careful definition of the treatment, we will always end up having a somewhat complex definition. Ignoring the requirement for a clear and careful definition of the treatment does not make this complexity go away. It only leads to erroneous interpretations of the estimated effect and threatens the internal validity of the method. In the words of Rubin (1975, p. 234): “we need to be terribly precise in the definitions of treatments in order to avoid deceiving ourselves.”

Hernán (2016) and Hernán and Robins (2020) discuss how a well-defined treatment of the experiment and a correct interpretation of its effect increase the differences compared to the treatment and effect of interest. They remind us that this difference is, in fact, “a good thing” and continue: “Refining the causal question, until it is agreed that no meaningful vagueness remains, is a fundamental component of causal inference” (Hernán and Robins, 2020, p. 34).

We can never know that we have been specific enough in our definition of the treatment (Hernán, 2016; Hernán and Robins, 2020, p. 34). We might learn later that there are other relevant, overlooked factors that create relevant variations in the treatment effect. What we can and should do is to be transparent with the definition that we choose and argue why this definition is sufficient for making any remaining variations of treatment irrelevant.

One useful tool for specifying clear and careful definitions of the treatment for natural experiments is to think in terms of “manipulability”. In the potential outcomes framework, and when estimating a LATE, “there exists no causation without manipulation” (Rubin, 1975, p. 238, see also, e.g., 1978, pp. 39–40). What we define as the treatment needs to be something that can be manipulated, or changed, by the assignment mechanism. When we are estimating a LATE, the need for manipulability is, therefore, inherent to the method.¹⁶

To estimate well-defined causal effects, we must rely on the stable unit treatment value assumption (SUTVA, see, e.g., Imbens and Rubin, 2015, pp. 9–12, 517, 589; Hernán and Robins, 2020, chaps. 1–3). The SUTVA includes two parts (Cox, 1992[1958], pp. 17–21; Rubin, 1990, p. 475; Heckman, 2005, pp. 11–12, 35–36, 43; VanderWeele and Hernan, 2013; Small et al., 2017). The first part requires that there should be no interference between the studied units; the treatment of one unit should, therefore, not affect another. This is also stated as there being no “equilibrium effects” (e.g., Heckman, 2005, p. 11).

The second part of SUTVA requires that there be only one version of the treatment; “there are no versions of treatments” (Rubin, 1980, p. 591). Cox (1992[1958], p. 17) termed this assumption “constancy of treatment effects.” It is, as mentioned, also called the consistency assumption.¹⁷

¹⁶ The need for manipulability is, for the same reasons, inherent also in, for example, regression discontinuity and difference-in-difference designs.

¹⁷ Hernán and Robins (2020, p. 26) summarize the consistency assumption like this: “the values of treatment under comparison correspond to well-defined interventions that, in turn, correspond to the versions of the treatment in the data”.

Thus, another indispensable tool for specifying a clear and careful definition of the treatment for natural experiments is to evaluate whether or not there are versions of the treatment. There cannot be any relevant versions of the treatment we define. We can only have versions of the treatment if these versions have the same effect on the outcome (i.e., constant treatment effects on the treated).¹⁸ If the effect of the treatment on the outcome is not the same for the different versions of the treatment, then the estimated effect is “an artificial quantity” without any clear interpretation (Cox, 1992 [1958], pp. 15–19). The estimated effect does not necessarily correspond to the average effect across the groups receiving the different versions of the treatment. Cox (1992 [1958]) provides simple numerical examples illustrating that in the presence of different versions of the treatment, the estimated average effect does not even have to be in the range of the effects for the different versions of the treatment.¹⁹

Next, I will provide an example of a definition of the treatment that leads to there being versions of the treatment that are relevant to the outcome.

¹⁸ We can have versions of the treatment that have different effects on the outcome if 1) we know which units receive which version, or 2) we have a complete and true causal model of what is determining which version of the treatment a unit receives as well as perfectly measured variables for all factors included in this model that we can use to adjust our estimates (VanderWeele and Hernan, 2013). Both of these situations will be exceedingly rare in practice.

¹⁹ The problems created by having versions of the treatment are similar to the issues created by misclassification (i.e., measurement error of the treatment) (Edwards et al., 2015). But, when we have an insufficiently specific definition of the treatment, it is the meaning and interpretation of the estimated effect that becomes uncertain rather than the point estimate.

Example 2. IVs based on the US draft lottery: An example of the presence of relevant versions of the treatment

Studies on how serving in the military affects later life earnings face problems because people choose to serve or not partly based on their civilian earnings potential. Joshua Angrist (1990) proposed using the US Vietnam-era draft lottery as an IV to obtain unbiased estimates despite issues with self-selection. There have since been a large number of studies following up on Angrist's (1990) results.

The draft lottery IV allows us to estimate the “effects of military service on men who served because they were draft eligible but who would not otherwise have served” (Angrist and Pischke, 2009, p. 155). The potential outcomes motivating a causal interpretation is that we are comparing the later life earnings of a person after having served in the military with his later life earnings after not serving in the military. Both of these outcomes must be hypothetically possible. We are, therefore, estimating the effect for physically and mentally fit men that did pass, or would have passed, the inspections before and after induction. Because both potential outcomes must be hypothetically possible, we are also estimating the effect for persons who were induced to serve because of being draft eligible. This excludes volunteers that served regardless of their draft eligibility. Any changes to behavior as an attempt to avoid serving because of being draft-eligible will cause problems for the analysis (e.g., Deuchert and Huber, 2017).

The definition of the treatment, in this case, is serving in the military because of being draft-eligible (with additional, context-specific specifications). When we think more closely about the differences among the men who “served because they were draft eligible but who would not otherwise have served”, we realize that there will be versions of this treatment that are relevant for the outcome.

The effect on a person's later life earnings (the outcome) from interrupting civilian opportunities by serving in the military will depend on what the person's civilian opportunities and earnings potential were. Among the men who served because they

were draft-eligible will have been men with different civilian opportunities and earnings potential. Some men perceived better opportunities in the civilian labor market than in the military and were reluctant to serve. These men were forced to serve against their will because of being draft-eligible—they are “forced” compliers. Other men were more or less indifferent towards their opportunities in the military and the civilian labor market—they are “indifferent” compliers. They still served because they were draft eligible but would not otherwise have served. Thus, we have two groups with different expected effects of the treatment.

Adding to the difficulties is the fact that the compliers will include some volunteers. Some men volunteered because they were draft eligible to improve their terms of service (Angrist, 1991). As long as they would not have volunteered without being draft eligible, they are also compliers (if we define the treatment of the experiment as having served in the military because of being draft eligible). The volunteer compliers could be expected to have yet another expected effect of the treatment.²⁰

In this case, we therefore have a range of treatments: serving in the military with different civilian opportunities and earning potentials. Having a range of treatments has the consequence of us estimating an unreliable effect that we cannot interpret.

Defining the treatment: The reception mechanism

When we are analyzing natural experiments in which the assignment mechanism affects different units differently (i.e., we have heterogenous effects of the instrument on the treatment), we need to consider how and why this is the case. We

²⁰ If we exclude the volunteers from the analysis, we introduce selection bias by selecting on the treatment (Swanson et al., 2015; Swanson 2019).

need to have, at least, a conceptual idea of “why observationally identical people make different choices” (Heckman, 2005, p. 5).²¹

I refer to the how and why the assignment mechanism affects different units differently as the “reception mechanism”. It is really a part of a well-defined treatment (i.e., a part of the “series of well-defined actions” that make up the treatment). It is still worth thinking about it as a separate part of the treatment because it is kept separate in the alternative structural approach, the selection model. Vytlacil (2002) and Kline and Walters (2019) show that estimating a LATE is equivalent to estimating a structural “control function” model (such as a Roy selection model) in which the change in treatment is modeled using an index variable and a threshold. The defining feature of such a model is that it explicitly models the reception mechanism.

The assignment mechanism in natural experiments in which the assignment mechanism affects different units differently can provide an as-good-as-random assignment. However, the assignment will only be as good as random for the subpopulation determined by the reception mechanism. To evaluate which group has had their treatment changed by being assigned to treatment (i.e., the compliers), we must determine how and why the assignment changes the treatment. This how and why determines which units are compliers in the (natural) experiment. The reception mechanism is, therefore, how we link the treatment to the compliers.

The reception mechanism for natural experiments incorporates two parts. One part determines which units do receive the treatment they are assigned to and which do not. For this part, we need to consider the reaction and behavior of the units. Even if we choose not to model the reception mechanism structurally, we must consider “the

²¹ Another way of putting it is: “A good instrument is correlated with the endogenous regressor for reasons the researcher can verify and explain” (Angrist and Krueger, 2001, p. 73).

incentives and restrictions faced by units assigned and not assigned to receive the treatment” (Imbens and Rubin, 2015, p. 550).²²

The other part of the reception mechanism for natural experiments determines which units are eligible to be assigned to treatment and which not. This part of the reception mechanism is similar to “the sampling mechanism which determines the experimental units to be studied” in a controlled experiment (Rubin, 1978, p. 37). In this part, we need to consider which units could potentially react to the assignment and which could not.

Only when we have clear and careful definitions of the treatment of the experiment, including the reception mechanism, can we truly evaluate the assumptions necessary for estimating a causal effect using the LATE (relevant IV, independence, exogeneity, monotonicity, and common support) (Angrist et al., 1996; Hernán and VanderWeele, 2011, p. 373; Imbens and Rubin, 2015, p. 550; Morgan and Winship, 2015, chap. 9; Huber and Wüthrich, 2019; Hernán and Robins, 2020, p. 38; Imbens, 2020, p. 1157) and interpret our results correctly.

The indispensable need for a clear and careful definitions of the treatment when using IVs has the consequence that we must consider all IVs as causal. “Non-causal IVs” are just poorly understood causal IVs. If we do not understand how, why, and for who the IV causes a change of the treatment, we can not evaluate the necessary assumptions. We can therefore not claim that we are estimating a reliable effect. We also can not interpret the estimated effect (e.g., Swanson and Hernán, 2018; Hernán

²² Cole and Frangakis (2009, p. 3) provide a useful tool for thinking about this issue: “Methods for causal inference require that the exposure is defined unambiguously. Specifically, one needs to be able to explain how a certain level of exposure could be hypothetically assigned to a person exposed to a different level.”

and Robins, 2020, p. 203). In conclusion, we can only use “non-causal IVs” if we are content with claiming that an unreliable and uninterpretable effect is causal.²³

Defining the compliers

The definitions of the treatment, including the reception mechanism, define which units are the compliers in the specific design of our natural experiment. This fact has been mentioned in the context of if and how we can compare LATEs estimated using different IVs. It is, in theory, well known that different IVs will have different compliers, which, in turn, will lead to estimates with different values, meaning, and interpretation (e.g., Heckman, 1997, pp. 451–452; Angrist and Pischke, 2009, p. 166; Angrist and Fernández-Val, 2013, p. 401; Lewbel, 2019, p. 870). But, as mentioned, I think there has been confusion regarding the possibility of knowing which units the compliers are conceptually and empirically. Angrist and Pischke (2009, p. 166), for example, write that we would “like to learn as much as we can about the compliers” but do not stress the importance for the research design of specifying clear and careful conceptual definitions of who they are and why.

Because we are estimating the effect for the group that has their treatment level changed by being assigned to treatment, we are estimating the effect of that change. When we are estimating a LATE, we are, therefore, always estimating the effect of a change.²⁴ Notably, the effect of a change is different from the effect of a level of the

²³ In the words of Hernán and Robins (2020, p. 39): “The problems generated by unspecified treatments cannot be dealt with by applying sophisticated statistical methods. All analytic methods for causal inference from observational data ... yield effect estimates that are only as good as the treatments that are being compared.”

²⁴ Imbens and Angrist (1994, p. 470) compare the LATE to a fixed-effects estimator because both estimate the effect of change.

treatment. The change in treatment is the result of the mechanisms through which the instrument affects the treatment of interest (i.e., the assignment and reception mechanisms of the treatment). These mechanisms define the specific treatment for which we estimate the effect.

To satisfy the exclusion restriction, the assignment to treatment (through the IV) should be the only thing that makes the compliers change their treatment or not. The idea behind natural experiments is that this kind of assignment mechanism is created by random events or non-universal, inevitable changes. IVs are typically based on random events, such as twin births, the sex of the first-born children, or draft lottery numbers. A valid assignment mechanism (i.e., a valid IV) should cause a change in the treatment that is independent of the potential outcomes of the outcome studied (e.g., Heckman, 1997, p. 452; Imbens and Rubin, 2015, p. 520). Using the LATE, we then estimate the effect for the units that have their treatment changed by this random event.

Imbens and Rubin (2015, p. 528) note that, when we estimate a LATE, even if the assignment mechanism needs to be independent of the potential outcomes, the change in the treatment does not have to be achieved independently of the potential outcomes. This means that the reception mechanism can include the expectations and preferences of the units so that compliers choose their treatment after assignment. I argue that, while the LATE estimator (quite naturally) will produce results also when the change in the treatment is not independent of the potential outcomes, there is no longer any support for claims of estimating a reliable and interpretable causal effect. Ebenstein (2009) and Goetghebeur et al. (2020) both reach a similar conclusion.

It is not enough to have a source of random variation. One needs to consider how the units react to this random variation. The different ways units react to an exogenous, random event is not automatically random (or exogenous) just because the event that initiated the reaction is. Therefore, we also need to have a clear conceptual idea of the reasons how and why the compliers react to the assignment (i.e., how and why

they change their treatment according to their assignment).²⁵ We cannot have versions of the treatment. Thus, we need to assume that all compliers react for the same reasons (i.e., we must assume a constant reception mechanism). We can allow for different reception mechanisms only if we assume that the reason why the assignment mechanism affects the treatment does not matter for the outcome.

There are several alternatives for how we can define how and why the compliers change their treatment according to their assignment. I argue that they can be usefully summarized in three categories. Either the compliers are a) indifferent towards the potential outcomes regardless of the change in the treatment (i.e., they should not have any expectations or preferences regarding the outcomes with or without the treatment), or b) not indifferent but are unable to act on their expectations or preferences (i.e., they are being forced to change the treatment), or c) are self-serving and choose their treatment based on their expectations or preferences. We will, thus, have compliers that are indifferent, forced, or self-serving, but we still need to get to know them well!

The compliers in the twin birth IV example are forced to change their treatment because of their assignment. Some of the families that only intended to have two children are “forced” to have a third, unintended child because of the twin birth. The compliers in the US draft lottery example were induced to serve either by force or were indifferent towards serving or not. Therefore, we have relevant versions of the treatment if we do not make the implausible assumption that it does not make a difference for their long-term income if they were forced or indifferent compliers.

²⁵ This is well-known in theory in epidemiology (e.g., Craig et al., 2017; Maciejewski and Brookhart, 2019, p. 2125). But it is not implemented in practice in epidemiology either as is demonstrated by a wide-spread use of “non-causal instruments” (see Swanson and Hernán, 2018, for a discussion).

Below, I will discuss the example of IVs based on the sex of the first-born children and how the compliers then can be characterized as self-serving.

Example 3. IVs based on the sexes of the first-born children: An example of self-serving compliers

Studies on how the number of children affects the parents face problems because the number of children in the family is endogenous in the model. Angrist and Evans (1998) proposed using the sexes of the first two children as an IV because the sex of a child born is as good as random for the parents. The reason this approach could work as an IV for the number of children is that some parents prefer having children of both sexes. Having the first two children born being of the same sex will, therefore, make some of these parents have another child to increase the probability of having children of both sexes (as demonstrated by, for example, Bisbee et al., 2017). This IV is defined by parity, just as the twin birth IVs. We, for example, study the effect of having a third child after having the first two children being of the same sex.

The intended assignment mechanism—the sexes of the first two children—is as good as random. The reaction of the parents to this assignment is not. By thinking closely about how and why the parents react the way they do, we can easily see that this creates issues for any claim of estimating a causal effect.

Because the parents choose to have a third child or not based on the sex mix of the first two children, the compliers are not forced in this case. It is not unimaginable that some parents are indifferent towards having another child (with the chance of having children of both sexes) or not. But it does seem more likely that the parents make a self-serving choice based on their expectations and preferences.

It is obvious that it is only the parents that prefer having children of both sexes that will react to this assignment mechanism.²⁶ If we think about their choice regarding having a third child or not as a benefit-cost valuation, we can conclude that the parents that choose to have a third child will have a strong preference for having children of both sexes, and/or expect other benefits from having a third child, and/or expect low costs of having a third child.

The strong preference for having children of both sexes makes the compliers in this case different from many other parents. But the fact that they needed only their first two children to be of the same sex as the reason to choose to have a third child makes them similar to parents that choose to have three children regardless of the sexes of their first two children. This approach can, therefore, increase the external validity of the effect estimated for these compliers.

The downside of having self-serving compliers is that it is unclear how the IV estimates are an improvement compared to a naïve OLS. The reason for using an IV is that we expect confounding from factors related to the parents' expectations and preferences affecting both their labor force participation and the number of children they choose to have (e.g., Angrist and Evans, 1998, p. 451). Effects estimated using IVs based on the sexes of the first two children incorporate rather than solve this issue (see also Ebenstein, 2009, e.g., p. 974). The grounds for claims of estimating a causal effect are, therefore, weak in this case. (See Öberg [2021] for further discussion of this IV.)

²⁶ Huber (2015) discusses that we cannot rule out the presence of defiers when using an IV based on the sexes of the firstborn children because it is not unthinkable that there are parents that want two children of the same sex and therefore go on to have another child if the first two are of different sexes.

Defining the causal models

Specifying clear and careful definitions of the treatment (including the reception mechanism and the compliers) is important for yet another reason. It allows us to consider the causal models underlying our estimator. If we truly trust our IV to be independent and exogenous, we do not need to adjust our estimates for any other variables. With a perfect IV, we can estimate the causal effect (including the LATE) by dividing the effect of the instrument on the outcome by the effect of the instrument on the endogenous variable using the Wald estimator (e.g., Angrist, Imbens, and Rubin, 1996; Morgan and Winship, 2015, chap. 9).

However, we always need to specify the causal model underlying our research design. Specifying the model is the only way we can evaluate the IV and check for sources of bias. Swanson (2019) describe a number of potential sources of selection bias when using IV methods that we can only evaluate once we have specified the model. Gallen and Raymond (2021) raise another potential problem that require that we specify the underlying causal model before evaluating our IV. Gallen and Raymond highlight that a widespread use of an instrument increases the risk of violations of the exclusion restriction. Claims that an instrument is relevant and valid for different endogenous variables and/or different outcomes make it unlikely that it is in fact valid in some (or all) of these cases.²⁷ The different endogenous variables and outcomes are likely to be closely related when they are analyzed using the same IV. If the IV is indeed relevant for one variable this will have to be considered in the model of its influence on another, related variable. Gallen and Raymond show how such models, with plausible interrelatedness of variables, quickly become impossible to use as the basis for an IV estimation model. When an IV is used for many different variables and/or outcomes, this creates a paradoxical

²⁷ Morck and Yeung (2011:49–50) call this the “econometric tragedy of the commons”. Bazzi and Clemens (2013:154–155) also discuss this issue.

situation where most claims of the IV being relevant and valid must be wrong or else all claims are wrong.

Almost all applications using IVs adjust the models for other variables. One reason for including covariates in the models is that this may reduce the variability of the dependent variable (Angrist and Pischke, 2009, p. 176). But the most important reason is that we might think that “the conditional independence and exclusion restrictions underlying IV estimation may be more likely to be valid after conditioning on covariates” (Angrist and Pischke, 2009, p. 175, see also 62, 122). If we do choose to include other variables in the models, this means that we change one unverifiable assumption for another; we substitute perfect independence and exogeneity with the assumption of no unmeasured confounding (Hernán and Robins 2020, p. 205). Adjusting the IV models for other variables thereby brings back some of the issues we hoped to avoid when using the IV in the first place.

When we estimate a LATE, we need to be able to consider the results from both the first and the second stage of the two-stage least squares to be causal estimates (e.g., Angrist et al., 1996, p. 446; Rosenzweig and Wolpin, 2000, pp. 841, 872; Angrist and Pischke, 2009, pp. 151–152). This requires that both models are good enough approximations of the true causal model underlying the processes we are studying (e.g., Deuchert and Huber, 2017). Claiming that we can specify and estimate good enough approximations of the true causal models is not an assumption we can make lightly.

Specifying a good enough causal model includes choosing the right variables to adjust the model for.²⁸ We can only decide on which factors could be confounding the association once we have clear and careful definitions of the treatment, the

²⁸ It also includes choosing an adequate functional form. The assumption of a linear effect from the number of siblings on the older child(ren) has, as an example, worked to hide the true effect (Mogstad and Wiswall, 2016).

reception mechanism, and the compliers (i.e., we have clear ideas of what exactly the treatment is and which units it affects, how, and why) (Hernán and VanderWeele, 2011; Hernán, 2016, p. 678; Kaufman, 2019).

It is self-evident that we need the second stage model of the effect of the treatment on the outcome to be a good enough approximation of the true causal model to consider the estimated results causal. However, although it is known in theory, it has not been the practice to think about the first stage as also being a causal model that we need to consider carefully. The routines used in statistical software programs are designed to include the variables from the second stage in the first stage while simply adding the IV(s). This practice does not guarantee that the first-stage model is a good enough approximation of the true causal model of the effect from the assignment mechanism on the share of the population that is treated.

Angrist and Pischke (2009, pp. 189–190) highlight not including the same covariates in the first and second stage as a “mistake”. They state that “[i]f a covariate is good enough for the second stage, it’s good enough for the first” (Angrist and Pischke, 2009, p. 190). This statement is misleading, considering that the goal is to formulate a good enough approximation of the true causal model for the first stage as well (e.g., Angrist and Pischke, 2009, p. 152).²⁹

²⁹ Angrist and Pischke (2017) provide another illustration of their relaxed attitude towards specifying a causal model. They do acknowledge that the problem of choosing the variables to adjust for “is one of the most challenging in empirical practice” (p. 134). But they also, again, recommend a mechanical, non-causal, view of the issue: “As far as the control variables go, regressor-residual orthogonality is *assured* rather than assumed; that is, regression algebra makes this happen. At the same time, while the controls are surely uncorrelated with the residuals, it’s unlikely that the regression coefficients multiplying the controls have a causal interpretation. We don’t imagine that the controls are as good as

The perceived benefit of including the same variables in the first and second stage models is that it guarantees that their residuals will not be correlated. This mechanically guaranteed “independence” of the residuals is not sufficient if we want to specify causal models to estimate causal effects. We need the residuals from the first and second stage models to be independent when both models are good enough approximations of the true causal models (e.g., Rosenzweig and Wolpin, 2000; Deaton, 2010).³⁰ That might require including different variables in the different stages.

We can learn about the IV, the population studied, as well as the compliers from the first stage results (e.g., Morgan and Winship, 2015, chap. 9). However, what we learn will only be true if the first stage model is a good enough approximation of the true causal model.

The first stage is modeling the difference in the average level of the treatment in the group indicated by the instrument compared to the group not indicated. Because we have a binary instrument, we also have a design with a binary treatment of the experiment. The difference in the average can, therefore, be interpreted as the difference in the share treated among units assigned to treatment and not (i.e., the share treated among units indicated by the instrument minus the share treated among units not indicated by the instrument).

Because we are modeling a difference in the share treated, the first stage model is not necessarily a model of the behavior of the units. We still need to consider the behaviors of the units assigned to treatment or not (i.e., the reception mechanism) to

randomly assigned and we needn’t care whether they are” (Angrist and Pischke, 2017, p. 132).

³⁰ Deaton (2010) discusses this as the difference between external and exogenous factors.

be able to choose a good enough specification of the model. The shares treated will depend on the behavior of the units; hence, this behavior will determine which covariates we need to adjust for.

The natural experiments literature has, “[e]xcept for warnings about including covariates that are themselves influenced by the treatment, ... not been very helpful” with regards to specifying causal models and choosing covariates (Imbens and Wooldridge, 2009, p. 50). Imbens (2020, pp. 1163–1166) discusses this issue but concludes that it is not a substantial problem. However, in line with my argument here, he writes that: “In practice the confusion often concerns what the treatment is and what that implies for the classification as pre- or posttreatment variables” (Imbens, 2020, p. 1164).

Deuchert and Huber (2017) and Lewbel (2019, p. 873) suggest using directed acyclic graphs (DAGs, e.g., Pearl and Mackenzie, 2018) when deciding on which variables to include (see also Swanson, 2019; Gallen and Raymond, 2020; see Imbens, 2020, for a discussion of the usefulness of DAGs). To me, it is obvious the tools for formulating a causal model and choosing covariates that are presented in, for example, Pearl and Mackenzie (2018; see also Morgan and Winship, 2015) are a clear improvement over the discussion in, for example, Angrist and Pischke (2009, chap. 3).

It should be noted that the decisions on the specification of the model (including the choice of variables to adjust for) need to be made before we begin estimating anything regarding the outcome (e.g., Rubin, 1974, p. 699; Imbens and Rubin, 2010[2008], p. 230). Brodeur et al. (2020) find indications that this is not always

how it is done today; they find clear signs of so-called *p*-hacking in the natural experiments literature, not the least in studies using IVs.³¹

Other, more general, issues that are also important to consider when estimating causal effects include checking for pretreatment balance or common support of the variables we include in our models (e.g., Huber and Wüthrich, 2019)³² and considering the measurement error in our variables. Measurement error seriously limits our ability to adjust our models, for example, to remove confounding (Phillips and Smith, 1993; Shadish et al., 2002, p. 49; Gelman, 2011; Schennach, 2016, pp. 344–345; Westfall and Yarkoni, 2016; Loken and Gelman, 2017; Hernán and Robins, 2020, chap. 9; see also Pei, Pischke, and Schwandt, 2019), and these problems arise even if only one or some of the variables have measurement errors (Shear and Zumbo, 2013).

Another general issue, closely related to the core argument of this paper, is that “the interpretation of what has been estimated and for whom depends on how one models behavior” (Rosenzweig and Wolpin, 2000, p. 841). Adjusting an estimate for another variable changes the substantive interpretation of the estimate.

³¹ Practices that lead to *p*-hacking include “continuing to collect data, *strategically selecting covariates*, or imposing sample restrictions until a significance threshold is met” (Brodeur et al., 2020, p. 3635n1, emphasis added).

³² Atanasov and Black (2021) provide examples from published studies of how the analysis can produce misleading results if we do not have common support for the included variables. They indeed show that the design of the studies they analyzed are invalidated when they limited the samples to the units having common support in the pretreatment variables.

Discussion

The empirical natural experiments literature, not the least in the social sciences (including economics), have overlooked the necessity of clear and careful (conceptual) definitions of the treatment, the reception mechanism, and the compliers. This has had the consequence that most studies have misinterpreted and over-generalized the substantive meaning of their results. This is a huge problem which can only be solved by reevaluating previous results and improving the conceptual definitions in future studies.

Others have recently made similar calls for improving the ways we estimate causal effects specifically and use quantitative methods in general. Gibson (2019) shows that several of the issues discussed here are not unique to the natural experiments literature. He asks us to consider whether we are “estimating the right thing”. He provides many examples where published papers report misinterpretations of their own results. He reminds us that we need to choose an estimator and model specification that actually provides the answers we seek. We can only make this choice if we carefully consider the connection between the estimator and the data by providing substantive definitions of the arguments and a correct interpretation of the estimates.³³

Lundberg et al. (2021) argue that sociologists need to get better at defining their research questions and their effects of interest. One of the main reasons why this is important is to enable a transparent discussion of the differences between the effect of interest and the effect estimated in the study. The empirical natural experiments literature is oftentimes clear on what the effect of interest is but has failed to acknowledge the difference between the effect of interest and the effect of the

³³ Gibson (2019) also shows that researchers often do not know enough about the data they use—how they were collected, what the variables are measuring, measurement errors etc.—and, therefore, fail to analyze them correctly.

experiment. Hence, although I fully sympathize with what Lundberg et al. are trying to achieve, I do not think their solution is sufficient for all disciplines.

Esterling, Brady, and Schwitzgebel (2021) show that internal validity is not sufficient for estimating useful and reliable causal effects. They argue convincingly as to why the empirical natural experiments literature must begin considering both construct and external validity to claim to estimate credible causal effects. Bo and Galiani (2021) also argue that more attention should be paid to external validity. They define external validity as constant treatment effects across populations and propose a mathematical test for testing this.

I argue that the key to improving causal estimates is to improve the conceptual definitions and substantive interpretations. Researchers doing applied, empirical work cannot rely on the mathematical clarity and inherent logic of the estimators we use (compare Dawid, 2000b, p. 446; Greenland, 2017, p. 16). We must carefully determine the correspondence between the estimator and the empirical aspects we are investigating. This process involves having a clear understanding of “what real world quantities are represented by variables appearing in the model” (Dawid, 2000a, p. 407).³⁴

We also need to check whether the assumptions required for the estimator we choose are reasonable given the data we have access to. This means connecting empirical, substantive meaning to the arguments of the equations. Using abstractions or vague attributions to label the arguments of the estimator does not help us to

³⁴ Jeffrey Wooldridge (2020 [2018], p. 40) provides us with an inspirational quote: “Whereas the mechanics of simple regression do not depend on how y and x are defined, the interpretation of the coefficients does depend on their definitions. For successful empirical work, it is much more important to become proficient at interpreting coefficients than to become efficient at computing formulas.”

clarify the substantive meaning of the assumptions. To clarify the assumptions, we need clear and careful, substantively meaningful definitions.

To be able to evaluate all necessary assumptions, we first need to know all the necessary assumptions. It can be difficult to realize the less obvious assumptions that can be built into an estimator. The notation of equations will often in itself imply assumptions (VanderWeele, 2009, p. 881) that can be particularly difficult to realize for an applied researcher that is not a specialist in econometrics. It is, therefore, important that the empirical examples presented by specialist econometricians describe all assumptions, including the implied, less obvious ones. It is also essential that they provide careful examples of the correct interpretation of the results, including how the estimator in itself changes this interpretation.

However, researchers doing applied, empirical work cannot rely solely on the examples of specialist econometricians. We need to understand the methods we apply well, at least their fundamental principles. We can rely on the mathematical clarity and inherent logic of the estimators we use in the sense that the methods work. What we must understand is why (i.e., the assumptions made and the requirements on the data) and how (i.e., how the data are used and what that means for the substantive interpretations).

What is always important to remember is that although a clever estimator can help us tease out information from data, it can and should never create information that is not already in the data. Holland (1986) provides an early summary of the potential outcomes framework. In it, he writes that the empirical samples of units studied and the empirical variables measured on these units are the primitives of the potential outcomes framework (Holland 1986, p. 945). Whereas Holland's statement is not true for the potential outcomes framework in general, it is true for any one empirical study defining a causal effect using the framework by, for example, estimating a LATE.

Holland's statement is, therefore, useful for reminding us that any empirical causal effect is defined and estimated based on the empirical data available to us. The effect estimated by a (natural) experiment is only informative about the effect of interest to the extent that the empirically measured variables correspond to the outcome and treatment of interest (see also Esterling et al., 2021). Whereas this issue can be less problematic in controlled experiments, applications using observational data (e.g., natural experiments) need to carefully describe how the empirically measured variables relate to the outcome and treatment of interest. To assume that the effect estimated by the experiment is the effect of interest, we need to assume that the empirical variables are perfect measures of the outcome and treatment of interest. We must also assume that the estimator has no effect on the interpretation of the estimated effect. The empirical natural experiments literature has made both of these assumptions, but neither assumption corresponds to reality.

Future empirical natural experiment studies need to include detailed definitions of all aspects of the study design, not least the treatment and reception mechanism. They should also interpret the estimated effects correctly. If these changes are made, these studies will produce useful and reliable results. But, as we have seen, improving the way we carry out and present our natural experiments will reveal that the estimated effects are much more specific than has been claimed so far. This revelation will, quite naturally, lead to a change in the role of results from natural experiments in the literature. We can still learn a lot from credible, specific results. The specificity of an estimated LATE can be "beneficial because of its targeted clarity" (Morgan and Winship, 2015, p. 315) because the compliers can give us useful insights into the behaviors and reactions underlying the processes that we are studying.

There will be many other benefits from improving how we describe our research designs, and how subtle differences can have important consequences for the results and how they should be interpreted. Clear and detailed definitions and accurate interpretations of results will improve opportunities for scientific discussion and

progress. Transparency about differences in design and interpretation of results could also help reduce confusion when results are being communicated to parties outside academia.

Funding:

The author gratefully acknowledges financial support from the Swedish Research Council (grant 2015-00961, PI: Christer Lundh).

Conflict of interest:

The author confirms that there are no conflicts of interest to report.

References

- Angrist, J.D., 1990. Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *American Economic Review* 80(3), 313–336. <https://www.jstor.org/stable/2006669>
- Angrist, J.D., 1991. The Draft Lottery and Voluntary Enlistment in the Vietnam Era. *Journal of the American Statistical Association* 86(415), 584–595. <https://doi.org/10.1080/01621459.1991.10475083>
- Angrist, J.D., Evans, W.N., 1998. Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *American Economic Review* 88(3), 450–477. <http://www.jstor.org/stable/116844>
- Angrist, J.D., Fernández-Val, I., 2013. ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework. In: *Advances in Economics and Econometrics Tenth World Congress*, Acemoglu, D., Arellano, M., Dekel, E. (eds.), pp. 401–434. Cambridge University Press.

- Angrist, J.D., Imbens, G.W., 1995. Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. *Journal of the American Statistical Association* 90(430), 431–442.
<https://doi.org/10.1080/01621459.1995.10476535>
- Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 91(434), 444–455. <https://doi.org/10.2307/2291629>
- Angrist, J.D., Krueger, A.B., 2001. Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives* 15(4), 69–85. <https://doi.org/10.1257/jep.15.4.69>
- Angrist, J.D., Lavy, V., Schlosser, A., 2005. *New Evidence on the Causal Link Between the Quantity and Quality of Children* (Working Paper No. 11835). National Bureau of Economic Research. <https://doi.org/10.3386/w11835>
- Angrist, J.D., Lavy, V., Schlosser, A., 2010. Multiple Experiments for the Causal Link between the Quantity and Quality of Children. *Journal of Labor Economics* 28(4), 773–824. <https://doi.org/10.1086/653830>
- Angrist, J.D., Pischke, J.-S., 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton.
- Angrist, J.D., Pischke, J.-S., 2010. The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives* 24(2), 3–30.
<https://doi.org/10.1257/jep.24.2.3>
- Angrist, J.D., Pischke, J.-S., 2017. Undergraduate Econometrics Instruction: Through Our Classes, Darkly. *Journal of Economic Perspectives* 31(2), 125–144. <https://doi.org/10.1257/jep.31.2.125>

- Atanasov, V., Black, B., 2021. The Trouble with Instruments: The Need for Pretreatment Balance in Shock-Based Instrumental Variable Designs. *Management Science* 67(2), 1270–1302. <https://doi.org/10.1287/mnsc.2019.3510>
- Bazzi, S., Clemens, M.A., 2013. Blunt Instruments: Avoiding Common Pitfalls in Identifying the Causes of Economic Growth. *American Economic Journal: Macroeconomics* 5(2), 152–186. <https://doi.org/10.1257/mac.5.2.152>
- Bhalotra, S., Clarke, D., 2020. The Twin Instrument: Fertility and Human Capital Investment. *Journal of the European Economic Association* 18(6), 3090–3139. <https://doi.org/10.1093/jeea/jvz058>
- Bisbee, J., Dehejia, R., Pop-Eleches, C., Samii, C., 2017. Local Instruments, Global Extrapolation: External Validity of the Labor Supply–Fertility Local Average Treatment Effect. *Journal of Labor Economics* 35(S1), S99–S147. <https://doi.org/10.1086/691280>
- Black, S.E., Devereux, P.J., Salvanes, K.G., 2005. The More the Merrier? The Effect of Family Size and Birth Order on Children’s Education. *The Quarterly Journal of Economics* 120(2), 669–700. <https://doi.org/10.1093/qje/120.2.669>
- Bo, H., Galiani, S., 2021. Assessing External Validity. *Research in Economics* 75(3), 274–285. <https://doi.org/10.1016/j.rie.2021.06.005>
- Brodeur, A., Cook, N., Heyes, A., 2020. Methods Matter: *p*-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review* 110(11), 3634–3660. <https://doi.org/10.1257/aer.20190687>

- Cole, S.R., Frangakis, C.E., 2009. The Consistency Statement in Causal Inference: A Definition or an Assumption? *Epidemiology* 20(1), 3–5.
<https://doi.org/10.1097/EDE.0b013e31818ef366>
- Cox, D.R., 1992 [1958]. *Planning of experiments*. John Wiley, New York.
- Craig, P., Katikireddi, S.V., Leyland, A., Popham, F., 2017. Natural Experiments: An Overview of Methods, Approaches, and Contributions to Public Health Intervention Research. *Annual Review of Public Health* 38, 39–56.
<https://doi.org/10.1146/annurev-publhealth-031816-044327>
- Dawid, A.P., 2000a. Causal Inference without Counterfactuals. *Journal of the American Statistical Association* 95(450), 407–424.
<https://doi.org/10.1080/01621459.2000.10474210>
- Dawid, A.P., 2000b. Causal Inference without Counterfactuals: Rejoinder. *Journal of the American Statistical Association* 95(450), 444–448.
<https://doi.org/10.1080/01621459.2000.10474218>
- Deaton, A., 2010. Instruments, Randomization, and Learning about Development. *Journal of Economic Literature* 48(2), 424–455.
<https://doi.org/10.1257/jel.48.2.424>
- Deuchert, E., Huber, M., 2017. A Cautionary Tale About Control Variables in IV Estimation. *Oxford Bulletin of Economics and Statistics* 79(3), 411–425.
<https://doi.org/10.1111/obes.12177>
- Dunning, T., 2008. Improving Causal Inference: Strengths and Limitations of Natural Experiments. *Political Research Quarterly* 61(2), 282–293.
<https://doi.org/10.1177/1065912907306470>

- Ebenstein, A., 2009. When is the Local Average Treatment Close to the Average? Evidence from Fertility and Labor Supply. *Journal of Human Resources* 44(4), 955-975. <https://doi.org/10.3368/jhr.44.4.955>
- Edwards, J.K., Cole, S.R., Westreich, D., 2015. All Your Data are Always Missing: Incorporating Bias due to Measurement Error into the Potential Outcomes Framework. *International Journal of Epidemiology* 44(4), 1452–1459. <https://doi.org/10.1093/ije/dyu272>
- Esterling, K.M., Brady, D., Schwitzgebel, E., 2021. *The Necessity of Construct and External Validity for Generalized Causal Claims*. OSF Preprints. <https://doi.org/10.31219/osf.io/2s8w5>
- Gallen, T. and Raymond, B. (2021). *Broken Instruments*. Social Science Research Network. <http://dx.doi.org/10.2139/ssrn.3671850>
- Gelman, A., 2011. Causality and Statistical Learning. *American Journal of Sociology* 117(3), 955–966. <https://doi.org/10.1086/662659>
- Gibson, J., 2019. Are You Estimating the Right Thing? An Editor Reflects. *Applied Economic Perspectives and Policy* 41(3), 329–350. <https://doi.org/10.1093/aep/pz009>
- Goetghebuer, E., le Cessie, S., De Stavola, B., Moodie, E.E.M., Waernbaum, I., 2020. Formulating Causal Questions and Principled Statistical Answers. *Statistics in Medicine* 39(30), 4922–4948. <https://doi.org/10.1002/sim.8741>
- Greenland, S., 2017. For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates. *European Journal of Epidemiology* 32(1), 3–20. <https://doi.org/10.1007/s10654-017-0230-6>

- Greenland, S., Robins, J.M., 1986. Identifiability, Exchangeability, and Epidemiological Confounding. *International Journal of Epidemiology* 15(3), 413–419. <https://doi.org/10.1093/ije/15.3.413>
- Haavelmo, T., 1944. The Probability Approach in Econometrics. *Econometrica* 12, iii–115. <https://doi.org/10.2307/1906935>
- Hearst, N., Newman, T.B., 1988. Proving Cause and Effect in Traumatic Stress: The Draft Lottery as a Natural Experiment. *Journal of Traumatic Stress* 1(2), 173–180. <https://doi.org/10.1002/jts.2490010205>
- Heckman, J.J., 1997. Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations. *The Journal of Human Resources* 32(2), 441–462. <https://doi.org/10.2307/146178>
- Heckman, J.J., 2005. The Scientific Model of Causality. *Sociological Methodology* 35, 1–97. <https://doi.org/10.1111/j.0081-1750.2006.00164.x>
- Heckman, J.J., 2010. Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy. *Journal of Economic Literature* 48(2), 356–398. <https://doi.org/10.1257/jel.48.2.356>
- Hernán, M.A., 2016. Does Water Kill? A Call for Less Casual Causal Inferences. *Annals of Epidemiology* 26(10), 674–680. <https://doi.org/10.1016/j.annepidem.2016.08.016>
- Hernán, M.A., Robins, J.M., 2020. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton. Available at: https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2021/03/ciwhatif_hernanrobins_30mar21.pdf
- Hernán, M.A., VanderWeele, T.J., 2011. Compound Treatments and Transportability of Causal Inference. *Epidemiology* 22(3), 368–377. <https://doi.org/10.1097/EDE.0b013e3182109296>

- Holland, P.W., 1986. Statistics and Causal Inference. *Journal of the American Statistical Association* 81(396), 945–960.
<https://doi.org/10.1080/01621459.1986.10478354>
- Huber, M., 2015. Testing the Validity of the Sibling Sex Ratio Instrument. *Labour* 29(1), 1–14. <https://doi.org/10.1111/labr.12045>.
- Huber, M., Wüthrich, K., 2019. Local Average and Quantile Treatment Effects Under Endogeneity: A Review. *Journal of Econometric Methods* 8(1), [1–28]. <https://doi.org/10.1515/jem-2017-0007>
- Imbens, G.W., 2020. Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature* 58(4), 1129–1179. <https://doi.org/10.1257/jel.20191597>
- Imbens, G.W., Angrist, J.D., 1994. Identification and Estimation of Local Average Treatment Effects. *Econometrica* 62(2), 467–475.
<https://doi.org/10.2307/2951620>
- Imbens, G.W., Rubin, D.B., 1997. Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance. *The Annals of Statistics* 25(1), 305–327. <https://www.jstor.org/stable/2242722>
- Imbens, G.W., Rubin, D.B., 2010 [2008]. Rubin Causal Model, in: Durlauf, S.N., Blume, L.E. (Eds.), *Microeconometrics*. Palgrave Macmillan UK, London, pp. 229–241.
- Imbens, G.W., Rubin, D.B., 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York.

- Imbens, G.W., Wooldridge, J.M., 2009. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature* 47(1), 5–86.
<https://doi.org/10.1257/jel.47.1.5>
- Kaufman, J.S., 2019. Commentary: Causal Inference for Social Exposures. *Annual Review of Public Health* 40, 7–21. <https://doi.org/10.1146/annurev-publhealth-040218-043735>
- Kline, P., Walters, C.R., 2019. On Heckits, LATE, and Numerical Equivalence. *Econometrica* 87(2), 677–696. <https://doi.org/10.3982/ECTA15444>
- Lewbel, A., 2019. The Identification Zoo: Meanings of Identification in Econometrics. *Journal of Economic Literature* 57(4), 835–903.
<https://doi.org/10.1257/jel.20181361>
- Loken, E., Gelman, A., 2017. Measurement Error and the Replication Crisis. *Science* 355(6325), 584–585. <https://doi.org/10.1126/science.aal3618>
- Lundberg, I., Johnson, R., Stewart, B., 2021. What is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review*. <https://doi.org/10.1177/00031224211004187>
- Maciejewski, M.L., Brookhart, M.A., 2019. Using Instrumental Variables to Address Bias from Unobserved Confounders. *JAMA* 321(21), 2124–2125.
<https://doi.org/10.1001/jama.2019.5646>
- Mogstad, M., Torgovitsky, A., 2018. Identification and Extrapolation of Causal Effects with Instrumental Variables. *Annual Review of Economics* 10, 577–613. <https://doi.org/10.1146/annurev-economics-101617-041813>
- Mogstad, M., Wiswall, M., 2016. Testing the Quantity–Quality Model of Fertility: Estimation Using Unrestricted Family Size Models. *Quantitative Economics* 7(1), 157–192. <https://doi.org/10.3982/QE322>

- Morgan, S.L., Winship, Christopher., 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd ed. Cambridge University Press, New York, NY.
- Morck, R., Yeung, B., 2011. Economics, History, and Causation. *Business History Review* 85(1), 39–63. <https://doi.org/10.1017/S000768051100002X>
- Öberg, S., 2019. *The casual effect of fertility: The multiple problems with instrumental variables for the number of children in families*. OSF Preprints. <https://doi.org/10.31235/osf.io/peuvz>
- Pearl, J., Mackenzie, D., 2018. *The Book of Why: The New Science of Cause and Effect*. Allen Lane.
- Pei, Z., Pischke, J.-S., Schwandt, H., 2019. Poorly Measured Confounders are More Useful on the Left than on the Right. *Journal of Business & Economic Statistics* 37(2), 205–216. <https://doi.org/10.1080/07350015.2018.1462710>
- Phillips, A.N., Smith, G.D., 1993. The Design of Prospective Epidemiological Studies: More Subjects or Better Measurements? *Journal of Clinical Epidemiology* 46(10), 1203–1211. [https://doi.org/10.1016/0895-4356\(93\)90120-P](https://doi.org/10.1016/0895-4356(93)90120-P)
- Rehkopf, D.H., Glymour, M.M., Osypuk, T.L., 2016. The Consistency Assumption for Causal Inference in Social Epidemiology: When a Rose Is Not a Rose. *Current Epidemiology Reports* 3(1), 63–71. <https://doi.org/10.1007/s40471-016-0069-5>
- Robins, J.M., Greenland, S., 2000. Comment [on Dawid 2000]. *Journal of the American Statistical Association* 95(450), 431–435. <https://doi.org/10.1080/01621459.2000.10474214>

- Rosenzweig, M.R., Wolpin, K.I., 1980a. Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment. *Econometrica* 48(1), 227–240. <https://doi.org/10.2307/1912026>
- Rosenzweig, M.R., Wolpin, K.I., 1980b. Life-Cycle Labor Supply and Fertility: Causal Inferences from Household Models. *Journal of Political Economy* 88(2), 328–348. <http://www.jstor.org/stable/1837294>
- Rosenzweig, M.R., Wolpin, K.I., 2000. Natural “Natural Experiments” in Economics. *Journal of Economic Literature* 38(4), 827–874. <https://doi.org/10.1257/jel.38.4.827>
- Rubin, D.B., 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66(5), 688–701. <https://doi.org/10.1037/h0037350>
- Rubin, D.B., 1975. Bayesian Inference for Causality: The Importance of Randomization. *The Proceedings of the social statistics section of the American Statistical Association* 233–239. Available at: <http://www.asasrms.org/Proceedings/y1975/Bayesian%20Inference%20for%20Causality%20-%20The%20Importance%20of%20Randomization.pdf>
- Rubin, D.B., 1978. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics* 6(1), 34–58. <http://www.jstor.org/stable/2958688>
- Rubin, D.B., 1980. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association* 75(371), 591–593. <https://doi.org/10.2307/2287653>

- Rubin, D.B., 1990. [On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.] Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. *Statistical Science* 5(4), 472–480. <http://www.jstor.org/stable/2245383>
- Schennach, S.M., 2016. Recent Advances in the Measurement Error Literature. *Annual Review of Economics* 8, 341–377. <https://doi.org/10.1146/annurev-economics-080315-015058>
- Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston.
- Shear, B.R., Zumbo, B.D., 2013. False Positives in Multiple Regression Unanticipated Consequences of Measurement Error in the Predictor Variables. *Educational and Psychological Measurement* 73(5), 733–756. <https://doi.org/10.1177/0013164413487738>
- Small, D.S., Tan, Z., Ramsahai, R.R., Lorch, S.A., Brookhart, M.A., 2017. Instrumental Variable Estimation with a Stochastic Monotonicity Assumption. *Statistical Science* 32(4), 561–579. <https://doi.org/10.1214/17-STS623>
- Swanson, S.A., 2019. A Practical Guide to Selection Bias in Instrumental Variable Analyses. *Epidemiology* 30(3), 345–349. <https://doi.org/10.1097/EDE.0000000000000973>
- Swanson, S.A., Robins, J.M., Miller, M., Hernán, M.A., 2015. Selecting on Treatment: A Pervasive Form of Bias in Instrumental Variable Analyses. *American Journal of Epidemiology* 181(3), 191–197. <https://doi.org/10.1093/aje/kwu284>

- Tan, H.R., 2019. More Is Less? The Impact of Family Size on Education Outcomes in the United States, 1850–1940. *The Journal of Human Resources* 54(4), 1154–1181. <https://doi.org/10.3368/jhr.54.4.0517.8768R1>
- VanderWeele, T.J., 2009. Concerning the Consistency Assumption in Causal Inference. *Epidemiology* 20(6), 880–883. <https://doi.org/10.1097/EDE.0b013e3181bd5638>
- VanderWeele, T.J., Hernan, M.A., 2013. Causal Inference Under Multiple Versions of Treatment. *Journal of Causal Inference* 1(1), 1–20. <https://doi.org/10.1515/jci-2012-0002>
- Vytlacil, E., 2002. Independence, Monotonicity, and Latent Index Models: An Equivalence Result. *Econometrica* 70(1), 331–341. <https://doi.org/10.1111/1468-0262.00277>
- Westfall, J., Yarkoni, T., 2016. Statistically Controlling for Confounding Constructs Is Harder than You Think. *PLOS ONE* 11(3), e0152719. <https://doi.org/10.1371/journal.pone.0152719>
- Wooldridge, J.M., 2020 [2018]. *Introductory Econometrics: A Modern Approach*, 7th ed. Cengage, Boston, MA.