The casual effect of fertility: The multiple problems with instrumental variables for the number of children in families

Author: Stefan Öberg (ORCID iD 0000-0003-1943-3693), University of Gothenburg, Unit for Economic History, Gothenburg, Sweden (email: stefan.oberg@econhist.gu.se).

Sept. 15, 2021

Abstract

Studies investigating how the number of children in a family affects the parents or the children face problems because the variable of interest is endogenous in the model. The currently accepted solution to this problem is to use instrumental variables (IVs), for example, based on twin births. In this paper, I review and add to the critique of IVs based on twin births and show that that there are so many issues—major and minor—with these IVs that results based on them are not reliable or interpretable. I also review other IVs used in the literature, for example IVs based on the sexes of the firstborn children, and conclude that there are, as of yet, no credible IVs for the number of children. We need to disregard results from studies applying these IVs, reevaluate the current state of knowledge, and develop new, more credible methods.

Keywords: causal inference; natural experiments; local average treatment effect; family size; quantity–quality trade-off; sibling sex composition

JEL codes: C26, D13, J13

1 Introduction

Studies investigating how the number of siblings affects the children or the number of children affects the parents face problems because the variable of interest—the number of children in a family—is endogenous in the model (i.e., is correlated with the residuals). The number of children that parents desire is linked to other characteristics, such as their level of ambition in their careers, their lifestyle, their expectations regarding the benefits and disadvantages of having a(nother) child, and the balance in their preferences for child "quality" and quantity (see, e.g., Deaton and Stone 2014; Kravdal 2014, 2019).[1] The desired number of children is one of the most important determinants of the achieved number of children (Schoen et al. 1999; Philipov et al. 2015, p. e.g., 168; Cleland et al. 2020; Yeatman et al. 2020), and the number of children in a family will therefore be related to, most often unobserved or even unobservable, characteristics of the parents that affect the life chances of both the children and the parents themselves. These characteristics are confounders in the model and will bias naïve attempts to estimate the effect.

So-called "natural experiments" are in widespread use in both economics and other disciplines, including demography, and are heralded as a way to circumvent issues of endogeneity and get estimates of causal effects from observational data (e.g., Rutter 2007; Dunning 2008; Imbens 2010; Bollen 2012). The idea behind natural experiments is "to exploit situations where the forces of nature or government policy have conspired to produce an

[1] Gong, Stinebrickner, and Stinebrickner (2020) and Wiswall and Zafar (2021) show that people make plans and have realistic expectations regarding their future work and family life.

environment somewhat akin to a randomized experiment" (Angrist and Krueger 2001, p. 73). These environments are then thought to be able to "assign the variable of interest randomly"[2] (Angrist and Krueger 2001, p. 72; see also, e.g., Angrist and Pischke 2009, pp. 21, 151–152). When true, 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 often used as the basis for instrumental variables (IVs). The birth of twins is a well-known and widespread example of a natural experiment used as the basis for instrumental variables in demography. The fact that twin births are "as good as randomly assigned" (Angrist and Pischke 2009, p. 160) have been thought enough to use them as IVs for the number of children in families (see Clarke 2018 for an overview of this literature). Using parity-specific twin births as IVs has been considered the "gold standard" method to estimate the causal effect of the number of siblings on the children or the number of children on the parents.

Recently, a number of papers have been published that criticize this "gold standard" method of using parity-specific twin births as the basis for IVs (Braakmann and Wildman 2016; Mogstad and Wiswall 2016; Guo et al. 2017; Farbmacher et al. 2018; Bhalotra and Clarke 2019, 2020; see also: Rosenzweig and Wolpin 2000; Rosenzweig and Zhang 2009).

---

[2] The statement that natural experiments can "assign the variable of interest randomly" is misleading in two ways. What we can find (in all but exceptional cases) is a natural experiment that assigns a treatment that is more or less closely related to the variable of interest as good as randomly for a specific subpopulation. I will explain these points further in this paper.

Importantly, several of these critical studies show that violations of the necessary assumptions that are both plausible and mild lead to substantively important biases of the results.[3] Further, these biases work to hide the true effect from the number of children or siblings on the parents or children. The biases will therefore have contributed to the pattern in previous results of finding a negative association but no negative effect when using a twin birth IV (see, for example, Black et al. 2005, 2010; Cáceres-Delpiano 2006, 2012a; Angrist et al. 2010; Åslund and Grönqvist 2010; Marteleto and de Souza 2012; Ponczek and Souza 2012; Baranowska-Rataj et al. 2017; Bhalotra and Clarke 2020).

In this paper, I will summarize and add to the critique of using parity-specific twin births as the basis for IVs for the number of children in the family. I will also, more briefly, criticize other IVs used in the literature, including the sexes of the first-born children. Because it is clear to me that many researchers that have applied this method (and/or have evaluated other researchers' use of this method) do not understand IVs properly, I will begin with a brief description of the method, its foundations, and the necessary assumptions. I then use these descriptions in my evaluation of IVs used for the number of children in the family. The reader

[3] Betz, Cook, and Hollenbach (2020) raise a very similar critique against common spatial instrumental variables, such as rainfall and natural disasters, because there will often be spatial correlations. They show that even relatively weak spatial correlations lead to substantively different results in these cases. Van Kippersluis and Rietveld (2018) show that also studies utilizing "Mendelian randomization" (i.e., using specific genes as instrumental variables) are subject to these biases. Adjusting for biologically plausible correlations between different genetic traits, again, lead to substantively different results.

that knows the foundations for estimating causal effects using IVs in general and the LATE estimator in particular can proceed to Section 3.

2 Background: An introduction to using IVs to estimate causal effects

Almost all applications of IVs in the social sciences in general only allow us to estimate the local average treatment effect (LATE, Imbens and Angrist 1994; Angrist and Imbens 1995; Angrist et al. 1996). The LATE is, as I will discuss further below, different from the effect that we set out to estimate but can still be of interest because of its strong internal validity (e.g., Imbens 2010, 2014). We are limited to estimating the LATE because the exogenous events or circumstances we use as the basis for our IVs affect different units of observation differently. (In econometrics lingo, we have less than perfect compliance with the assignment or a heterogenous effect of the IV on the treatment.)[4]

As social scientists, we are used to that people can react in many different ways to (seemingly) identical events or circumstances. Social scientists can still use natural experiments because we can identify subpopulations that will be affected in the same way by

[4] The LATE is also suitable for situations in which the effect of the treatment varies across units (i.e., heterogeneous treatment effects). When we are using the LATE, we estimate the effect for the subpopulation of units which have their treatment changed by the assignment mechanism (i.e., the compliers). This effect can (and will most often) be different from the effect for other groups. It is important to note that whereas the effect of the treatment can be different for compliers compared to other groups in the population, the effect of the treatment cannot vary among the compliers (see Öberg 2021 for further discussion).

an exogenous event or circumstance. Such events or circumstances can assign the treatment as good as randomly for this subpopulation which allows us to estimate a causal effect for this subpopulation.[5] This effect is what we can use the LATE estimator to estimate.

Any claim to estimate a causal effect is conditioned on a framework for thinking about what constitutes a cause and its effect. The claims that the LATE estimator can retrieve causal effects is (as are many other estimators of causal effects) based on the potential outcomes framework (Angrist et al. 1996; Imbens and Rubin 2015). There is no single framework that is suitable for answering all types of scientific questions (e.g., Heckman 2005, 2010; Imbens 2010; Krieger and Davey Smith 2016). How children and parents are affected by the number of children in the family is a substantive policy question with relevance for scientific theories that are of the type "effect of causes". They can therefore be successfully analyzed using the potential outcomes framework (Holland 1986; see also Heckman 2010, p. 361). Below, I will introduce the potential outcomes framework and how it is used when estimating causal LATEs using IVs.

2.1 The potential outcomes framework terminology and the design of (natural) experiments

The potential outcomes framework conceptualizes the estimation of a causal effect in terms of a designed experiment (Morgan and Winship 2015; Imbens and Rubin 2015). This conceptualization does not mean that the framework is valid only for designed experiments.

[5] There are several other requirements and assumptions that need to be met for finding a valid natural experiment or IV. I will discuss these requirements and assumptions further below.

The arguments are applicable to most attempts to estimate causal effects, including in social sciences in which experiments are frequently impossible or unethical. Conceptualizing the research design as an experiment is useful for highlighting the, often implicit, assumptions made when we estimate causal effects. When we are using natural experiments to estimate causal effects it is self-evident that we need to think about our research design in terms of an experiment.[6]

The terminology of designed experiments has not been widely used in the social sciences. This is not because social scientists are doing something fundamentally different, but rather just that we have used different words for the same things.

We want to estimate the effect some factor has on an "outcome". The outcome is more commonly called the dependent or explained variable in social sciences. The factor that we want to estimate the effect of is most often called the independent or explanatory variable of interest. This factor is called the "treatment" in the potential outcomes framework. In all cases we start out with having an idea on how the outcome is related to the factor that we want to

---

[6] Some authors indeed argue that all observational studies should be conceptually designed to mimic a hypothetical designed experiment of the issue that is investigated (Hernán 2016, p. 676, 2018; Bind and Rubin 2021; Moreno-Betancur 2021). Indeed, Bind and Rubin (2021) argue that it is necessary to mimic a conceptual designed experiment for "valid Fisherian or Neymanian inferences" (i.e., using "randomization-based *p*-values or confidence intervals"). Defining a hypothetical designed experiment for one's observational study is a good idea but does not lessen the need for conceptual definitions and substantive interpretations.

estimate the effect of. We make this idea explicit in our model.[7] To estimate a causal effect, we must assume that the independent or explanatory variable is independent(!), or exogenous, in the model. The reason why we can assume that the treatment is determined independently, or exogenously, is called the "assignment mechanism" in the potential outcomes framework. Randomization of the treatment in a designed experiment is an example of an assignment mechanism. But studies using observational data must also have an assignment mechanism that works in a similar way to be able to estimate causal effects.[8] Meyer (1995, p. 153) summarizes this fact nicely: "Without the ability to experimentally vary the relevant variables, researchers should seek to find variation that is driven by factors that are clearly identified and understood."

Processes including people are, as mentioned, rarely deterministic. We must therefore allow for some variability in how people react to being assigned to treatment. I call this the "reception mechanism" (see Öberg 2021 for further discussion). The reception mechanism is our description of how different groups of units will react to their assignment and why.

---

[7] The model does not have to be expressed using equations. A useful alternative is to present the model graphically (Pearl and Mackenzie 2018; see also Morgan and Winship 2015).

[8] The assignment mechanism is not always discussed explicitly in social sciences outside the natural experiments literature. Claims that the independent variable of interest is independent, or exogenous, is instead often based on less explicit assumptions of conditional independence (i.e., that the independent variable is independent, or exogenous, when conditioned on the control variables).

In summation, a (designed or natural) experiment consists of (at least) four different aspects:

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

To analyze data from any study, we need to have a clear definition of all aspects of the experiment design, including the treatment, as well as the assignment to and receipt of treatment. The need for clear definitions of all aspects of our design is a requirement for any attempt to estimate any effect. Here I will focus on how to accomplish this when analyzing natural experiments.

2.2 The potential outcomes framework as the foundation for causal estimates

To estimate a causal effect, we must both define it conceptually and estimate it empirically (e.g., Holland 1986; Heckman 2005, p. 50, 2010; Imbens and Rubin 2015, chap. 1; Pearl and Mackenzie 2018). There is a, sometimes unrecognized, distinction between the two, but both are necessary. Conceptual definitions without any empirical estimation quickly turn into speculation about hypotheticals. However, an empirical estimation without a well-defined counterfactual situation for which the causal effect is estimated also risks becoming less than productive. The definition of the causal effect is conceptual but still has important implications for the empirical estimation of the effect. It defines the specific treatment of which we estimate the effect, and therefore, it also has consequences for how to think about the assumptions underlying the empirical estimation.

In the potential outcomes framework, the causal effect is conceptually defined for a single unit of observation as the difference in outcome between a situation in which the unit was treated and a situation in which the unit was not treated. At least one of these situations is in

practice a hypothetical, or potential outcome. However, even if we do not observe all outcomes for all units, we can (and should) think about their substantive meaning, and this will depend on what we define as the treatment.[9] The definition of the treatment is therefore a crucial part of using the potential outcomes framework to estimate causal effects (Öberg 2021).

To move from this unit-specific definition of the causal effect to the corresponding effect for the population, we substitute other comparable units for the missing observation of the single unit. To estimate the effect, we compare the treated units with the untreated units while assuming these groups of units are truly comparable.

There are different ways of making the assumption of comparability plausible. The most common way to do this in the social sciences is to adjust the estimates for the relevant background characteristics. It is a strong (and often even strenuous) assumption that we can and do adjust the estimates for all relevant background characteristics. Adding to the difficulty is the fact that we need to have perfectly measured variables for all characteristics. If the variables include measurement errors (including and especially proxy variables), they do not remove all confounding (Phillips and Smith 1993; Gelman 2011; Shear and Zumbo 2013; Schennach 2016, pp. 344–345; Westfall and Yarkoni 2016; Loken and Gelman 2017).

The idea behind using natural experiments is that the as good as random assignment to treatment makes the treated and untreated units comparable (on average and in large samples). As discussed above, the assignment mechanism will only be as good as random for a subpopulation (i.e., we need to use the LATE estimator or one of its equivalents). We can not

[9] This is called "defining the potential outcomes".

identify this subpopulation empirically. We therefore compare outcomes across the units assigned to treatment and not assigned to treatment. Some of the (un)treated units will be (not) treated because of the assignment mechanism (i.e., some of these units have had their treatment assigned as good as randomly).[10] We then rely on strong (and often even strenuous) assumptions to assume that all other units are comparable across the levels of the assignment mechanism. Comparing the outcomes across the levels of the assignment mechanism thereby corresponds to comparing the outcomes of treated and untreated units that are comparable because the treatment was determined as good as randomly.

2.3 Defining the treatment of the (natural) experiment in the potential outcomes framework

The treatment is, as mentioned, that which we are estimating the effect of. The treatment is related to the endogenous variable that we are interested in. However, the treatment will not be the same as the variable of interest (i.e., the endogenous variable) (Öberg 2021). The treatment of the (natural) experiment always includes a number of specifications and qualifications determined by the research question, the study's context, the data used, and, last but not least, the design of the (natural) experiment (see, e.g., Hernán 2016).

A sufficiently specific definition of the treatment of a (natural) experiment actually includes not only the treatment in itself but also the assignment mechanism as well as the reception mechanism. The treatment that we are estimating the effect of when using a (natural)

[10] Even if we can not observe these units empirically, we must have a clear conceptual definition of who they are, and how and why they are affected the way they are by the assignment mechanism. I will discuss this further below.

experiment design is the treatment as experienced by the units studied. This means that what we are estimating the effect of is a treatment assigned and received through specific mechanisms. It is not sufficient to rely on a (possibly confirmed) assumption that the IV affects the endogenous variable of interest. We must also consider the reasons how and why there is such an effect because both the 'how' and the 'why' are parts of the definition of the treatment that we are estimating the effect of.

There are two further assumptions that the treatment needs to fulfill, namely the two parts of the stable unit treatment value assumption (SUTVA) (e.g., Imbens and Rubin 2015, pp. 9–12, 517, 589). The first part of SUTVA requires that:

- the treatment of one unit should not influence the effect of treatment of another unit (i.e., there should be no interaction effects).

This assumption is standard in most methods and, for example, corresponds to assuming that residuals of different units are independent in an ordinary least squares regression.

The second part of SUTVA requires that:

- there should be no hidden variation in (i.e., different versions of) the treatment.

This assumption is less well known and has rarely been considered in the social sciences (for further discussion, see Öberg 2021). This assumption requires that what we specify to be the treatment of the (natural) experiment has the same effect on all

treated units.[11] If the treatment does not have the same effect on all treated units, we cannot interpret or rely on the estimated effect. With variation of the treatment effect, the estimated effect is "an artificial quantity" (Cox 1992[1958], pp. 15–19).

2.4 The potential outcomes framework and the LATE estimator

When it comes to the LATE estimator specifically, it seems that many could benefit from the insight that the estimator is, at its core, quite simple.[12] What this estimator does is comparing the outcome of the units assigned to treatment to the outcome of the units not assigned to treatment. This difference will be downwardly biased because not all units do as we intended them to in the design (i.e., not all units [not] assigned to treatment are [not] treated). We therefore use the difference in the share treated between those assigned to treatment and not to scale up the difference in outcome. What I describe here is called the Wald estimator (Angrist et al. 1996; Morgan and Winship 2015, chap. 9). We can always evaluate the design of our natural experiment conceptually using the Wald estimator. Adjusting the models for other

[11] The literature defining and describing the LATE estimator has added to the confusion on this point. The LATE is sometimes described as allowing for that the treatment has a different effect on different units (i.e., heterogenous treatment effects). The LATE estimator, indeed, allows for the treatment effect to be different for the compliers compared to other treated units. However, the effect of the treatment must be the same among the compliers (i.e., we need homogenous treatment effects among the compliers).

[12] 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 needed to evaluate its validity.

variables or interpreting the results as a LATE does not change the underlying logic of the estimator.

The LATE is defined through assumptions regarding the four possible combinations of being assigned to treatment or not, and being treated or not.[13] It is, as described above, estimated by comparing the groups defined by these four combinations of being assigned to treatment or not, and being treated or not. The four groups with different combinations of assignment and treatment status are called "types", in turn called "compliers", "always-takers", "never-takers", and "defiers" (Angrist et al. 1996; Morgan and Winship 2015, chap. 9; Imbens and Rubin 2015, chaps. 23–24).[14]. Their combinations of assignment and treatment status can be nicely summarized in a table (Table 1).

Table 1. The four types as defined by the assignment to and receipt of treatment

		Not assigned to treatment	
		Not treated	Treated
Assigned to treatment	Not treated	Never-takers	Defiers
	Treated	Compliers	Always-takers

[13] The most common case, including the examples discussed further below, is that the assignment mechanism is binary. A unit is either assigned to treatment or not. The LATE can be estimated also with other than a binary assignment (Angrist and Imbens 1995).

[14] With one-sided non-compliance, there are only three types (Imbens and Rubin 2015, chap. 23). There is only one-sided non-compliance (i.e., there are never-takers but no always-takers) in the three examples provided in this paper.

Units that behave as we intend in our design are the compliers. The compliers are the group which are (not) treated because they are (not) assigned to treatment (i.e., they comply with their assigned treatment).

The other types do not comply with their assigned treatment in different ways. The defiers do the opposite of what is intended. They choose the treatment because they are not assigned to it and choose not to receive treatment because they are assigned to it. If there are units that behave in this way, we can not estimate an interpretable and reliable LATE without invoking even further assumptions (Morgan and Winship 2015, chap. 9).[15] It may sound implausible that there are units behaving in this contrarian way, but we need to evaluate if there are any possibilities for such units in each study design. The assumption we use to claim that there are no defiers is called the monotonicity assumption. This is called the monotonicity assumption because the assignment mechanism has a monotone, one-directional, effect on the treatment (when it has an effect).

The always-takers and never-takers are partially defined based on how they would have behaved had their assignment been different. Always-takers are treated regardless of whether they were assigned to treatment or not. The always-takers who are assigned to treatment are different from compliers because they would have been treated even if they had not been assigned to it. Never-takers are the opposite. They are not treated regardless of whether they

[15] de Chaisemartin (2017) has recently proposed a new set of assumptions that can be added to proceed with an analysis when it is likely that there are defiers. Small et al. (2017) also recently introduced a new type of causal effect that can be estimated despite a presence of defiers.

were assigned to treatment or not. The never-takers who are not assigned to treatment are different from compliers because they would not have been treated even if they had been assigned to it.

Because the classification of the types depends on how the units would have behaved had their assignment been different, we cannot observe empirically which type a unit is. We observe the same combination of assignment and treatment for two different types. If a unit is assigned to treatment and is treated, then it can, for example, be either a complier or an always-taker. In practice, we are left with another, simpler cross-table but with two types in each cell (Table 2).

Table 2. The four groups within the population with different assigned treatments and different receptions of their assigned treatments

		Assigned to treatment?	
		No	Yes
Treated?	No	Compliers and Never-takers	(Defiers and) Never-takers
	Yes	(Defiers and) Always-takers	Compliers and Always-takers

The LATE is estimated by comparing the (possibly conditional on the variables we adjust for) expected values of the share treated and the outcome across the two columns of Table 2, that is, across the two levels of the instrument. We can estimate the causal effect of the treatment from the levels of the instrument because there are observations that comply with their assignment, namely, the compliers. For this group, the instrument determines the treatment, and therefore, they appear on different rows, treated and not treated.

The monotonicity assumption is often the least problematic of the necessary assumptions. In empirical work, we always have to evaluate how likely it is that there are no defiers. But, for this explanation of how the method works, we will accept the monotonicity assumption as valid and ignore the defiers in Table 2.

In addition to the compliers, there are also always-takers and never-takers in both columns. They are also ignored in this method. Morgan and Winship (2015, p. 308fn26) provide a nice summary of the assumption allowing us to ignore them:

> "In a sense, the outcomes of always takers and never takers represent a type of background noise that is ignored by the IV estimator. More precisely, always takers and never takers have a distribution of outcomes, but the distribution of these outcomes is balanced across the values of the instrument".

The observed differences across the columns will correspond to the differences between compliers provided that there are no systematic differences between the always-takers and never-takers who are indicated by the IV or not. We assume that there are no such systematic differences between the units indicated by the IV or not by relying on the independence assumption and the exclusion restriction. (As noted above, by relying on the monotonicity assumption, we assume there are no defiers.) The only acceptable systematic difference between units that are indicated by the IV or not is that a larger share of the units that are indicated are treated (because of the compliers). The independence assumption and the exclusion restriction (as well as all the other requirements and assumptions) must hold for the IV and the estimated effect to be valid. Naturally, this method (as any other method) will produce an estimate even if the requirements and assumptions are not met, but this estimate will be biased (and often uninterpretable).

We should always be able to conceptually specify the behaviors that characterizes units as one of these (three or) four types. This enables us to substantiate the necessary assumptions and evaluate them against the available knowledge. Is it reasonable to assume that there is no contrarian behavior (i.e., is it reasonable that there are no defiers)? It also becomes easier to evaluate if it is reasonable to assume that there are no systematic differences between always- and never-takers that are assigned to treatment and not when we can describe how these units behave and react conceptually. The same goes for thinking about systematic differences between treated and untreated compliers that are not related to the treatment. When designing our study, we can define the types using different definitions of the treatment to evaluate if the necessary assumptions are reasonable (Cole and Frangakis 2009, p. 5; see Öberg 2019 for an example).

2.5 Estimating a LATE using instrumental variables

IVs are used to isolate exogenous variation when we have reasons to believe that the independent or explanatory variable of interest in endogenous, e.g., because it is related to unobserved factors in the model.

The idea when using IVs is to only analyze the part of the variation in the endogenous variable that is related to the exogenous influence from the IV. If we have a binary instrument, the estimated effect is based on comparing the average level of the endogenous variable across the levels of the instrument (i.e., across units assigned to treatment and not assigned to treatment). Hence, if we have a binary instrument, we also reduce the variation in the endogenous variable to be binary, for example, a higher or a lower level of the endogenous variable.

The interpretation of the estimated effect is close to being the causal effect of the endogenous variable if the IV influences all units the same way and uniformly across the distribution of

the endogenous variable. When we estimate a LATE, the interpretation of the estimated effect is a lot more specific because the IV only influences a subpopulation of units. The interpretation of the effect is therefore specific to these units, to what characterizes these units, and to the reason they are influenced by the IV. [16]

As is well-known, when using any IV, there are strict requirements and assumptions on the model, the instrument itself, as well as the treatment. IVs used to estimate a LATE must fulfill even stricter requirements and assumptions. The requirements and assumptions are even stricter because the natural experiments for which we are limited to estimating a LATE are even less like designed experiments than (the exceptional) situations when we can estimate a more general effect using IVs.

Regarding any IV, we need it to be:

- relevant, i.e., have a substantial causal effect on the instrumented variable.[17]

We choose an IV because we think it can cause a change of the treatment for some units. For an IV to be relevant, this must be true so that there is a substantive difference in the share of

[16] The interpretation of the effect when we are not limited to the LATE will also include the reasons why the IV influences the endogenous variable.

[17] This assumption is made using different wordings in some sources, for example, "Nonzero Average Causal Effect of Z on D" (Angrist et al. 1996, p. 447), "First stage" (Angrist and Pischke 2009, p. 155), and "First-stage (population of compliers have positive probability)" (Henderson et al. 2008, p. 172).

units that are treated in the groups of units that are assigned to treatment and not assigned to treatment (i.e., a difference in the share treated between the levels of the instrument).[18]

As a special requirement to be able to estimate a LATE, we need the IV to be:

- affecting the level of the instrumented variable in only one direction (i.e., monotonicity).

The monotonicity assumption is qualifying the requirement that the instrument must be relevant. To estimate a LATE, there cannot be any units that are not treated *because* they are assigned to treatment or units that are treated *because* they are not assigned to treatment. Units with this kind of contrarian behavior are, as mentioned, called "defiers" because they defy the intended design of the study.

We also need the IV to be:

- randomly assigned (which is also called the assumption of independence).

---

[18] We test if the instrument(s) add sufficient variation to the model of the endogenous variable by conducting an *F*-test of the coefficient(s) in the first-stage model. We want to test if the unique variation added by the instrument(s) is sufficient, not the whole first-stage model. The *F*-test should therefore be conducted only on the instrument(s) (i.e., not including any potential variables we adjust the first-stage model for). A well-known rule-of-thumb is that the *F*-statistic of the instrument(s) should be above 10 to reduce the risk of issues associated with "weak instruments" (Staiger and Stock 1997; for more on weak instruments, see Andrews et al. 2019).

The other reason we choose an IV is because it is an external factor that is independent of the units studied. Ideally, the instrument works as a random assignment to treatment. As discussed above, when we study people, we must (almost) always take into consideration that different people will react differently to the assignment to treatment through the IV. What we can find are IVs that create an assignment to treatment that is "as good as random at least for a subpopulation" (Imbens 2020, p. 1152) and use this to estimate a LATE. As was also discussed above, we must have a clear conceptual idea of who is affected by the IV, how, and why (i.e., we must define the reception mechanism).

The IV must also be:

- affecting the outcome only through its effect on the treatment (the exclusion restriction).

This requirement is oftentimes the most difficult to meet. And whereas we can test some of the other assumptions, the exclusion restriction is untestable empirically. It can sometimes be plausible that there is no direct effect from the instrument on the outcome. However, what we also need is that there are no systematic differences between the units that are assigned to treatment and not by the instrument. This should be made plausible through the independence assumption. In this sense, the exclusion restriction and the independence assumption are closely related and can have similar implications when violated. If the assignment mechanism is not truly random (independence assumption is violated), we risk having relevant, systematic differences between the units that are assigned to treatment and not assigned to treatment. If the treated units are relevantly and systematically different from untreated units, the exclusion restriction is violated. On the other hand, if there are no direct effect from the IV on the outcome nor any systematic differences, we can exclude the instrument from the causal model of the outcome, hence the exclusion restriction.

The most commonly used technique to estimate a LATE using IVs is a two-stage least squares estimator. This method consists of two parts, or "stages", or "reduced form" models: one, the first stage, in which we estimate the *causal* effect of the instrument on the endogenous variable of interest, and two, the second stage, in which we estimate the *causal* effect on the outcome from the treatment. When we estimate the LATE with a binary assignment mechanism, the treatment is the change in the endogenous variable that is caused by the assignment mechanism (i.e., the IV). The first stage is then estimating the share of units that are treated because they were assigned to treatment. We use the predicted values for the endogenous variable (or share of units treated) from the first-stage model instead of the original values in the second-stage model of the outcome. These predicted values are a linear combination of the variables we use to adjust our first-stage model and the unique variation added by the instrument. Ideally, this should leave only the exogenous variation related to the IV. What is often overlooked is that it also changes the substantive interpretation of the effect that is estimated in the second stage (see Öberg 2021 for further discussion).

To be able to claim that we estimate causal effects using this two-stage procedure, both the first and the second stage need to be good enough approximations of the true causal models. Naturally, we need to be able to claim that we are using a good enough approximation of the true causal model when we estimate the causal effect of the treatment on the outcome in the second stage. But, we must also estimate a causal effect in the first stage. If the first stage is not a good enough approximation of the true causal model of the effect of the IV on the share

treated among units assigned to treatment and not, the effect estimated by the second stage will be biased and uninterpretable.[19]

To be able to claim that we are using good enough approximations of both of these true causal models is a major challenge for any social scientist. Any attempt at specifying such a model will, almost always, include other variables that we need to adjust our models for. If we must include other variables to adjust our models, we also must make sure that these variables have:

- overlapping distributions in the groups indicated by the instrument and not.[20]

Not having overlapping distributions for the variables we use to adjust the models can lead to false and/or biased results (for examples, see Atanasov and Black 2020).

In the textbooks and introductory texts describing how to estimate causal effects using IVs (that are known to me) there is nothing that require us to substantively define the effect we are estimating. The implicit assumption is that we are estimating the effect of the endogenous variable of interest. This is the case only if we can also assume that everyone reacts the same way to being assigned to treatment (or not), the assignment affects the variable of interest

---

[19] The fact that the first stage model also needs to be a good enough approximation of the true causal model is often overlooked in the literature using this method. Indeed, the routines for two-stage least squares in statistical software programs are designed to just include the variables we use to adjust the second stage model estimates plus the IV in the first stage.

[20] This is also discussed as there being common support (e.g., Huber and Wüthrich 2019).

uniformly across its distribution, and the effect of the treatment is the same for everyone. This is, as mentioned, not the case in (almost any) social science applications. What we can find is a subpopulation for which the assignment mechanism creates an as good as random variation in the treatment. The definition of this treatment includes the way that the treatment is made to change, for whom, and why (i.e., the assignment and reception mechanisms). Because of these additional specifications, the treatment of the (natural) experiment will never be the same as the endogenous variable of interest. The treatment also needs to fulfil the requirements on a treatment in the potential outcomes framework that were discussed in Section 2.3. How to define the treatment for a natural experiment is discussed further in Öberg (2021).

3 Evaluating IVs based on twin births

IVs for the number of children based on twin births were initially proposed by Rosenzweig and Wolpin in two papers published in 1980 (1980b, a), and have since been used in many (75+) studies (Table 3). Rosenzweig and Wolpin's idea was that twin births led to families getting "exogenously distributed children" (or "exogenous extra children") (see also Rosenzweig and Wolpin 1980b, p. 227, a, p. 329). This is still the motivation in the literature for using IVs based on twin births. Bhalotra and Clarke (2020, p. 3097) summarize the thinking nicely: "The idea is that children … are randomly assigned either one sibling (the control group) or two siblings (the treatment group) … Comparing these allows us to estimate causal impacts of the additional birth…".[21]

[21] Research reported in demography journals also interpret IVs based on twin births as providing "exogenous variation" in (or an "exogenous increase" of) family size, the

Thus, when we use IVs based on twin births, we think of all twin births as an attempt to assign an exogenous increase in the number of children (by one child) to a set of randomly chosen families.[22] However, the reaction to a twin birth will vary, as some families foil the mad, scientific dream by having always wanted to have (at least) one more child. For such families, the assignment of another child will only lead to them having the children they want unexpectedly fast. (And as I will discuss further below, when using this method, we assume that there are no consequences from having children unexpectedly fast.) Because families react in different ways to the birth of twins, this is a case where we are limited to estimating a LATE.[23]

---

number of siblings, or fertility (e.g., Kolk 2015; Baranowska-Rataj et al. 2016, 2017; Cools and Hart 2017; Chan et al. 2019; Hart and Cools 2019). This exogenous variation leads the researchers to claim that they are estimating "causal effects" of family size, the number of siblings, fertility, etc.

[22] For simplicity, I ignore higher order multiple births.

[23] Not all studies using IVs based on twin births are explicit about interpreting their estimated effects as LATEs. After Angrist, Lavy, and Schlosser (2005, 2010) and Angrist and Pischke (2009, pp. 160–161) followed Angrist and Evans (1998) in interpreting the estimated effect as a LATE, it has been adopted by some other studies (Cáceres-Delpiano 2006, 2012a, b; Åslund and Grönqvist 2010; Cáceres-Delpiano and Simonsen 2012; Moschion 2013; Baranowska-Rataj et al. 2016; Baranowska-Rataj and Matysiak 2016; Braakmann and Wildman 2016; Silles 2016, 2019; Brinch et al. 2017; Cools and Hart 2017; Fontaine 2018; Gabel et al. 2018; Guo et al. 2018; Ajefu 2019; Bhalotra and Clarke 2019; Diaz and Fiel 2020; Priebe 2020; Aaronson et al. 2021). The other (two-

The conceptual model underlying the use of IVs based on twin births, including the treatment, has been insufficiently discussed in the previous literature. It has not been made clear why a twin birth leads to "exogenously distributed children". The importance of these definitions is obvious in the case of twin birth IVs. In practice, what we actually do when using this method is to compare families that experienced a (parity-specific) twin birth to all other families. The mechanics of the method are really that simple. It is also quite far removed from estimating the causal effect of the number of children on the parents or siblings. The basis for such a leap in interpretation is our model and conceptual definitions. But, in the case of twin birth IVs, that leap has been made prematurely, without thinking about these definitions.

3.1 The relevance of IVs based on twin births

The idea is, as we just saw, that the twin birth leads to an exogenous increase in the number of children in a family. To be relevant as the basis for IVs for the number of children, a twin birth must cause some families to have a larger number of children than what they would have had without the twin birth. Researchers using IVs based on twin births have found support for this in the fact that families that experience twin births do, on average, have a larger number of children (e.g., Bhalotra and Clarke 2020, Tables 6 and 7). Twin births are relatively rare

---

thirds of the) studies in the literature (Table 3) have implicitly assumed that all families react in the same way to a twin birth, that a twin birth increase the number of children equally much in all families regardless of how many children they have and want to have, and that the effect of increasing the number of children have the same consequences in all families. As should be obvious, none of these assumptions are realistic. The reason we are using twin birth IVs in the first place is that we think that parents' preferences matter.

and they therefore explain only a miniscule share of the variation in the number of children in the population in general. Twin births are still relevant as the basis for IVs because they have a substantive impact on the subpopulation of families that have a larger number of children *because* of the twin birth (i.e., the compliers).

3.2 The monotonicity of twin births

The number of children will, in practice, (almost) always be larger in families indicated by a twin birth IV compared to those not indicated. The average *net* effect of the IV on the share treated will therefore be positive. However, the monotonicity assumption requires that the effect be non-negative for everyone (i.e., a twin birth must have a zero or a positive effect on the number of children for all families). There should be no parents who change their mind about wanting children when they have a singleton instead of a twin birth or parents whose fertility preferences are fundamentally changed when they experience a twin birth. The former group is not very likely even if an early death of one twin could result in a similar situation. The latter group is less unlikely, but it is difficult to evaluate how common this reaction is. The estimated effect does not have any well-defined causal interpretation if there are such exceptions in the population. In conclusion, it is not unthinkable that there are violations of the monotonicity assumption in the case of IVs based on twin birth, but these groups of defiers are likely to be small. The bias they create for the estimate of the causal effect are therefore also likely to be relatively minor. The substantive interpretation of the estimated effect is still a challenge in the presence of defiers.

3.3 The questionable randomness of twin births

Rosenzweig and Wolpin (1980b, a) explained that because the likelihood of experiencing any twin birth clearly increases with the number of births, it is necessary to standardize for the number of births (see also Rosenzweig and Wolpin 2000, p. 862fn48). Their two 1980

publications used different specifications to try to achieve this standardization. They also used the IVs to study different types of outcomes. The methodological variation has since increased rather than decreased over time (Table 3). The different specifications of the IV imply different conceptual models of what is being estimated, and the different IVs are also more or less plausible as valid instruments (for a critical discussion of invalid versions of this instrument, see Section 3.4).

The specification using parity-specific twin births was alluded to in Rosenzweig and Wolpin (1980b), but it was further elaborated in Angrist and Evans (1998). The form of the specification was then found in Angrist, Lavy, and Schlosser (2005, 2010), Black, Devereux, and Salvanes (2005), and Cáceres-Delpiano (2006). These studies use parity-specific twin births as the IV, for example, a twin birth as the second birth. The analysis is then conducted on a sample that includes families with at least that many births, for example, two or more births. These samples are therefore called *n*+ samples, where *n* is the parity used to define the instrument. In practice, this method allows only the impact of younger siblings to be studied. If, for example, we use a twin birth as the second birth as the IV, then we study how the first-born child is affected by having another younger sibling. Twins are almost always excluded from the analysis because of their special characteristics (e.g., Silventoinen et al. 2013). If we use a twin birth as the second birth as the IV, families that had twins as the first birth are therefore also excluded from the analysis. Using parity-specific twin births and *n*+ samples (including twins as first birth) is the only way that twin births can be used to estimate a valid and interpretable effect (or valid and interpretable if we choose to accept the other necessary assumptions).

Table 3. A methodological summary of published studies using twin births as instrumental variables

Reference	**Specification of the twin birth instrument**	**Complete fertility history?**	**Studying the effect on…**
Rosenzweig and Wolpin (1980a)	Twin as first birth	Not only complete families	Mothers
Rosenzweig and Wolpin (1980b)	Share of twin births	Not only complete families	Children
Bronars and Grogger (1994)	Twin as first birth	Not only complete families	Mothers
Angrist and Evans (1998)	Parity-specific twin births and *n*+ samples	Not only complete families	Mothers
Jacobsen, Pearce III, and Rosenbloom (1999)	Twin as first birth	Not only complete families	Mothers
Black, Devereux, and Salvanes (2005)	Parity-specific twin births and *n*+ samples	Only (or mostly) complete families	Children
Cáceres-Delpiano (2006)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Glick, Marini, and Sahn (2007)	Twin as first birth	Not only complete families	Children
Li, Zhang, and Zhu, Y. (2008)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Dayioğlu, Kirdar, and Tansel (2009)	Share of twin births	Only (or mostly) complete families	Children
Lu (2009)	Being part of a twin birth[3]	Not only complete families	Children
Rosenzweig and Zhang (2009)	Parity-specific twin births and *n*+ samples	Only (or mostly) complete families	Children
Angrist, Lavy, and Schlosser (2010)	Parity-specific twin births and *n*+ samples	Only (or mostly) complete families	Children
Åslund and Grönqvist (2010)	Parity-specific twin births and *n*+ samples	Only (or mostly) complete families	Children
Black, Devereux, and Salvanes (2010)	Parity-specific twin births and *n*+ samples	Only (or mostly) complete families	Children
de Haan (2010)	Twin as last birth	Not only complete families	Children
Hatton and Martin (2010)	Twin as last birth	Not only complete families	Children
Moschion (2010)	Twin birth in the first two parities	Not only complete families	Parents
Frenette (2011a)	Twin as second or subsequent birth	Not only complete families	Parents
Frenette (2011b)	Twin as second or subsequent birth	Not only complete families	Children
Vere (2011)	Parity-specific twin births and *n*+ samples	Not only complete families	Mothers
Cáceres-Delpiano (2012a)	Parity-specific twin births and *n*+ samples	Not only complete families	Mothers and children
Cáceres-Delpiano (2012b)	Parity-specific twin births and *n*+ samples	Not only complete families	Mothers
Cáceres-Delpiano and Simonsen (2012)	Parity-specific twin births and *n*+ samples	Not only complete families	Mothers
Marteleto and de Souza (2012)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Ponczek and Souza (2012)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Holmlund, Rainer, and Siedler (2013)	Parity-specific twin births and *n*+ samples	Only complete families	Parents and children

Marteleto and de Souza (2013)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Moschion (2013)	Parity-specific twin births and *n*+ samples	Not only complete families	Mothers
Kruk and Reinhold (2014)	Parity-specific twin births and *n*+ samples	Only complete families	Parents
Abdul-Razak, Abd Karim, and Abdul-Hakim (2015)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Kolk (2015)	Parity-specific twin births and *n*+ samples	Only complete families	Children
Baranowska-Rataj, de Luna, and Ivarsson (2016)	Parity-specific twin births and *n*+ samples	Only (or mostly) complete families	Children
Baranowska-Rataj and Matysiak (2016)	Twin as first birth	Not only complete families	Mothers
Braakmann and Wildman (2016)	Any twin birth	Not only complete families	Mothers
He and Zhu, R. (2016)	Twin as first birth	Not only complete families	Mothers
Karbownik and Myck (2016)	Parity-specific twin births and *n*+ samples	Not only complete families	Mothers
Mogstad and Wiswall (2016)	Parity-specific twin births and *n*+ samples	Only complete families	Children
Oliveira (2016a)	Twin as first birth	Only complete families	Parents and children
Oliveira (2016b)	Twin as first birth	Not only complete families	Mothers
Silles (2016)	Parity-specific twin births and *n*+ samples	Only complete families	Mothers
Arouri, Ben-Youssef, and Nguyen, C. V. (2017)	Twin as first birth	Not only complete families	Parents
Baranowska-Rataj, Barclay, and Kolk (2017)	Parity-specific twin births and *n*+ samples	Only (or mostly) complete families	Children
Brinch, Mogstad, and Wiswall (2017)	Parity-specific twin births and *n*+ samples	Only (or mostly) complete families	Children
Chen, Q. (2017)	Twin birth in the first two parities	Not only complete families	Children
Cools and Hart (2017)	Parity-specific twin births and *n*+ samples	Only complete families	Parents and children
de Jong, Smits, and Longwe (2017)	Any twin birth	Not only complete families	Mothers
Lundborg, Plug, and Rasmussen (2017)	Twin as first birth	Not only complete families	Mothers/Women
Nguyen, C. V., and Tran, A. (2017)	Parity-specific twin births and *n*+ samples	Not only complete families	Mothers
Shen, Zou, J., and Liu, X. (2017)	Any twin birth	Only (or mostly) complete families	Children
Zhang, Junchao (2017a)	Parity-specific twin births and *n*+1 samples	Not only complete families	Mothers
Bonner and Sarkar (2018)	Being part of a twin birth	Not only complete families	Children
Dasgupta and Solomon (2018)	Twins among the younger siblings	Not only complete families	Children
Farbmacher, Guber, and Vikström (2018)	Parity-specific twin births and *n*+ samples	Not only complete families	Mothers
Fontaine (2018)	Parity-specific twin births and *n*+ samples	Not only complete families	Mothers
Gabel et al. (2018)	Parity-specific twin birth and *n*+ sample	Only complete families	Parents
Guo et al. (2018)	Twin as first birth	Only complete families	Parents

Ajefu (2019)	Any twin birth	Not only complete families	Mothers
Alidou and Verpoorten (2019)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Bhalotra and Clarke (2019)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Chan, Henderson, and Stuchbury (2019)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Chen, S.-H., Chen, Y.-C., and Liu, J.-T. (2019)	Parity-specific twin birth and *n*+ sample	Only complete families	Children
Fletcher and Kim (2019)	Twins among the younger siblings (parity-specific)	Not only complete families	Children
Hart and Cools (2019)	Twin as second birth	Not only complete families	Adult sibling of parent
Majbouri (2019)	Twin as first birth	Not only complete families	Mothers
Silles (2019)	Twin as first birth	Not only complete families	Mothers
Tan (2019)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Bhalotra and Clarke (2020)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Diaz and Fiel (2020)	Parity-specific twin births and *n*+ samples	Only (or mostly) complete families	Children
Feng (2020)	Any twin birth	Not only complete families	Children
Jeong and Kim (2020)	Twin as first birth	Not only complete families	Parents
Majbouri (2020)	Twin as first birth	Not only complete families	Mothers
Mont, Nguyen, C. V., and Tran, A. (2020)	Twins among the younger siblings	Not only complete families	Children
Nguyen, C. V., and Tran, A. (2020)	Twin as first birth	Not only complete families	Families
Priebe (2020)	Parity-specific twin births and *n*+ samples	Not only complete families	Mothers
Aaronsson et al. (2021)	Twin as second birth and *n*+ samples	Not only complete families	Mothers
Bagger et al. (2021)	Twin as last birth/Any twin birth	Only (or mostly) complete families	Children
Chen, C., et al. (2021)	Twin as first birth	Not only complete families	Parents
Vu and Tran, T. Q. (2021)	Parity-specific twin births and *n*+ samples/ Twin as last birth	Not only complete families	Children

Note: This summary is not necessarily complete. I have carried out an as systematic search as possible but have, most likely, missed some relevant studies. I have excluded one study published in French that do apply these instruments because I don't know French (Moschion 2009). Fitzsimons and Malde (2014, p. 36fn6) and Myrskylä and Margolis (2014, p. 1863fn12) estimated models using IVs based on twin births but did not report the details.

As mentioned, when using IVs based on parity-specific twin births, we can only study the effect that having another younger sibling has on the oldest child(ren). This has been recognized as a limitation of the method and Mogstad and Wiswall (2016) and Guo, Yi, and Zhang (2017) show that the effect of having another sibling varies between parities. Most studies investigating the effect on children use twins as the second birth as the basis for the IV. This means that these studies are estimating the effect on the firstborn child. This is potentially problematic since several previous studies have found effects on children from their birth order (e.g., Myrskylä et al. 2013; Jayachandran and Pande 2017).

In the paper published in the *Journal of Political Economy*, Rosenzweig and Wolpin (1980a) used twins as the first birth as the instrument when studying how women's labor force participation is affected by the number of children. Experiencing a twin birth as the first birth is an event that is as random as twin births ever are. However, how the parents behave after experiencing a twin birth or a single birth as the first birth is not random but determined by their desired number of children. This instrument will therefore be a poor predictor of the final number of children if it is common to desire two or more children (with the risk of it being a so-called weak instrument). Or, as Rosenzweig and Wolpin (2000, p. 863) put it, having a twin as first birth "correspond mainly to a difference in the timing of births" (see also Rosenzweig and Wolpin 1980a, p. 341).

The supposed randomness of twin births has been the most important argument for using twin birth as IVs. In the literature applying this method, it has always been known that twin births are not completely random events. Rosenzweig and Wolpin (1980b, p. 233, a, p. 336) noted that the chance of a twin birth increases with parity. Angrist and Evans (1998, pp. 458, 469) found systematic differences in race, age, and education between mothers giving births to twins or not. To date, however, it has been viewed as relatively unproblematic to assume that

they are random enough to be ignorable or "as good as randomly assigned" (Angrist and Pischke 2009, p. 160). However, recently, there have been some studies challenging this assumption by showing that even weak systematic influences on the likelihood of a twin birth have substantively important consequences for the results (Braakmann and Wildman 2016; Farbmacher et al. 2018; Bhalotra and Clarke 2019, 2020).

The randomness of twin births is, for example, threatened by the higher rate of twinning among mothers using fertility treatments. This issue is thoroughly discussed in Braakman and Wildman (2016). But knowing of the use of fertility treatments does not solve all problems. Farbmacher, Guber, and Vikström (2018) and Bhalotra and Clarke (2019) show that twin births are related to a large number of different characteristics of the mother, including her health and health-related behaviors, also in populations not using fertility treatments. Farbmacher, Guber, and Vikström (2018) argue that it is only the chance of dizygotic twins that are affected by the characteristics of the mother whereas monozygotic twins are truly random occurrences. They, therefore, propose using same-sex twins as a more robust instrument.[24]

The results in Bhalotra and Clarke (2019) contradict the solution proposed by Farbmacher, Guber, and Vikström (2018). Bhalotra and Clarke (2019) hypothesize that the reason that mothers giving birth to twins are different from other mothers is the higher demands on the female body from a twin pregnancy compared to a singleton pregnancy. The higher demands

[24] Monozygotic (or identical) twins will always be of the same sex whereas dizygotic twins have an equal chance of being of either the same sex or opposite sexes. The share of dizygotic twins is therefore lower among twins of the same sex.

on the female body during a twin pregnancy lead to higher rates of miscarriage among women that are less fit. They conclude that the consequence for using twin births as a natural experiment is that even if the conception of twins is as good as random, carrying a twin pregnancy to term is not. The women that are able to carry a twin pregnancy to term and giving birth to twins are positively selected with regard to a wide range of health-related factors.

This is an example of how the independence assumption and the exclusion restriction are closely connected. If the women giving birth to twins are positively selected, this risks affecting the outcome and therefore violate the exclusion restriction.

3.4 Critique of invalid versions of IVs based on twin births

Because the likelihood of experiencing any twin birth clearly increases with the number of births, any IV based on twin births needs to compensate for this. There are still a number of versions of the IV in use that fail to compensate for the increasing likelihood. (I have found 18 studies published from 2009 onwards that use versions of twin birth IVs that are clearly not valid, such as an indicator for any twin birth in the family [Table 3].)

The most obviously invalid IV in use is an indicator of whether the family has experienced any twin birth (as used in Braakmann and Wildman 2016; de Jong et al. 2017; Shen et al. 2017; Ajefu 2019; Feng 2020; Bagger et al. 2021). Defining the IV as an indicator of whether the study person is part of a twin birth (as used in Lu 2009; Bonner and Sarkar 2018) is a similar, and equally invalid, version of the IV. These are clearly not valid instruments because the chance of a twin birth increases with the number of births. The positive association between this IV and the realized number of children will also make it positively associated with the desired number of children and, therefore, with other important confounding factors. For the same reason, versions of this specification, such as any twin birth as a second or

subsequent birth (Frenette 2011a, b) or any twin birth among younger siblings (Dasgupta and Solomon 2018; Fletcher and Kim 2019; Mont et al. 2020), will also not be valid. Defining the IV as an indicator for a twin birth in the first two parities (as used in Moschion 2010; Chen 2017) does not produce an interpretable effect and is therefore also invalid.

Some studies have used a twin as the last birth as the instrument (de Haan 2010; Hatton and Martin 2010; Bagger et al. 2021; Vu and Tran 2021). This instrument is not plausible because it is associated with the desired number of children in two opposing directions. As always, the likelihood of experiencing a twin birth increases with the number of births. However, the likelihood of ending with a twin is simultaneously dependent on the number of children the parents want. Both parents who wanted as many children as they have with the twin birth and parents who would have preferred one instead of two more children stop having children with the twin birth. If parents want even more children, then they will instead proceed to have another birth after the twin birth. There is no clear substantive interpretation of the effect estimated using this IV.

In their 1980 paper published in *Econometrica*, Rosenzweig and Wolpin (1980b) used the share of twin births among all (completed) pregnancies as the instrument when studying how the schooling of children is affected by the number of siblings. (This specification is also used in Dayioğlu et al. 2009.) Naturally, the share of twin births will take on values within the same range regardless of the number of births. However, these values will have a different substantive meaning, and different values will be more or less common for different numbers of births. There is therefore no meaningful substantive interpretation of the effect estimated using this IV. Rosenzweig and Wolpin (2000, p. 838) also criticized this instrument as not being determined completely randomly.

3.5 The definitions of the treatment and the compliers, and the interpretation of the estimated effect when using IVs based on twin births

The conceptual model underlying the use of IVs based on twin births has, as mentioned, been insufficiently discussed in the previous literature. This is not the least true regarding the treatment. There are several careful econometric discussions of IVs based on twin births that all fail to specify the treatment or define how and why IVs based on twin births help identify causal effects (e.g., Rosenzweig and Wolpin 2000; Rosenzweig and Zhang 2009; Angrist et al. 2010; Chesher and Rosen 2013; Mogstad and Wiswall 2016; Brinch et al. 2017). Rosenzweig and Wolpin (1980b, pp. 232–233) briefly discuss why twin births can be thought to create an exogenous increase of the number of children, but fail to follow-through with a definition of the treatment.

The (endogenous) variable of interest when we are using twin birth IVs is the number of children in the family. We want to estimate the effect of the number of children on the parents or siblings. This is *not* the effect we are estimating when we use twin birth IVs. What we would ideally like to be estimating in this case is the effect of the increase in the number of children caused by the twin birth. But, we need to be a lot more specific about how we think about this increase when we define the treatment for the estimated effect.

We, as just mentioned, use IVs based on twin births because we are interested in the causal effect of the number of children in the family. The number of children is a discrete variable taking on positive integer values. However, this variable is reduced to a binary variable when we use IVs based on twin births. IVs based on (parity-specific) twin births are binary; either a family experienced a twin birth at the studied parity or not. When using binary twin birth IVs, the variation in the number of children is therefore also reduced to two different values; families that do not experience a (parity-specific) twin birth are assigned the average number

of children in that group, and families that experience a (parity-specific) twin birth are assigned the (slightly higher) average of that group. Therefore, the only variation in the number of children that is used for the analyses is that families that experience a (parity-specific) twin birth, on average, have a larger number of children than families that, instead, have a single birth at the studied parity.

The reason *why* twin births lead to an exogenous increase of the number of children in some families is the basis for the treatment we estimate the effect of (and the reason why the average number of children is slightly higher in the group of families experiencing twin births). This has been overlooked in previous research applying this method. None of the studies included in Table 1 state a clear and unambiguous definition of the treatment. The most explicit discussions of the treatment end up being ambiguous and self-contradictory.

The most unfortunate definitions of the treatment in the literature are the ones proposed by Angrist, Lavy, and Schlosser (2010). At one point in their paper, they claim that the "treatment is defined as a dummy for having another child" (Angrist et al. 2010, p. 788). This confuses a technical operationalization for the conceptual definition of the treatment. It is also not correct to define the treatment as "having another child". Their de facto definition of the treatment is also not correct. They argue that IVs based on twin births are a special case among IVs because all families that experience a twin birth comply with their assignment to treatment, i.e., there is "perfect compliance" (Angrist et al. 2010, p. e.g., 776, 788).[25] There

[25] This argument is also developed in the handbook by Angrist and Pischke (2009, pp. 160–161) and has been adopted by some other researchers (Cáceres-Delpiano 2012a, p. 156, b, p. 8; Cáceres-Delpiano and Simonsen 2012, p. 754; Baranowska-Rataj et al. 2016, p.

can only be perfect compliance with IVs based on twin births if the treatment is defined as the "extra" child being born at the studied parity because of the twin birth (i.e., the treatment is defined as experiencing a twin birth rather than a singleton birth). This interpretation thus uses (or confuses) the assignment mechanism for the treatment. This definition of the treatment does not meet the requirements on a treatment for the potential outcomes framework.

Angrist and Evans (1998, p. 452) claim that IVs based on twin births "identify the impact of moving from the second to the third child". That is also not true. Later in the paper they provide another interpretation that comes closer to being accurate: "the IV estimates reflect the effect of children on labor supply for those who have had more children than they otherwise would have because of twinning" (Angrist and Evans 1998, p. 458). This is basically correct even if it has to be specified further.[26]

Rosenzweig and Wolpin (1980a, b), as mentioned, did not clearly define the treatment when they introduced the idea of using twin birth for IVs. However, from what they write, it is still clear that they interpret the birth of an "unwanted" child in some families as the treatment (Rosenzweig and Wolpin 1980a, p. 338, b, pp. 232–233).[27] An "unwanted" child is not the

---

1267; Baranowska-Rataj and Matysiak 2016, p. 350; Fontaine 2018, pp. 881, 896; Priebe 2020, p. 856fn20).

[26] Cools and Hart (2017) stand out in the literature by providing a reasonable interpretation of the estimated effect: "The twin instrument captures the effect of an unintended third birth…" (Cools and Hart 2017, p. 34, see also 30).

[27] In the words of Rosenzweig and Wolpin (1980b, p. 232, italics added): "To the extent that multiple births from one pregnancy are unanticipated and children cannot readily be

same as any other child, but such a situation still basically amounts to being an exogenous increase in the number of children.

The entire reason why we use the IVs based on twin births, as noted above, is that we think that there is a difference between having an intended and an unintended child (i.e., that the preferences of the parents are an important confounder). When we use IVs based on twin births, we estimate the effect of having an unintended (or "unwanted") child. This limits the generalizability of the estimated effect and is one of the ways that the estimated effect is different from the original effect of interest. Several studies have also shown that unintended (or "unwanted") children have worse life chances than other children (e.g., Gipson et al. 2008; Hall et al. 2017; Lin et al. 2020).

The treatment and estimated effect also need to be defined for a specific parity. For example, if we use a twin birth as the first birth as our IV, we are estimating the effect of having an unintended second child for families that only wanted to have one child. The same goes if we specify the IV for higher parities. Another basis for claiming that twin births can lead to an exogenous increase of the number of children in some families is thus that parents have a fixed desired number of children (Rosenzweig and Wolpin 1980b, p. 232; see also, e.g., Black et al. 2005, p. 681, 2010, p. 37; Cáceres-Delpiano and Simonsen 2012, p. 754). This is not a realistic assumption (e.g., Bachrach and Morgan 2013; Gray et al. 2013; Philipov et al. 2015; Hruschka et al. 2019). Many parents, for example, tend to rationalize unintended pregnancies after the birth so that they don't report the child as being unintended or unwanted (i.e. they

---

> bought or sold, *some* households with twins will have experienced an exogenous increase in [the number of children] *N* above the level [of the desired number of children] *N** which would otherwise have been achieved."

update their fertility preferences) (Cleland et al. 2020). But the fact that the average number of children is higher in families experiencing (a parity-specific) twin birth than in other families, means that some parents must behave *as if* they have such fixed preferences.[28]

When we use IVs based on twin births, we are estimating the effect of having an unintended child at a specific parity because of the twin birth at that parity.[29] We are not estimating the effect of having an unintended child from a singleton birth. At the same time, it is not realistic to assume that there are no unintended pregnancies and singleton births (Singh et al. 2010; Alkema et al. 2013; Bachrach and Morgan 2013; Bearak et al. 2018; Lin et al. 2020; Brzozowska et al. 2021). If we allow for the possibility of unintended singleton births, we need to incorporate them into the definition of the treatment and estimated effect. We must also assume that the likelihood of an unintended pregnancy is not related to the number of children in the family. If it is more likely to have an unintended pregnancy in families with fewer children, we are estimating the effect of having an unintended child at the studied parity because of a twin birth *net* of the effect of having an unintended child born through a

[28] When we rely on the assumption that parents have a fixed desired number of children we also, in practice, assume that all parent couples stay together or, at least, that the parent couples also have the same desired number of children after one of the partners in a couple changes.

[29] We are therefore implicitly assuming that all parents are willing to risk surpassing their desired number of children to reach their desired number (i.e., they are willing to take the "risk" of having twins when they decide to get pregnant).

singleton birth.[30] The estimated effect will be unreliable and have no straightforward substantive interpretation.

---

[30] When we use IVs based on twin births, the compliers are the families that do (not) experience a twin birth at the studied parity and therefore do (not) have an unintended child being born because of this twin birth. Families that do not experience a twin birth (at the studied parity) but have an unintended child from a singleton birth are still compliers not assigned to treatment. Families that do experience a twin birth (at the studied parity) but later have an unintended child from a singleton birth are still compliers assigned to treatment. We therefore need to assume that unintended singleton births are equally likely in families assigned to treatment and not (i.e., equally likely in families which have three or two children). This is, most likely, not a realistic assumption because Lin, Pantano, and Sun (2020) show that the risk of a birth being unintended increases with birth order. If it is indeed more likely for families having two children to experience an unintended singleton birth, there will be systematic differences across families assigned to treatment and not. The estimated effect will be biased towards zero and is challenging to interpret.

We further need to assume that there are no systematic differences between families that have an "unwanted" child as a result of singleton and twin births, which is not very realistic (Gipson et al. 2008; Hall et al. 2017). It is plausible that "unwanted" *singleton* births are, on average, more common among parents with unobserved characteristics associated with worse outcomes for the children. If this plausibility is the case, then the effect that we estimate using the twin birth IVs will be positively biased (i.e., tending to hide the effect if the true effect is negative) (Angrist et al. 1996, p. 451).

There are also widespread problems with involuntary childlessness and infertility worldwide (Gurunath et al. 2011; Mascarenhas et al. 2012; Beaujouan and Berghammer 2019). If we do not assume that these problems do not exist, we need to incorporate them into our definition of the treatment and interpretation of the estimated effect. If some parents have fewer children than they would have liked, a smaller share of the (parity-specific) twin births will result in the birth of an unintended child (i.e., it will result in fewer families being compliers). We also need to assume that having fewer children than desired does not affect the outcome studied. If parents that have fewer children than they desire behave in a different way and/or treat their children differently, there will be relevant systematic differences between the units assigned to treatment and not.

3.6 Violations of the exclusion restriction when using IVs based on twin births

The (parity-specific) twin birth should be the only thing that creates systematic differences in the number of children between parents who do and do not experience it.[31] All variation in the number of children should therefore be *at* the studied parity. If, for example, we use a twin as the second birth as the IV, the only variation in the number of children related to a (valid) IV should be that some families that wanted two children had three because of the twin birth.

The LATE is an estimate of "the average causal effect of treatment for those whose treatment status is affected by the instrument" (Angrist and Imbens 1995, p. 434). In the twin IV case, these are the parents who, for example, wanted two children but had three because of the twin birth. These are the only people whose treatment status, i.e., number of children, is affected

[31] Remember that we, for example, assume that the parents we compare (i.e., the compliers) have the same fixed desired number of children.

by the instrument, i.e., the parity-specific twin birth. The treatment when using (valid) twin birth IVs is therefore reduced to being binary (i.e., having or not having an unintended child because of the [parity-specific] twin birth).

A potential objection to the fact that the treatment is binary is that the endogenous variable of interest—the number of children—is not a binary variable but rather an integer variable (taking on non-negative values). But, when we use valid IVs based on twin births, all exogenous variation in the number of children is *at* the studied parity. The only families whose treatment status (i.e., number of children) is affected by the instrument (i.e., the parity-specific twin birth) are those with parents who, for example, wanted two children but had three because of the twin birth. The families either experience a parity-specific twin birth or not. This has the consequence of them having one child more than they had intended (the treatment) or them having their intended number of children unexpectedly fast (which we assume is irrelevant). The treatment when using (valid) twin birth IVs is therefore reduced to being binary. IVs based on twin births can therefore never provide an estimate of the effect of the number of children. What they can (potentially) do is provide an estimate of, for example, the effect of having a third unintended child because of the twin birth in families that only intended to have two children. This is an example of how the substantive interpretation of the estimated effect is different from the original endogenous variable of interest when we use IVs (see Öberg 2021 for further discussion).

IVs based on (parity-specific) twin births are not valid if the twin birth has effects on families other than leading to the birth of an unintended child in some families. The instrument is, for example, not valid if a twin birth as the second birth induces some families to have four children instead of the (two or) three children they originally intended. This situation could only occur if a twin birth as the second birth changed the preferences of the parents, the costs

of fertility control, or the cost of rearing the children. Any of these factors would make families experiencing a twin birth systematically different from families not experiencing a twin birth as the second birth. Such systematic differences make IVs violate the exclusion restriction and would bias the estimated effect.

Rosenzweig and Wolpin (1980b, p. 234) discussed the possibility that families that experience a twin birth are also affected in ways other than having an "extra" child being born at the studied parity (see also Angrist and Evans 1998, p. 473). There are also examples in later studies where researchers describe indications of violations of the exclusion restriction. In their study on Israeli populations, Angrist, Lavy, and Schlosser (2010) discuss how a parity-specific twin birth affects the number of children at the studied parity but with the important difference that it has an effect "only (*or mostly*) at the parity of occurrence" (Angrist et al. 2010, p. 776, italics added). They proceed to discuss the reasons why (parity-specific) twin birth IVs are also associated with a larger number of children at higher parities (Angrist et al. 2010, p. 788fn15). These discussions are explanations of how their IVs violate the exclusion restriction. Bhalotra and Clarke (2020, pp. 3118–3120) find a very similar pattern of parity-specific twin births increasing the number of children both at the studied parity and later parities when they investigate this in data from 72 countries. Again, they therefore also find indications that the exclusion restriction is violated meaning that the estimated effect will be biased and without a clear substantive interpretation.

Rosenzweig and Zhang (2009) further argue that twin birth IVs violate the exclusion restriction because parents change their behaviors and will shift resources to non-twin siblings in the case of a twin birth. They argue that effects estimated using IVs based on twin births will be "confounded by inter-child allocation effects because of the endowment deficit and close spacing of twins" (Rosenzweig and Zhang 2009, p. 1149). This would have the consequence of making the effect estimated for older (non-twin) siblings a lower-bound

estimate of the true effect. (Rosenzweig and Zhang propose using the effect on the twins themselves as the corresponding upper-bound.)

The exclusion restriction requires that the IV (i.e., the assignment mechanism) does not have a direct effect on the outcome. When we use IVs for the number of children based on twin births, we therefore also need to assume that it makes no difference to the children or parents that two children are being born at once instead of with some time in between.[32] This is discussed in terms of the "timing" of births in the literature using these IVs. If the "timing" of births makes a difference for the children and/or the parents, this will create systematic differences between families that do and do not experience the (parity-specific) twin birth. The estimated effect would be biased and without any meaningful substantive interpretation. There are plenty of empirical evidence that the timing and spacing of births do affect both the children and the parents (Gipson et al. 2008; Conde-Agudelo et al. 2012; Kozuki et al. 2013;

---

[32] Rosenzweig and Wolpin (2000, p. 832) elaborated on this assumption and write that when we use twin birth IVs to study effects on women, it is "necessary to assume that … having twins has no effect on the costs of children for identification to be achieved". ("Costs" here include both monetary and non-monetary costs, such as parental time or energy.) We also can not allow for the possibility that having two children born at once instead of with some time in between affects the "marginal utility of leisure" (Rosenzweig and Wolpin 2000, p. 867). (As any parent will know, the assumption that the "marginal utility of leisure" would not be affected by having two children born at once is highly unlikely to be true.) These assumptions are also necessary when using twin birth IVs to study how children are affected by their number of siblings (Rosenzweig and Zhang 2009).

Hall et al. 2017; Molitoris et al. 2019; Bucher-Koenen et al. 2020; see also Rosenzweig and Zhang 2009). It is therefore likely that there are issues with effects estimated using IVs based on twin births because this assumption is violated.

There is another issue related to the timing of births that we need to consider when using IVs based on twin births. When we use this method, we can only analyze data on "complete families" (i.e., we can only study families that have reached [or surpassed] their desired number of children). As discussed above, some families experiencing the (parity-specific) twin birth will always have wanted (at least) one more child. For these families, the only consequence of the twin birth is that they have their desired number of children unexpectedly fast. (And, as we discussed above, we must assume that the timing and spacing of births makes no difference to the parents or children.) If we analyze incomplete families, some of the difference in the number of children between families that do and do not experience a (parity-specific) twin birth will be due to the fact that some families that have twins (at the studied parity) have their children unexpectedly fast.

That the observed difference in the number of children between families having or not having twins (at the studied parity) will vary and decline over time has been discussed in a number of studies in the literature (e.g., Bronars and Grogger 1994, p. 1143; Jacobsen et al. 1999, p. 456; Cáceres-Delpiano 2006, pp. 749-751fn13; Vere 2011; Braakmann and Wildman 2016; Fontaine 2018). But none of these studies have been explicit on the need to include only complete families in the study population when using IVs based on twin births. This requirement has to date been overlooked in the literature using IVs based on twin births and a

majority (c.75%) of the studies in Table 3 include incomplete families in their study populations.[33]

If we include incomplete families in our study population, we create an (assumed irrelevant) systematic difference in the number of children between families having or not having twins at the studied parity. This has the consequence of biasing the estimated effect towards zero. The only thing that should create differences in the number of children between families that are assigned or not assigned to treatment is that some families have an unintended child because of the (parity-specific) twin birth (i.e., the only useful difference is that between complete families that have reached or surpassed their desired number of children). We need to assume that this is the only difference between these groups of families that affects the outcome studied. As discussed, this includes that we must assume that the timing and spacing of births makes no difference to the parents or children. When we include incomplete

[33] The early studies after Rosenzweig and Wolpin's 1980 papers all studied the effects of the number of children on women using the twin as first birth instrument (Bronars and Grogger 1994; Angrist and Evans 1998; Jacobsen et al. 1999). All these studies acknowledge that the twin as first birth instrument mainly affects the timing of the births rather than the final achieved family size. When studying a (short-term) effect on mothers, it makes some sense to study the effect of a temporary increase, the "timing failure" of a twin birth (e.g., Bronars and Grogger 1994, p. 1142; Angrist and Evans 1998, p. 452; Jacobsen et al. 1999, p. 457). That these early "follower" studies all studied mothers using the twin as first birth instrument can be part of the explanation for why later studies overlooked the need for including only complete families in the study population.

families, we dilute the causal effect on the outcome by overestimating the difference in the number of children. The estimated effect will therefore be biased toward zero (provided that the other assumptions hold) and will not have any clear substantive interpretation.

A number of recent studies have pointed out even more reasons to expect that the exclusion restriction is violated when we use IVs based on twin births. Mothers experiencing twin births seem to be systematically different from other mothers. One issue, thoroughly discussed in Braakman and Wildman (2016), is that twin births are much more common among mothers undergoing fertility treatments. Mothers undergoing fertility treatments are likely to be (on average) more motivated to become parents, and to have more financial, social, and emotional resources to invest in getting pregnant and in parenting. But the reason for why the women need help to get pregnant might be something that affect them in other ways as well. It is therefore unlikely that mothers undergoing fertility treatments (as a group) will be comparable to (the group of) mothers not undergoing fertility treatments. Because twin births are more common among mothers undergoing fertility treatments, there will be relevant systematic differences between mothers having or not having the (parity-specific) twin birth.

Farbmacher, Guber, and Vikström (2018) and Bhalotra and Clarke (2019) show that mothers giving birth to twins are different to mothers having only singleton births in a number of potentially relevant ways also in populations not using fertility treatments. They show that mothers giving birth to twins (on average) have better health as well as health-related behaviors (before experiencing the twin birth). This is likely to have the consequence of them being able to create more positive outcomes for themselves and for their children. The effect

estimated using IVs based on (parity-specific) twin births will be biased if mothers giving birth to twins are a positively selected sample.[34]

Bhalotra and Clarke (2020) estimate the effect on older children from having one more, unintended sibling because of the (parity-specific) twin birth with and without adjusting for the health of the mother. Adjusting the estimates for the characteristics in which the twin mothers are positively selected makes the estimates slightly more negative. This is support for the idea that mothers giving birth to twins are a positively selected sample. Farbmacher, Guber, and Vikström (2018), as mentioned, argue that monozygotic twins are truly random occurrences and propose using same-sex twins as a more robust instrument. They find a stronger negative effect on labor force participation and earnings of the mothers when using the same-sex twins instrument compared to the more common any (parity-specific) twin birth instrument; thus, also finding support for positive selection into having a twin. The positive selection of mothers giving birth to twins will therefore tend to conceal any negative effect from having a larger number of children.

[34] Robson and Smith (2011, 2012) and Ekamper and van Poppel (2021) show for two different historical populations that mothers of twins had an (on average) lower age at first birth, older age at last birth, and shorter inter-birth intervals. This resulted in a higher overall number of children being born even after accounting for the multiple births.

3.7 The need for true causal models and common support when using IVs based on twin births

There are no established models of how siblings or parents are affected by the number of children in the family. No study (known to me) argues for why they have specified a good enough approximation of the true causal models. Few studies discuss their first-stage model at all. The de facto models used in this literature include a varied range of different factors. There is agreement on the need to adjust the model for the socioeconomic status of the parents, but this is done in many different and more or less adequate ways. We should always presume that our measures for socioeconomic status are inadequate because of measurement error and unobserved differences (e.g., Dickerson and Popli 2018). No study in this literature (known to me) discusses how their results would be affected by residual confounding from socioeconomic status. If the models we use are not good enough approximations of the true causal models, the estimated effects will be biased and have no meaningful substantive interpretation.

Rosenzweig and Wolpin (2000) stress that there is also a distinction between a causal effect and a behavioral effect. They show that additional assumptions about preferences and (household) technology are required to interpret the estimated causal effect as the behavioral effect we are interested in (Rosenzweig and Wolpin 2000, p. 864).

Even if we would choose to accept that the exclusion restriction is plausible in the specific case we are studying, we need to also consider all the other uses of the IV in the literature. 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

different outcomes make it unlikely that it is in fact valid in some (or all) of these cases.[35] 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 (2021) analyze, among other examples, IVs for the number of children in a family. They 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 outcomes, this creates a paradox where most claims of the IV being relevant and valid must be wrong or else all claims are wrong.

4 Other IVs for the number of children

4.1 The sexes of the first-born children (the SameSex IV)

The most common alternative to IVs based on twin births is IVs based on the sexes of the firstborn children as proposed by Angrist and Evans (1998).[36] Angrist and Evans explained that the idea behind this IV is based on "the widely observed phenomenon of parental

---

[35] Morck and Yeung (2011, pp. 49–50) call this the "econometric tragedy of the commons". Bazzi and Clemens (2013, pp. 154–155) also discuss this issue.

[36] Recent applications include: Cools and Hart (2017), Cools et al. (2017), Kugler and Kumar (2017), Li et al. (2017), Dasgupta and Solomon (2018), Angelov et al. (2019), Chan et al. (2019), Hart and Cools (2019), Kleven et al. (2019), Nguyen (2019), Briole et al. (2020), Chen et al. (2020), Diaz and Fiel (2020), Priebe (2020), van den Broek (2020), van den Broek and Tosi (2020), Aaronson et al. (2021), Chen (2021), Dehejia et al. (2021), and Vu et al. (2021).

preferences for a mixed sibling-sex composition. In particular, parents of same-sex siblings are significantly and substantially more likely to go on to have an additional child. Because sex mix is virtually randomly assigned, a dummy for whether the sex of the second child matches the sex of the first child provides a plausible instrument for further childbearing among women with at least two children" (Angrist and Evans 1998, p. 451).

There is plenty of support for the phenomenon that some families desire to have children of both sexes (or behave as if they do). Therefore, if their first-born children are of the same sex, some of them choose to have another child even if they had not originally intended to. Norling (2018) modeled parents' preferences for 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, p. 209). Notably, this type of preference is not only present in lower-income countries (Mills and Begall 2010). 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. (2017) 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).

The results of Norling (2018) and Bisbee et al. (2017) seem to be contradictory. If the fertility decisions of more than half of all couples are affected by the sexes of their children, there should be a larger difference in the number of children between families with same-sex and mixed-sex firstborn children. What is important to remember is that the SameSex IV also is specified at a certain parity. The only way the sexes of the firstborn children can influence the final number of children in a family is if a) the parents have a preference for having children of both sexes, and b) the parents intended to have a certain number of children, and c) the children the parents intended to have are born with the same sex, and d) having children of both sexes are more important to the parents than having only as many children as they had originally intended so that they decide to have (at least) one more child to have a chance of having children of both sexes. If we use a SameSex IV defined for the first two children, it is only parents that intended to have two children but change their mind when these two are of the same sex that are induced by the instrument to change their number of children (i.e., they are the compliers of this IV).[37]

Angrist and Evans (1998) proposed the SameSex IV based on a well-established association. But to be able to estimate an interpretable effect, an IV cannot just be associated with the

[37] Just like with the twin birth IVs, the SameSex IV thus also relies on parents having a desired number of children. But, there is an important difference in how we assume that parents behave when we use twin birth and SameSex IVs. With twin birth IVs we assume that parents have a fixed desired number of children so that some parents are forced to have another child because of the twin birth. With SameSex IVs, some parents change their mind and update their desired number of children when their firstborn children are of the same sex.

variable of interest, but rather needs to have a causal effect on the variable of interest.[38] Angrist and Evans (1998) did not consider enough the reason *why* the SameSex IV causes a change in the (endogenous) variable of interest, the number of children. If they had considered the reason why the SameSex IV can cause a change in the number of children, they would have realized that the reason is that some parents *choose* to have another child even if they had not originally intended to.[39] Therefore, the SameSex IV does not create an exogenous change in the number of children but rather a specific, endogenous change.[40]

Ebenstein (2009) analyzed the SameSex IV and reach the same conclusion. He argues that it will cause a change in the number of children that is related to the parents' preferences and abilities, as well as the perceived cost of a(nother) child (see also Brinch et al. 2017, p. 1019). The SameSex IV therefore does not remove the confounding from these factors. Ebenstein concludes that "instruments may not identify the parameter of interest, especially when the

---

[38] It is a common misconception that the IV can be just associated with the variable of interest. See Goetghebeur et al. (2020, p. 4937) for a recent example of this misconception.

[39] Öberg (2021) calls this type of compliers "self-serving compliers" and concludes that they cannot be used to estimate credible causal effects.

[40] Huber (2015) used the test proposed by Huber and Mellace (2014) to test if the SameSex IV is invalid (in the data used by Angrist and Evans [1998]). The test does not reject the IV being valid. This is a reminder that there are no technical tests for if an IV makes sense conceptually.

effects of treatment are heterogeneous and parents are able to observe the costs and benefits to treatment" (Ebenstein 2009, p. 974).

There are reasons to question results based on SameSex IVs even if we would be satisfied with estimating results that cannot be interpreted as causal effects. The results in Raley and Bianchi (2006) provide reasons to question the exclusion restriction. Families are affected differently and parents behave differently depending on the sexes of their children. Further, Huber (2015) discusses that we cannot rule out the presence of defiers when using the SameSex IV 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. Violations of the exclusion restriction as well as the presence of defiers make results estimated using SameSex IVs unreliable and without any clear substantive interpretation.

4.2 Examples of other IVs proposed in the literature

Policies to control fertility have been used as the basis for IVs for the number of children in a family. The most well-known and widely used example is the one-child policy in China (e.g., Rosenzweig and Zhang 2009; Chen 2017; Li and Zhang 2017; Huang et al. 2020; Chen et al. 2021). Zhang (2017b) provides an insightful review of issues with using China's one-child policy as a source of exogenous variation in fertility. He, for example, describes how regional variation in the implementation of the policy is related to variations in economic conditions and pre-existing differences in fertility preferences.

There is yet another issue with using policies limiting the number of children as a basis for IVs. Such policies force (many or most) families to have fewer children than they would have liked to have. Such policies do not remove differences between parents with regard to how many children they would have liked and felt that they would have been able to care for. Some parents are forced to have fewer children than they would have liked. China's one-child

policy sometimes allowed parents *the choice* to have another child if the first-born was a girl. Because the effect of this policy is based on the choice of the parents, it cannot remove any associations with the parents' preferences or desired number of children. China's one-child policy is sometimes also combined with twin births for IVs. Twins as the first birth were, in some cases, the only way to have more than one child without going against the policy. What is important to remember in these cases is that the policy forced (many or most) families to have fewer children than they would have liked. The only parents that are forced by the twin birth to have two children instead of one are the parents that desired to have only one child. The other twin parents might reach or get closer to their desired number of children because of the twin birth. It is therefore not self-evident that there is any gain from IVs over a standard OLS using the observed number of children.

Several studies have also used miscarriages and/or infertility as an IV for the number of children (e.g., Hotz et al. 2005; Agüero and Marks 2008, 2011; Bratti et al. 2020; Markussen and Strøm 2020). What is being estimated when using these IVs, is the effect of having fewer children than the parents would have liked and felt that they would have been able to care for. It is *not* the effect of sibship size (as it is erroneously interpreted as in, for example, Bratti, Fiore, and Mendola 2020, e.g., p. 511).[41] How the parents and children are affected by the parents having fewer children than they would have liked will partly depend on the preferences and abilities of the parents. These IVs will therefore not remove the confounding

[41] Some studies using IVs based on miscarriages or infertility have been explicit about aiming to estimate a "timing effect" rather than an effect from the number of children (i.e., having fewer children than intended for a period of time) (e.g., Hotz et al. 2005; Markussen and Strøm 2020).

from these preferences and abilities or other related factors. The estimated effect when using these IVs will be a combination of the effects from the (from the parents' perspective) suboptimal number of children, the parents' ability and preferences (associated with a larger number of children), the actual realized number of children, and the interactions between these factors. There is, therefore, no clear substantive interpretation to the estimated effects. Further, these IVs all depend on the strenuous assumption that a miscarriage or infertility does not influence the life chances of the parents or children in any other way than reducing the number of children.

Lundborg, Plug, and Rasmussen (2017) used the success (or failure) of the first IVF treatment with the aim of estimating a "timing effect" of how labor market outcomes for women are affected by having a child. The rationale for being able to estimate this effect is that women for whom the first IVF treatment was successful had their first child earlier than the women in the comparison group. Lundborg, Plug, and Rasmussen (2017) argue for why these groups of women are comparable. The problem is that even if the outcome of the first IVF treatment is indeed as good as random[42], the comparison group will be heterogenous thus making it difficult to interpret the estimated effect. The women who got pregnant at the first IVF treatment are compared to women having one (or more) child(ren) from a later attempt, as well as women not having any children because they did not undergo any more treatments, and women not having any children because all following attempts were unsuccessful.

[42] To be compliers, all women must have been able to get pregnant from the first IVF treatment. We need to assume that they did not get pregnant because of random circumstances that are not related to the women themselves.

Yet another part of the literature proposes using network effects as IVs. For example, Blaabæk, Jæger, and Molitoris (2020) propose using the number of siblings of one of the parents as the IV for the number of children in the family. They also adjust their models with an "extended family fixed effect" (i.e., the number of children in the family of the other parent's sibling[s]). They combine this with the IV that is the in-married spouse's number of siblings "to improve causal interpretations". The first-stage coefficient on their instrument (i.e., the in-married spouse's number of siblings) is positive. As discussed in the paper, this means that the compliers among the studied families have a larger (smaller) number of children because the in-married spouse have many (few) siblings. Blaabæk, Jæger, and Molitoris also provide a reasonable substantive interpretation of the estimated effect and hypothesize that it includes the effect of "discordant fertility preferences between partners" (Blaabæk et al. 2020, p. 592).[43] This is an example of how we can make use of the specificity of the substantive meaning of the LATE to add to our knowledge (even if there are more straightforward ways of learning about this). The downside of the IV they propose is that it does not remove the confounding from the factors associated with the desired number of children. The estimated effect is the effect of a change in the number of children that is

[43] They can interpret the effect in this way because the compliers have a larger (smaller) number of children *because* of the number of siblings the in-married spouse has. The number of siblings of the in-married spouse should therefore have overridden other priorities and preferences. They do not know or adjust for the fertility preferences or the number of siblings of the spouse and therefore do not estimate the effect from differences in fertility preferences. They do adjust their models with an "extended family fixed effect" (i.e., the number of children in the family of a first cousin).

directly related to one of the parents' preferences (or possibly fecundity). It is therefore not self-evident why we should interpret the estimated effect as any more causal than that from a standard OLS. (This critique also goes for, for example, Buyukkececi et al. 2020.)

5 Summary and conclusions

When we analyze any natural experiment, we need to provide clear and carefully conceptual and substantive definitions of all aspects of the experiment as well as all the necessary assumptions. This is especially important when we are estimating a LATE rather than a general causal effect (i.e., almost all natural experiments used in social sciences). When we estimate a LATE, we need to consider for who, how, and why the IV causes a change in the variable of interest (for some units). We should then use these insights to provide a correct interpretation of the experiment's treatment and the estimated effect. By considering for who, how, and why the IV causes a change in the variable of interest we can also make sure that the IV is sensical, for example that it does not rely on self-serving compliers as the SameSex IV.

I hope to have shown that there are, as of yet, no convincing IVs for the number of children in use in the literature. To me, the twin birth IVs are the ones that are most likely to cause an exogenous change to the number of children and therefore to be useful for estimating a causal effect. But, as we have seen, there are a multitude of issues with IVs based on twin births (even in the cases using a valid version of the instrument and data on complete families only). To be able to claim that we estimate a reliable and interpretable causal effect using IVs based on twin births, we need to rely on a large number of implausible assumptions. For example, we need to assume that parents have a fixed desired number of children and that having two children being born at once instead of with some time in between has no consequences for the parents or the children. We also need to ignore all the available results indicating violations of the independence assumption and the exclusion restriction. Even if we could reasonably

assume that all the necessary assumptions are fulfilled, we would be estimating the effect on from having an unintended child being born at a specific parity because of the twin birth. We are not estimating the effect of any "extra" child, and definitely not the effect of the number of children.

A further issue when estimating any causal effect is that we need to assume to be able to specify a good-enough approximation of the true causal model and that we have perfectly measured variables of all the factors included in this model. This is at least as important when we use IVs because we then have to assume that we can specify good-enough approximations of the true causal models for both the IV's effect on the variable of interest and the variable of interest's effect on the outcome. Such claims are made even less likely to be true when we use IVs that have been used for different variables of interest as well as a large number of different outcomes, which is the case with IVs based on twin births and the SameSex IV.

Extraordinary claims require extraordinary support. Being able to estimate causal effects from observational data is an extraordinary claim. The bases for these extraordinary analyses should therefore stand up to even close scrutiny, ideally conducted by the researchers applying the method. This goes for any claim of being able to estimate a causal effect regardless of which IV or other method that is used. A similar method as the one I have applied in this paper can be used to evaluate also applications of, for example, a difference-in-difference method. Data might not be smart in and by itself, but there are also no ways to outsmart data into producing something that is not in it to begin with.

I think that previous applications have overlooked the problems I raise in this paper partly because they have not written about their analyses enough using words. It also seems as though many researchers applying this method or evaluating others' applications of it do not fully understand it. Some of them are people just like me who struggle to grasp the empirical

implications of, for example, the assumptions that the covariance is equal to zero. The potential outcomes framework provides useful tools we can use to evaluate the assumptions made using this method while maintaining a closer connection to the empirical question. We can use it to interpret the methods that we use and the assumptions that we make verbally to explain their empirical implications. I think that doing so will help us when we work to develop new methods to find exogenous variation in the number of children.

Results from studies applying twin birth, SameSex, or other IVs contribute to the current state of knowledge on how siblings and parents are affected by the number of children. We need to reevaluate the current state of the knowledge while disregarding the results from studies applying these IVs. The effect of the number of children on the parents or siblings is a policy relevant and theoretically important issue. The scientific record should therefore also be corrected to not lead to misguided decisions.

6 References

Aaronson D, Dehejia R, Jordan A, et al (2021) The Effect of Fertility on Mothers' Labor Supply over the Last Two Centuries. Econ J 131:1–32. https://doi.org/10.1093/ej/ueaa100

Abdul-Razak NA, Abd Karim MZ, Abdul-Hakim R (2015) Does Trade-Off Between Child Quantity and Child Quality Exist in Malaysia? Singap Econ Rev 60:1550021. https://doi.org/10.1142/S0217590815500216

Agüero JM, Marks MS (2008) Motherhood and Female Labor Force Participation: Evidence from Infertility Shocks. Am Econ Rev 98:500–504. https://doi.org/10.1257/aer.98.2.500

Agüero JM, Marks MS (2011) Motherhood and Female Labor Supply in the Developing World: Evidence from Infertility Shocks. J Hum Resour 46:800–826. https://doi.org/10.1353/jhr.2011.0002

Ajefu JB (2019) Does having children affect women's entrepreneurship decision? Evidence from Nigeria. Rev Econ Household 17:843–860. https://doi.org/10.1007/s11150-019-09453-2

Alidou S, Verpoorten M (2019) Family size and schooling in sub-Saharan Africa: testing the quantity-quality trade-off. J Popul Econ 32:1353–1399. https://doi.org/10.1007/s00148-019-00730-z

Alkema L, Kantorova V, Menozzi C, Biddlecom A (2013) National, regional, and global rates and trends in contraceptive prevalence and unmet need for family planning between 1990 and 2015: a systematic and comprehensive analysis. Lancet 381:1642–1652. https://doi.org/10.1016/S0140-6736(12)62204-1

Andrews I, Stock JH, Sun L (2019) Weak Instruments in Instrumental Variables Regression: Theory and Practice. Annu Rev Econ 11:727–753. https://doi.org/10.1146/annurev-economics-080218-025643

Angelov N, Johansson P, Lee M (2019) Practical causal analysis for the treatment timing effect on doubly censored duration: effect of fertility on work span. J R Stat Soc A Stat 182:1561–1585. https://doi.org/10.1111/rssa.12474

Angrist JD, Evans WN (1998) Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. Am Econ Rev 88:450–477. http://www.jstor.org/stable/116844

Angrist JD, Imbens GW (1995) Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. J Am Stat Assoc 90:431–442. https://doi.org/10.1080/01621459.1995.10476535

Angrist JD, Imbens GW, Rubin DB (1996) Identification of Causal Effects Using Instrumental Variables. J Am Stat Assoc 91:444–455. https://doi.org/10.2307/2291629

Angrist JD, Krueger AB (2001) Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. J Econ Perspect 15:69–85. https://doi.org/10.1257/jep.15.4.69

Angrist JD, Lavy V, Schlosser A (2010) Multiple Experiments for the Causal Link between the Quantity and Quality of Children. J Labor Econ 28:773–824. https://doi.org/10.1086/653830

Angrist JD, Lavy V, Schlosser A (2005) New Evidence on the Causal Link Between the Quantity and Quality of Children. National Bureau of Economic Research Working Paper. https://doi.org/10.3386/w11835

Angrist JD, Pischke J-S (2009) Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton

Arouri M, Youssef AB, Nguyen CV (2017) The More Children You Have the More Likely You Are to Smoke? Evidence from Vietnam. Oxford Dev St 45:260–275. https://doi.org/10.1080/13600818.2016.1193129

Åslund O, Grönqvist H (2010) Family size and child outcomes: Is there really no trade-off? Labour Econ 17:130–139. https://doi.org/10.1016/j.labeco.2009.05.003

Atanasov V, Black B (2020) The Trouble with Instruments: The Need for Pretreatment Balance in Shock-Based Instrumental Variable Designs. Manage Sci 67:1270–1302. https://doi.org/10.1287/mnsc.2019.3510

Bachrach CA, Morgan SP (2013) A Cognitive–Social Model of Fertility Intentions. Popul Dev Rev 39:459–485. https://doi.org/10.1111/j.1728-4457.2013.00612.x

Bagger J, Birchenall JA, Mansour H, Urzúa S (2021) Education, Birth Order and Family Size. Econ J 131:33–69. https://doi.org/10.1093/ej/ueaa089

Baranowska-Rataj A, Barclay KJ, Kolk M (2017) The Effect of the Number of Siblings on Adult Mortality: Evidence from Swedish Registers. Pop Stud-J Demog 71:43–63. https://doi.org/10.1080/00324728.2016.1260755

Baranowska-Rataj A, de Luna X, Ivarsson A (2016) Does the number of siblings affect health in midlife? Evidence from the Swedish Prescribed Drug Register. Demogr Res 35:1259–1302. https://doi.org/10.4054/DemRes.2016.35.43

Baranowska-Rataj A, Matysiak A (2016) The Causal Effects of the Number of Children on Female Employment - Do European Institutional and Gender Conditions Matter? J Labor Res 37:343–367. https://doi.org/10.1007/s12122-016-9231-6

Bazzi S, Clemens MA (2013) Blunt Instruments: Avoiding Common Pitfalls in Identifying the Causes of Economic Growth. Am Econ J-Macroecon 5:152–86. https://doi.org/10.1257/mac.5.2.152

Bearak J, Popinchalk A, Alkema L, Sedgh G (2018) Global, regional, and subregional trends in unintended pregnancy and its outcomes from 1990 to 2014: estimates from a Bayesian hierarchical model. Lancet Glob Health 6:e380–e389. https://doi.org/10.1016/S2214-109X(18)30029-9

Beaujouan E, Berghammer C (2019) The Gap Between Lifetime Fertility Intentions and Completed Fertility in Europe and the United States: A Cohort Approach. Popul Res Policy Rev 38:507–535. https://doi.org/10.1007/s11113-019-09516-3

Betz T, Cook SJ, Hollenbach FM (2020) Spatial interdependence and instrumental variable models. Political Science Research and Methods 8:646–661. https://doi.org/10.1017/psrm.2018.61

Bhalotra SR, Clarke D (2019) Twin Birth and Maternal Condition. Rev Econ Stat 101:853–864. https://doi.org/10.1162/rest_a_00789

Bhalotra SR, Clarke D (2020) The Twin Instrument: Fertility and Human Capital Investment. J Eur Econ Assoc 18:3090–3139. https://doi.org/10.1093/jeea/jvz058

Bind M-AC, Rubin DB (2021) The importance of having a conceptual stage when reporting non-randomized studies. Biostatistics & Epidemiology 5:9–18. https://doi.org/10.1080/24709360.2021.1913707

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. J Labor Econ 35:S99–S147. https://doi.org/10.1086/691280

Blaabæk EH, Jæger MM, Molitoris J (2020) Family Size and Educational Attainment: Cousins, Contexts, and Compensation. Eur J Popul 36:575–600. https://doi.org/10.1007/s10680-019-09543-y

Black SE, Devereux PJ, Salvanes KG (2005) The More the Merrier? The Effect of Family Size and Birth Order on Children's Education. Q J Econ 120:669–700. https://doi.org/10.1093/qje/120.2.669

Black SE, Devereux PJ, Salvanes KG (2010) Small Family, Smart Family? Family Size and the IQ Scores of Young Men. J Hum Resour 45:33–58. https://doi.org/10.3368/jhr.45.1.33

Bollen KA (2012) Instrumental Variables in Sociology and the Social Sciences. Annu Rev Sociol 38:37–72. https://doi.org/10.1146/annurev-soc-081309-150141

Bonner S, Sarkar D (2018) The quality-quantity trade-off among Australian children. Econ Model 70:383–389. https://doi.org/10.1016/j.econmod.2017.08.010

Braakmann N, Wildman J (2016) Reconsidering the effect of family size on labour supply: the twin problems of the twin birth instrument. J R Stat Soc A Stat 179:1093–1115. https://doi.org/10.1111/rssa.12160

Bratti M, Fiore S, Mendola M (2020) The impact of family size and sibling structure on the great Mexico–USA migration. J Popul Econ 33:483–529. https://doi.org/10.1007/s00148-019-00754-5

Brinch CN, Mogstad M, Wiswall M (2017) Beyond LATE with a discrete instrument. J Polit Econ 125:985–1039. https://doi.org/10.1086/692712

Briole S, Le Forner H, Lepinteur A (2020) Children's socio-emotional skills: Is there a quantity–quality trade-off? Labour Econ 64:101811. https://doi.org/10.1016/j.labeco.2020.101811

Bronars SG, Grogger J (1994) The Economic Consequences of Unwed Motherhood: Using Twin Births as a Natural Experiment. Am Econ Rev 84:1141–1156. https://www.jstor.org/stable/2117765

Brzozowska Z, Buber-Ennser I, Riederer B (2021) Didn't Plan One but got One: Unintended and sooner-than-intended Parents in the East and the West of Europe. Eur J Popul 37:727–767. https://doi.org/10.1007/s10680-021-09584-2

Bucher-Koenen T, Farbmacher H, Guber R, Vikström J (2020) Double Trouble: The Burden of Child-rearing and Working on Maternal Mortality. Demography 57:559–576. https://doi.org/10.1007/s13524-020-00868-6

Buyukkececi Z, Leopold T, van Gaalen R, Engelhardt H (2020) Family, Firms, and Fertility: A Study of Social Interaction Effects. Demography 57:243–266. https://doi.org/10.1007/s13524-019-00841-y

Cáceres-Delpiano J (2006) The Impacts of Family Size on Investment in Child Quality. J Hum Resour XLI:738–754. https://doi.org/10.3368/jhr.XLI.4.738

Cáceres-Delpiano J (2012a) Can We Still Learn Something From the Relationship Between Fertility and Mother's Employment? Evidence From Developing Countries. Demography 49:151–174. https://doi.org/10.1007/s13524-011-0076-6

Cáceres-Delpiano J (2012b) Impacts of Family Size on the Family as a Whole: Evidence from the Developing World. BE J Econ Anal Poli 12:17. https://doi.org/10.1515/1935-1682.2850

Cáceres-Delpiano J, Simonsen M (2012) The toll of fertility on mothers' wellbeing. J Health Econ 31:752–766. https://doi.org/10.1016/j.jhealeco.2012.05.006

Chan TW, Henderson M, Stuchbury R (2019) Family size and educational attainment in England and Wales. Pop Stud-J Demog 73:165–178. https://doi.org/10.1080/00324728.2019.1577479

Chen C, Terrizzi S, Chou S-Y, Lien H-M (2020) The effect of sibship size on educational attainment of the first born: evidence from three decennial censuses of Taiwan. Empir Econ. https://doi.org/10.1007/s00181-020-01930-3

Chen C, Zhao W, Chou S-Y, Lien H-M (2021) The effect of family size on parents' labor supply and occupational prestige: Evidence from Taiwan and Mainland China. China Econ Rev 66:101596. https://doi.org/10.1016/j.chieco.2021.101596

Chen Q (2017) Relaxed population policy, family size and parental investments in children's education in rural Northwestern China. Int J Educ Dev 54:39–50. https://doi.org/10.1016/j.ijedudev.2017.03.009

Chen Q (2021) Population policy, family size and child malnutrition in Vietnam – Testing the trade-off between child quantity and quality from a child nutrition perspective. Econ Hum Biol 41:100983. https://doi.org/10.1016/j.ehb.2021.100983

Chen SH, Chen Y-C, Liu J-T (2019) The Impact of Family Composition on Educational Achievement. J Hum Resour 54:122–170. https://doi.org/10.3368/jhr.54.1.0915.7401R1

Chesher A, Rosen AM (2013) What Do Instrumental Variable Models Deliver with Discrete Dependent Variables? Am Econ Rev 103:557–562. https://doi.org/10.1257/aer.103.3.557

Clarke D (2018) Children and Their Parents: A Review of Fertility and Causality. J Econ Surv 32:518–540. https://doi.org/10.1111/joes.12202

Cleland J, Machiyama K, Casterline JB (2020) Fertility preferences and subsequent childbearing in Africa and Asia: A synthesis of evidence from longitudinal studies in 28 populations. Pop Stud-J Demog 74:1–21. https://doi.org/10.1080/00324728.2019.1672880

Cole SR, Frangakis CE (2009) The Consistency Statement in Causal Inference: A Definition or an Assumption? Epidemiology 20:3–5. https://doi.org/10.1097/EDE.0b013e31818ef366

Conde-Agudelo A, Rosas-Bermudez A, Castaño F, Norton MH (2012) Effects of Birth Spacing on Maternal, Perinatal, Infant, and Child Health: A Systematic Review of Causal Mechanisms. Stud Family Plann 43:93–114. https://doi.org/10.1111/j.1728-4465.2012.00308.x

Cools S, Hart RK (2017) The Effect of Childhood Family Size on Fertility in Adulthood: New Evidence From IV Estimation. Demography 54:23–44. https://doi.org/10.1007/s13524-016-0537-z

Cools S, Markussen S, Strøm M (2017) Children and Careers: How Family Size Affects Parents' Labor Market Outcomes in the Long Run. Demography 54:1773–1793. https://doi.org/10.1007/s13524-017-0612-0

Cox DR (1992[1958]) Planning of experiments. John Wiley, New York

Dasgupta K, Solomon KT (2018) Family size effects on childhood obesity: Evidence on the quantity-quality trade-off using the NLSY. Econ Hum Biol 29:42–55. https://doi.org/10.1016/j.ehb.2018.01.004

Dayioğlu M, Kirdar MG, Tansel A (2009) Impact of Sibship Size, Birth Order and Sex Composition on School Enrolment in Urban Turkey. Oxford B Econ Stat 71:399–426. https://doi.org/10.1111/j.1468-0084.2008.00540.x

de Chaisemartin C (2017) Tolerating defiance? Local average treatment effects without monotonicity. Quant Econ 8:367–396. https://doi.org/10.3982/QE601

de Haan M (2010) Birth order, family size and educational attainment. Econ Edu Rev 29:576–588. https://doi.org/10.1016/j.econedurev.2009.10.012

de Jong E, Smits J, Longwe A (2017) Estimating the Causal Effect of Fertility on Women's Employment in Africa Using Twins. World Dev 90:360–368. https://doi.org/10.1016/j.worlddev.2016.10.012

Deaton A, Stone AA (2014) Evaluative and hedonic wellbeing among those with and without children at home. P Natl Acad Sci USA 111:1328–1333. https://doi.org/10.1073/pnas.1311600111

Dehejia R, Pop-Eleches C, Samii C (2021) From Local to Global: External Validity in a Fertility Natural Experiment. J Bus Econ Stat 39:217–243. https://doi.org/10.1080/07350015.2019.1639407

Diaz CJ, Fiel JE (2020) When Size Matters: IV Estimates of Sibship Size on Educational Attainment in the U.S. Popul Res Policy Rev. https://doi.org/10.1007/s11113-020-09619-2

Dickerson A, Popli G (2018) The Many Dimensions of Child Poverty: Evidence from the UK Millennium Cohort Study. Fisc Stud 39:265–298. https://doi.org/10.1111/1475-5890.12162

Dunning T (2008) Improving Causal Inference: Strengths and Limitations of Natural Experiments. Polit Res Quart 61:282–293. 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. J Hum Resour 44:955–975. https://doi.org/10.3368/jhr.44.4.955

Ekamper P, van Poppel FWA (2021) Maternal Life-Histories of Multiple Birth Mothers Compared to Singleton Only Mothers in 19th and Early 20th Century Netherlands. Historical Life Course Studies 10:101–105. https://doi.org/10.51964/hlcs9576

Farbmacher H, Guber R, Vikström J (2018) Increasing the credibility of the twin birth instrument. J Appl Economet 33:457–472. https://doi.org/10.1002/jae.2616

Feng N (2020) The Effect of Sibling Size on Children's Educational Attainment: Evidence from Indonesia. ECNU Rev Educ 2096531120921703. https://doi.org/10.1177/2096531120921703

Fitzsimons E, Malde B (2014) Empirically probing the quantity–quality model. J Popul Econ 27:33–68. https://doi.org/10.1007/s00148-013-0474-8

Fletcher JM, Kim J (2019) The effect of sibship size on non-cognitive Skills: Evidence from natural experiments. Labour Econ 56:36–43. https://doi.org/10.1016/j.labeco.2018.11.004

Fontaine I (2018) L'effet causal du nombre d'enfants sur l'offre de travail des mères : le cas de la France métropolitaine et de ses départements d'outre-mer [The causal effect of family size on mothers' labor supply: Evidence from France and its overseas regions]. Revue Economique 69:869–898. https://doi.org/10.3917/reco.695.0869 [English version: https://www.cairn-int.info/article-E_RECO_695_0869--the-causal-effect-of-family-size-on.htm]

Frenette M (2011a) How does the stork delegate work? Childbearing and the gender division of paid and unpaid labour. J Popul Econ 24:895–910. https://doi.org/10.1007/s00148-010-0307-y

Frenette M (2011b) Why do larger families reduce parental investments in child quality, but not child quality per se? Rev Econ Household 9:523–537. https://doi.org/10.1007/s11150-010-9115-0

Gabel F, Jürges H, Kruk KE, Listl S (2018) Gain a child, lose a tooth? Using natural experiments to distinguish between fact and fiction. J Epidemiol Commun H 72:552–556. https://doi.org/10.1136/jech-2017-210210

Gallen T, Raymond B (2021) Broken Instruments. Social Science Research Network. http://dx.doi.org/10.2139/ssrn.3671850

Gelman A (2011) Causality and Statistical Learning. Am J Sociol 117:955–966. https://doi.org/10.1086/662659

Gipson JD, Koenig MA, Hindin MJ (2008) The Effects of Unintended Pregnancy on Infant, Child, and Parental Health: A Review of the Literature. Stud Family Plann 39:18–38. https://doi.org/10.1111/j.1728-4465.2008.00148.x

Glick PJ, Marini A, Sahn DE (2007) Estimating the Consequences of Unintended Fertility for Child Health and Education in Romania: An Analysis Using Twins Data. Oxford B Econ Stat 69:667–691. https://doi.org/10.1111/j.1468-0084.2007.00476.x

Goetghebeur E, Cessie S le, Stavola BD, et al (2020) Formulating causal questions and principled statistical answers. Stat Med 39:4922–4948. https://doi.org/10.1002/sim.8741

Gong Y, Stinebrickner R, Stinebrickner T (2020) Marriage, children, and labor supply: Beliefs and outcomes. J Econometrics. https://doi.org/10.1016/j.jeconom.2020.03.023

Gray E, Evans A, Reimondos A (2013) Childbearing desires of childless men and women: When are goals adjusted? Adv Life Course Res 18:141–149. https://doi.org/10.1016/j.alcr.2012.09.003

Guo R, Li H, Yi J, Zhang J (2018) Fertility, household structure, and parental labor supply: Evidence from China. J Comp Econ 46:145–156. https://doi.org/10.1016/j.jce.2017.10.005

Guo R, Yi J, Zhang J (2017) Family size, birth order, and tests of the quantity–quality model. J Comp Econ 45:219–224. https://doi.org/10.1016/j.jce.2016.09.006

Gurunath S, Pandian Z, Anderson RA, Bhattacharya S (2011) Defining infertility—a systematic review of prevalence studies. Hum Reprod Update 17:575–588. https://doi.org/10.1093/humupd/dmr015

Hall JA, Benton L, Copas A, Stephenson J (2017) Pregnancy Intention and Pregnancy Outcome: Systematic Review and Meta-Analysis. Matern Child Health J 21:670–704. https://doi.org/10.1007/s10995-016-2237-0

Hart R, Cools S (2019) Identifying interaction effects using random fertility shocks. Demogr Res 40:261–278. https://doi.org/10.4054/DemRes.2019.40.10

Hatton TJ, Martin RM (2010) The effects on stature of poverty, family size, and birth order: British children in the 1930s. Oxford Econ Pap 62:157–184. https://doi.org/10.1093/Oep/Gpp034

He X, Zhu R (2016) Fertility and Female Labour Force Participation: Causal Evidence from Urban China. Manch Sch 84:664–674. https://doi.org/10.1111/manc.12128

Hearst N, Newman TB (1988) Proving Cause and Effect in Traumatic Stress: The Draft Lottery as a Natural Experiment. J Trauma Stress 1:173–180. https://doi.org/10.1002/jts.2490010205

Heckman JJ (2005) The Scientific Model of Causality. Sociol Methodol 35:1–97. https://doi.org/10.1111/j.0081-1750.2006.00164.x

Heckman JJ (2010) Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy. J Econ Lit 48:356–398. https://doi.org/10.1257/jel.48.2.356

Henderson DJ, Millimet DL, Parmeter CF, Wang L (2008) Fertility and the health of children: A nonparametric investigation. In: Fomby T, Hill RC, Millimet DL, et al. (eds) Modelling and evaluating treatment effects in econometrics. Emerald Group Publishing Limited, Bingley, pp 167–195.

Hernán MA (2016) Does water kill? A call for less casual causal inferences. Ann Epidemiol 26:674–680. https://doi.org/10.1016/j.annepidem.2016.08.016

Hernán MA (2018) The C-Word: Scientific Euphemisms Do Not Improve Causal Inference from Observational Data. Am J Public Health 108:616–619. https://doi.org/10.2105/AJPH.2018.304337

Holland PW (1986) Statistics and Causal Inference. J Am Stat Assoc 81:945–960. https://doi.org/10.1080/01621459.1986.10478354

Holmlund H, Rainer H, Siedler T (2013) Meet the Parents? Family Size and the Geographic Proximity Between Adult Children and Older Mothers in Sweden. Demography 50:903–931. https://doi.org/10.1007/s13524-012-0181-1

Hotz VJ, McElroy SW, Sanders SG (2005) Teenage Childbearing and Its Life Cycle Consequences Exploiting a Natural Experiment. J Hum Resour XL:683–715. https://doi.org/10.3368/jhr.XL.3.683

Hruschka DJ, Sear R, Hackman J, Drake A (2019) Worldwide fertility declines do not rely on stopping at ideal parities. Pop Stud-J Demog 73:1–17. https://doi.org/10.1080/00324728.2018.1513164

Huang W, Lei X, Sun A (2020) Fertility Restrictions and Life-Cycle Outcomes: Evidence from the One-Child Policy in China. Rev Econ Stat. https://doi.org/10.1162/rest_a_00921

Huber M (2015) Testing the Validity of the Sibling Sex Ratio Instrument. Labour 29:1–14. https://doi.org/10.1111/labr.12045

Huber M, Mellace G (2014) Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints. Rev Econ Stat 97:398–411. https://doi.org/10.1162/REST_a_00450

Huber M, Wüthrich K (2019) Local Average and Quantile Treatment Effects Under Endogeneity: A Review. Journal of Econometric Methods 8:20170007. https://doi.org/10.1515/jem-2017-0007

Imbens GW (2010) Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). J Econ Lit 48:399–423. https://doi.org/10.1257/jel.48.2.399

Imbens GW (2014) Instrumental Variables: An Econometrician's Perspective. Stat Sci 29:323–358. https://doi.org/10.1214/14-STS480

Imbens GW (2020) Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. J Econ Lit 58:1129–1179. https://doi.org/10.1257/jel.20191597

Imbens GW, Angrist JD (1994) Identification and Estimation of Local Average Treatment Effects. Econometrica 62:467–475. https://doi.org/10.2307/2951620

Imbens GW, Rubin DB (2015) Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, New York

Jacobsen JP, Pearce JW, Rosenbloom JL (1999) The Effects of Childbearing on Married Women's Labor Supply and Earnings: Using Twin Births as a Natural Experiment. J Hum Resour 34:449–474. https://doi.org/10.2307/146376

Jayachandran S, Pande R (2017) Why Are Indian Children So Short? The Role of Birth Order and Son Preference. Am Econ Rev 107:2600–2629. https://doi.org/10.1257/aer.20151282

Jeong SY, Kim J (2020) Asset or burden? Impact of children on parents' retirement. J Asian Econ 71:101251. https://doi.org/10.1016/j.asieco.2020.101251

Karbownik K, Myck M (2016) For some mothers more than others. How children matter for labour market outcomes when both fertility and female employment are low. Econ Transit 24:705–725. https://doi.org/10.1111/ecot.12104

Kleven H, Landais C, Søgaard JE (2019) Children and Gender Inequality: Evidence from Denmark. Am Econ J-Appl Econ 11:181–209. https://doi.org/10.1257/app.20180010

Kolk M (2015) The causal effect of an additional sibling on completed fertility: An estimation of intergenerational fertility correlations by looking at siblings of twins. Demogr Res 32:1409–1420. https://doi.org/10.4054/DemRes.2015.32.51

Kozuki N, Sonneveldt E, Walker N (2013) Residual confounding explains the association between high parity and child mortality. BMC Public Health 13:S5. https://doi.org/10.1186/1471-2458-13-S3-S5

Kravdal Ø (2014) The Estimation of Fertility Effects on Happiness: Even More Difficult than Usually Acknowledged. Eur J Population 30:263–290. https://doi.org/10.1007/s10680-013-9310-9

Kravdal Ø (2019) Research note: What kind of individual-level effects of childbearing would we ideally be interested in learning about? The important distinction between expected, unexpected, varying and general effects. J Popul Res 36:1–12. https://doi.org/10.1007/s12546-018-9218-7

Krieger N, Davey Smith G (2016) The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. Int J Epidemiol 45:1787–1808. https://doi.org/10.1093/ije/dyw114

Kruk KE, Reinhold S (2014) The effect of children on depression in old age. Soc Sci Med 100:1–11. https://doi.org/10.1016/j.socscimed.2013.09.003

Kugler AD, Kumar S (2017) Preference for Boys, Family Size, and Educational Attainment in India. Demography 54:835–859. https://doi.org/10.1007/s13524-017-0575-1

Li B, Zhang H (2017) Does population control lead to better child quality? Evidence from China's one-child policy enforcement. J Comp Econ 45:246–260. https://doi.org/10.1016/j.jce.2016.09.004

Li H, Zhang J, Zhu Y (2008) The quantity-quality trade-off of children in a developing country: Identification using Chinese twins. Demography 45:223–243. https://doi.org/10.1353/dem.2008.0006

Li J, Dow WH, Rosero-Bixby L (2017) Education Gains Attributable to Fertility Decline: Patterns by Gender, Period, and Country in Latin America and Asia. Demography 54:1353–1373. https://doi.org/10.1007/s13524-017-0585-z

Lin W, Pantano J, Sun S (2020) Birth order and unwanted fertility. J Popul Econ 33:413–440. https://doi.org/10.1007/s00148-019-00747-4

Loken E, Gelman A (2017) Measurement error and the replication crisis. Science 355:584–585. https://doi.org/10.1126/science.aal3618

Lu Y (2009) Sibship size and education in South Africa: Black–White variations. Res Soc Strat Mobil 27:110–125. https://doi.org/10.1016/j.rssm.2009.04.002

Lundborg P, Plug E, Rasmussen AW (2017) Can Women Have Children and a Career? IV Evidence from IVF Treatments. Am Econ Rev 107:1611–1637. https://doi.org/10.1257/aer.20141467

Majbouri M (2019) Twins, family size and female labour force participation in Iran. Appl Econ 51:387–397. https://doi.org/10.1080/00036846.2018.1497853

Majbouri M (2020) Fertility and the puzzle of female employment in the Middle East and North Africa. Econ Transit 28:225–244. https://doi.org/10.1111/ecot.12243

Markussen S, Strøm M (2020) Children and labor market outcomes: separating the effects of the first three children. J Popul Econ. https://doi.org/10.1007/s00148-020-00807-0

Marteleto LJ, de Souza LR (2012) The Changing Impact of Family Size on Adolescents' Schooling: Assessing the Exogenous Variation in Fertility Using Twins in Brazil. Demography 49:1453–1477. https://doi.org/10.1007/s13524-012-0118-8

Marteleto LJ, de Souza LR (2013) The Implications of Family Size for Adolescents' Education and Work in Brazil: Gender and Birth Order Differences. Soc Forces 92:275–302. https://doi.org/10.1093/sf/sot069

Mascarenhas MN, Flaxman SR, Boerma T, et al (2012) National, Regional, and Global Trends in Infertility Prevalence Since 1990: A Systematic Analysis of 277 Health Surveys. PLOS Medicine 9:e1001356. https://doi.org/10.1371/journal.pmed.1001356

Meyer BD (1995) Natural and Quasi-Experiments in Economics. J Bus Econ Stat 13:151–161. https://doi.org/10.1080/07350015.1995.10524589

Mills M, Begall K (2010) Preferences for the sex-composition of children in Europe: A multilevel examination of its effect on progression to a third child. Pop Stud-J Demog 64:77–95. https://doi.org/10.1080/00324720903497081

Miranda V, Dahlberg J, Andersson G (2018) Parents' Preferences for Sex of Children in Sweden: Attitudes and Outcomes. Popul Res Policy Rev 37:443–459. https://doi.org/10.1007/s11113-018-9462-8

Mogstad M, Wiswall M (2016) Testing the quantity–quality model of fertility: Estimation using unrestricted family size models. Quant Econ 7:157–192. https://doi.org/10.3982/QE322

Molitoris J, Barclay K, Kolk M (2019) When and Where Birth Spacing Matters for Child Survival: An International Comparison Using the DHS. Demography 56:1349–1370. https://doi.org/10.1007/s13524-019-00798-y

Mont D, Nguyen CV, Tran A (2020) The Effect of Sibship Size on Children's Outcomes: Evidence from Vietnam. Child Indic Res 13:147–173. https://doi.org/10.1007/s12187-019-09673-z

Morck R, Yeung B (2011) Economics, History, and Causation. Bus Hist Rev 85:39–63. https://doi.org/10.1017/S000768051100002X

Moreno-Betancur M (2021) The Target Trial: A Powerful Device Beyond Well-defined Interventions. Epidemiology 32:291–294. https://doi.org/10.1097/EDE.0000000000001318

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

Moschion J (2013) The Impact of Fertility on Mothers' Labour Supply in Australia: Evidence from Exogenous Variation in Family Size. Econ Rec 89:319–338. https://doi.org/10.1111/1475-4932.12042

Moschion J (2010) Reconciling Work and Family Life: The Effect of the French Paid Parental Leave. Annals of Economics and Statistics 217–246. https://doi.org/10.2307/41219166

Moschion J (2009) Offre de travail des mères en France: l'effet causal du passage de deux à trois enfants. Economie et Statistique 422:51–78. https://doi.org/10.3406/estat.2009.8018

Myrskylä M, Margolis R (2014) Happiness: Before and After the Kids. Demography 51:1843–1866. https://doi.org/10.1007/s13524-014-0321-x

Myrskylä M, Silventoinen K, Jelenkovic A, et al (2013) The association between height and birth order: evidence from 652 518 Swedish men. J Epidemiol Commun Health 67:571–577. https://doi.org/10.1136/jech-2012-202296

Nguyen CV, Tran A (2017) The Effect of Having Children on Women's Marital Status: Evidence from Vietnam. J Dev Stud 53:2102–2117. https://doi.org/10.1080/00220388.2016.1269887

Nguyen CV, Tran A (2020) Are children an incentive or a disincentive for migration? Evidence from Vietnam. Econ Transit 28:467–485. https://doi.org/10.1111/ecot.12246

Nguyen G (2019) Sibling-sex composition, childbearing and female labour market outcomes in Indonesia. J Pop Research 36:13–34. https://doi.org/10.1007/s12546-018-9210-2

Norling J (2018) Measuring heterogeneity in preferences over the sex of children. J Dev Econ 135:199–221. https://doi.org/10.1016/j.jdeveco.2018.07.004

Öberg S (2021) Treatment for natural experiments: How to improve causal estimates using conceptual definitions and substantive interpretations. SocArXiv Papers. https://doi.org/10.31235/osf.io/pkyue

Öberg S (2019) Instrumental Variables Based on Twin Births are By Definition Not Valid (v.3). SocArXiv Papers. http://doi.org/10.17605/osf.io/zux9s

Oliveira J (2016a) Fertility, Migration, and Maternal Wages: Evidence from Brazil. J Hum Capital 10:377–398. https://doi.org/10.1086/687416

Oliveira J (2016b) The value of children: Inter-generational support, fertility, and human capital. J Dev Econ 120:1–16. https://doi.org/10.1016/j.jdeveco.2015.12.002

Pearl J, Mackenzie D (2018) The Book of Why: The New Science of Cause and Effect. Allen Lane

Philipov Dimiter, Liefbroer AC, Klobas JE (2015) Reproductive Decision-Making in a Macro-Micro Perspective. Springer Netherlands, Dordrecht. http://dx.doi.org/10.1007/978-94-017-9401-5

Phillips AN, Smith GD (1993) The Design of Prospective Epidemiological Studies: More Subjects or Better Measurements? Journal of Clinical Epidemiology 46:1203–1211. https://doi.org/10.1016/0895-4356(93)90120-P

Ponczek V, Souza AP (2012) New Evidence of the Causal Effect of Family Size on Child Quality in a Developing Country. J Hum Resour 47:64–106. https://doi.org/10.3368/jhr.47.1.64

Priebe J (2020) Quasi-experimental evidence for the causal link between fertility and subjective well-being. J Popul Econ 33:839–882. https://doi.org/10.1007/s00148-020-00769-3

Robson SL, Smith KR (2011) Twinning in humans: maternal heterogeneity in reproduction and survival. P Roy Soc B-Biol Sci 278:3755–3761. https://doi.org/10.1098/rspb.2011.0573

Robson SL, Smith KR (2012) Parity progression ratios confirm higher lifetime fertility in women who bear twins. P Roy Soc B-Biol Sci 279:2512–2514. https://doi.org/10.1098/rspb.2012.0436

Rosenzweig MR, Wolpin KI (2000) Natural "Natural Experiments" in Economics. J Econ Lit 38:827–874. https://doi.org/10.1257/jel.38.4.827

Rosenzweig MR, Wolpin KI (1980a) Life-Cycle Labor Supply and Fertility: Causal Inferences from Household Models. J Polit Econ 88:328–348. https://www.jstor.org/stable/1837294

Rosenzweig MR, Wolpin KI (1980b) Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment. Econometrica 48:227–240. https://doi.org/10.2307/1912026

Rosenzweig MR, Zhang J (2009) Do Population Control Policies Induce More Human Capital Investment? Twins, Birth Weight and China's "One-Child" Policy. Rev Econ Stud 76:1149–1174. https://doi.org/10.1111/j.1467-937X.2009.00563.x

Rutter M (2007) Proceeding from Observed Correlation to Causal Inference The Use of Natural Experiments. Perspect Psychol Sci 2:377–395. https://doi.org/DOI 10.1111/j.1745-6916.2007.00050.x

Schennach SM (2016) Recent Advances in the Measurement Error Literature. Annu Rev Econ 8:341–377. https://doi.org/10.1146/annurev-economics-080315-015058

Schoen R, Astone NM, Kim YJ, et al (1999) Do Fertility Intentions Affect Fertility Behavior? J Marriage Fam 61:790–799. https://doi.org/10.2307/353578

Shear BR, Zumbo BD (2013) False Positives in Multiple Regression Unanticipated Consequences of Measurement Error in the Predictor Variables. Educ Psychol Meas 73:733–756. https://doi.org/10.1177/0013164413487738

Shen G, Zou J, Liu X (2017) Economies of scale, resource dilution and education choice in developing countries: Evidence from Chinese households. China Econ Rev 44:138–153. https://doi.org/10.1016/j.chieco.2017.03.003

Silles MA (2016) The impact of children on women's labour supply and earnings in the UK: evidence using twin births. Oxford Econ Pap 68:197–216. https://doi.org/10.1093/oep/gpv055

Silles MA (2019) The Labor Market Consequences of Teenage Childbearing. Contemp Econ Policy 37:694–713. https://doi.org/10.1111/coep.12417

Silventoinen K, Myrskylä M, Tynelius P, et al (2013) Social modifications of the multiple birth effect on IQ and body size: a population-based study of young adult males. Paediatr Perinat Epidemiol 27:380–387. https://doi.org/10.1111/ppe.12054

Singh S, Sedgh G, Hussain R (2010) Unintended Pregnancy: Worldwide Levels, Trends, and Outcomes. Stud Family Plann 41:241–250. https://doi.org/10.1111/j.1728-4465.2010.00250.x

Small DS, Tan Z, Ramsahai RR, et al (2017) Instrumental Variable Estimation with a Stochastic Monotonicity Assumption. Stat Sci 32:561–579. https://doi.org/10.1214/17-STS623

Staiger D, Stock JH (1997) Instrumental Variables Regression with Weak Instruments. Econometrica 65:557–586. https://doi.org/10.2307/2171753

Tan HR (2019) More Is Less? The Impact of Family Size on Education Outcomes in the United States, 1850–1940. J Hum Resour 54:1154–1181. https://doi.org/10.3368/jhr.54.4.0517.8768R1

van den Broek T (2020) Is having more children beneficial for mothers' mental health in later life? Causal evidence from the national health and aging trends study. Aging Ment Health. https://doi.org/10.1080/13607863.2020.1774739

van den Broek T, Tosi M (2020) The More the Merrier? The Causal Effect of High Fertility on Later-Life Loneliness in Eastern Europe. Soc Indic Res 149:733–748. https://doi.org/10.1007/s11205-019-02254-1

van Kippersluis H, Rietveld CA (2018) Pleiotropy-robust Mendelian randomization. Int J Epidemiol 47:1279–1288. https://doi.org/10.1093/ije/dyx002

Vere JP (2011) Fertility and parents' labour supply: new evidence from US census data. Oxford Econ Pap 63:211–231. https://doi.org/10.1093/oep/gpr003

Vu LH, Tran TQ (2021) Sibship composition, birth order and education: Evidence from Vietnam. Int J Educ Dev 85:102461. https://doi.org/10.1016/j.ijedudev.2021.102461

Vu LH, Tran TQ, Phung TD (2021) Children and female labor market outcomes in Vietnam. Heliyon 7:e07508. https://doi.org/10.1016/j.heliyon.2021.e07508

Westfall J, Yarkoni T (2016) Statistically Controlling for Confounding Constructs Is Harder than You Think. PLOS ONE 11:e0152719. https://doi.org/10.1371/journal.pone.0152719

Wiswall M, Zafar B (2021) Human Capital Investments and Expectations about Career and Family. J Polit Econ 129:1361–1424. https://doi.org/10.1086/713100

Yeatman S, Trinitapoli J, Garver S (2020) The enduring case for fertility desires. Demography 57:2047–2056. https://doi.org/10.1007/s13524-020-00921-4

Zhang J (2017a) A dilemma of fertility and female labor supply: Identification using Taiwanese twins. China Econ Rev 43:47–63. https://doi.org/10.1016/j.chieco.2016.12.005

Zhang J (2017b) The evolution of China's one-child policy and its effects on family outcomes. J Econ Perspect 31:141–160. https://doi.org/10.1257/jep.31.1.141