Göteborg Papers in Economic History

No. 25. October 2019 ISSN: 1653-1000

Too LATE for Natural Experiments: A Critique of Local Average Treatment Effects Using the Example of Angrist and Evans (1998)

Stefan Öberg

DEPARTMENT OF ECONOMY AND SOCIETY

UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

UNIT FOR ECONOMIC HISTORY

Too LATE for Natural Experiments: A Critique of Local Average Treatment Effects Using the Example of Angrist and Evans (1998)*

Stefan Öberg

stefan.oberg@econhist.gu.se

Abstract:

There has been a fundamental flaw in the conceptual design of many natural experiments used in the economics literature, particularly among studies aiming to estimate a local average treatment effect (LATE). When we use an instrumental variable (IV) to estimate a LATE, the IV only has an indirect effect on the treatment of interest. Such IVs do not work as intended and will produce severely biased and/or uninterpretable results. This comment demonstrates that the LATE does not work as previously thought and explains why using the natural experiment proposed by Angrist and Evans (1998) as the example.

JEL: C21, C90, J13

Keywords: causal inference, natural experiment, local average treatment effect, complier average causal effect, instrumental variable

ISSN: 1653-1000 *online version*
ISSN: 1653-1019 *print version*

University of Gothenburg
School of Business, Economics and Law
Department of Economy and Society
Unit for Economic History
P.O. Box 625
SE-405 30 GÖTEBORG
http://es.handels.gu.se/english/units/unit-for-economic-history/

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

An experiment is a scientific investigation that estimates the effect that a well-defined treatment has on a well-defined and pre-selected outcome. To estimate the effect of this treatment, the groups of treated and untreated units must be comparable. The most powerful method of achieving this (on average and in large samples) is to randomize which units receive treatment and which do not.[1] The randomization then serves as the assignment mechanism (Imbens and Rubin 2015, chap. 3). When studying people, we must often also allow for the possibility that not all people behave as we had intended them to behave in the design of the experiment. Then, it is necessary to analyze who complies with their assigned treatment and who does not, i.e. the mechanism determining reception of the treatment.

In summation, an experiment consists of (at least) four different aspects:

- a well-defined outcome,
- a well-defined treatment,
- a mechanism for assignment to the treatment,
- and the mechanism determining reception of the treatment.

Although the use of controlled experiments is increasing in economics[2], the use of so-called "natural experiments" is far more common. The idea behind natural experiments is "to exploit situations where the forces of nature or government policy have conspired to produce an environment somewhat akin to a randomized experiment" (Angrist and Krueger 2001, 73). These environments are then thought to be able to "assign the variable of interest randomly" (Angrist and Krueger 2001, 72; see also, for example, Angrist and Pischke 2008, 21, 151–152). In turn, this makes it possible to treat the data as if they were the result of a randomized experiment, thereby allowing us to estimate causal effects.

Natural experimental situations are most often used as the basis for instrumental variables (IVs). Most applications of such instruments in economics—and social sciences in general—only allow us to estimate the local average treatment effect (LATE, Imbens and Angrist 1994; Angrist and Imbens 1995; Angrist, Imbens, and Rubin 1996). This can—and often will—differ from the average treatment effect for the population; however, it can still be of interest (for example, Imbens 2010). We can only estimate the LATE because, most often, we have

1 Randomized controlled trials (RCTs) have recently been criticized for both inherent weaknesses and common shortcomings in applications. Despite these issues, they are an indispensable tool for science (for criticism and defense of RCTs, see the contributions in Kawachi, Subramanian, and Mowat (2018).
2 As exemplified by the work of the 2019 laureates of the Nobel Memorial Prize in Economic Sciences: Esther Duflo, Abhijit Banarjee, and Michael Kremer.

less than perfect compliance with the assignment and because the effect of the treatment varies across units (i.e. heterogeneous treatment effects).

It is well known that, when we estimate a LATE, we estimate the effect for the group for which the treatment level is changed from their assignment to treatment by the instrument. I argue that an important consequence of this fact has been overlooked in the literature.

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 (for example, Angrist and Krueger 2001, 73). This requires a clear and unambiguous definition of the treatment that is based on the mechanism through which the instrument affects the treatment (or variable) of interest. However, such definitions are missing from most of the literature using natural experiments. This is problematic because such definitions make it clear that there is a difference between the effect that we are estimating, the LATE, and the effect of the treatment (or variable) of interest.

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. Notably, the effect of the change is different from the effect of the level of the treatment. The change in treatment is the result of the mechanism through which the instrument affects the treatment of interest, i.e. the mechanism determining reception of the treatment. This mechanism defines the specific treatment for which we estimate the effect. This specific treatment will be different from the treatment (or variable) of interest. Therefore, we cannot expect their effects to be the same.

The mechanism creating the change in the treatment completely mediates the association between the instrument and the treatment of interest. However, this has the unfortunate consequence of delinking the instrument from the treatment of interest, causing IVs with this type of indirect effect on the treatment of interest to not assign it "as good as random". Therefore, they will also be unable to solve any issues of endogeneity and results based on indirect IVs, such as those used to estimate LATEs, will be severely biased and/or uninterpretable.

This issue has been overlooked in the existing literature, an issue that has led to many studies being fundamentally flawed. The literature has been overly focused on the assignment mechanism (see, for example, Imbens 2019, 30–31; see also, Dunning 2008). Finding an as-good-as-random assignment mechanism has been considered sufficient for a valid natural experimental situation despite it being well known (for example, Angrist, Imbens,

and Rubin 1996, 447) that a random assignment mechanism is not enough. Additional assumptions are required to estimate a LATE, such as the exclusion restriction and monotonicity. I argue that we must also be more careful when defining the experiment's treatment[3] and the process underlying the receipt of this treatment. In the words of Angrist and Krueger (2001, 73), a "good instrument [is] correlated with the endogenous regressor for reasons the researcher can verify and explain".

In this work, I use the example of the "*Same sex* instrument" proposed in Angrist and Evans (1998). This IV relates to the treatment of interest for reasons we can verify and explain. In this case, the reason is that some parents desire to have children of both sexes and have another child in an attempt to achieve this. This reason is an example of a completely mediating step between the IV and the treatment of interest. Moreover, because this IV has only an indirect effect on the treatment of interest, it does not solve the issues of endogeneity. In Section IV, I demonstrate this point using the potential outcomes framework and simulated data.

To enable that analysis, we need a much clearer definition of the experiment's treatment than the aforementioned reason. Furthermore, we must not only consider the reason but also the how and why this reason exists. There will always be several different possible interpretations of this. However, the different interpretations of the experiment's treatment do not in any way change how the method works mathematically (for example, Greenland 2017, 9). The method is quite simple in practice; we compare the units that are indicated by the instrument with the units that are not. Moving from that to estimating a causal effect is based on both the necessary assumptions *and* the interpretations. The different possible interpretations of the experiment's treatment determine which units comply with their assignment. Therefore, the different interpretations also have varying consequences for how well the necessary assumptions are fulfilled (for example, Imbens 2019, 37).

In Section IV, I will discuss four different interpretations of the experiment's treatment for the *Same sex* IV and demonstrate how well the natural experiment works under the different interpretations. To precede the analysis, the conclusion is that the IV violates both the exclusion restriction and the independence assumption and/or produces meaningless and uninterpretable results.

3 The requirement to have a clear and unambiguous definition of the treatment is also not anything new (for example, Rubin 1978, 39–40; Angrist, Imbens, and Rubin 1996, 446). However, detailed conceptual interpretations remain missing from the literature applying natural experiments.

1. The *Same sex* Instrument as a Source of Exogenous Variation in the Number of Children in a Family

Microeconomic studies of families examining the effect of the number of children on the parents or siblings represent examples of insufficient conceptual definitions leading to the use of natural experiments that are not valid. It has been assumed that these natural experiments provide estimates of the effect we are interested in rather than the effect of the variation created by the (random) assignment mechanism used as the basis for the natural experiment.

This literature has heavily relied on two "natural natural experiments" (Rosenzweig and Wolpin 2000) to identify exogenous variation in the number of children: (1) the occurrence of multiple births (for example, Rosenzweig and Wolpin 1980; Black, Devereux, and Salvanes 2005) and (2) the sex of the first-born child(ren) (Angrist and Evans 1996; 1998; Angrist, Lavy, and Schlosser 2010). The idea behind the first natural experiment is that a multiple birth leads to an exogenous increase in the number of children in the family. The idea behind the second is that, as previously mentioned, some families desire to have children of both sexes. Therefore, if their first-born child(ren) are of the same sex, they will have another child even if they had not originally intended to. An alternative specification of this instrument that is often used in Asian populations is that families desire a boy and will go on to have another child if their firstborn is a girl.

Generally, people have been found to behave as though they have this type of gender preference for their children. Norling (2018) modeled parents' preferences over the sex of their children using a large number of birth history surveys from Africa, Asia, and the Americas. He concludes that sex preferences are more common than previously established and "influence the decision to have additional children for more than half of couples" (Norling 2018, 209). Notably, this type of preference is not only present in lower-income countries. For example, Miranda, Dahlberg, and Andersson (2018) found that parents in Sweden tend to prefer having children of both sexes. They show this using both revealed preferences from register data and responses to survey questions.

Using census data from 51 countries, Bisbee et al. (2018) applied the *Same sex* instrument (first two children) to estimate the "the relationship between having a third child and a mother's labor force participation". They found that having two first-born children of the same sex is associated with a higher probability of having another child in 132 out of 139 available country-year observations. They also found that the instrument has sufficient

predictive power in almost all cases, even though the average difference in the number of children is not large (the global average of the first-stage coefficients is 0.041 children).

2. The Potential Outcomes Framework Applied to Instrumental Variables

The potential outcomes framework provides ways to think about the defining features of experiments (Imbens and Rubin 2015; Morgan and Winship 2015; Imbens 2019). This framework can easily be extended to natural experiments and instrumental variables (e.g. Angrist, Imbens, and Rubin 1996).

An evaluation of a natural experiment using the potential outcomes framework is based on the four types of observations that are created by combinations of their (binary) assignment to treatment and (binary) treatment status.[4] The four types are the compliers, the never-takers, the always-takers, and the defiers. When using empirical data, it is often impossible to conclusively determine which type a unit is. However, we can determine that they belong to one of two types (Table 1). Compliers are units that behave as we expect and intend in the experiment. They are treated when assigned to treatment and are not treated when not assigned. Conversely, units that do not behave as expected and intended are known as non-compliers. These are either treated even if not assigned to treatment (always-takers) or are not treated even if assigned to treatment (never-takers). Moreover, it is possible that there are units—referred to as defiers—that behave in the opposite manner to what is expected and intended. They are either treated because they are not assigned to treatment or are not treated because they are assigned to treatment. Estimation of LATE relies on the monotonicity assumption, in which we assume that there are no defiers. It is unlikely that defiers exist in the case of *Same sex* IVs, and I will ignore them in the following analyses.

TABLE 1. THE FOUR TYPES OF OBSERVATIONS CREATED BY COMBINATIONS OF THEIR ASSIGNMENT TO TREATMENT AND TREATMENT STATUS

		Indicated by the instrument?	
		Yes	No
Treated?	Yes	Compliers and Always-takers	Always-takers (and Defiers)
	No	Never-takers (and Defiers)	Compliers and Never-takers

[4] Although the number of children in a family is not binary, the specific treatment investigated when using *Same sex* IVs is. A family either has or does not have another child in an attempt to have children of both sexes.

Due to the fact that I am using simulated data in this study, I know how the units would have behaved in other possible situations and I can therefore assign each unit to a type. While this is not the typical situation, it has the advantage of enabling me to evaluate the underlying assumptions.

In the analyses, I compare the desired number of children and the prevalence of a preference for children of both sexes among the types as defined under the different interpretations of the experiment's treatment. If the exclusion restriction and independence assumption are fulfilled, there should be no systematic differences between compliers (or non-compliers) that are assigned to treatment and those that are not assigned to treatment. This was evaluated for the four different interpretations of the experiment's treatment in Tables 2, 4, 5, and 6. The independence assumption implies that units assigned to treatment and those that are not should not differ in their (hypothetical) reactions to different assignments ("unconfounded types") (Henderson et al. 2008, 171). Therefore, it should be equally likely for a unit to be a complier (or non-complier) whether or not it is assigned to treatment. These results are presented in Table 3 for all interpretations of the experiment's treatment.

3. Description of the Simulated Data

As previously mentioned, I used a simulated population to evaluate the *Same sex* instrument. The starting point for the simulated data is a population of families with at least two births. For simplicity, I ignored multiple births. The families varied in how many children they want, originally desiring between two and nine children. The distribution of these preferences is a stylized mid-twentieth-century Scandinavian population based on Black, Devereux, and Salvanes (2005, tab. II), and Åslund and Grönqvist (2010, tab. 1).[5] The majority of families (60 percent) in the population desired children of both sexes, whereas the rest were neutral. For 20 percent of the families desiring children of both sexes, this desire is so strong that it makes them have one more child than they had originally intended if the first two are of the same sex. For simplicity, the families in the simulation only had one more child in an attempt to have children of both sexes. This creates 141 different possible combinations of events and preferences leading to the families ending up with between two and ten children

5 Half of the families (50.0 percent) desire two children, 33.3 percent desire three, 11.0 percent desire four, 4.0 percent desire five, 1.2 percent desire six, 0.5 percent desire seven, 0.2 percent desire eight, and 0.1 percent desire nine (or more).

(average 2.8). These combinations are associated with a probability based on the distribution of preferences and events. For simplicity, it was equally likely for a family in the simulation to have a boy or a girl. This simulation allows me to present the originally desired number of children, the realized number of children, the share desiring children of both sexes, the share with a strong desire for children of both sexes, and the share of the population for each type using different interpretations of the experiment's treatment. The results are the average for units with different combinations of events and preferences within each type that are weighted by the probability of each combination of events and preferences. The simulation and results are presented in full in the online appendix spreadsheet (available at: https://osf.io/84vs3/).

4. Evaluating the Same sex IV Under Different Definitions of the Experiment's Treatment

The results of evaluating this and other natural experiments differ depending on what we define the experiment's treatment to be. In the following section, I present the results for four different interpretations of this treatment for the *Same sex* IV: the first two children born being of the same sex (that is, using the assignment mechanism as the experiment's treatment); having a third child in an attempt to have children of both sexes; having another child at any parity in an attempt to have children of both sexes; and having another child at any parity for any reason.

A. The First Two Children Born are of the Same Sex as the Experiment's Treatment

If we interpret the experiment's treatment as the first two children born being of the same sex, we are using the assignment mechanism as the treatment. In this case, we have a situation with perfect compliance because there can be neither never-takers nor always-takers. In this case, the natural experiment works well in the sense that the preferences are balanced between compliers assigned to treatment and those not assigned (Table 2); therefore, the exclusion restriction is met. The compliers assigned to treatment also end up with a slightly larger number of children, thus implying that the instrument is relevant. Units that are assigned to treatment and not assigned to treatment have the same likelihood of being compliers (p = 1.000 in both cases) (Table 3). Therefore, the independence assumption is also fulfilled.

The problem with using the assignment mechanism as the experiment's treatment is that we are then estimating the intention-to-treat effect. This effect has little in common with the causal effect of the number of children that we are interested in estimating. It is also not a valid natural experiment because the consequences of the experiment's treatment are not the same for everyone. Having the first two children of the same sex will have different consequences for different families. This violates the second part of the stable unit treatment value assumption (SUTVA, Imbens and Rubin 2015, 11; see also, Cox 1958, 17–21). This part of the SUTVA is an assumption related to the treatment itself. However, Imbens and Rubin (2015) suggest that it should be evaluated by thinking about whether the outcome would be the same regardless of which version of the treatment a unit receives. The outcome would not be the same for the different versions of the treatment if we use this interpretation of the experiment's treatment.

B. Having a Third Child to Try to Mix Sexes as the Experiment's Treatment

A more empirically interesting interpretation of the experiment's treatment is to use compliance as the treatment; that is, having a third child in an attempt to have children of both sexes if the first two children born are of the same sex. However, when we make this change in the interpretation, the natural experiment is no longer valid because the IV violates the exclusion restriction due to systematic differences in the preferences of the compliers assigned to treatment and those not assigned (Table 4). Moreover, the realized number of children is larger among the units assigned to treatment due to the differences between the groups of compliers *and* because the never-takers have a larger realized number of children than compliers not assigned to treatment. Furthermore, the independence assumption is also no longer met (Table 3).

TABLE 2. EVALUATION OF THE *SAME SEX* IV WHEN THE EXPERIMENT'S TREATMENT IS DEFINED AS HAVING THE FIRST TWO CHILDREN BORN OF THE SAME SEX

Panel A. The parents' desired number of children

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	2.76	—	2.76
First two children of the same sex?	No	—	2.76	2.76
Column average		2.76	2.76	Overall average = 2.76

Panel B. The parents' realized number of children

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	2.85	—	2.85
First two children of the same sex?	No	—	2.76	2.76
Column average		2.85	2.76	Overall average = 2.80

Panel C. Share of parents desiring children of both sexes

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	0.600	—	0.600
First two children of the same sex?	No	—	0.600	0.600
Column average		0.600	0.600	Share of population = 0.600

Panel D. Share of parents with a strong desire to have children of both sexes

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	0.120	—	0.120
First two children of the same sex?	No	—	0.120	0.120
Column average		0.120	0.120	Share of population = 0.120

Panel E. Share of the population in each category

		Indicated by the instrument?		
		Yes	No	Row sum
Treated?:	Yes	0.500	—	0.500
First two children of the same sex?	No	—	0.500	0.500
Column sum		0.500	0.500	Total = 1.000

Source: Author's calculations based on the simulated population.

TABLE 3. THE LIKELIHOOD OF BELONGING TO EACH TYPE WITHIN THE LEVELS OF THE INSTRUMENT

	Compliers Indicated by the instrument?		Never-takers Indicated by the instrument?		Always-takers Indicated by the instrument?	
Treatment definition:	Yes	No	Yes	No	Yes	No
First two children of the same sex	1.000	1.000	—	—	—	—
Having a third child to have both sexes	0.060	1.000	0.940	0.000	—	—
Having another child at any parity to have both sexes	0.084	1.000	0.916	0.000	—	—
Having another child after the first two	0.560	0.500	0.440	0.000	0.000	0.500

Source: Author's calculations based on the simulated population.

C. Having Another Child at Any Parity to Try to Mix Sexes as the Experiment's Treatment

We can also use the behavior that the natural experiment is based on as the experiment's treatment. In this case, we interpret the treatment as having another child at any parity in an attempt to have children of both sexes. However, this interpretation again leads to a natural experiment that is not valid because both the exclusion restriction (Table 5) and the independence assumption (Table 3) are violated.

D. Having Another Child at Any Parity for Any Reason as the Experiment's Treatment

We use the *Same sex* IV because we are interested in the causal effect of the number of children, or at least the causal effect of having another child. Thus, we aim to interpret the experiment's treatment as having another child at any parity for any reason. Notably, we cannot use this interpretation of the treatment with the *Same sex* IV. The natural experiment is not valid because both the exclusion restriction (Table 6) and the independence assumption (Table 3) are violated.

TABLE 4. EVALUATION OF THE *SAME SEX* IV WHEN THE EXPERIMENT'S TREATMENT IS DEFINED AS HAVING A THIRD CHILD IN AN ATTEMPT TO HAVE CHILDREN OF BOTH SEXES

Panel A. The parents' desired number of children

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	2.00	—	2.00
Have a third child to have both sexes?	No	2.81	2.76	2.79
	Column average	2.76	2.76	Overall average = 2.76

Panel B. The parents' realized number of children

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	3.00	—	3.00
Have a third child to have both sexes?	No	2.84	2.76	2.80
	Column average	2.85	2.76	Overall average = 2.80

Panel C. Share of parents desiring children of both sexes

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	1.000	—	1.000
Have a third child to have both sexes?	No	0.574	0.600	0.588
	Column average	0.600	0.600	Share of population = 0.600

Panel D. Share of parents with a strong desire to have children of both sexes

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	1.000	—	0.120
Have a third child to have both sexes?	No	0.064	0.120	0.093
	Column average	0.120	0.120	Share of population = 0.120

Panel E. Share of the population in each category

		Indicated by the instrument?		
		Yes	No	Row sum
Treated?:	Yes	0.030	—	0.030
Have a third child to have both sexes?	No	0.470	0.500	0.970
	Column sum	0.500	0.500	Total = 1.000

Source: Author's calculations based on the simulated population.

TABLE 5. EVALUATION OF THE *SAME SEX* IV WHEN THE EXPERIMENT'S TREATMENT IS DEFINED AS HAVING ANOTHER CHILD AT ANY PARITY IN AN ATTEMPT TO HAVE CHILDREN OF BOTH SEXES

Panel A. The parents' desired number of children

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	2.34	—	2.34
Another child at any parity to have both sexes?	No	2.80	2.76	2.78
Column average		2.76	2.76	Overall average = 2.76

Panel B. The parents' realized number of children

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	3.34	—	3.34
Another child at any parity to have both sexes?	No	2.80	2.76	2.78
Column average		2.85	2.76	Overall average = 2.80

Panel C. Share of parents desiring children of both sexes

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	1.000	—	1.000
Another child at any parity to have both sexes?	No	0.563	0.600	0.583
Column average		0.600	0.600	Share of population = 0.600

Panel D. Share of parents with a strong desire to have children of both sexes

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	1.000	—	1.000
Another child at any parity to have both sexes?	No	0.039	0.120	0.082
Column average		0.120	0.120	Share of population = 0.120

Panel E. Share of the population in each category

		Indicated by the instrument?		
		Yes	No	Row sum
Treated?:	Yes	0.042	—	0.042
Another child at any parity to have both sexes?	No	0.458	0.500	0.958
Column sum		0.500	0.500	Total = 1.000

Source: Author's calculations based on the simulated population.

TABLE 6. EVALUATION OF THE *SAME SEX* IV WHEN THE EXPERIMENT'S TREATMENT IS DEFINED AS HAVING ANOTHER CHILD AT ANY PARITY FOR ANY REASON

Panel A. The parents' desired number of children

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	3.36	3.52	3.44
Another child at any parity to have both sexes?	No	2.00	2.00	2.00
Column average		2.76	2.76	Overall average = 2.76

Panel B. The parents' realized number of children

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	3.51	3.52	3.52
Another child at any parity to have both sexes?	No	2.00	2.00	2.00
Column average		2.85	2.76	Overall average = 2.80

Panel C. Share of parents desiring children of both sexes

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	0.643	0.600	0.623
Another child at any parity to have both sexes?	No	0.545	0.600	0.574
Column average		0.600	0.600	Share of population = 0.600

Panel D. Share of parents with a strong desire to have children of both sexes

		Indicated by the instrument?		
		Yes	No	Row average
Treated?:	Yes	0.214	0.120	0.170
Another child at any parity to have both sexes?	No	0.000	0.120	0.064
Column average		0.120	0.120	Share of population = 0.120

Panel E. Share of the population in each category

		Indicated by the instrument?		
		Yes	No	Row sum
Treated?:	Yes	0.280	0.250	0.530
Another child at any parity to have both sexes?	No	0.220	0.250	0.470
Column sum		0.500	0.500	Total = 1.000

Source: Author's calculations based on the simulated population.

5. Explaining why the LATE does not Work as Intended

My results show that the *Same sex* instrument is only valid if we use the assignment mechanism as the experiment's treatment. Under this interpretation, the natural experiment still does not produce meaningful estimates of any effect because the experiment's treatment then violates the second part of the SUTVA. Using other interpretations, the instrument violates both the exclusion restriction and the independence assumption. Notably, this is the same result as Öberg (2019) found when analyzing IVs based on multiple births. Both of these IVs have an indirect effect on the treatment of interest and represent examples of how this type of instrument does not work as previously thought.

It might come as a surprise to the reader that estimates of the LATE do not function as intended, it was certainly a surprise to the author initially. However, this result is less surprising when we consider the causal structure of the problem. This is a case for which a directed acyclical graph (DAG, for example, Pearl and Mackenzie 2018) is a useful tool.

Because we are interested in estimating the causal effect of a treatment of interest (X) on an outcome (Y), we require an instrumental variable (IV) to do this due to an unobserved confounder, $\mathbb{U}_1$. A relevant IV (Z) is related to the treatment of interest, X (edge f in Figure 1 would not be crossed out). A valid IV is only related to the outcome through its effect on the treatment (edge d in Figure 1 is, and should be, crossed out). Moreover, there should not be any unobserved confounding of the instrument's effects on the treatment or the outcome (as illustrated in Figure 1 by the absence of edges).

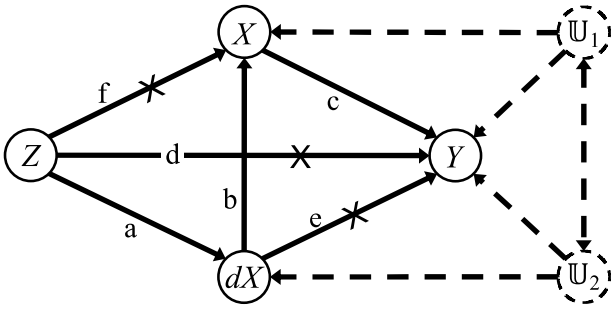

FIGURE 1. A DIRECTED ACYCLICAL GRAPH ILLUSTRATING WHY LOCAL AVERAGE TREATMENT EFFECTS DO NOT WORK AS INTENDED

The thing that creates problems when we are estimating LATEs is that the IV, *Z*, does not have a direct effect on the treatment of interest, *X*. The IV is related to the treatment of interest through a change caused by a specific mechanism, *dX*, which is what affects the level of the treatment of interest. This specific mechanism, *dX*, completely mediates the effect of the IV on the treatment of interest. This has the consequence that, although the IV is valid as an instrument for *dX*, it is not valid for the treatment of interest. The intermediary, *dX*, delinks *Z* from *X* so that the effect of $X|Z$ on *Y* will still be confounded by $\mathbb{U}_1$.

The purpose of the IV is to estimate the effect of *X* on *Y*, not the effect of *dX* on *Y*. In fact, we do not want there to be any direct effect of *dX* on *Y* since such an effect would violate the exclusion restriction. Using the Same sex IV as an example, a direct effect of *dX* on *Y* could emerge if there was an effect on the outcome from the parents' preferences for having children of both sexes.[6]

The most controversial part of my argument is that there is no direct effect of the instrument on the treatment of interest (edge f is crossed out in Figure 1). In my opinion, the *Same sex* and multiple birth IVs are good illustrations of why we cannot consider the instrument to have a direct effect on the treatment of interest. Notably, neither having the first child(ren) born being of the same sex nor experiencing a multiple birth directly affects the number of children in the family. These events do not assign the value of this treatment of interest as good as random; however, what they can do is induce a change that can subsequently affect the value of the treatment of interest. *Some* families will have one more child than they had originally intended, which will affect the number of children in the family.

Whereas the number of children in the family is a discrete variable that can take on several different values, these IVs are binary; either a family experienced the event used as the basis for the IV or not. This binary IV is related to a change that occurs in some families while not in others; in other words, a binary event. This change is the manipulation that we use as the basis for estimating a causal effect. However, because it is separate from the treatment of interest, we must take into account the extended causal structure illustrated in Figure 1 when considering this type of IV.

The mechanisms of change will be different for different IVs, although all IVs used to estimate a LATE will have this intermediary step between the IV and the treatment of

6 In the case of IVs based on multiple births, a direct effect of dX on Y could emerge if there was an effect on the outcome from having two (or more) children being born at once instead of with some time in between.

interest. By definition, the LATE is the causal effect for the group that is influenced by the instrument and undergoes a change due to being assigned to treatment. Any other IVs that are only indirectly related to the treatment of interest, such as the one in Acemoglu, Johnson, and Robinson (2001), will also have to be evaluated using the extended causal structure presented in Figure 1.

6. Discussion and Conclusions

When designing a natural experiment, it is equally important to have a clear and unambiguous definition of the experiment's treatment as having an as-good-as-random assignment mechanism. A (natural) experiment is not defined by its assignment mechanism, but rather by all of its features. Different natural experiments can be designed based on the same (random) allocation mechanism.

Although I criticize many previous applications of natural experiments, I certainly do not intend to invalidate the idea. The idea of utilizing natural experiments is a good one, though we must carefully define all aspects of such experiments.

Both Angrist and Krueger (2001, 80) and Angrist and Pischke (2008, 7–8) quoted Haavelmo's (1944) distinction between controlled and natural experiments.[7] I find what follows immediately after the quoted distinction to be useful when considering all aspects of a natural experiment.

> A design of experiments[...] is an essential appendix to any quantitative theory. And we usually have some such experiment in mind when we construct the theories, although—unfortunately—most economists do not describe their designs of experiments explicitly. If they did, they would see that the experiments they have in mind may be grouped into two different classes, namely, (1) experiments that *we should like to make* to see if certain real economic phenomena—when *artificially isolated* from "other influences"—would verify certain hypotheses, and (2) the stream of experiments that Nature is steadily turning out from her enormous laboratory, and which we merely watch as passive

[7] Haavelmo's (1944) definition of experiments is broader than how we most often think about them. He considers experiments and the design of (hypothetical or real) experiments as being the link between theory and reality. They are what forces us to consider empirical definitions and measurements of the variables we include in our models. In his own words: "What makes a piece of mathematical economics not only mathematics but also economics is, I believe, this: When we set up a system of theoretical relationships and use economic names for the otherwise purely theoretical variables involved, we have in mind some actual *experiment*, or some *design of an experiment*, which we could at least imagine arranging, in order to measure those quantities in real economic life that we think might obey the laws imposed on their theoretical namesakes" (Haavelmo 1944, 6).

> observers. In both cases the aim of theory is the same, namely, to become master of the happenings of real life. But our approach is a little different in the two cases. [...] In the second case [natural experiments] we can only try to adjust our theories to reality as it appears before us. And what is the meaning of a design of experiments in this case? It is this: 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. (Haavelmo 1944, 14)

The last sentence covers both one of the most fundamental objections to estimates based on natural experiments—that they are not producing the results that are of theoretical interest (for example, Rosenzweig and Wolpin 2000; Heckman 2005; Mogstad and Torgovitsky 2018)—*and* that which I argue has been missing from the literature: an explicit discussion of all aspects of the design of the experiment, (i.e. the assignment mechanism, the treatment, and the mechanism for the reception of treatment).

We can and should continue to use natural experiments. However, we must use natural experiments that produce meaningful results even after we have adjusted our models and the (ex post) design of the experiments "in such a way that the resulting data *would be* those which we get by passive observation of reality" (Haavelmo 1944, 14).

When we consider this (ex post) design, we must ensure that the IV has a direct effect on the treatment of interest. This precludes situations for which we are limited to estimating a LATE. Although this will certainly limit the number of valid natural experiments available, we will still be significantly helped by those that we are able to find. However, there remains a need to reconsider and correct the multitude of published studies that have estimated LATEs or have used IVs with an indirect effect on the treatment of interest.

REFERENCES

Acemoglu, Daron, Simon H. Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *The American Economic Review* 91 (5): 1369–1401. https://doi.org/10.1257/aer.91.5.1369

Angrist, Joshua D., and William N. Evans. 1996. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size.". National Bureau of Economic Research Working Paper 5778. https://doi.org/10.3386/w5778

Angrist, Joshua D., and William N. Evans. 1998. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *The American Economic Review* 88 (3): 450–77. https://www.jstor.org/stable/116844

Angrist, Joshua D., and Guido W. Imbens. 1995. "Two-stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association* 90 (430): 431–42. https://doi.org/10.1080/01621459.1995.10476535

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–55. https://doi.org/10.2307/2291629

Angrist, Joshua D., and Alan B. Krueger. 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, Joshua D., Victor Lavy, and Analia Schlosser. 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, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Åslund, Olof, and Hans Grönqvist. 2010. "Family Size and Child Outcomes: Is There Really No Trade-off?" *Labour Economics* 17 (1): 130–39. https://doi.org/10.1016/j.labeco.2009.05.003

Bisbee, James, Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii. 2017. "Local Instruments, Global Extrapolation: External Validity of the Labor Supply–Fertility Local Average Treatment Effect." *Journal of Labor Economics* 35 (S1): S99–147. https://doi.org/10.1086/691280

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. 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

Cox, David Roxbee. 1958. *Planning of Experiments*. New York: John Wiley.

Dunning, Thad. 2008. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61 (2): 282–293. https://doi.org/10.1177/1065912907306470

Greenland, Sander. 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

Haavelmo, Trygve. 1944. "The Probability Approach in Econometrics." *Econometrica* 12 (supplement): iii–115. https://doi.org/10.2307/1906935

Heckman, James J. 2005. "The Scientific Model of Causality." *Sociological Methodology* 35 (1): 1–97. https://doi.org/10.1111/j.0081-1750.2006.00164.x

Henderson, Daniel J., Daniel L. Millimet, Christopher F. Parmeter, and Le Wang. 2008. "Fertility and the Health of Children: A Nonparametric Investigation." In *Modelling and Evaluating Treatment Effects in Econometrics*, edited by Tom Fomby, R. Carter Hill, Daniel L. Millimet, Jeffrey A. Smith, and Edward J. Vytlacil, 167–95. Bingley: Emerald Group Publishing Limited. https://doi.org/10.1016/S0731-9053(07)00007-2

Imbens, Guido W. 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48 (2): 399–423. https://doi.org/10.1257/jel.48.2.399

Imbens, Guido W. 2019. "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics." National Bureau of Economic Research Working Paper 26104. https://doi.org/10.3386/w26104

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75. https://doi.org/10.2307/2951620

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.

Kawachi, Ichiro, S. V. Subramanian, and Ryan Mowat eds. 2018. *Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue. Social Science & Medicine* 210.

Miranda, Vitor, Johan Dahlberg, and Gunnar Andersson. 2018. "Parent's Preferences for Sex of Children in Sweden: Attitudes and Outcomes." *Population Research and Policy Review* 37 (3): 443–59. https://doi.org/10.1007/s11113-018-9462-8

Mogstad, Magne, and Alexander Torgovitsky. 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

Morgan, Stephen L., and Christopher Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd ed. New York, NY: Cambridge University Press.

Norling, Johannes. 2018. "Measuring Heterogeneity in Preferences Over the Sex of Children." *Journal of Development Economics* 135: 199–221. https://doi.org/10.1016/j.jdeveco.2018.07.004

Öberg, Stefan. 2019. "Instrumental Variables Based on Twin Births Are by Definition Not Valid (v.3)." SocArXiv Papers. http://doi.org/10.17605/OSF.IO/ZUX9S

Rosenzweig, Mark R., and Kenneth I. Wolpin. 1980. "Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment." *Econometrica* 48 (1): 227–40. https://doi.org/10.2307/1912026

Rosenzweig, Mark R., and Kenneth I. Wolpin. 2000. "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature* 38 (4): 827–74. https://doi.org/10.1257/jel.38.4.827

Göteborg Papers in Economic History

Available online at S-WOPEC: (http://swopec.hhs.se/gunhis/)

1. Jan Bohlin: Tariff protection in Sweden 1885-1914. 2005
2. Svante Larsson: Globalisation, inequality and Swedish catch up in the late nineteenth century. Williamson's real wage comparisons under scrutiny. 2005
3. Staffan Granér: Thy Neighbour's Property. Communal property rights and institutional change in an iron producing forest district of Sweden 1630-1750. 2005
4. Klas Rönnbäck: Flexibility and protectionism. Swedish trade in sugar during the early modern era. 2006
5. Oskar Broberg: Verkstadsindustri i globaliseringens tidevarv. En studie av SKF och Volvo 1970-2000. 2006
6. Jan Bohlin: The income distributional consequences of agrarian tariffs in Sweden on the eve of World War I. 2006
7. Jan Bohlin and Svante Larsson: Protectionism, agricultural prices and relative factor incomes: Sweden's wage-rental ratio, 1877–1926. 2006
8. Jan Bohlin: Structural Change in the Swedish economy in the late nineteenth and early twentieth century – The role of import substitution and export demand. 2007
9. Per Hallén Levnadsstandarden speglad i boupptеckningar. En undersökning av två metoder att använda svenska boupptеckningar för en levnadsstandards undersökning samt en internationell jämförelse. 2007
10. Klas Rönnbäck: The price of sugar in Sweden. Data, source & methods. 2007
11. Klas Rönnbäck: From extreme luxury to everyday commodity – sugar in Sweden, 17th to 20th centuries. 2007
12. Martin Khan: A decisive intelligence failure? British intelligence on Soviet war potential and the 1939 Anglo-French-Soviet alliance that never was. 2008
13. Bengt Gärdfors: Bolagsrevisorn. En studie av revisionsverksamheten under sent 1800-tal och tidigt 1900-tal. Från frivillighet till lagreglering och professionalisering. 2010
14. Ann-Sofie Axelsson, Oskar Broberg och Gustav Sjöblom (red.): Internet, IT-boomen och reklambranschen under andra hälften av nittiotalet. Transkript av ett vittnesseminarium på ABF-huset i Stockholm den 17 februari 2010. 2011
15. Staffan Granér and Klas Rönnbäck: Economic Growth and Clean Water in the Göta River. A Pilot Study of Collective Action and the Environmental Kuznets Curve, 1895-2000. 2011
16. Ulf Olsson: En värdefull berättelse. Wallenbergarnas historieprojekt. 2013
17. Irene Elmerot: Skrivhandledning för doktorander i ekonomisk historia vid Göteborgs universitet.2015

18. Tobias Karlsson and Christer Lundh: The Gothenburg Population Panel 1915–1943. GOPP Version 6.0. 2015
19. Tobias Karlsson: Pushed into unemployment, pulled into retirement: Facing old age in Gothenburg, 1923–1943. 2015
20. Stefan Öberg and Klas Rönnbäck: Mortality among European settlers in pre-colonial West Africa: The "White Man's Grave" revisited. 2016
21. Dimitrios Theodoridis: The ecological footprint of early-modern commodities. Coefficients of land use per unit of product. 2017
22. Stefan Öberg: An introduction to using twin births as instrumental variables for sibship size. 2017
23. Stefan Öberg: Instrumental variables based on twin births are by definition not valid. 2018
24. Jesper Hamark and Kristoffer Collin: Industrial wages in mid-1880s Sweden: estimations beyond Bagge's *Wages in Sweden*. Data, source and methods
25. Stefan Öberg: Too LATE for natural experiments: A critique of Local Average Treatment Effects using the example of Angrist and Evans (1998)