Instrumental variables based on twin births are by definition not valid

Author: Stefan Öberg, Unit for Economic History, University of Gothenburg

Address: P.O.Box 625, 405 30 Göteborg, Sweden

Phone: +46 31 786 4736

Email: stefan.oberg@econhist.gu.se

Acknowledgments:

The author gratefully acknowledges financial support from the Jan Wallanders and Tom Hedelius Foundation in the form of a Wallander Postdoc (W2014-0396:1). The author thanks James J. Heckman and three anonymous reviewers for the Journal of Political Economy for their useful comments. The author also thanks Niklas Vahlne for helpful comments and discussions.

Abstract

Twin births are a well-known and widespread example of a so-called “natural experiment”. Instrumental variables based on twin births have been used in many studies to estimate the causal effect of the number of children on the parents or siblings. I use the potential outcomes framework to show that these instrumental variables do not work as intended. They are fundamentally flawed and will always lead to severely biased estimates without any meaningful interpretation. This has been overlooked in previous research because too little attention has been paid to defining the treatment in this natural experiment. I analyze three different possible interpretations of the treatment and show that they all lead to inherent violations of the necessary assumptions. The effect of the number of on the parents or siblings is a policy relevant and theoretically important issue. The scientific record should therefore be corrected to not lead to misguided decisions.

1. Introduction

So-called “natural experiments” are in widespread use in both economics and other disciplines. They are heralded as a way to get estimates of causal effects from observational data (e.g. Angrist and Krueger 2001; Rutter 2007; Dunning 2008; Imbens 2010; Bollen 2012). To analyze data from any experiment, we need to have a clear definition of all aspects of the experiment design, including the treatment, and the assignment to and receipt of treatment. The literature analyzing natural experiments using, for example, instrumental variables (IVs) or difference-in-difference methods have primarily paid attention to the assignment to treatment. This is the random occurrence or other way—e.g. policies, weather, multiple births etc.—through which an event becomes exogenous to the population studied. A random assignment to treatment has oftentimes been thought enough to be able to estimate causal effects. In this paper, I show that we need to pay just as much attention to the other aspects of a natural experiment.

The birth of twins is a well-known and widespread example of a natural experiment. The fact that twin births are “as good as randomly assigned” (Angrist and Pischke 2008:160) 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). Studies investigating how the number of siblings affects the children or the number of children affects the parents face problems because 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). The desired number of children is one of the most important determinants of the achieved number of children, 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.

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. The use of the method is expanding rapidly also outside economics. In this paper, I critically assess this method focusing on the definition of the treatment that we are estimating the effect of when using this method. By analyzing the method using the potential outcomes framework (Imbens and Rubin 2015; Morgan and Winship 2015) I reveal that IVs based on twin births do not work as intended. Studies applying this method will always produce severely biased and/or ill-defined estimates that cannot be interpreted. The method is fundamentally flawed and cannot be fixed. This has been overlooked in previous research because not enough attention has been paid to the conceptually defining features of this natural experiment. It is time we stop using this method and return to these important issues using other methods.

The structure of the paper is as follows. Section 2 introduces twin birth IVs and discusses previous work criticizing them. Section 3 introduces the assumptions and conceptual definitions needed when using IVs in general and twin birth IVs in particular as well as the potential outcomes framework applied in this paper. I also present the simulated populations I use to illustrate the problems with these IVs. Section 4 then discusses the necessary conceptual definitions for estimating a causal effect using twin birth IVs, focusing on why we think that twin births increase the number of children in a family. This is less straightforward than one might think and is crucial for our estimation because it defines the treatment we are estimating the effect of. Section 5 use the conceptual definitions discussed in Section 4 and evaluates the assumptions needed to use twin birth IVs. Section 6 concludes.

2. Background: Previous work using and criticizing twin birth IVs

IVs for the number of children based on twin births were initially proposed by Rosenzweig and Wolpin in two papers published in 1980, and they have been used in many (60+) studies since (Table 1). Rosenzweig and Wolpin (1980a, 1980b) 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, e.g., Rosenzweig and Wolpin 2000:862fn48). 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 1). 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. I have found ten 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 (for a critical discussion of other versions of this instrument, see Appendix, Section A1).

A number of later studies tried to refine the method to disassociate the instrument from the desired number of children (Angrist and Evans 1998; Angrist, Lavy, and Schlosser 2005, 2010; Black, Devereux, and Salvanes 2005). 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). In these cases, families that had twins as the first birth are therefore also excluded from the analysis.

Table 1. 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 ¹	Not only complete families	Children
Li, Zhang, and Zhu (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)	Any twin birth	Only (or mostly) 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 as first or second birth	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
Marteletto 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
Marteleteo 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 ³	Not only complete families	Mothers
Braakmann and Wildman (2016)	Any twin birth	Not only complete families	Mothers
He and Zhu (2016)	Twin as first birth ⁴	Not only complete families	Mothers
Karbownik and Myck (2016)	Parity-specific twin births and $n+$ samples	Not only complete families	Mothers
Oliveira (2016a)	Twin as first birth	Only complete families	Parents and children
Oliveira (2016b)	Twin as first birth ²	Not only complete families	Mothers
Mogstad and Wiswall (2016)	Parity-specific twin births and $n+$ samples	Only complete families	Children
Silles (2016)	Parity-specific twin births and $n+$ samples	Only complete families	Mothers
Arouri, Ben-Youssef, and Nguyen Viet (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 complete families	Children
Chen (2017)	Twin birth in first two parities	Not only complete families	Children
de Jong, Smits, and Longwe (2017)	Any twin birth ⁵	Not only complete families	Mothers
Nguyen and Tran (2017)	Any twin birth	Not only complete families	Mothers
Shen, Zou, and Liu (2017)	Any twin birth	Only (or mostly) complete families	Children
Zhang (2017)	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
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
Silles (2018)	Twin as first birth	Not only complete families	Mothers
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
Chan, Henderson, and Stuchbury (2019)	Parity-specific twin births and $n+$ samples	Not only complete families	Children
Chen, Chen, and Liu (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

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 two studies published in French that do apply these instruments because I don't read French (Moschion 2009; Fontaine 2018).

¹ Including cases in which one of the twins had died. ² The oldest resident children were of the same age. ³ Including also children born in the same year. ⁴ Any twin birth among mothers with one or two children. ⁵ Among children under 6 years old.

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 et al. (2005, 2010), Black et al. (2005), and Cáceres-Delpiano (2006). Using parity-specific twin births as the instrument in $n+$ samples has since been considered the “gold standard” method for investigating the effect of the number of children on parents or siblings. Rather, it was considered the “gold standard” method until a number of critical papers recently emerged in the literature. IVs based on twin births are the most convincing when defined using parity-specific events and $n+$ samples, including twin as the first birth. Therefore, this type of specification is the specification that I discuss in this paper. I argue that despite its seeming robustness, it does not work because it, by definition, violates several necessary assumptions. Because the violations arise from the fundamental design of the method, they cannot be fixed.

Twin births are rare events. This fact means that when we use IVs based on twin births, we can study only the effect of the number of children for a small group of families that may or may not be completely representative. The external validity of IVs in general and IVs based on twin births in particular has been questioned by some (e.g., Heckman 1997, 2010; Moffitt 2005). The strength of this method is its supposed high degree of internal validity (e.g., Imbens 2010, 2014). However, this internal validity has been called into question in a number of different studies with mounting criticism in recent years (Rosenzweig and Wolpin 2000; Rosenzweig and Zhang 2009; Braakmann and Wildman 2016; Bhalotra and Clarke 2018a, 2018b; Farbmacher, Guber, and Vikström 2018).

To work as valid IVs, twin births cannot affect families in any other way than increasing the number of children in some families. They can, for example, not affect the cost of raising the children (see, e.g., Rosenzweig and Wolpin 2000:832, 867). Possibilities for such violations were discussed already by Rosenzweig and Wolpin (1980b:234, see also Angrist

and Evans 1998:473). Angrist et al. (2010) show indications of violations of this assumption in their empirical data and discuss the reasons why (parity-specific) twin birth IVs are also associated with a larger number of children at higher parities (Angrist et al. 2010:788fn15). These discussions are explanations of how their IVs violate the exclusion restriction. Rosenzweig and Zhang (2009) provide another reason why twin birth IVs violate the exclusion restriction. They argue that parents change their behaviors and will shift resources to non-twin siblings in the case of a twin birth. (Previous arguments that twin birth IVs violate the exclusion restriction are discussed in the Appendix, Section A2).

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 been known for a long period of time that twin births are not completely random events (e.g., Angrist and Evans 1998:469). 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 2008:160). 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; Bhalotra and Clarke 2018a; Farbmacher et al. 2018).

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. Bhalotra and Clarke (2018b) and Farbmacher et al. (2018) 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. They, for example, show that mothers giving birth to twins, on average, are healthier before the twin birth. This has the consequence of making estimates from using twin birth IVs downwardly biased because there

is positive selection into giving birth to twins. Farbmacher et al. (2018) 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 when using the same-sex twins instrument compared to the more common any (parity-specific) twin birth instrument; therefore finding support for positive selection into having a twin.

Yet another issue that has worked against finding the true effect is that the effect has been assumed to be linear. Mogstad and Wiswall (2016) and Guo, Yi, and Zhang (2017) argue that the effect of the number of siblings varies between parities. Assuming a linear effect leads to us estimating the average effect across parities. Mogstad and Wiswall (2016) confirm that the assumption of linear effects conceals the true and varying effects at different parities.

In summary, there is already a considerable amount of criticism against twin birth IVs. 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.ⁱ 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, Devereux, and Salvanes 2005, 2010; Cáceres-Delpiano 2006, 2012b; Angrist et al. 2010; Åslund and Grönqvist 2010; Marteleto and de Souza 2012; Ponczek and Souza 2012; Baranowska-Rataj, Barclay, and Kolk 2017).

Despite the criticism accumulated this far, there is still widespread confidence in the reliability of the twin birth IV method. Some of the critical studies also propose ways to try to improve on the method (e.g. Rosenzweig and Zhang 2009; Mogstad and Wiswall 2016; Farbmacher et al. 2018). My paper goes beyond the previous criticism and show that there are

even more serious issues related to the internal validity of IVs based on twin births. I think that the challenges of twin birth IVs raised previously are extremely important to consider for the literature applying these particular IVs, as well as IVs in general. Still, I ignore all the problem mentioned above in this paper and, for example, assume that twin births occur completely at random and don't affect the cost of children or contraception. The results and issues that I raise in this paper are, therefore, problems beyond what is being argued in these other challenges to the validity of IVs based on twin births.

I argue that twin birth IVs are by definition not valid and will always produce ill-defined and biased results. The method is fundamentally flawed and cannot be fixed. I think that the reason this has been overlooked in previous research is that not enough attention has been paid to the conceptual definitions of what is being analyzed. There are several thorough econometric discussions of this method (e.g. Rosenzweig and Wolpin 1980a, 1980b, 2000; Rosenzweig and Zhang 2009; Angrist et al. 2010; Mogstad and Wiswall 2016). None of them have realized that the method is nonsensical, probably because there is nothing in econometric specifications of instrumental variables that require a careful definition of what the treatment is (compare Pearl and Mackenzie 2018:86).

3. Method: The potential outcomes framework for causal analyses, assumptions, conceptual definitions, and simulations

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 2010:361).

The potential outcomes framework conceptualizes the estimation of a causal effect in terms of a designed experiment (Imbens and Rubin 2015; Morgan and Winship 2015). This conceptualization does not mean that the framework is valid only for experiments. The arguments are applicable to all attempts to estimate causal effects, including in social sciences in which experiments are frequently impossible or unethical. Conceptualizing the model as an experiment is useful for highlighting the, often implicit, assumptions made when we estimate causal effects.

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 causal effect of something; an independent or explanatory variable of interest. This is called the treatment in the potential outcomes framework. We estimate its effect on an outcome. This is the dependent or explained variable. We assume that the independent or explanatory variable is independent, or exogenous, because it is determined independently of the other factors in the model. The reason why this factor is determined independently is called the assignment to treatment in the potential outcomes framework. This is created by randomization in a controlled experiment. But even in studies using observational data we must have something in the data generating process that works in a similar way. Otherwise we cannot estimate a causal effect. Processes including humans are rarely deterministic. We must therefore allow for some variability in how people react to being assigned to treatment. This is called the receipt of treatment in the potential outcomes framework.

3.1 Estimating causal effects using instrumental variables

Starting with the assumptions is actually a poor practice for empirical work. The meaning of these assumptions depends on the how we define our model conceptually. But, for this

methodological paper I will first present the familiar assumptions needed when using any instrumental variable before moving on the specific set needed when using twin birth IVs.

IVs are used to isolate exogenous variation when we have reasons to believe that the independent or explanatory variable of interest is endogenous, i.e. related to other factors in the model (including the residuals). To use an IV, it must be certain that a number of assumptions regarding both the instrument(s) and the effect of interest are fulfilled. What assumptions are needed depends on what we can assume about the model that we estimate. (See the Appendix, Section A6, for a detailed discussion and references.)

Regarding any instrument(s), we need it to be:

- relevant, i.e., have a substantial influence on the instrumented variable.
- randomly assigned, which is also called the assumption of independence.
- affecting the outcome only through its effect on the treatment, the exclusion restriction.

If we must include other variables to adjust our models, we also must assume that these variables have:

- overlapping distributions in the groups indicated by the instrument or not.

Regarding the treatment of interest, we need it to fulfill:

- the two parts of the stable unit treatment value assumption (SUTVAs I and II):
 - the treatment of one unit should not influence the effect of treatment of another unit (i.e. no interaction effects), and
 - there should be no hidden variation in (i.e. different versions of) the treatment.

When both the IV and the model fulfill all these necessary assumptions, we can estimate the causal effect. The most common method of doing so is a two-stage least squares regression. 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 explanatory variable, and two, the second stage, in which we estimate the causal effect of the treatment on the outcome. We use the predicted values for the endogenous variable 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.

Please note that there is nothing in the specification of an IV presented above that requires us to define the effect we are estimating. The implicit assumption is that we are estimating the effect of the endogenous variable of interest. If we can also assume that everyone reacts the way to being assigned to treatment, the assignment affects the variable of interest uniformly across its distribution, and the effect of the treatment also has to be the same for everyone. In the case of twin birth IVs, this means that everyone needs to react in the same way to a twin birth, a twin birth needs to increase the number of children equally much in all families regardless of how many children they have and want to have, and the effect of increasing the number of children must 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.

But all is not lost yet. Angrist and Evans (1998:458) suggest that we estimate a so-called local average treatment effect (LATE) when we use IVs based on twin births. The LATE is a causal effect of treatment for a subset of the population, which was introduced by Imbens and Angrist (1994) and further elaborated in Angrist and Imbens (1995) and Angrist, Imbens, and Rubin (1996). The advantage of the LATE is that, given some additional requirements, we

can allow for both heterogeneous treatment effects and variation in the effect of the instrument on the treatment. This is, as we have just seen, what we should expect with twin birth IVs. More importantly, we can also allow for systematic sorting into treatment, seemingly allowing the parents' preferences to play a role in these differences. Estimating the LATE comes at the cost of requiring additional behavioral assumptions (Heckman 1997), an additional assumption of the instrument, so-called monotonicity, and a stricter version of the exclusion restriction (see, e.g., Heckman, Urzua, and Vytlačil 2006:391). See the Appendix, Section A6, for references and further discussion.ⁱⁱ

The instrument should be:

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

The number of children will (almost) always be larger in families indicated by a twin birth IV compared to those not indicated (See, Appendix, table A2). The average *net* effect will therefore be positive. This is what makes the instruments relevant. However, the monotonicity assumption requires that the effect be positive for everyone. 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 (Morgan and Winship 2015:Ch. 9).ⁱⁱⁱ 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 and difficulty that they create for the estimate of the causal effect and its

interpretation are therefore also likely to be relatively minor. I will proceed assuming that the monotonicity assumption is fulfilled in the case of twin birth IVs.

When we estimate a LATE we also need to reduce our claims of estimating a generalizable causal effect. The LATE is the causal effect of the treatment on the subset of the population that reacts as intended to being assigned to treatment or not (they are therefore compliers). This effect is not what we are primarily interested in but can still be interesting. It does require us to have a clear definition of the treatment so that we can know who is behaving as intended, thus being compliers. This last requirement has been overlooked in previous research applying this method.

3.2 Conceptual definitions for estimating causal effects

To estimate a causal effect, we must both define it conceptually and estimate it empirically (e.g., Holland 1986; Heckman 2005:50, 2010; Imbens and Rubin 2015:Ch. 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.

The conceptual model underlying the use of IVs based on twin births, including the treatment, has been insufficiently discussed in the previous literature. To borrow the harshly formulated judgment of James Heckman, the twin IV literature is an example of when analysts have wanted “something for nothing: a definition of causality without a clearly

articulated model of the phenomenon being described” (Heckman 2005:2). The estimated causal effect has not been carefully defined with regard to the treatment or the model determining the treatment, which has prevented critical discussions of the implicit assumptions made when using twin births as IVs (compare Rosenzweig and Wolpin 2000:828–29).^{iv} The importance of 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 conceptual definitions. But, in the case of twin birth IVs, that leap has been made prematurely, without thinking about these definitions.

The treatment 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. It is important to decide if we think of the increase as being a result of an “unwanted”, “another”, or “extra” child. I will discuss this thoroughly below.

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.

Families either experience a parity-specific twin birth or not—the assignment to treatment—and have an “unwanted”/“another”/“extra” child or not—the receipt of treatment. We therefore have a situation with a binary instrument and a binary treatment. Angrist et al. (1996) explained how we can use the potential outcomes framework to analyze such situations through the four types created by combining the binary instrument and treatment (See also Imbens and Rubin 2015; Morgan and Winship 2015. Greenland and Robins 1986 provide a similar discussion using different terminology.). I discuss twin birth IVs further below, assuming that both the instrument and the treatment are binary.^v

3.3 The potential outcomes framework for causal analyses applied to instrumental variables

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 families, we can think about their substantive meaning, and this will depend on what we define as the treatment. (See Appendix, Section A5, for such interpretations.) I examine the three definitions introduced above, the birth of an “unwanted” child, the birth of

“another” (e.g., third) child, and the birth of an “extra” child at the studied parity because of a twin birth.

To move from this unit-specific definition of the causal effect to the corresponding effect for the population population, we must ensure that the units we compare across are truly comparable. The potential outcomes framework presents the four types that arise from the (binary) assignment to treatment and the (binary) receipt of the assigned treatments. The four types are called compliers, always-takers, never-takers, and defiers. Their combinations of assignment and reception of treatment can be nicely summarized in a table (Table 2).

Table 2. 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

Families that behave as we intend are compliers. They have an “unwanted”/“another”/“extra” child if they experience a (parity-specific) twin birth and do not have any “unwanted”/“another”/“extra” child if they do not experience a (parity-specific) twin birth. The compliers are the group for which the treatment is affected by the assignment 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 was expected. They choose the treatment if they are not assigned to it and choose not to receive treatment if they are assigned to it. As noted above, it is not unthinkable that there are small groups of defiers in the twin birth IV case. I nonetheless accept the monotonicity assumption that there are none and ignore them in the following.

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 received the treatment even if they had not been assigned to it. Never-takers are the opposite. They are not treated regardless of whether they were assigned to treatment or not. The never-takers who are not assigned to treatment are different from compliers because they would not have received treatment even if they had been assigned to it.

Most often, therefore, we cannot observe empirically which type a unit is because we can 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. A family that does not experience a (parity-specific) twin birth and that does not have an “unwanted”/“another”/“extra” child born is either a family of either compliers or never-takers. In practice, we are left with another, simpler cross-table but with two types in each cell (Table 3).

Table 3. The four groups within the population with different assigned treatments and different receptions of their assigned treatments

		Twin birth at the studied parity?	
		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 (conditional on the variables we adjust for) expected values of the number of children and the outcome studied across the two columns of

Table 3, 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. (Again, note that we estimate the effect by comparing families that experience a (parity-specific) twin birth to all other families. Everything beyond that relies on the assumptions we make.)

In addition to the compliers, there are also always-takers and never-takers in both columns. They are ignored in this method. Morgan and Winship (2015:308fn26) 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 did and did not experience a (parity-specific) twin birth. This aspect is part of the exclusion restriction, which requires that there be (after variable adjustment) no systematic differences between the units indicated by the IV or not. The only acceptable systematic difference is that some of the families that experienced a (parity-specific) twin birth had an (exogenous) increase in their number of children, which, in turn, might have had consequences for the outcome studied. The exclusion restriction must hold for the IV to be valid. Naturally, this method will produce an estimate even if the assumption does not hold, but it will be biased.

The types thus provide us with implications and comparisons that can be used to test the assumptions underlying the IV estimation. The advantage of working with simulated data is that I can know which type each unit belongs to and therefore test the assumptions. To proceed with these evaluations, we must have a clear definition of the treatment. This definition depends on why we think that twin births increase the number of children in a family. First, I therefore discuss this issue in Section 4.1 before turning to the definition of the treatment in Section 4.2.

3.4 The simulations used to illustrate the problems with twin birth IVs

Before moving to the definitions, I need to describe the simulated populations I use to illustrate the problems with twin birth IVs. It is impossible to evaluate most of the assumptions required for IVs empirically. It is possible using a simulated population in which we can control factors that are usually unobserved or unobservable. The simulated populations are constructed using a number of assumptions that we make, more or less explicitly, when using twin birth IVs (see Appendix, Section A4, for a detailed discussion). I do so to illustrate that the problems are present also in the unlikely case that real people did behave as we assume they do when we use this method. All actual deviations from these assumptions will make the estimated effect even less well defined.

The starting point of the simulation model is a population of prospective parents desiring between one and nine children. The relative distribution of their desired number of children follows the distributions of the realized number of children in four different empirical populations with different levels of fertility (Sweden, people born 1972–1979: Åslund and Grönqvist 2010, the distribution is presented in Table 1; Norway, average year of birth 1962: Black et al. 2005:Tab. II; Saint Paul, MN, USA, children aged 0–17 years in 1920: Roberts and Warren 2017:Tab. 3; The Netherlands, men born 1944–1947: Stradford, van Poppel, and

Lumey 2017:Tab. 1). Using the realized number of children as a proxy for the desired number of children is less than perfect but has no consequences for the substantive results.

The parents in the simulated populations then experience different combinations of twin and single births until they reach or surpass their desired number of children (Figure 1). (Throughout the paper, for simplicity, I ignore higher-order multiple births.) Each combination is associated with a probability of occurring. (Also for simplicity, I assume a constant probability for a twin birth, $p = 0.0175$.^{vi}) The regressions and correlations are estimated across the 230 or 141 different combinations that are possible when using two different versions of the IV; twins as first or second birth. All different possible chains of decisions and events, as well as their respective probabilities are presented in the Appendix spreadsheet. These different possibilities, i.e. observations, are weighted by a combination of their probabilities and the likelihood that a family desired that many children. I simulated the effect of including incomplete families by only including the first five births in the data. See the Appendix, Section A4, for further details.

Figure 1. A flowchart of the process of having children

The simulation, as mentioned, relies on a number of simplifying assumptions that are presented and discussed in the Appendix, Section A4. The simplistic simulation model is a sufficiently accurate representation of reality and of the behaviors assumed when using twin births for IVs. Two of the most unrealistic assumptions in the model are that I assume that parents have, *ex ante*, a fixed number of children that they desire and that there are no unintended pregnancies. Both of these assumptions are made implicitly in the literature using IVs based on twin births. The assumption of a fixed desired number of children is unrealistic in real life (e.g. Bachrach and Morgan 2013; Philipov, Liefbroer, and Klobas 2015; Hruschka et al. 2019) but was used by Rosenzweig and Wolpin in their original derivation of the method (Rosenzweig and Wolpin 1980b:232; see also, e.g., Black et al. 2005:681). Assuming that there are no unintended pregnancies is also unrealistic (Singh, Sedge, and Hussain 2010; Alkema et al. 2013; Bachrach and Morgan 2013) but this assumption is also implicitly made in the literature.

4. Defining the treatment for IVs based on twin births

4.1 The effect of twin births on the number of children in the family

Twin births lead to two children being born at once, in contrast to the one child born in a single birth. Twin births therefore lead to an unexpected (or, at least unexpected until the first ultrasound during the pregnancy) increase in the number of children in the family. We use IVs based on twin births because we think that they can create *exogenous* variation in the number of children (e.g., Bronars and Grogger 1994:1142; Angrist and Evans 1998:474; Jacobsen, Pearce III, and Rosenbloom 1999:457; Black et al. 2005:670; Angrist et al. 2010:775; Gabel et al. 2018:555). This exogenous variation in the number of children is necessary to be able to estimate its causal effect on the parent(s) (Rosenzweig and Wolpin 1980a:335) or the children (Rosenzweig and Wolpin 1980b:232).

To create exogenous variation in the number of children, the twin birth must lead to a both unexpected *and* unintended increase in the number of children. The (parity-specific) twin birth will therefore create exogenous variation in the number of children only when combined with a specific desired number of children, which will consequently occur only in *some* families. Some families will always have intended to have (at least) one more child. For them, the twin birth only leads to them having this intended increase in their number of children faster than expected.^{vii} The method relies on the assumption that timing—i.e. having intended children faster than expected—is inconsequential, so this should not matter for the outcome. The fact that twin births lead to an exogenous increase in the number of children only in some families does not in and by itself make twin birth IVs invalid. It contributes to the fact that we could only, at best, estimate a LATE, not a general causal effect.

4.1.1 A “*timing failure*”

Rosenzweig and Wolpin (1980a) add the important insight that a twin birth will have different consequences for the number of children in the family depending on how much time has passed since the birth. A twin birth will lead to an exogenous increase in the number of children for some families, but it will lead to a “timing failure” for all “since two children appear simultaneously” (Rosenzweig and Wolpin 1980a:338). The twin birth will affect the final, realized number of children in the family only if the parents have one more child than intended because of the twin birth. However, the timing failure will give parents who experience a twin birth a head start even among parents who wanted at least as many children as they had through the twin birth. Parents who experience a twin birth will therefore have a larger number of children than other parents, at least for a while. However, some or most of this difference will vanish if other parents are given time to catch up. The effect of a twin birth on the number of children will therefore depend on whether we are studying *only*

“complete families”, which have all reached (or surpassed) their desired number of children, or also include other families, here called “incomplete families”.^{viii}

The “timing failure” effect on the number of children affects our estimates of the causal effect. We, as discussed above, assume that the timing has no effect on the outcome. The only thing that should create differences in the outcome is the difference in the final number of children created by the twin birth, i.e., the difference among complete families. If we include incomplete families in the study population, we will therefore bias the estimated effect on the outcome towards zero by overestimating the difference in the number of children in the first stage. (This is illustrated and discussed further in the Appendix, Section A3). If the assumption does not hold—if, for example, there are effects of the timing on the outcome—then we are not estimating the effect of the number of children but a sample-specific effect that is not well defined.

4.2 The possible definitions of the treatment for twin birth IVs

The reason why twin births lead to an increase of the number of children in families is the basis for the treatment we estimate the effect of. 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 ambiguity in the previous literature can be exemplified using Angrist and Evans (1998). In the introduction, they state that “both twinning and the sex-mix instrument identify the impact of moving from the second to the third child” (Angrist and Evans 1998:452). They then explain that “when [the instrument] z_i is an indicator of multiple births at the second pregnancy, *Twins-2*, 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 1998:458). These statements correspond to treatments with subtle but important differences. (These slight, but important, shifts of the

interpretations in previous literature are discussed further in Appendix, Section A9.) The latter—and original—interpretation of the treatment is that twin births lead to an exogenous increase in the number of children in *some* families (Black et al. 2005:681; see also, e.g., Black et al. 2010:37; Cáceres-Delpiano and Simonsen 2012:754). In the words of Rosenzweig and Wolpin (1980b:232, italics added):

“To the extent that multiple births from one pregnancy are unanticipated and children cannot readily be 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”.

Rosenzweig and Wolpin did not clearly define the treatment when they wrote about twin birth IVs. However, from what they write, it is still clear that they interpret the birth of an “unwanted” child as the treatment (Rosenzweig and Wolpin 1980a:338, 1980b:232–33). An “unwanted” child is not the same as any other child, but such a situation still basically amounts to being an exogenous increase in the number of children.

The other interpretation of the treatment mentioned in Angrist and Evans (1998:452) is having “another” child, for example, a third child if we use a twin birth as second birth as the IV. This interpretation is different from having one more child at any parity but could be a generalizable effect at the studied parity.

The last alternative interpretation—as argued by, for example, Angrist and Pischke (2008) and Angrist et al. (2010)—is that the treatment is the “extra” child born at the studied parity in all families experiencing a twin birth at that parity. Angrist et al. (2010) have one of the best discussions on the effect estimated by twin birth IVs in the literature. Nonetheless, the closest they come to an explicit definition of the treatment is the statement that the

“treatment is defined as a dummy for having another child” (Angrist et al. 2010:788).^{ix}

Below, I discuss the additional claims they make about the estimated effect that means that the treatment they have in mind is the “extra” child born at the studied parity in all families experiencing a twin birth at that parity. This treatment is very specific, and the causal effect of it is not applicable to the number of children in general or even to any twin birth.

I will show that none of these interpretations lead to valid estimates of any causal effect. To explain why, I discuss the conceptual definition of the estimated effects and evaluate each of the assumptions necessary for estimating them.

4.2.1 The perfect compliance interpretation, an “extra” child

Angrist et al. (2010:e.g., 776, 788) argue that IVs based on twin births are a special case among IVs because all families that experience a twin birth comply with their treatment, i.e., there is “perfect compliance”. This argument is also developed in the handbook by Angrist and Pischke (2008:160–61) and has been adopted by some other researchers (e.g., Cáceres-Delpiano 2012a:156, 2012b:8; Cáceres-Delpiano and Simonsen 2012:754; Baranowska-Rataj et al. 2016:1267; Baranowska-Rataj and Matysiak 2016:350).

As noted above, it is somewhat unclear what Angrist et al. (2010) define the treatment to be. However, the statement about perfect compliance makes it clear. There will be perfect compliance in the twin birth IV case only if the treatment is defined as the “extra” child being born at the studied parity because of the twin birth.

If we use the “extra” child being born at the studied parity because of the twin birth as the treatment, then there can be no never-takers. All families that experience a twin birth have the “extra” child. With this interpretation, the first-stage coefficient should therefore always be equal to one.^x No family that does not experience a twin birth at the studied parity will have the “extra” child born because of a twin birth at that parity. Therefore, we can also exclude

the always-takers. We are left with only compliers, i.e., perfect compliance (Table 4). This interpretation is attractive because it means that the estimated LATE is also the effect on the non-treated: “[s]pecifically, the subpopulation of compliers affected by the *twins2* [twin as second birth] instrument is the entire population with two children” (Angrist et al. 2010:788). This would, of course, be a very attractive feature. This benefit is lost because we are, in reality, estimating an effect that is not generalizable and, as we shall see, not even interpretable. It is most definitely not the causal effect of the number of children in the family.

Table 4. The four possible combinations of assignment and receipt of treatment of the “extra” child born at the studied parity because of the twin birth

Panel A Distribution based on: Black, Devereux, and Salvanes (2005)		Twin birth at the studied parity?	
		Yes	No
“Extra” child born at the studied parity because of the twin birth?	Yes	Cell I Compliers $N^* = 2.77$ $N = 3.27$ $\pi = 1.75\%$	————
	No	————	Cell IV Compliers $N^* = 2.77$ $N = 2.78$ $\pi = 98.25\%$

Panel B Distribution based on: Åslund and Grönqvist (2010)		Twin birth at the studied parity?	
		Yes	No
“Extra” child born at the studied parity because of the twin birth?	Yes	Cell I Compliers $N^* = 2.74$ $N = 3.26$ $\pi = 1.75\%$	————
	No	————	Cell IV Compliers $N^* = 2.74$ $N = 2.75$ $\pi = 98.25\%$

Panel C Distribution based on: Stradford, Van Poppel, and Lumey (2017)		Twin birth at the studied parity?	
		Yes	No
“Extra” child born at the studied parity because of the twin birth?	Yes	Cell I Compliers $N^* = 4.61$ $N = 4.80$ $\pi = 1.75\%$	_____
	No	_____	Cell IV Compliers $N^* = 4.61$ $N = 4.62$ $\pi = 98.25\%$

Panel D Distribution based on: Roberts and Warren (2017)		Twin birth at the studied parity?	
		Yes	No
“Extra” child born at the studied parity because of the twin birth?	Yes	Cell I Compliers $N^* = 3.60$ $N = 3.94$ $\pi = 1.75\%$	_____
	No	_____	Cell IV Compliers $N^* = 3.60$ $N = 3.61$ $\pi = 98.25\%$

It is important to note that the “extra” child interpretation of the treatment is not the same as the treatment being to experience any twin birth. With that interpretation there will be always-takers in the population (families experiencing twin births at other parities). We will therefore end up with a situation very similar to the one for the “another” child interpretation explained below. Studying the effect of a twin birth is also not at all the same as studying the effect of the number of children, but could be interesting in some cases (See, e.g., Grogger and Bronars 1993 or Bucher-Koenen et al. 2019 which specifically study the effects on mothers from a twin birth.).

4.2.2 The having “another” child interpretation

To have perfect compliance, the treatment cannot be defined as having “another” child, for example, a third child if we use a twin as second birth instrument. If we interpret the treatment as “another” child, then there will again be no never-takers, but there will be always-takers (Table 5). In the (parity-specific) twin birth IV case, the overwhelming majority are not assigned to treatment, that is, do not experience a twin birth at the parity studied. Nonetheless, there are many families that have “another” child without being assigned to have one through a twin birth. These families are therefore treated without being assigned to treatment. This creates a situation with one-sided noncompliance (Imbens and Rubin 2015:Ch. 23). All families that do experience a twin birth at the studied parity and also have “another” child are compliers. There are therefore no always-takers that are indicated by the instrument.

Table 5. The four possible combinations of assignment and receipt of the treatment of having “another” child

Panel A Distribution based on: Black, Devereux, and Salvanes (2005)		Twin birth at the studied parity?	
		Yes	No
“Another” child?	Yes	Cell I Compliers $N^* = 2.77$ $N = 3.27$ $\pi = 1.75\%$	Cell II Always-takers $N^* = 3.53$ $N = 3.55$ $\pi = 49.29\%$
	No	————	Cell IV Compliers $N^* = 2.00$ $N = 2.00$ $\pi = 48.96\%$

Panel B Distribution based on: Åslund and Grönqvist (2010)		Twin birth at the studied parity?	
		Yes	No
“Another” child?	Yes	Cell I Compliers $N^* = 2.74$ $N = 3.26$ $\pi = 1.75\%$	Cell II Always-takers $N^* = 3.52$ $N = 3.54$ $\pi = 47.97\%$
	No	————	Cell IV Compliers $N^* = 2.00$ $N = 2.00$ $\pi = 50.28\%$

Panel C Distribution based on: Stradford, Van Poppel, and Lumey (2017)		Twin birth at the studied parity?	
		Yes	No
“Another” child?	Yes	Cell I Compliers $N^* = 4.61$ $N = 4.80$ $\pi = 1.75\%$	Cell II Always-takers $N^* = 5.18$ $N = 5.20$ $\pi = 80.52\%$
	No	————	Cell IV Compliers $N^* = 2.00$ $N = 2.00$ $\pi = 17.73\%$

Panel D Distribution based on: Roberts and Warren (2017)		Twin birth at the studied parity?	
		Yes	No
“Another” child?	Yes	Cell I Compliers $N^* = 3.60$ $N = 3.94$ $\pi = 1.75\%$	Cell II Always-takers $N^* = 4.39$ $N = 4.41$ $\pi = 65.65\%$
	No	————	Cell IV Compliers $N^* = 2.00$ $N = 2.00$ $\pi = 32.60\%$

4.2.3 The “unwanted” children interpretation

The interpretation of the treatment that comes closest to an exogenous increase in the number of children is the interpretation of the birth of an “unwanted” child as the treatment. One way to think about this interpretation is to 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. However, the reception of the treatment 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 will only lead to them having the children they want unexpectedly fast.

If we interpret the treatment as being an “unwanted” child born in some families, then as a result, there are both always-takers and never-takers in the population (Table 6). The estimated LATE will therefore be no more generalizable than usual: it will be the effect for the subpopulation that has its treatment status changed by its assignment to the treatment—in other words, the compliers.

Table 6. The four possible combinations of assignment and receipt of the treatment of an “unwanted” child

Panel A Distribution based on: Black, Devereux, and Salvanes (2005)		Twin birth at the studied parity?	
		Yes	No
Having an “unwanted” child?	Yes	Cell I Compliers $N^* = 2.02$ $N = 3.02$ $\pi = 0.88\%$	Cell II Always-takers $N^* = 3.52$ $N = 4.52$ $\pi = 0.86\%$
	No	Cell III Never-takers $N^* = 3.52$ $N = 3.52$ $\pi = 0.87\%$	Cell IV Compliers $N^* = 2.76$ $N = 2.76$ $\pi = 97.39\%$

Panel B Distribution based on: Åslund and Grönqvist (2010)		Twin birth at the studied parity?	
		Yes	No
Having an “unwanted” child?	Yes	Cell I Compliers $N^* = 2.01$ $N = 3.01$ $\pi = 0.90\%$	Cell II Always-takers $N^* = 3.51$ $N = 4.51$ $\pi = 0.83\%$
	No	Cell III Never-takers $N^* = 3.51$ $N = 3.51$ $\pi = 0.85\%$	Cell IV Compliers $N^* = 2.73$ $N = 2.73$ $\pi = 97.42\%$

Panel C Distribution based on: Stradford, Van Poppel, and Lumey (2017)		Twin birth at the studied parity?	
		Yes	No
Having an “unwanted” child?	Yes	Cell I Compliers $N^* = 2.22$ $N = 3.22$ $\pi = 0.33\%$	Cell II Always-takers $N^* = 5.18$ $N = 6.18$ $\pi = 1.39\%$
	No	Cell III Never-takers $N^* = 5.18$ $N = 5.18$ $\pi = 1.42\%$	Cell IV Compliers $N^* = 4.60$ $N = 4.60$ $\pi = 96.86\%$

Panel D Distribution based on: Roberts and Warren (2017)		Twin birth at the studied parity?	
		Yes	No
Having an “unwanted” child?	Yes	Cell I Compliers $N^* = 2.07$ $N = 3.07$ $\pi = 0.59\%$	Cell II Always-takers $N^* = 4.38$ $N = 5.38$ $\pi = 1.14\%$
	No	Cell III Never-takers $N^* = 4.38$ $N = 4.38$ $\pi = 1.16\%$	Cell IV Compliers $N^* = 3.59$ $N = 3.59$ $\pi = 97.11\%$

The compliers that are not indicated by the instrument (Table 6: A–D, Cell IV) consist of families that did not experience a twin birth at the studied parity and that also do not have an

“unwanted” child. Some of these families will proceed to have more children after the studied parity, and some of these births will be twins. By definition, we know that they always intended to have as many children as they end up with.

The group in Cell III of Table 6: A–D—the never-takers—consists of families that experienced a twin birth at the studied parity but that still do not have an “unwanted” child. Again, some of these families will proceed to have more children after the studied parity, and some of these births will be twins. Again, by definition, we know that for the families in this group, they always intended to have as many children as they end up with. All families that do not experience a twin birth at the studied parity and also do not have any “unwanted” children are compliers. There are therefore no never-takers that are not indicated by the instrument.

The group in Cell II of Table 6: A–D—the always-takers—consists of families that did not experience a twin birth at the studied parity but that still have an “unwanted” child. Who we think belongs in this groups depends on whether we allow for the possibility of unintended pregnancies or not, that is, whether if a singleton birth can lead to an “unwanted” child or not.

Allowing for unintended pregnancies is more realistic, given the empirical evidence (Singh et al. 2010; Alkema et al. 2013; Bachrach and Morgan 2013). However, the exclusion restriction then requires us 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, Koenig, and Hindin 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. 1996:451).

If we assume that there are no unintended pregnancies, then it is only twin births that can lead to “unwanted” children. The only group that can be always-takers in Cell II of Table 6: A–D consists of parents who wanted at least one more child after the studied parity and who experienced a twin birth at one of these subsequent births. This subsequent twin birth leads to the birth of an “unwanted” child. All families that do experience a twin birth at the studied parity and also have an “unwanted” child are compliers. There are therefore no always-takers that are indicated by the instrument.

The group in Cell I of Table 7: A–D—compliers indicated by the instrument—consists of families that experienced a twin birth at the studied parity and that have an “unwanted” child. If we do not allow for any unintended pregnancies, all “unwanted” children are born from twin births. The “unwanted” child for the families in Cell I can therefore only be the result of the twin birth at the studied parity or a subsequent twin birth. Not allowing for unintended pregnancies also means that all parents will stop having children when they reach or surpass their desired number. This assumption thus implies that parents who experienced a twin birth at the studied parity and who have an “unwanted” child wanted only N^* children but had $N^* + 1$ because of the twin birth. This group consists of compliers that had their treatment status, here, an “unwanted” child, changed by the assignment to the treatment, here, the twin birth. These families are therefore the families for which we estimate the LATE when using this interpretation of the treatment.

5. Results: Evaluating some of the necessary assumptions for twin birth IVs

5.1 The random assignment of twin births

For twin birth IVs to work, there cannot be anything that creates systematic differences in the outcome between families that do and do not experience a (parity-specific) twin birth except

the effect on the number of children from the twin birth. To that end, we rely on the randomness of twin births. Provided that twin births are random events, there should be no systematic differences between families at a specific parity that do and do not experience a twin birth (on average and in large samples). Recent studies have, as mentioned, showed that twin births are not completely random and that these systematic patterns have substantive effects on the results. As also mentioned, I will ignore this important critique and proceed assuming that twin births are indeed random. This is also how I generated my simulated populations.

If twin births indeed occur at random, then there should be no association between the type and the chance of experiencing a twin birth (Morgan and Winship 2015:311–12). The implication is that the probability of being a complier, never-taker, or always-taker should be the same regardless of whether a unit is indicated by the instrument or not. Henderson et al. (2008:171) call this phenomenon an “unconfounded type”.

The randomness is actually supported when we use the “extra” child being born at the studied parity because of the twin birth as the treatment (Table 7, panel A). Everyone is a complier in this case, so there are no never- or always-takers. The probability of being a complier is therefore equal to one within both levels of the instrument. The results are less reassuring if we instead use the “another” or “unwanted” child interpretation of the treatment (Table 7, panels B and C). The probabilities of the types vary between the two different specifications of the treatment and between populations. In some instances, the probabilities are almost balanced in the populations based on Black et al. (2005) and Åslund and Grönqvist (2010). However, even in these cases, it is clear that twin births are not random when conditioning on the type.

Table 7. The probability of belonging to a type within levels of the instrument, here: having twin as second birth

	Compliers		Never-takers		Always-takers	
	Indicated by instrument?		Indicated by instrument?		Indicated by instrument?	
	Yes	No	Yes	No	Yes	No
Panel A. The “extra” child born at the studied parity because of the twin birth as the treatment						
Black, Devereux and Salvanes (2005)	1.000	1.000	—	—	—	—
Åslund and Grönqvist (2010)	1.000	1.000	—	—	—	—
Stradford, van Poppel and Lumey (2017)	1.000	1.000	—	—	—	—
Roberts and Warren (2017)	1.000	1.000	—	—	—	—
Panel B. “Another” child, here: a third child, as the treatment						
Black, Devereux and Salvanes (2005)	1.000	0.498	—	—	0.000	0.502
Åslund and Grönqvist (2010)	1.000	0.512	—	—	0.000	0.488
Stradford, van Poppel and Lumey (2017)	1.000	0.180	—	—	0.000	0.820
Roberts and Warren (2017)	1.000	0.332	—	—	0.000	0.668
Panel C. An “unwanted” child as the treatment						
Black, Devereux and Salvanes (2005)	0.501	0.991	0.499	0.000	0.000	0.009
Åslund and Grönqvist (2010)	0.515	0.992	0.485	0.000	0.000	0.008
Stradford, van Poppel and Lumey (2017)	0.191	0.986	0.809	0.000	0.000	0.014
Roberts and Warren (2017)	0.339	0.988	0.661	0.000	0.000	0.012

5.2. Twin births should affect the outcome only by affecting the number of children

Random assignment to treatment is not enough to have an instrument that satisfies the necessary assumptions. To be able to estimate a LATE, we need both the random assignment and the exclusion restriction (e.g., Imbens and Angrist 1994:468; Angrist and Pischke 2008:153). The birth of twins instead of a single birth at the studied parity should be the only reason why families that have twins (at the studied parity) have a larger number of children than do other families. This is what we assume in the exclusion restriction, that there should be no systematic differences between families that do and do not experience a (parity-specific) twin birth. If there are such systematic differences, they will bias the estimated effect.

The implication of this is that the potential outcomes among the units that have received the treatment should be balanced across those that were assigned to treatment and those that were not assigned to treatment, here, experiencing a (parity-specific) twin birth or not. We, of course, need this to be true for the compliers. It is also the assumption we use to be able to ignore the always-takers and never-takers. In equations 1 and 2 below, this is defined for a single unit. When we use the method empirically, we use a stochastic version of the exclusion restriction in which the distribution of potential outcomes is balanced between the two groups within each type.

Always-takers:

$$Y_i(Z_i = 0, W_i(Z_i = 0) = 1) = Y_i(Z_i = 1, W_i(Z_i = 1) = 1) \quad (1)$$

Never-takers:

$$Y_i(Z_i = 1, W_i(Z_i = 1) = 0) = Y_i(Z_i = 0, W_i(Z_i = 0) = 0) \quad (2)$$

If we use the “extra” child being born at the studied parity because of the twin birth as the treatment, there are neither always- nor never-takers (Table 4, A–D). They therefore don’t need to be ignored, they do not exist with this interpretation. But, as we have seen, as soon as we use either the “another” or the “unwanted” child interpretation of the treatment, there are always-takers, or always- and never-takers. More problematic is that these groups exist only within one level of the instrument. This does *not* mean that we can ignore them. They are still included in the analysis, but without any chance of balancing themselves out across the levels of the instrument.

If we use the “another” child interpretation of the treatment, there is one-sided noncompliance. The definition of the estimated effect (following the notation of Imbens and Rubin 2015:525) is then as follows:

(3)

$$\hat{\beta} = \pi_{co} \cdot \frac{1}{N_{co}} \sum_{i:G_i=co} (Y_i(Z_i = 1, W_i(Z_i = 1) = 1) - Y_i(Z_i = 0, W_i(Z_i = 0) = 0)) +$$

$$\pi_{at} \cdot \frac{1}{N_{at}} \sum_{i:G_i=at} (Y_i(Z_i = 1, W_i(Z_i = 1) = 1) - Y_i(Z_i = 0, W_i(Z_i = 0) = 1))$$

Our estimate of the effect of interest, $\hat{\beta}$, is a weighted sum of the average difference between the compliers that are indicated by the instrument and not and the average difference between the always-takers that are indicated by the instrument and not. The weights are defined by the shares of compliers, π_{co} , and always-takers, π_{at} , in the population.

If the exclusion restriction holds, then. This part of Equation 3—the second line—can therefore be ignored. However, if the exclusion restriction does not hold, then we can no longer estimate the LATE.

The first line of equation 3 is the comparison of the compliers that are assigned to treatment and treated with the compliers that are not assigned and not treated. This is the causal effect that we are looking for.

But, there is a second argument in the equation because, with this interpretation of the treatment, there are also always-takers (subscripted with at). We would like to be able to ignore the second line based on the exclusion restriction, i.e. the assumption that there are no systematic differences between noncompliers that are or are not indicated by the instrument. But, in the twin birth IV case, all always-takers are, by definition, in the group not indicated by the instrument, i.e. $Z_i = 0$. There is no corresponding group of always-takers that are indicated by the instrument that could cancel out the other group. The first part of the summation on the second line of equation 3 is therefore not used (Equation 4). This has the consequence that we subtract the average outcome of the always-takers with $Z_i = 0$ from the

difference in outcome between the two groups of compliers. The estimate of the causal effect will therefore be severely biased. (4)

$$\hat{\beta} = \pi_{co} \cdot \frac{1}{N_{co}} \sum_{i:G_i=co} (Y_i(Z_i = 1, W_i(Z_i = 1) = 1) - Y_i(Z_i = 0, W_i(Z_i = 0) = 0)) +$$

$$\pi_{at} \cdot \frac{1}{N_{at}} \sum_{i:G_i=at} (-Y_i(Z_i = 0, W_i(Z_i = 0) = 1))$$

We can also see in Table 5, A–D, that there is another, even more serious, inherent source of bias with the “another” child interpretation of the treatment. There are systematic differences between the compliers at the different levels of the instrument. The compliers that do not experience a (parity-specific) twin birth and also do not have “another” child, will always, by definition, want and also have two children. This is not the case for compliers that do experience a (parity-specific) twin birth and have (at least) “another” child. We therefore do not have comparable groups of compliers to estimate the effect from.

If we use the “unwanted” child interpretation of the treatment, there are both always- and never-takers in the population. The definition of the estimated effect therefore has three arguments, one each for compliers (subscripted *co*), always-takers (subscripted *at*), and never-takers (subscripted *nt*) (Imbens and Rubin 2015:553): (5)

$$\hat{\beta} = \pi_{co} \cdot \frac{1}{N_{co}} \sum_{i:G_i=co} (Y_i(Z_i = 1, W_i(Z_i = 1) = 1) - Y_i(Z_i = 0, W_i(Z_i = 0) = 0)) +$$

$$\pi_{nt} \cdot \frac{1}{N_{nt}} \sum_{i:G_i=nt} (Y_i(Z_i = 1, W_i(Z_i = 1) = 0) - Y_i(Z_i = 0, W_i(Z_i = 0) = 0)) +$$

$$\pi_{at} \cdot \frac{1}{N_{at}} \sum_{i:G_i=at} (Y_i(Z_i = 1, W_i(Z_i = 1) = 1) - Y_i(Z_i = 0, W_i(Z_i = 0) = 1))$$

Again, all always-takers are in the group not indicated by the instrument, $Z_i = 0$. All never-takers are instead in the group that is indicated, $Z_i = 1$. Ideally these two groups should cancel each other out (Equation 6). But, as we can see in table 6, A–D, the two groups are not comparable. While they do desire the same number of children, N^* , the always-takers have a larger number of children, N . If there is a negative effect from the number of children, there will be a negative bias in the estimated effect. (6)

$$\begin{aligned}\hat{\beta} = & \pi_{co} \cdot \frac{1}{N_{co}} \sum_{i:G_i=co} (Y_i(Z_i = 1, W_i(Z_i = 1) = 1) - Y_i(Z_i = 0, W_i(Z_i = 0) = 0)) + \\ & \pi_{nt} \cdot \frac{1}{N_{nt}} \sum_{i:G_i=nt} (Y_i(Z_i = 1, W_i(Z_i = 1) = 0)) + \\ & \pi_{at} \cdot \frac{1}{N_{at}} \sum_{i:G_i=at} (-Y_i(Z_i = 0, W_i(Z_i = 0) = 1))\end{aligned}$$

But, again, this bias is not the most serious problem with the estimator when using the “unwanted” child interpretation of the treatment. The compliers that are and are not indicated by the instrument are not comparable (Table 6, A–D). The compliers that are not indicated both desire and have a larger number of children. The estimated effect is therefore not a causal effect of anything sensible.

5.3 The stable unit treatment value assumptions (SUTVAs I and II)

To estimate well-defined causal effects, we must rely on the SUTVAs (Imbens and Rubin 2015:9–12; see also Angrist et al. 1996:446). The SUTVA includes two parts, SUTVAs I and II (Cox 1958:17–21; Rubin 1990:475; Heckman 2005:11–12, 35–36, 43; Small et al. 2017). The first part requires that there should be no interference between the studied units; the treatment of one unit should therefore not affect another. This requirement is the part of the SUTVA

that is most often acknowledged in the literature, when it is discussed at all (e.g., Morgan and Winship 2015:48–52).^{xi} This is unproblematic for twin birth IVs in most populations.

The second part of SUTVA requires that there be only one version of the treatment; there should be “No Hidden Variations of Treatments” (Imbens and Rubin 2015:11).^{xii} Rather, there can be different versions of the treatment, W and W' , provided that the potential outcomes are the same regardless of which version one receives, i.e., $Y_i(Z_i, W_i(Z_i)) = Y_i(Z_i, W'_i(Z_i))$.

SUTVA II is an assumption related to the treatment in itself. Imbens and Rubin (2015) still suggest that we should evaluate it by thinking about if the outcome would be the same regardless of which version of the treatment a unit receives.

If the effect of the treatment is not the same for the different versions of the treatment, then the estimated effect is “an artificial quantity” without any clear interpretation (Cox 1958:15–19). It does not necessarily correspond to the average effect across the groups receiving the different versions of the treatment. Cox (1958) provides simple numerical examples illustrating that in the presence of different versions of the treatment, the estimated average effect does not even have to be in the range of the effects for the different versions of the treatment. An important reason why there can be different versions of the treatment is when it depends on some (observable or unobservable) differences in the characteristics of the units (Cox 1958:18). If we do not or cannot adjust our analyses for these differences in characteristics, then the estimated effect cannot tell us anything about the true effect (but see Small et al. 2017).

This situation is what occurs in the twin birth IV case if we consider the “extra” or “another” child born at the studied parity because of a twin birth as the treatment. With these definitions of the treatment, there are two versions of the treatment: one version if the twin birth leads to an unexpected and unintended, that is, exogenous, increase in the number of

children and another if the twin birth leads to only an unexpectedly short birth interval of otherwise wanted children. The version of the treatment that the families receive depends on their desired number of children—a, most often, unobserved (or unobservable) characteristic of the families. We could accommodate for these versions of the treatment if we could observe and adjust for the presence of “unwanted” children, which corresponds to adding a variable indicating “unwanted” children, U , to the definition of the potential outcomes (Rubin 1990:475):

$$Y_i(Z_i = 1, W_i(Z_i = 1) = 1, U_i = 0) \neq Y_i(Z_i = 1, W_i(Z_i = 1) = 1, U_i = 1) \quad (7)$$

The entire reason why we use twin birth IVs, 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. Therefore, we cannot expect these potential outcomes to be the same. The second part of the SUTVA is therefore not satisfied with these interpretations of the treatment. One way to fix this particular problem could be to instead interpret the “unwanted” child as the treatment. However, as we have observed, doing so does not work either because the independence assumption and exclusion restriction are violated.

6. Conclusions

IVs based on twin births are fundamentally flawed and will always produce results that are meaningless. Depending on how we choose to interpret the treatment, these IVs violate different combinations of the exclusion restriction, the independence assumption, and the second part of the SUTVAs. These violations are inherent to the method and cannot be fixed. All studies using twin births as IVs have estimated a casual rather than a causal effect of the number of children in the family.

This has been overlooked in previous research because these researchers have paid too little attention to the defining features of this natural experiment, and especially to the

interpretation of the treatment. With “natural experiments” we need to carefully conceptualize all aspects of the experiment, i.e. the assignment and reception of treatment, as well as the treatment. 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.

I think that previous applications have overlooked these problems because they have not written about it enough in 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 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 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.

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.

7. References

- Abdul-Razak, Nor Azam., Mohd Zaini Abd Karim, and Roslan Abdul-Hakim. 2015. "Does Trade-Off Between Child Quantity and Child Quality Exist in Malaysia?" *Singapore Economic Review* 60(4):Article 1550021. <https://doi.org/10.1142/S0217590815500216>
- Ajefu, Joseph Boniface. *Forthcoming* 2019. "Does Having Children Affect Women's Entrepreneurship Decision? Evidence from Nigeria." *Review of Economics of the Household*. <https://doi.org/10.1007/s11150-019-09453-2>
- Alidou, Sahawal, and Marijke Verpoorten. *Forthcoming* 2019. "Family Size and Schooling in Sub-Saharan Africa: Testing the Quantity-Quality Trade-Off." *Journal of Population Economics*. <https://doi.org/10.1007/s00148-019-00730-z>
- Alkema, Leontine, Kantorova Vladimira, Menozzi Clare, and Ann Biddlecom. 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." *The Lancet* 381(9878):1642–52. [https://doi.org/10.1016/S0140-6736\(12\)62204-1](https://doi.org/10.1016/S0140-6736(12)62204-1)
- Angrist, Joshua D. 2004. "Treatment Effect Heterogeneity in Theory and Practice." *The Economic Journal* 114(494):C52–C83. <https://doi.org/10.1111/j.0013-0133.2003.00195.x>
- Angrist, Joshua D., and William N. Evans. 1998. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *The American Economic Review* 88(3):450–77. <http://www.jstor.org/stable/116844>
- Angrist, Joshua D., and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment

- Intensity.” *Journal of the American Statistical Association* 90(430):431–42.
<https://doi.org/10.1080/01621459.1995.10476535>
- Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. “Identification of Causal Effects Using Instrumental Variables.” *Journal of the American Statistical Association* 91(434):444–55. <https://doi.org/10.2307/2291629>
- Angrist, Joshua D., and Alan B. Krueger. 2001. “Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments.” *Journal of Economic Perspectives* 15(4):69–85.
<https://doi.org/10.1257/jep.15.4.69>
- Angrist, Joshua D., Victor Lavy, and Analia Schlosser. 2005. “New Evidence on the Causal Link Between the Quantity and Quality of Children.” Working Paper. National Bureau of Economic Research. <https://doi.org/10.3386/w11835>
- Angrist, Joshua D., Victor Lavy, and Analia Schlosser. 2010. “Multiple Experiments for the Causal Link between the Quantity and Quality of Children.” *Journal of Labor Economics* 28(4):773–824. <https://doi.org/10.1086/653830>
- Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton, NJ: Princeton University Press.
- Arouri, Mohamed, Adel Ben Youssef, and Cuong Nguyen-Viet. 2017. “The More Children You Have the More Likely You Are to Smoke? Evidence from Vietnam.” *Oxford Development Studies* 45(3):260–75.
<https://doi.org/10.1080/13600818.2016.1193129>
- Åslund, Olof, and Hans Grönqvist. 2010. “Family Size and Child Outcomes: Is There Really No Trade-off?” *Labour Economics* 17(1):130–39.
<https://doi.org/10.1016/j.labeco.2009.05.003>

- Bachrach, Christine A., and S. Philip Morgan. 2013. "A Cognitive–Social Model of Fertility Intentions." *Population and Development Review* 39(3):459–85.
<https://doi.org/10.1111/j.1728-4457.2013.00612.x>
- Baranowska-Rataj, Anna, Kieron J. Barclay, and Martin Kolk. 2017. "The Effect of the Number of Siblings on Adult Mortality: Evidence from Swedish Registers." *Population Studies* 71(1):43–63. <https://doi.org/10.1080/00324728.2016.1260755>
- Baranowska-Rataj, Anna, Xavier de Luna, and Anneli Ivarsson. 2016. "Does the Number of Siblings Affect Health in Midlife? Evidence from the Swedish Prescribed Drug Register." *Demographic Research* 35(Article 43):1259–1302.
<https://doi.org/10.4054/DemRes.2016.35.43>
- Baranowska-Rataj, Anna, and Anna Matysiak. 2016. "The Causal Effects of the Number of Children on Female Employment—Do European Institutional and Gender Conditions Matter?" *Journal of Labor Research* 37(3):343–67.
<https://doi.org/10.1007/s12122-016-9231-6>
- Betz, Timm, Scott J. Cook, and Florian M. Hollenbach. *Forthcoming* 2019. "Spatial interdependence and instrumental variable models." *Political Science Research and Methods*. <https://doi.org/10.1017/psrm.2018.61>
- Bhalotra, Sonia R., and Damian Clarke. 2018a. "The Twin Instrument: Fertility and Human Capital Investment." IZA Discussion Papers 11878. Retrieved November 2, 2018 (<https://www.iza.org/publications/dp/11878/the-twin-instrument-fertility-and-human-capital-investment>).
- Bhalotra, Sonia R., and Damian Clarke. *Forthcoming* 2018b. "Twin Birth and Maternal Condition." *The Review of Economics and Statistics*.
https://doi.org/10.1162/rest_a_00789

- Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. .2005. "The More the Merrier? The Effect of Family Size and Birth Order on Children's Education." *The Quarterly Journal of Economics* 120(2):669–700.
<https://doi.org/10.1093/qje/120.2.669>
- Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. 2010. "Small Family, Smart Family? Family Size and the IQ Scores of Young Men." *Journal of Human Resources* 45(1):33–58. <https://doi.org/10.3368/jhr.45.1.33>
- Bollen, Kenneth A. 2012. "Instrumental Variables in Sociology and the Social Sciences." *Annual Review of Sociology* 38:37–72.
<https://doi.org/10.1146/annurev-soc-081309-150141>
- Bonner, Suzanne, and Dipanwita Sarkar. 2018. "The Quality-Quantity Trade-off among Australian Children." *Economic Modelling* 70:383–89.
<https://doi.org/10.1016/j.econmod.2017.08.010>
- Braakmann, Nils, and John Wildman. 2016. "Reconsidering the Effect of Family Size on Labour Supply: The Twin Problems of the Twin Birth Instrument." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179(4):1093–1115.
<https://doi.org/10.1111/rssa.12160>
- Brinch, Christian N., Magne Mogstad, and Matthew Wiswall. 2017. "Beyond LATE with a Discrete Instrument." *Journal of Political Economy* 125(4):985–1039.
<https://doi.org/10.1086/692712>.
- Bronars, Stephen G., and Jeff Grogger. 1994. "The Economic Consequences of Unwed Motherhood: Using Twin Births as a Natural Experiment." *The American Economic Review* 84(5):1141–56. <http://www.jstor.org/stable/2117765>

- Bucher-Koenen, Tabea, Helmut Farbmacher, Raphael Guber, and Johan Vikström. 2019. "Double Trouble: The Burden of Child Rearing and Working on Maternal Mortality," Retrieved May 24, 2019 (http://www.farbmacher.de/working_papers/Bucher_Koenen_etal_2019.pdf).
- Cáceres-Delpiano, Julio. 2006. "The Impacts of Family Size on Investment in Child Quality." *Journal of Human Resources* 41(4):738–54. <https://doi.org/10.2307/40057289>
- Cáceres-Delpiano, Julio. 2012a. "Impacts of Family Size on the Family as a Whole: Evidence from the Developing World." *The B.E. Journal of Economic Analysis & Policy* 12(1):Article 17. <https://doi.org/10.1515/1935-1682.2850>
- Cáceres-Delpiano, Julio. 2012b. "Can We Still Learn Something from the Relationship Between Fertility and Mother's Employment? Evidence from Developing Countries." *Demography* 49(1):151–74. <https://doi.org/10.1007/s13524-011-0076-6>
- Cáceres-Delpiano, Julio, and Marianne Simonsen. 2012. "The Toll of Fertility on Mothers' Wellbeing." *Journal of Health Economics* 31(5):752–66. <https://doi.org/10.1016/j.jhealeco.2012.05.006>
- Chan, Tak Wing, Morag Henderson, and Rachel Stuchbury. *Forthcoming* 2019. "Family Size and Educational Attainment in England and Wales." *Population Studies*. <https://doi.org/10.1080/00324728.2019.1577479>.
- Chen, Stacey H., Yen-Chien Chen, and Jin-Tan Liu. 2019. "The Impact of Family Composition on Educational Achievement." *Journal of Human Resources* 54(1):122–70. <https://doi.org/10.3368/jhr.54.1.0915.7401R1>.

- Chen, Qihui. 2017. “Relaxed Population Policy, Family Size and Parental Investments in Children’s Education in Rural Northwestern China.” *International Journal of Educational Development* 54(Supplement C):39–50.
<https://doi.org/10.1016/j.ijedudev.2017.03.009>
- Clarke, Damian. 2018. “Children and Their Parents: A Review of Fertility and Causality.” *Journal of Economic Surveys* 32(2):518–40.
<https://doi.org/10.1111/joes.12202>
- Cox, David Roxbee. 1958. *Planning of Experiments*. New York: John Wiley.
- D’Addato, Agata V. 2007. “Secular Trends in Twinning Rates.” *Journal of Biosocial Science* 39(1):147–51. <https://doi.org/10.1017/S0021932006001337>
- Dasgupta, Kabir, and Keisha T. Solomon. 2018. “Family size effects on childhood obesity: Evidence on the quantity–quality trade-off using the NLSY.” *Economics and Human Biology* 29:42–55. <https://doi.org/10.1016/j.ehb.2018.01.004>
- Dayioğlu, Meltem, Murat G. Kirdar, and Aysit Tansel. 2009. “Impact of Sibship Size, Birth Order and Sex Composition on School Enrolment in Urban Turkey.” *Oxford Bulletin of Economics and Statistics* 71(3):399–426.
<https://doi.org/10.1111/j.1468-0084.2008.00540.x>
- de Chaisemartin, Clément. 2017. “Tolerating Defiance? Local Average Treatment Effects without Monotonicity.” *Quantitative Economics* 8(2):367–96.
<https://doi.org/10.3982/QE601>
- de Haan, Monique. 2010. “Birth Order, Family Size and Educational Attainment.” *Economics of Education Review* 29(4):576–88.
<https://doi.org/10.1016/j.econedurev.2009.10.012>

- de Jong, Eelke, Jeroen Smits, and Abiba Longwe. 2017. “Estimating the Causal Effect of Fertility on Women’s Employment in Africa Using Twins.” *World Development* 90:360–68. <https://doi.org/10.1016/j.worlddev.2016.10.012>
- Dunning, Thad. 2008. “Improving Causal Inference: Strengths and Limitations of Natural Experiments.” *Political Research Quarterly* 61(2):282–93. <https://doi.org/10.1177/1065912907306470>
- Farbmacher, Helmut, Raphael Guber, and Johan Vikström. 2018. “Increasing the Credibility of the Twin Birth Instrument.” *Journal of Applied Econometrics* 33(3):457–72. <https://doi.org/10.1002/jae.2616>
- Fitzsimons, Emla, and Bansi Malde. 2014. “Empirically Probing the Quantity–Quality Model.” *Journal of Population Economics* 27(1):33–68. <https://doi.org/10.1007/s00148-013-0474-8>
- Fletcher, Jason M., and Jinho Kim. 2019. “The Effect of Sibship Size on Non-Cognitive Skills: Evidence from Natural Experiments.” *Labour Economics* 56:36–43. <https://doi.org/10.1016/j.labeco.2018.11.004>.
- Fontaine, Idriss. 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(5):869–98. <https://doi.org/10.3917/reco.695.0869>.
- Frenette, Marc. 2011a. “How Does the Stork Delegate Work? Childbearing and the Gender Division of Paid and Unpaid Labour.” *Journal of Population Economics* 24(3):895–910. <https://doi.org/10.1007/s00148-010-0307-y>

- Frenette, Marc. 2011b. "Why Do Larger Families Reduce Parental Investments in Child Quality, but Not Child Quality per Se?" *Review of Economics of the Household* 9(4):523–37. <https://doi.org/10.1007/s11150-010-9115-0>
- Gabel, Frank, Hendrik Jürges, Kai E. Kruk, and Stefan Listl. 2018. "Gain a Child, Lose a Tooth? Using Natural Experiments to Distinguish between Fact and Fiction." *Journal of Epidemiology and Community Health* 72(6):552–56. <https://doi.org/10.1136/jech-2017-210210>
- Gipson, Jessica D., Michael A. Koenig, and Michelle J. Hindin. 2008. "The Effects of Unintended Pregnancy on Infant, Child, and Parental Health: A Review of the Literature." *Studies in Family Planning* 39(1):18–38. <https://doi.org/10.1111/j.1728-4465.2008.00148.x>
- Glick, Peter J., Alessandra Marini, and David E. Sahn. 2007. "Estimating the Consequences of Unintended Fertility for Child Health and Education in Romania: An Analysis Using Twins Data." *Oxford Bulletin of Economics and Statistics* 69(5):667–91. <https://doi.org/10.1111/j.1468-0084.2007.00476.x>
- Greenland, Sander, and James M. Robins. 1986. "Identifiability, Exchangeability, and Epidemiological Confounding." *International Journal of Epidemiology* 15(3):413–19. <https://doi.org/10.1093/ije/15.3.413>
- Grogger, Jeff, and Stephen Bronars. 1993. "The Socioeconomic Consequences of Teenage Childbearing: Findings from a Natural Experiment." *Family Planning Perspectives* 25(4):156–74. <https://doi.org/10.2307/2135923>.
- Guo, Rufei, Hongbin Li, Junjian Yi, and Junsen Zhang. 2018. "Fertility, Household Structure, and Parental Labor Supply: Evidence from China." *Journal of Comparative Economics* 46(1):145–56. <https://doi.org/10.1016/j.jce.2017.10.005>.

- Guo, Rufei, Junjian Yi, and Junsen Zhang. 2017. "Family Size, Birth Order, and Tests of the Quantity-Quality Model." *Journal of Comparative Economics* 45(2):219–24. <https://doi.org/10.1016/j.jce.2016.09.006>
- Hall, Jennifer A., Lorna Benton, Andrew Copas, and Judith Stephenson. 2017. "Pregnancy Intention and Pregnancy Outcome: Systematic Review and Meta-Analysis." *Maternal and Child Health Journal* 21(3):670–704. <https://doi.org/10.1007/s10995-016-2237-0>
- Hart, Rannveig, and Sara Cools. 2019. "Identifying Interaction Effects Using Random Fertility Shocks." *Demographic Research* 40(Article 10):261–78. <https://doi.org/10.4054/DemRes.2019.40.10>.
- Hatton, Timothy J., and Richard M. Martin. 2010. "The Effects on Stature of Poverty, Family Size, and Birth Order: British Children in the 1930s." *Oxford Economic Papers* 62(1):157–84. <https://doi.org/10.1093/Oep/Gpp034>
- He, Xiaobo, and Rong Zhu. 2016. "Fertility and Female Labour Force Participation: Causal Evidence from Urban China." *The Manchester School* 84(5):664–74. <https://doi.org/10.1111/manc.12128>
- Heckman, James J. 1997. "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." *Journal of Human Resources* 32(3):441–62. <https://doi.org/10.2307/146178>
- Heckman, James J. 2005. "The Scientific Model of Causality." *Sociological Methodology* 35(1):1–97. <https://doi.org/10.1111/j.0081-1750.2006.00164.x>
- Heckman, James J. 2010. "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy." *Journal of Economic Literature* 48(2):356–98. <https://doi.org/10.1257/jel.48.2.356>

- Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. “Understanding Instrumental Variables in Models with Essential Heterogeneity.” *The Review of Economics and Statistics* 88(3):389–432. <https://doi.org/10.1162/rest.88.3.389>
- Henderson, Daniel J., Daniel L. Millimet, Christopher F. Parmeter, and Le Wang. 2008. “Fertility and the Health of Children: A Nonparametric Investigation.” In: *Modelling and Evaluating Treatment Effects in Econometrics*, edited by Tom Fomby, R. Carter Hill, Daniel L. Millimet, Jeffrey A. Smith, and Edward J. Vytlacil, 167–95. Bingley: Emerald Group Publishing Limited.
- Holland, Paul W. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81(396):945–60. <https://doi.org/10.1080/01621459.1986.10478354>
- Holmlund, Helena, Helmut Rainer, and Thomas Siedler. 2013. “Meet the Parents? Family Size and the Geographic Proximity Between Adult Children and Older Mothers in Sweden.” *Demography* 50(3):903–31. <https://doi.org/10.1007/s13524-012-0181-1>
- Hruschka, Daniel J., Rebecca Sear, Joseph Hackman, and Alexandria Drake. 2019. “Worldwide Fertility Declines Do Not Rely on Stopping at Ideal Parities.” *Population Studies* 73(1):1–17. <https://doi.org/10.1080/00324728.2018.1513164>
- Imbens, Guido W. 2010. “Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009).” *Journal of Economic Literature* 48(2):399–423. <https://doi.org/10.1257/jel.48.2.399>
- Imbens, Guido W. 2014. “Instrumental Variables: An Econometrician’s Perspective.” *Statistical Science* 29(3):323–58. <https://doi.org/10.1214/14-STS480>

- Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2):467–75.
<https://doi.org/10.2307/2951620>
- Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.
- Jacobsen, Joyce P., James Wishart Pearce III, and Joshua L. Rosenbloom. 1994. "The Effects of Childbearing on Married Women's Labor Supply and Earnings: Using Twin Births as a Natural Experiment." *Journal of Human Resources* 34(3):449–74. <https://doi.org/10.2307/146376>
- Kolk, Martin. 2015. "The Causal Effect of an Additional Sibling on Completed Fertility: An Estimation of Intergenerational Fertility Correlations by Looking at Siblings of Twins." *Demographic Research* 32(Article 51):1409–20.
<https://doi.org/10.4054/DemRes.2015.32.51>
- Krieger, Nancy, and George Davey Smith. 2016. "The Tale Wagged by the DAG: Broadening the Scope of Causal Inference and Explanation for Epidemiology." *International Journal of Epidemiology* 45(6):1787–1808.
<https://doi.org/10.1093/ije/dyw114>
- Kruk, Kai Eberhard, and Steffen Reinhold. 2014. "The Effect of Children on Depression in Old Age." *Social Science & Medicine* 100(Supplement C):1–11.
<https://doi.org/10.1016/j.socscimed.2013.09.003>
- Li, Hongbin, Junsen Zhang, and Yi Zhu. 2008. "The Quantity–Quality Trade-off of Children in a Developing Country: Identification Using Chinese Twins." *Demography* 45(1):223–43. <https://doi.org/10.1353/dem.2008.0006>

- Lu, Yao. 2009. "Sibship Size and Education in South Africa: Black–White Variations." *Research in Social Stratification and Mobility* 27(2):110–25.
<https://doi.org/10.1016/j.rssm.2009.04.002>
- Majbouri, Mahdi. 2019. "Twins, Family Size and Female Labour Force Participation in Iran." *Applied Economics* 51(4):387–97.
<https://doi.org/10.1080/00036846.2018.1497853>.
- Marteleteo, Letícia J., and Laetícia R. de Souza. 2012. "The Changing Impact of Family Size on Adolescents' Schooling: Assessing the Exogenous Variation in Fertility Using Twins in Brazil." *Demography* 49(4):1453–77.
<https://doi.org/10.1007/s13524-012-0118-8>
- Marteleteo, Letícia J., and Laetícia R. de Souza. 2013. "The Implications of Family Size for Adolescents' Education and Work in Brazil: Gender and Birth Order Differences." *Social Forces* 92(1):275–302. <https://doi.org/10.1093/sf/sot069>
- Moffitt, Robert. 2005. "Remarks on the Analysis of Causal Relationships in Population Research." *Demography* 42(1):91–108. <https://doi.org/10.1353/dem.2005.0006>
- Mogstad, Magne, and Matthew Wiswall. 2016. "Testing the Quantity–Quality Model of Fertility: Estimation Using Unrestricted Family Size Models." *Quantitative Economics* 7(1):157–92. <https://doi.org/10.3982/QE322>
- Morgan, Stephen L., and Christopher Winship. 2015. *Counterfactuals and Causal Inference. Methods and Principles for Social Research*. (2nd ed.) New York, NY: Cambridge University Press.
- Moschion, Julie. 2009. "Offre de travail des mères en France: l'effet causal du passage de deux à trois enfants. [Labour supply of mothers in France: Causal effect of the

- shift from two to three children.]” *Economie et Statistique* 422(1):51–78.
<https://doi.org/10.3406/estat.2009.8018>.
- Moschion, Julie. 2010. “Reconciling Work and Family Life: The Effect of the French Paid Parental Leave.” *Annals of Economics and Statistics* 99/100:217–46.
<https://doi.org/10.2307/41219166>.
- Moschion, Julie. 2013. “The Impact of Fertility on Mothers’ Labour Supply in Australia: Evidence from Exogenous Variation in Family Size.” *Economic Record* 89(286):319–38. <https://doi.org/10.1111/1475-4932.12042>.
- Nguyen, Cuong, and Anh Tran. 2017. “The Effect of Having Children on Women’s Marital Status: Evidence from Vietnam.” *The Journal of Development Studies* 53(12):2102–17. <https://doi.org/10.1080/00220388.2016.1269887>
- Oliveira, Jaqueline. 2016a. “The Value of Children: Inter-Generational Support, Fertility, and Human Capital.” *Journal of Development Economics* 120(Supplement C):1–16. <https://doi.org/10.1016/j.jdeveco.2015.12.002>
- Oliveira, Jaqueline. 2016b. “Fertility, Migration, and Maternal Wages: Evidence from Brazil.” *Journal of Human Capital* 10(3):377–98. <https://doi.org/10.1086/687416>
- Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. London, UK: Allen Lane.
- Philipov, Dimitar, Aart C. Liefbroer, and Jane E. Klobas (eds.). 2015. *Reproductive Decision-Making in a Macro-Micro Perspective*. Dordrecht: Springer Netherlands.
<https://doi.org/10.1007/978-94-017-9401-5>
- Ponczek, Vladimir, and Andre Portela Souza. 2012. “New Evidence of the Causal Effect of Family Size on Child Quality in a Developing World.” *Journal of Human Resources* 47(1):64–106. <https://doi.org/10.1353/jhr.2012.0006>

- Roberts, Evan, and John Robert Warren. 2017. "Family Structure and Childhood Anthropometry in Saint Paul, Minnesota in 1918." *The History of the Family* 22(2–3):258–90. <https://doi.org/10.1080/1081602X.2016.1224729>
- Rosenzweig, Mark R., and Kenneth I. Wolpin. 1980a. "Life-Cycle Labor Supply and Fertility: Causal Inferences from Household Models." *Journal of Political Economy* 88(2):328–48. <https://doi.org/10.1086/260868>
- Rosenzweig, Mark R., and Kenneth I. Wolpin. 1980b. "Testing the Quantity–Quality Fertility Model: The Use of Twins as a Natural Experiment." *Econometrica* 48(1):227–40. <https://doi.org/10.2307/1912026>
- Rosenzweig, Mark R., and Kenneth I. Wolpin. 2000. "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature* 38(4):827–74. <https://doi.org/10.1257/jel.38.4.827>
- Rosenzweig, Mark R., and Junsen Zhang. 2009. "Do Population Control Policies Induce More Human Capital Investment? Twins, Birth Weight and China's 'One-Child' Policy." *The Review of Economic Studies* 76(3):1149–74. <https://doi.org/10.1111/j.1467-937X.2009.00563.x>
- Rubin, Donald B. 1990. "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science* 5(4):472–80. <http://www.jstor.org/stable/2245383>
- Rutter, Michael. 2007. "Proceeding From Observed Correlation to Causal Inference The Use of Natural Experiments." *Perspectives on Psychological Science* 2(4):377–95. <https://doi.org/10.1111/j.1745-6916.2007.00050.x>
- Shen, Guangjun, Jingxian Zou, and Xiaogunag Liu. 2017. "Economies of Scale, Resource Dilution and Education Choice in Developing Countries: Evidence from

- Chinese Households.” *China Economic Review* 44:138–53.
<https://doi.org/10.1016/j.chieco.2017.03.003>
- Silles, Mary A. 2016. “The Impact of Children on Women’s Labour Supply and Earnings in the UK: Evidence Using Twin Births.” *Oxford Economic Papers* 68(1):197–216. <https://doi.org/10.1093/oep/gpv055>
- Silles, Mary A. *Forthcoming* 2018. “The Labor Market Consequences of Teenage Childbearing.” *Contemporary Economic Policy*.
<https://doi.org/10.1111/coep.12417>.
- Silventoinen, Karri, Mikko Myrskylä, Per Tynelius, Yoshie Yokoyama, and Finn Rasmussen. 2013. “Social Modifications of the Multiple Birth Effect on IQ and Body Size: A Population-Based Study of Young Adult Males.” *Paediatric and Perinatal Epidemiology* 27(4):380–87. <https://doi.org/10.1111/ppe.12054>
- Singh, Susheela, Gilda Sedge, and Rubina Hussain. 2010. “Unintended Pregnancy: Worldwide Levels, Trends, and Outcomes.” *Studies in Family Planning* 41(4):241–50. <https://doi.org/10.1111/j.1728-4465.2010.00250.x>
- Small, Dylan S., Zhiqiang Tan, Roland R. Ramsahai, Scott A. Lorch, and M. Alan Brookhart. 2017. “Instrumental Variable Estimation with a Stochastic Monotonicity Assumption.” *Statistical Science* 32(4):561–79.
<https://doi.org/10.1214/17-STS623>
- Stradford, Laura, Frans W. A. Van Poppel, and L. H. Lumey. 2017. “Can Resource Dilution Explain Differences in Height by Birth Order and Family Size? A Study of 389,287 Male Recruits in Twentieth-Century Netherlands.” *The History of the Family* 22(2–3):214–35. <https://doi.org/10.1080/1081602X.2016.1230510>

- Van Kippersluis, Hans, and Cornelius A. Rietveld. 2018. "Pleiotropy-robust Mendelian randomization." *International Journal of epidemiology* 47(4):1279–88.
<https://doi.org/10.1093/ije/dyx002>
- Vere, James P. 2011. "Fertility and Parents' Labour Supply: New Evidence from US Census Data." *Oxford Economic Papers* 63(2):211–31.
<https://doi.org/10.1093/oep/gpr003>
- Zhang, Junchao. 2017. "A Dilemma of Fertility and Female Labor Supply: Identification Using Taiwanese Twins." *China Economic Review* 43:47–63.
<https://doi.org/10.1016/j.chieco.2016.12.005>

ⁱ Betz, Cook, and Hollenbach (2019) raise a very similar critique against common spatial instrumental variables, such as rainfall and natural disasters, in cases where it is likely there are 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.

ⁱⁱ After Angrist, Lavy, and Schlosser (2005, 2010) and Angrist and Pischke (2008:160–61) also suggested that the estimated effect when using twin birth IVs should be interpreted as a LATE, it has been adopted by some other studies (Cáceres-Delpiano 2006, 2012b; Åslund and Grönqvist 2010; Cáceres-Delpiano and Simonsen 2012; Moschion 2013; Baranowska-Rataj, de Luna, and Ivarsson 2016; Baranowska-Rataj and Matysiak 2016; Braakmann and Wildman 2016; Silles 2016; Brinch, Mogstad, and Wiswall 2017; Gabel et al. 2018; Guo et al. 2018; Silles 2018; Ajefu 2019). The other studies in the literature have implicitly assumed homogenous reception of treatment, a uniform effect of the instrument on the endogenous variable, and homogenous treatment effects. This is, as we have seen, not at all plausible.

ⁱⁱⁱ 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 families that deviate from the expected reaction to twin and single births. Small et al. (2017) also recently introduced a new type of causal effect that can be estimated despite a presence of defiers.

^{iv} 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 2000:864).

^v A potential objection to this assumption is that the treatment of interest—the number of children—is not actually a binary variable. The implication could be that twin birth IVs should be discussed using the non-binary version of the LATE, the “average causal response” (Angrist and Imbens 1995). However, doing so would only lead us back to the fact that the treatment in the twin birth IV case is binary. (See Appendix, Section A8, for further discussion.)

^{vi} The probability was chosen to be realistic (D’Addato 2007), but changing its level does not change any of the conclusions in the paper.

^{vii} The group that have their children faster than expected are, as we will see below, called always-takers or never-takers depending on how we define the treatment. They are ignored in the estimation based on assumptions.

^{viii} It has been more common in the literature studying the effect of the number of children on the mothers’ labor force participation to recognize that the effect of a multiple birth will vary (decline) over time (e.g., Bronars and Grogger 1994:1143; Jacobsen, Pearce III, and Rosenbloom 1999:456; Vere 2011; Braakmann and Wildman 2016). The “timing effect” is less well recognized in the literature studying the effects of their number of siblings on children (but see, e.g., Cáceres-Delpiano 2006:749–51fn13).

^{ix} There is some ambiguity in what Angrist, Lavy, and Schlosser (2010) think about the definition of the treatment. Even if they define the treatment as being “a dummy for having another child” (p. 788), they also discuss how “the subpopulation of compliers consists of individuals who switch from having fewer than j to at least j children because of the

instrument” (p. 787). The clause “because of the instrument” acknowledges that some families would have had another child even without the twin birth. This acknowledgment is also reflected in their discussion of their first-stage coefficients, where they explain why these are lower than the coefficients estimated by Angrist and Evans (1998): “Multiple births result in a smaller increase in family size when families would have been large even in the absence of a multiple birth” (Angrist, Lavy, and Schlosser 2010:785). Bringing the desired number of children into the definition moves the interpretation in the direction of “unwanted” children.

^x If we use the “extra” child interpretation of the treatment there is therefore no need to use a two-stage estimator. Its “causal” effect is equal to the reduced form effect of a twin birth on the outcome (as in, for example, Baranowska-Rataj et al. 2017). The estimated effect is still unreliable and meaningless.

^{xi} The SUTVAs are not mentioned in any of the papers applying twin birth IVs, and it is never evaluated. Angrist and Evans (1998), Angrist, Lavy, and Schlosser (2010), Baranowska-Rataj, de Luna, and Ivarsson (2016), Baranowska-Rataj and Matysiak (2016), Braakmann and Wildman (2016), Silles (2016), and Brinch, Mogstad, and Wiswall (2017) cite Angrist, Imbens, and Rubin (1996) for the definition of compliers and the LATE. (Åslund and Grönqvist (2010) cite Angrist (2004), and Cáceres-Delpiano (2012b) and Cáceres-Delpiano and Simonsen (2012) cite Angrist and Pischke (2008) for the same purpose.) None of the other papers listed in Table 1 cites Angrist, Imbens, and Rubin (1996), which is the only text defining the LATE that explicitly mentions the SUTVAs (before Imbens and Rubin 2015).

^{xii} In the words of Small et al. (2017:567), there should be “no unrepresented versions of the IV”.

Appendix: Instrumental variables based on twin births are by definition not valid

Section A1: Critique of versions of the twin birth IV

Because the likelihood of experiencing any twin birth clearly increases with the number of births, an indicator of whether the family has experienced any twin birth (as used in, e.g., Lu 2009; Braakmann and Wildman 2016; de Jong, Smits, and Longwe 2017; Nguyen and Tran 2017; Shen, Zou, and Liu 2017; and Ajefu 2019) is clearly not a valid instrument (e.g. Rosenzweig and Wolpin 2000:862fn48). The positive association between the instrument 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, 2011b) or any twin birth among younger siblings (Dasgupta and Solomon 2018; Fletcher and Kim 2019), will also not be valid.

Some studies have used a twin as the last birth as the instrument (de Haan 2010; Hatton and Martin 2010). This instrument will also be 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. As shown below in Table A1, the net association between this instrument and the desired number of children will not always be strong in practice. It is still not a plausible instrument.

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 used also in Dayioğlu, Kirdar, and Tansel 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. Interpreting the estimated effect will therefore be a severe challenge. Rosenzweig and Wolpin (2000:838) also criticized this instrument as not being determined completely randomly.

Table A1. Correlations between the twin birth instrument and the parents' desired and realized number of children and the number of births

	Desired	Realized	Births
Panel A. Distribution based on: Black, Devereux and Salvanes (2005)			
Twin as first birth	+0.000	+0.020	−0.096
Twin as second birth	+0.000	+0.065	−0.067
Twin as last birth	+0.022	+0.107	−0.048
Share twin births	+0.029	+0.095	−0.076
Any twin birth	+0.093	+0.166	−0.014
Panel B. Distribution based on: Åslund and Grönqvist (2010)			
Twin as first birth	+0.000	+0.013	−0.106
Twin as second birth	+0.000	+0.066	−0.065
Twin as last birth	+0.015	+0.101	−0.062
Share twin births	+0.023	+0.088	−0.092
Any twin birth	+0.089	+0.162	−0.027
Panel C. Distribution based on: Stradford, van Poppel and Lumey (2017)			
Twin as first birth	+0.000	+0.003	−0.055
Twin as second birth	+0.000	+0.011	−0.049
Twin as last birth	+0.007	+0.048	−0.032
Share twin births	+0.020	+0.046	−0.063
Any twin birth	+0.139	+0.166	+0.048
Panel D. Distribution based on: Roberts and Warren (2017)			
Twin as first birth	+0.000	+0.011	−0.06
Twin as second birth	+0.000	+0.025	−0.051
Twin as last birth	+0.018	+0.069	−0.026
Share twin births	+0.029	+0.067	−0.048
Any twin birth	+0.133	+0.172	+0.048

Note: The 230 (all but twin birth as the second birth) or 141 (twin birth as the second birth) observations were weighted by the probabilities of each outcome and the number of children desired. Each panel presents figures in which the distribution of the desired number of children is based on the study in the heading of the panel.

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. Or, as Rosenzweig and Wolpin (2000:863) put it, having a twin as first birth “correspond mainly to a difference in the timing of births”.

Table A1 presents how different specifications of the twin birth IV are associated with the parents’ desired and realized number of children and the number of births. I have chosen to illustrate the problematic associations using the desired number of children because it is more easily comprehensible than an invented proxy for the parents’ preferences. The simulated populations used to make these calculations are presented below in Section A4. (See also the Appendix Spreadsheet.) Observably, the non-parity-specific IVs—a twin birth as the last birth, the share of twin births, and any twin birth—are all associated with the desired number of children. The associations are not strong, as measured by Pearson’s correlation coefficient, but the presence of these associations will nonetheless affect our results. Reassuringly, we observe that the parity-specific IVs—a twin birth as the first birth and a twin birth as the second birth—are not associated with the desired number of children. This fact comes at the cost of them being worse predictors of the realized number of children. However, as IVs, they are clearly superior because they are not by definition associated with one of the confounding factors in our analyses, the desired number of children.

All twin birth IVs are negatively associated with the number of births, which is natural because a twin birth consists of two births at once. This association is not automatically a problem, but it could be if we have reason to believe that the effect of the number of siblings varies between parities, as recently argued in a couple of papers discussing twin birth IVs (Mogstad and Wiswall 2016; Guo et al. 2017).

Section A2: Previous discussions of violations of the exclusion restriction

IVs based on (parity-specific) twin births are not valid if the twin birth has effects on families other than increasing the number of children for 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 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: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 1998:473). Rosenzweig and Wolpin (2000:832) clarified 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”. We also cannot allow for the possibility that “having two children simultaneously may directly affect both the marginal utility of leisure and child costs” (Rosenzweig and Wolpin 2000:867). 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).

There are still examples in the literature where researchers describe indications of violations of these assumptions. Angrist, Lavy, and Schlosser (2010), for example, elaborate on this and 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. 2010:776, 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. 2010:788fn15). These discussions are explanations of how their IVs violate the exclusion restriction.

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. This would have the consequence of making the effect estimated for older (non-twin) siblings a lower-bound estimate of the true effect. (They propose using the effect on the twins themselves as the corresponding upper-bound.) Rosenzweig and Zhang stress that even estimates using parity-specific twin birth IVs in $n+$ samples will still be “confounded by inter-child allocation effects because of the endowment deficit and close spacing of twins” (Rosenzweig and Zhang 2009:1149).

Section A3: The effect of including “incomplete” families in our study population

When we use twin births as IVs for the number of children, we, as always for IVs, assume that there is nothing else associated with a twin birth that is affecting the outcome other than the fact that the families that experience a twin birth have more children than those that do not. Included in this assumption is that it makes no difference to the children if they have two siblings being born at once instead of with some time in between; in other words, “timing” should not matter. Provided that we study incomplete families, some of the difference in the number of children between families that do and do not experience a twin birth will be due to the timing effect. We assume that the timing has no effect on the outcome. The only thing that should create differences in the outcome is the difference in the final number of children created by the twin birth, i.e., the difference among complete families. When we include incomplete families, we will therefore dilute the causal effect on the outcome by overestimating the difference in the number of children in the first stage. When we use twin birth IVs in samples including incomplete families, the effect will therefore be biased toward zero provided that the assumptions hold. This fact has to date been overlooked in the literature using IVs based on twin births. If the assumption does not hold—if, for example, there are effects of the timing on the outcome—then we are not estimating the effect of the number of children but a sample-specific effect that is not well defined and therefore has no clear interpretation.

The effect of a twin birth on the number of children is estimated by the coefficient on the instrument in the first-stage regression. More specifically, we estimate the effect of experiencing a twin birth at the parity we use to define our instrument. The size of the effect will depend on which parity we are studying and the distribution of the desired number of children in the population. As discussed above, it will also depend on whether we are studying only complete families or not. The results in Table A2 illustrate these influences on the first-stage coefficient. The simulated populations used to make these calculations are presented below in Section A4. (See also the Appendix Spreadsheet.)

Table A2. The size and statistical significance of the first-stage coefficients for different versions of instrumental variables based on twin births across four different populations

Distribution of desired number of children based on:	Instrument	Complete families			Incomplete families			N
		$b_{complete}$	$F_{complete}$	$R^2_{complete}$	$b_{incomplete}$	$F_{incomplete}$	bias in $\hat{\beta}$ (%)	
Black, Devereux, and Salvanes (2005)	Twin as first birth	+0.172	0.0909	0.00040	+0.188	0.1246	-8.5	230
	Twin as second birth	+0.493	0.5857	0.00420	+0.512	0.7840	-3.7	141
	Twin as last birth	+0.690	2.6193	0.01136	+0.685	2.9817	+0.7	230
	Share twin births	+1.521	2.0869	0.00907	+1.541	2.4769	-1.3	230
	Any twin birth	+0.932	6.4279	0.02742	+0.912	7.1318	+2.2	230
Åslund and Grönqvist (2010)	Twin as first birth	+0.108	0.0380	0.00017	+0.125	0.0624	-13.6	230
	Twin as second birth	+0.506	0.6131	0.00439	+0.526	0.8574	-3.8	141
	Twin as last birth	+0.622	2.3433	0.01017	+0.617	2.7968	+0.8	230
	Share twin births	+1.354	1.7901	0.00779	+1.377	2.2492	-1.7	230
	Any twin birth	+0.866	6.1185	0.02613	+0.840	6.9992	+3.1	230
Stradford, van Poppel, and Lumey (2017)	Twin as first birth	+0.048	0.0017	0.00001	+0.331	0.2443	-85.5	230
	Twin as second birth	+0.177	0.0157	0.00011	+0.477	0.3855	-62.9	141
	Twin as last birth	+0.599	0.5171	0.00226	+0.483	0.9825	+24.0	230
	Share twin births	+1.679	0.4792	0.00210	+2.175	2.3669	-22.8	230
	Any twin birth	+1.433	6.4461	0.02750	+1.111	11.5584	+29.0	230
Roberts and Warren (2017)	Twin as first birth	+0.150	0.0261	0.00011	+0.262	0.1410	-42.7	230
	Twin as second birth	+0.327	0.0861	0.00062	+0.461	0.3608	-29.1	141
	Twin as last birth	+0.725	1.0910	0.00476	+0.671	1.6704	+8.0	230
	Share twin births	+1.845	1.0253	0.00448	+1.980	2.1114	-6.8	230
	Any twin birth	+1.398	6.9358	0.02952	+1.215	9.4139	+15.1	230

Note for Table 4: The bias was calculated as $100 \times (1/(b_{incomplete}/b_{complete}) - 1)$.

The results in Table A2 show that IVs based on twin births are poor predictors of the number of children in the family even though they are always associated with a substantial difference in the number of children, i.e., a sizable first-stage coefficient. All versions of the instrument explain a miniscule amount of the variation in all populations. It is also only the coefficients for the any-twin-birth instrument that are statistically significant. All F -statistics are far from Staiger and Stock's (1997) often-cited rule-of-thumb value of ten that is used to indicate that the instrument is not "weak".¹ The statistical significance of the first-stage coefficients in empirical applications of IVs based on twin births is therefore mostly a result of the sample size. The inclusion of incomplete families in the analyses will also contribute to increasing both the size of the first-stage coefficient and its level of statistical significance. The increase in the size of the coefficient is exemplified in Table A2 by relating the difference to the true value estimated from only complete families. The bias varies in size depending on both the version of the instrument and the population studied. It is therefore difficult to predict how severe the bias will be in the many studies that have included incomplete families in their analyses. The bias is positive, meaning an overestimated first-stage coefficient, in all cases but one. We can therefore conclude that almost all these studies will underestimate any causal effect of the number of children.

As shown in Table A2, the size of the coefficient will depend on the parity and population studied but will always have the same substantive interpretation. If, for example, we use a twin birth as the second birth as the instrument, then the first-stage coefficient is the difference between the probability of having a third child when experiencing a twin birth as the second birth (probability equal to one) and the probability of having a third child when not experiencing a twin birth as the second birth (compare with Angrist and Pischke 2015:128, see also p. 118). The probability of having a third child when not experiencing a twin birth as the second birth depends on the distribution of the desired number of children in the population. In other words, the difference between the two probabilities, i.e., the first-stage coefficient, is therefore the share of families that desired two children but had three because of the twin birth. The first-stage coefficient is seldom interpreted in studies applying this method. However, Angrist and Pischke (2008:132, 2015:128) have discussed the results in Angrist and Evans (1998) and Angrist et al. (2005, 2010). Their interpretation is the same as that which I have just presented.

Even if the results are not always interpreted, most studies report the results from the first-stage regression. The first-stage coefficient on the twin birth IV is approximately 0.7–0.8 in

most present-day populations (Bhalotra and Clarke 2018a:18–19). That the coefficient is smaller than one illustrates that twin births lead only to an exogenous increase in the number of children in some families. We know that the coefficient must be below one in all reasonable applications of the method. There should be nothing, except the twin birth, that makes families that experience a (parity-specific) twin birth systematically different from other families. If this assumption—the exclusion restriction—is correct, then the largest possible difference in the number of children between families that do and do not experience a (parity-specific) twin birth is one. A coefficient of one would mean that all twin births (at the studied parity) result in unintended, “unwanted” births.

Section A4: A detailed description of the simulated populations

As always, we should, start with a clear definition of what we are studying. IVs based on twin births are used to study how parents are affected by their number of children and how children are affected by their number of siblings. The number of children/siblings is an endogenous explanatory variable in such a model, and IVs based on twin births are intended to provide exogenous variation.

A benefit of studying people and families is that we can know something about the process through which children are born, i.e., how families behave, and we can use this knowledge to create a simplified model of the underlying process of what we study (Figure A1 in Section A6). My aim here is to formulate a simplistic but logical theoretical model of events and behaviors related to the birth of children, not to formulate a complete behavioral model. I argue that the simplistic model is a sufficiently accurate representation of reality and of the behaviors assumed when using twin births for IVs. Other behavioral assumptions underlying the use of twin births for IVs are discussed in, for example, Rosenzweig and Wolpin (2000) and Rosenzweig and Zhang (2009).

The starting point of my model is a population of prospective parents. These parents desire different numbers of children (but all desire at least one). These parents then proceed to become pregnant and give birth to either one, i.e., single birth, or two, i.e., twin birth, children. (Throughout the paper, for simplicity, I ignore higher-order multiple births.) They go on to do so until they have reached or, through a twin birth, surpassed their desired number of children. Figure 1 (in the main article) shows the possible sequences of single and twin births through which families desiring two or three children can reach or surpass their desired

number of children. This simplistic model relies on a number of assumptions that are unlikely to be completely accurate; I discuss these assumptions further below. I argue that the model is nonetheless useful for discussing IVs based on twin births.

Based on the theoretical model, I created a simulated population of parents with different birth sequences. To do so, I extended the model in Figure 1 (in the main article) so that families are allowed to desire 1, 2, 3, 4, 5, 6, 7, 8, or 9 children. In reality, few families will desire more than nine children and setting this maximum value limits the number of possible sequences in the data. Allowing the families in the simulation to desire between one and nine children leads to 230 different combinations of single and twin births that they can experience before they have all reached or surpassed their desired number of children. When we use a twin as the second birth as the IV, we focus on families with at least two births. All families therefore desire at least two children, i.e., between two and nine children. If we, as is most common, exclude the twins themselves from the analyses, then we also exclude families that have a twin as the first birth. This exclusion leads to 141 different possible combinations of single and twin births to reach or surpass the desired number of children for this subpopulation.

Each sequence has a probability of occurring. Sequence 1 in Figure 1 (in the main article), for example, will have the probability $p_1 = (1 - p(\text{twin}))(1 - p(\text{twin}))$, with $p(\text{twin})$ being the likelihood of a twin birth. The complete list of sequences and their respective probabilities is presented in a spreadsheet available as an Appendix. The probability of each sequence is determined by the likelihood of a twin birth. More importantly, however, which sequence a family follows is determined by the number of children that the parents desire.

The desired number of children is rarely observed empirically. When we use twin births as IVs, we also assume that parents have a fixed number of children that they desire (Rosenzweig and Wolpin 1980b:232; see also, e.g., Black, Devereux, and Salvanes 2005:681). This assumption is, most likely, unrealistic in real life (e.g. Bachrach and Morgan 2013; Philipov, Liefbroer, and Klobas 2015), making the desired number of children even more difficult to observe empirically.

I used the complete set of different possible combinations of twin and single births as well as their respective probabilities to create the simulated populations of families. To do so, I also needed distributions of the desired number of children; thus, I used four different

empirical distributions of the relative frequencies of families having different numbers of children. The four distributions cover different populations and time periods (Sweden, people born 1972–1979: Åslund and Grönqvist 2010: the distribution is presented in Table 1; Norway, average year of birth 1962: Black et al. 2005:Tab. II; Saint Paul, MN, USA, children aged 0–17 years in 1920: Roberts and Warren 2017:Tab. 3; The Netherlands, men born 1944–1947: Stradford, van Poppel, and Lumey 2017:Tab. 1). I used these different distributions of realized numbers of children as proxies for hypothetical distributions of the desired numbers of children. This is not a perfect solution, but it allows me to investigate how the twin birth IVs are affected by changes in the behavior and preferences of populations. The fact that I am using distributions of realized number of children as a proxy for the desired number does not affect the substantive results. I could have used different mathematical distributions to the same end, but I chose to use the empirical distributions because I prefer to have some empirical basis for their shapes and levels.

When discussing the effect of including incomplete families in the analysis sample, I simulate cutting some fertility histories short by only including the first five births for all families. This is a crude adjustment because it changes the observed values only for families desiring many children. In empirical applications including incomplete families, the number of observed children would be affected for all families. This would lead to stronger biases also in the populations desiring fewer children, i.e. the populations based on the distributions in Black et al. 2005 and Åslund and Grönqvist 2010.

I make the following assumptions in the model creating the simulated population of birth histories:

- Everyone can and will reach (or surpass) their desired number of children. In other words, there is no involuntary childlessness, infertility, or other limitations on fertility decisions.
- Parents have, *ex ante*, a fixed number of children that they want. This assumption also implies 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.
- All parents are willing to risk surpassing their desired number of children to reach their desired number.

- There are no unintended pregnancies, and therefore, there can be no “unwanted” single births.
- Twin births occur completely at random with a constant probability ($p = 0.0175$).
- Multiple births occur only as twin births.
- The timing and spacing of births have no effects on the children or the parents.
- There is no effect of birth order on the outcome.

These assumptions are all wrong. There are widespread problems with involuntary childlessness and infertility worldwide (Gurunath et al. 2011; Mascarenhas et al. 2012). Fertility preferences are complex and are dynamically influenced by a large number of different factors (Bachrach and Morgan 2013; Philipov et al. 2015). There are many unintended pregnancies in populations all over the world (Singh, Sedge, and Hussain 2010; Alkema et al. 2013). Twin births do not occur completely at random, at least not dizygotic twin births (Braakmann and Wildman 2016; Bhalotra and Clarke 2018b; Farbmacher, Guber, and Vikström 2018). The timing (Gipson, Koenig, and Hindin 2008; Hall et al. 2017) and the spacing of births (Conde-Agudelo et al. 2012; Kozuki et al. 2013) can be expected to have a number of different effects on both the children and the parents (see also Rosenzweig and Zhang 2009). Several previous studies have found effects of birth order (e.g., Myrskylä et al. 2013; Jayachandran and Pande 2017). The linear specification, assuming no differences in effects between parities, has recently been shown not to work well when we investigate how children are affected by their number of siblings (Mogstad and Wiswall 2016; Guo, Yi, and Zhang 2017).

I make these assumptions, despite knowing that they are not actually accurate, to preserve the clarity of the model and to avoid tangential issues. Realistic processes for determining the fertility preferences of parents and the birth of a child would change my deterministic, probability-weighted outcomes into probabilistic outcomes. The strong assumption regarding parents having a fixed desired number of children is made implicitly in all studies using twin births as IVs and was used by Rosenzweig and Wolpin (1980b: e.g., 232) in their original derivation of the method. I maintain the assumption that twin births occur completely at random for clarity and to show that the issues that I raise here are independent of the issues raised by Braakmann and Wildman (2016), Bhalotra and Clarke (2018a; 2018b), and Farbmacher et al. (2018). The strong assumption that neither parents nor children are affected

by the timing and spacing of births is made in almost all studies using twin birth IVs (but see Rosenzweig and Zhang 2009).

Section A5: Further discussion of the potential outcomes framework and interpretations of the treatment

The potential outcomes framework conceptualizes the estimation of a causal effect in terms of a designed experiment. This conceptualization does not mean that the framework is valid only for experiments. The arguments are applicable to all attempts to estimate causal effects, including in social sciences in which experiments are frequently impossible or unethical. Conceptualizing the model as an experiment is useful for highlighting the often implicit assumptions made when we estimate causal effects.

Using the experimental terminology, we estimate the effect of a “treatment”, W , on the outcome, Y . The treatment must be something that we can, at least hypothetically, think of as being assigned as a treatment in an experiment. This criterion is one of the reasons why this framework is not suitable for all types of research questions. The treatment in the twin birth IV case is the number of children. The causal effect in the potential outcomes framework is defined as the difference between two potential outcomes defined for the same unit. In the twin birth IV case, the implication is that we, for example, compare the outcome for the child after varying the treatment, that is, the number of siblings. We can have one observation on the family and the fate of the child, which, following the notation in Imbens and Rubin (2015), is as follows: $Y_i^{obs} = Y_i(W_i)$. This observation is then compared with how the child would have fared with a different, for example, larger, number of siblings, $Y_i^{mis} = Y_i(W_i')$. Because in this comparison the same family and child are studied in two situations (of which one is hypothetical), everything except the number of children is kept constant. The effect of increasing the number of children on the outcome, meaning $W_i' > W_i$, is therefore $Y_i^{mis} - Y_i^{obs}$.

This discussion is easily extended to include an IV for the treatment. Again conceptualizing it as an experiment, the treatment is divided into two parts, the assignment to treatment—the instrument—and the receipt of treatment—the treatment we are estimating the effect of (Imbens and Rubin 2015:513). In the words of Angrist and Pischke (2015:120), “The IV causal chain begins with random assignment to treatment, runs through treatment delivered, and ultimately affects outcomes”. We use IVs when there are reasons to believe

that the units receiving treatment are systematically different from other units in unobservable ways. The IV, or assignment to treatment, should not be affected by this (unobservable) confounding and can therefore isolate exogenous variation in the receipt of treatment. In the twin birth IV case, the (supposedly) randomly occurring twin births constitute the assignment mechanism creating exogenous variation in the receipt of treatment, which would ideally be the number of children.

In observational data, we can seldom expect that the effect of the instrument on the treatment is the same for everyone. The effect of the treatment on the outcome will also vary. Furthermore, it is most likely the case that the treatment is not unique to those indicated by the instrument. In such common situations, we must include both the instrument and the treatment in the definition of the potential outcome. The potential outcome is then $Y_i(Z_i, W_i(Z_i))$, with Z_i indicating the level of the instrument and $W_i(Z_i)$ the level of the treatment at that value for the instrument. Naturally, the causal effect of interest remains the effect of the treatment, W , on the outcome, Y . However, to estimate the effect, we use only the variation in the treatment that is being caused by the instrument, $W(Z)$.

In the potential outcomes framework, the causal effect is conceptually defined for a single unit of observation. As noted above, the causal effect is 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 potential outcome. However, even if we do not observe all outcomes for all families, we can think about their substantive meaning, which will depend on what we define as the treatment.

Below, I examine the three definitions of the treatment discussed in the main article, the birth of an “unwanted” child, the birth of “another” (e.g., third) child, and the birth of an “extra” child at the studied parity because of a twin birth. I assume, for the moment, that there are no other variables that we must include to adjust our estimates.

The instrument, Z_i , will be equal to one if the family experienced a (parity-specific) twin birth and zero otherwise. The value for the treatment, W_i , will depend on the value of the instrument, i.e., $W_i(Z_i)$, and how the unit reacts to being assigned to the treatment. The LATE is estimated for people whose treatment is affected by the instrument. The others are ignored based on assumptions. The causal effect is therefore defined by comparing units that are both

indicated by the instrument and treated with units that are neither indicated nor treated. The corresponding unit-level definition of the causal effect is as follows:

$$\beta_i = Y_i^{obs/mis}(Z_i = 1, W_i(Z_i = 1) = 1) - Y_i^{mis/obs}(Z_i = 0, W_i(Z_i = 0) = 0) \quad (A6)$$

If we define the birth of an “unwanted” child as the treatment, then the two situations correspond to:

- The family experienced a twin birth at the studied parity and had an “unwanted” child. This could be the result of the twin birth at the studied parity or a later twin birth.
- The family did not experience a twin birth at the studied parity and did not have an “unwanted” child (at this or any other parity).

By comparing these two situations, we estimate the effect of the birth of an unintended birth, i.e., an “unwanted” child. This effect is not the same as the effect of an increase in the number of children, but it could still be interesting to estimate.

If we define having “another” child as the treatment, then the two situations correspond to:

- The family experienced a twin birth at the studied parity and had “another”, e.g., third, child. The family might then go on to have more children.
- The family did not experience a twin birth at the studied parity and did not have “another” child, e.g., only had two children. The preferences of these parents are therefore, by definition, different from the other group of compliers.

By comparing these two situations, we estimate the effect of wanting, and voluntarily or involuntarily having, at least, “another” child.

If, finally, we define the “extra” child born at the studied parity because of a twin birth at that parity as the treatment, then the two situations correspond to:

- The family experienced a twin birth at the studied parity and therefore had an “extra” child being born at that parity.
- The family did not experience a twin birth at the studies parity and therefore did not have an “extra” child born at that parity.

By comparing these two situations, we consequently estimate the causal effect of experiencing a twin birth at the studied parity. This effect is not generalizable as an effect of the number of children in the family. It is also not generalizable to experiencing a twin birth at any parity.

We can never estimate the causal effect based on the unit-specific potential outcomes. One of these outcomes will always be merely a potential, unobservable outcome. For this reason, “the Fundamental Problem of Causal Inference” is missing data (Holland 1986:947). We are forced to find other units that are comparable and that have different values of their treatment and outcomes, which is what we do when we change the definition of the causal effect to estimate it from populations. We need some additional assumptions to hold to make this change. Our estimates of this causal effect will be accurate only if the complier units that we choose to compare are truly comparable.

To be able to estimate the causal effect empirically by comparing compliers with different assignments to treatment, we also need to be able to ignore the always- and never-takers. To be able to do this there cannot be any systematic differences between the always- and never-takers that are indicated by the instrument or not. As we see in the main article, that is most often not the case with twin birth IVs.

Section A6: Further discussion of the underlying model and IV estimators

Here, I will discuss twin birth IVs as applied to studies of how children are affected by their number of siblings. Twin birth IVs have been applied in many studies of this issue as an attempt to solve problems with endogeneity (see Table 1 in the main article). These studies often try to test Becker’s proposition that parents make trade-offs between the quantity and “quality” of their children, i.e., how many children to have and how much to invest in each (Becker and Lewis 1973; Becker and Tomes 1976). I find it useful to summarize my thinking in a mind map, and therefore, Figure A1 provides a graphical summary of an example of a model for investigating this issue.ⁱⁱ It shows how the estimated effect of the number of siblings will be biased through confounding from, in this case, for example, the parents’ socioeconomic status, the presence of “unwanted” children, and the parents’ preferences for “child quality”. Most often, we can attempt to adjust our estimates for the parents’ socioeconomic status. However, we rarely have information on the presence of “unwanted” children. In addition, the parents’ preferences for “child quality” is almost always impossible

to measure. The model in Figure A1 therefore shows one example of why and how the number of siblings becomes an endogenous variable in the model.

Figure A1. A graphical summary of a model for investigating how the number of siblings affects the outcome of a child

I use the model in Figure A1 for my illustrations of how and why IVs based on twin births do not work as intended. I model the influence of the number of siblings on children using a linear model with additive effects. The linear specification and the corresponding assumption of constant effects across parities have recently been shown not to work well for this application (Mogstad and Wiswall 2016; Guo et al. 2017). This finding is a serious and important critique of previous literature; nonetheless, I use a linear model to preserve clarity. For the same reason, I sometimes assume that there are no other variables that we must include to adjust the model.

The model that I use includes the number of siblings, N , the number of “unwanted” children in the family, U , the parents’ socioeconomic status, S , and their preference for child “quality”, Q^* , as influences on the outcome for the child, Y .

The number of siblings, N , is the variable of interest, the “treatment” of which we would like to estimate the effect. We apply IVs when investigating the effect of the number of children because we do not think that the effect of having another child is the same for everyone, i.e., that intended and unintended children will be associated with different

outcomes for children. It is therefore reasonable to include a separate factor for the potential presence of “unwanted” children in the family in the model of the outcome.

The resources available to parents will affect both the opportunities to have children and the opportunities to invest in them. We summarize these resources as the parents’ socioeconomic status, S , which is therefore a factor for which we should and, most often, can attempt to adjust for in our models. However, it is important to remember that our empirical variables—for example, the parents’ educational level, occupational status or income—will never be able to fully capture all aspects or resources summarized as the parents’ “socioeconomic status” (e.g., Dickerson and Popli 2018). There will therefore be a measurement error in the empirical variable, which, in turn, will lead to residual confounding from the parents’ socioeconomic status. Measurement errors will also contribute to the fact that we cannot know whether we are recovering the true causal effect even if our model is, in theory, correctly specified.

Parents will also have different preferences regarding how to rear children and, for example, be more or less focused on optimizing the development of the child in different aspects. These preferences will vary across different aspects of parenting. The confounding from Q^* will therefore also be different for different outcomes. However, ideally, we would like to adjust our estimates for the relevant preferences. The model would therefore be as follows:

$$Y = \alpha_1 + \beta N + \mu_1 U + \delta_1 S + \rho_1 Q^* \quad (\text{A1})$$

The number of siblings, N , must be instrumented to obtain an unbiased estimate of its effect on the outcome, β . Self-evidently, I use an IV based on twin births. The model of the number of siblings includes the twin instrument, Z , and the parents’ socioeconomic status, S . In addition to these variables, we would like to include the parents’ desired number of children, N^* , because this is one of the best predictors for the realized number of children (e.g., Schoen et al. 1999; Philipov et al. 2015:e.g., 168). The following, then, is the model that is estimated in the first-stage regression:

$$N = \alpha_2 + \beta_{FS} Z + \delta_2 S + \eta_2 N^* \quad (\text{A2a})$$

To be able to estimate the first-stage regression, we need an instrument that fulfills a number of assumptions. We need it to be:

- relevant, i.e., have a substantial influence on the instrumented variable.

This assumption is made using different wordings in different sources, for example, “Nonzero Average Causal Effect of Z on D” (Angrist, Imbens and Rubin 1996:447), “First stage” (Angrist and Pischke 2008:155), and “First-stage (population of compliers have positive probability)” (Henderson et al. 2008:172). They all mean that there must be a substantive difference in the level of the treatment between the subjects assigned to the treatment and not. In the case of IVs based on twin births, this means that families that experience a parity-specific twin birth must, on average, have a larger number of children than families that instead have a single birth at the parity used.

Regarding the instrument, we also need it to be:

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

There is more agreement on the term monotonicity (e.g., Angrist et al. 1996:447; Angrist and Pischke 2008:154; Henderson et al. 2008:172; Imbens and Rubin 2015:551; Swanson and Hernán 2018), even if Heckman, Urzua and Vytlačil (2006:391–92) suggest the term “uniformity”. This assumption adds to the previous. Monotonicity is stricter than that the average number of children must be larger in families experiencing a twin birth than in other families. The monotonicity assumption requires that the instrument changes the level of the treatment in only one direction. *All* families that experience a twin birth have at least as many children as they had intended to have. None of them have fewer than intended because of the twin birth. This is relatively unproblematic in regard to IVs based on twin births even if it is not unthinkable that there are small groups violating this assumption.

Next, we need the IV to be:

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

This assumption is discussed using different names: “Random assignment” (Angrist et al. 1996:446), “independence” (Angrist and Pischke 2008:152), and “unconfounded type” (Henderson et al. 2008:171). In the case of IVs based on twin births it is most common to discuss it in terms of twin births being random, hence leading to a random assignment to treatment.

The next assumption is most-often called the “Exclusion restriction” (Angrist et al. 1996:447; Angrist and Pischke 2008:153), but is also discussed as, for example, “Mean independence within subpopulations” (Henderson et al. 2008:171). It requires that the instrument is:

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

The exclusion restriction is separate from the independence assumption (e.g., Imbens and Angrist 1994:468; Angrist and Pischke 2008:153). This is clearer when we use the potential outcomes framework to analyze IVs than in specifications based on structural equations.

If we must include other variables to adjust our models, we also must, as was noted by Henderson et al. (2008:172), assume that these variables have:

- overlapping distributions in the groups indicated by the instrument or not.

To be able to interpret the effect we estimate using the IV as a causal effect we need to make assumptions also regarding the treatment of interest. Specifically, we need it to fulfill:

- the two parts of the stable unit treatment value assumption (SUTVAs I and II).

The SUTVA really consists of two parts (see, e.g., Cox 1958:17–21; Rubin 1990:475; Imbens and Rubin 2015:9–12; see also Heckman 2005:11–12, 35–36, 43). The first part states that there should be no interference between units. The treatment of one unit should not affect any other unit. This first part of the SUTVA is sometimes discussed such that there should be no “equilibrium effects” (Heckman 2005:11). The second part of SUTVA requires that there are no hidden versions of the treatment. The treatment should be a well-defined event that can be expected to affect all units similarly. Small et al. (2017:567) write about this issue such that there should be “no unrepresented versions of the IV”.

If our instrument and the treatment of interest fulfill the assumptions listed above, we can proceed to estimate the causal effect of the treatment on the outcome. If we do not need to adjust the model for any other variables, then we can estimate the causal effect using the Wald estimator (Angrist and Pischke 2008:Ch. 4; see also Imbens and Rubin 2015:Ch. 23). In these cases, the second stage is the estimation of the causal effect of the instrument on the outcome, β'_{SS} . A plausible instrument is not related to the outcome other than through its

effect on the treatment. The effect of the instrument on the outcome is therefore the result solely of the differences in the level of the treatment between those indicated by the instrument and not. Thus, the Wald estimate of the causal effect of the treatment on the outcome is the ratio of the effect of the instrument on the outcome divided by the effect of the instrument on the treatment, $\hat{\beta}_{Wald} = \beta'_{SS}/\beta_{FS}$. The Wald estimator is important because it is easily comprehensible and therefore provides a way to gain an intuitive understanding of what is occurring when we use IVs.

In practice, we most often want to include other variables to adjust our model. The purpose of including other variables in the first stage is to estimate the causal effect of the instrument on the treatment. To do so, we must remove all confounding from other factors. We then use the predicted value for the endogenous variable from the first-stage model, $\hat{N}$, instead of the original values in the 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.

$$Y = \alpha_3 + \beta_{SS}\hat{N} + \mu_3U + \delta_3S + \rho_3Q^* \quad (A3a)$$

We can estimate the two stages using ordinary least squares. We rely on the usual assumptions needed for this method, importantly, for example, that the instrument is not associated with the error terms. In most cases, the model that we estimate in the regression will deviate from the theoretical model with which we commence. Some factors will be unobserved or unobservable. We will, for example, almost never have empirical information on how many children the parents desired, and therefore, we will also not have information on whether a child birth was “unwanted” or what the parents’ preferences are regarding the “quality” of children. These factors, as well as many others, end up in the error term because they are not included in the empirical model. The empirical versions of the first-and second-stage models that we end up estimating are the following:

$$N_i = \alpha_2 + \beta_{FS}Z_i + \delta_2S_i + \varepsilon_{2,i} \quad (A2b)$$

$$\varepsilon_{2,i} = a_2 + \eta_2N_i^* + e_{2,i} \quad (A4)$$

$$Y_i = \alpha_3 + \beta_{SS}\hat{N}_i + \delta_3S_i + \varepsilon_{3,i} \quad (A3b)$$

$$\varepsilon_{3,i} = a_3 + \rho_3Q_i^* + \mu_3U_i + e_{3,i} \quad (A5)$$

The error terms, $\varepsilon_{2,i}$ and $\varepsilon_{3,i}$, consist of both the unobserved factors and stochastic error terms, $e_{2,i}$ and $e_{3,i}$. The instrument must be independent of the error term, $\varepsilon_{2,i}$, after

conditioning on the included variables. Otherwise, our first-stage coefficient, β_{FS} , will be biased. It must also be independent of the error term in the model of the outcome, $\varepsilon_{3,i}$, or the estimated causal effect will be biased. In the models outlined above, the implication is that there cannot be systematic differences in the desired number of children or the parents' preferences for child "quality" between families that do and do not experience a parity-specific twin birth. They also should not be any more or less likely to have an "unwanted" child if one experiences a parity-specific twin birth. This last part is not compatible with the basic foundation of the method, that twin births increase the number of children exogenously in some families.

Section A7: Instruments based on twins as first birth are weak

Rosenzweig and Wolpin (1980a) show that a twin birth as the first birth is a poor predictor of the number of children in complete families. In their words, "the occurrence of twins in the first birth represents mainly a timing failure" (Rosenzweig and Wolpin 1980a:341). This aspect makes a twin birth as the first birth a weak instrument for the number of children. In incomplete families, it will capture a "timing" effect that will depend on the time since the birth.

There are a couple of studies that seem to contradict this. Oliveira (2016a) and Guo et al. (2018) study the effect of the number of children on families in China. Both studies use data including only complete families. Nonetheless, they find surprisingly large and significant differences in family size (c.+0.9) when comparing families that experienced a twin birth as the first birth to other families. One possible explanation for this large difference is the one-child policy. Guo et al. (2018:148) concludes that the family size policies "accounts for a substantial part of the variation in fertility". For the population studied in Oliveira (2016a), these policies were implemented after the first birth of most of the mothers in the sample, but it could still have reduced their propensity to proceed to have a second or third child.

The same argument goes for all other studies applying twin birth IVs to data on Chinese populations (Li, Zhang, and Zhu 2008; Rosenzweig and Zhang 2009; He and Zhu 2016; Chen 2017; Shen, You, and Liu 2017). All other studies using twins as the first birth use data on incomplete families (Rosenzweig and Wolpin 1980a; Bronars and Grogger 1994; Jacobsen, Pearce III and Rosenbloom 1999; Glick, Marani and Shan 2007; Baranowska-Rataj and Matysiak 2016; He and Zhu 2016; Oliveira 2016b; Arouri, Ben-Youssef and Nguyen Viet

2017; Silles 2018; Majbouri 2019). The timing effect will therefore contribute to artificially strengthening the twin as first birth instrument in these applications.

Section A8: Analysis of the “average causal response” (Angrist and Imbens 1995) in the case of IVs based on twin births

A 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. 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, then the only variation in the number of children related to the (valid) IV should be that some families that wanted two children had three because of the twin birth. Angrist and Imbens (1995) show that a two-stage least squares method “identifies a weighted average of per-unit treatment effects along the length of a causal response function” (p. 431). Because all exogenous variation is *at* the studied parity, the causal response function is reduced to a binary indicator. Regardless of whether the treatment is binary or not, the LATE “is the average causal effect of treatment for those whose treatment status is affected by the instrument” (Angrist and Imbens 1995:434). 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 by the instrument, i.e., the parity-specific twin birth. The treatment when using (valid) twin birth IVs is therefore reduced to being binary.

Section A9: The development of the definition of the treatment in previous research

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). When studying the 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 1994:1142; Angrist and Evans 1998:452; Jacobsen et al. 1999:457). 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. That the early “follower” studies all studied mothers using the twin as first birth instrument can be part of the explanation for the shift in the interpretation of the treatment from an “unwanted” birth to “another” or an “extra” child being born.

The literature using twin births as IVs to investigate the effect on children took off after a number of publications in 2005 (Angrist, Lavy and Schlosser 2005 [later published in 2010]; Black et al. 2005; see also Cáceres-Delpiano 2006). In many cases, the interpretation of the treatment then returned to being an “unwanted” birth. The discussion in, for example, Black, Devereux and Salvanes (2005, 2010) indicates that they consider the “unwanted” children to be the treatment. They also mention how Rosenzweig and Wolpin's (1980b) model assumes that “parents have an optimal number of children” and that the “birth of twins can vary the actual family size from the desired size” (Black et al. 2005:681; see also, e.g., Black et al. 2010:37; Cáceres-Delpiano and Simonsen 2012:754). Angrist et al. (2010) also investigate the effect on children while nonetheless arguing that the “extra” child born is the treatment.

References for the appendix

Alkema, Leontine, Kantorova Vladimira, Menozzi Clare, and Ann Biddlecom. 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.” *The Lancet* 381(9878):1642–52. [https://doi.org/10.1016/S0140-6736\(12\)62204-1](https://doi.org/10.1016/S0140-6736(12)62204-1)

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

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

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

Angrist, Joshua D., Victor Lavy, and Analia Schlosser. 2005. “New Evidence on the Causal Link Between the Quantity and Quality of Children.” Working Paper. National Bureau of Economic Research. <https://doi.org/10.3386/w11835>

Angrist, Joshua D., Victor Lavy, and Analia Schlosser. 2010. “Multiple Experiments for the Causal Link between the Quantity and Quality of Children.” *Journal of Labor Economics* 28(4):773–824. <https://doi.org/10.1086/653830>

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2015. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton, NJ: Princeton University Press.

Arouri, Mohamed, Adel Ben Youssef, and Cuong Nguyen-Viet. 2017. “The More Children You Have the More Likely You Are to Smoke? Evidence from Vietnam.” *Oxford Development Studies* 45(3):260–75. <https://doi.org/10.1080/13600818.2016.1193129>

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

- Bachrach, Christine A., and S. Philip Morgan. 2013. "A Cognitive–Social Model of Fertility Intentions." *Population and Development Review* 39(3):459–85.
<https://doi.org/10.1111/j.1728-4457.2013.00612.x>
- Baranowska-Rataj, Anna, and Anna Matysiak. 2016. "The Causal Effects of the Number of Children on Female Employment—Do European Institutional and Gender Conditions Matter?" *Journal of Labor Research* 37(3):343–67. <https://doi.org/10.1007/s12122-016-9231-6>
- Becker, Gary S., and H. Gregg Lewis. 1973. "On the Interaction between the Quantity and Quality of Children." *Journal of Political Economy* 81(2, Part 2):S279–S288.
<https://doi.org/10.1086/260166>
- Becker, Gary S., and Nigel Tomes. 1976. "Child Endowments and the Quantity and Quality of Children." *Journal of Political Economy* 84(4, Part 2):S143–S162.
<https://doi.org/10.1086/260536>
- Bhalotra, Sonia R., and Damian Clarke. 2018a. "The Twin Instrument: Fertility and Human Capital Investment." IZA Discussion Papers 11878. Retrieved November 2, 2018 (<https://www.iza.org/publications/dp/11878/the-twin-instrument-fertility-and-human-capital-investment>).
- Bhalotra, Sonia R., and Damian Clarke. *Forthcoming* 2018b. "Twin Birth and Maternal Condition." *The Review of Economics and Statistics*. https://doi.org/10.1162/rest_a_00789
- Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. .2005. "The More the Merrier? The Effect of Family Size and Birth Order on Children's Education." *The Quarterly Journal of Economics* 120(2):669–700. <https://doi.org/10.1093/qje/120.2.669>
- Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. 2010. "Small Family, Smart Family? Family Size and the IQ Scores of Young Men." *Journal of Human Resources* 45(1):33–58. <https://doi.org/10.3368/jhr.45.1.33>
- Braakmann, Nils, and John Wildman. 2016. "Reconsidering the Effect of Family Size on Labour Supply: The Twin Problems of the Twin Birth Instrument." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179(4):1093–1115.
<https://doi.org/10.1111/rssa.12160>

- Bronars, Stephen G., and Jeff Grogger. 1994. "The Economic Consequences of Unwed Motherhood: Using Twin Births as a Natural Experiment." *The American Economic Review* 84(5):1141–56. <http://www.jstor.org/stable/2117765>
- Cáceres-Delpiano, Julio. 2006. "The Impacts of Family Size on Investment in Child Quality." *Journal of Human Resources* 41(4):738–54. <https://doi.org/10.2307/40057289>
- Cáceres-Delpiano, Julio, and Marianne Simonsen. 2012. "The Toll of Fertility on Mothers' Wellbeing." *Journal of Health Economics* 31(5):752–66. <https://doi.org/10.1016/j.jhealeco.2012.05.006>
- Conde-Agudelo, Agustín, Anyeli Rosas-Bermudez, Fabio Castaño, and Maureen H. Norton. 2012. "Effects of Birth Spacing on Maternal, Perinatal, Infant, and Child Health: A Systematic Review of Causal Mechanisms." *Studies in Family Planning* 43(2):93–114. <https://doi.org/10.1111/j.1728-4465.2012.00308.x>
- Cox, David Roxbee. 1958. *Planning of Experiments*. New York: John Wiley.
- Dickerson, Andrew, and Gurleen Popli. 2018. "The Many Dimensions of Child Poverty: Evidence from the UK Millennium Cohort Study." *Fiscal Studies* 39(2):265–98. <https://doi.org/10.1111/1475-5890.12162>
- Farbmacher, Helmut, Raphael Guber, and Johan Vikström. 2018. "Increasing the Credibility of the Twin Birth Instrument." *Journal of Applied Econometrics* 33(3):457–72. <https://doi.org/10.1002/jae.2616>
- Gipson, Jessica D., Michael A. Koenig, and Michelle J. Hindin. 2008. "The Effects of Unintended Pregnancy on Infant, Child, and Parental Health: A Review of the Literature." *Studies in Family Planning* 39(1):18–38. <https://doi.org/10.1111/j.1728-4465.2008.00148.x>
- Glick, Peter J., Alessandra Marini, and David E. Sahn. 2007. "Estimating the Consequences of Unintended Fertility for Child Health and Education in Romania: An Analysis Using Twins Data." *Oxford Bulletin of Economics and Statistics* 69(5):667–91. <https://doi.org/10.1111/j.1468-0084.2007.00476.x>
- Guo, Rufe, Junjian Yi, and Junsen Zhang. 2017. "Family Size, Birth Order, and Tests of the Quantity-Quality Model." *Journal of Comparative Economics* 45(2):219–24. <https://doi.org/10.1016/j.jce.2016.09.006>

- Gurunath, S., Z. Pandian, R. A. Anderson, and S. Bhattacharya. 2011. "Defining Infertility—a Systematic Review of Prevalence Studies." *Human Reproduction Update* 17(5):575–88. <https://doi.org/10.1093/humupd/dmr015>
- Hall, Jennifer A., Lorna Benton, Andrew Copas, and Judith Stephenson. 2017. "Pregnancy Intention and Pregnancy Outcome: Systematic Review and Meta-Analysis." *Maternal and Child Health Journal* 21(3):670–704. <https://doi.org/10.1007/s10995-016-2237-0>
- He, Xiaobo, and Rong Zhu. 2016. "Fertility and Female Labour Force Participation: Causal Evidence from Urban China." *The Manchester School* 84(5):664–74. <https://doi.org/10.1111/manc.12128>
- Heckman, James J. 2005. "The Scientific Model of Causality." *Sociological Methodology* 35(1):1–97. <https://doi.org/10.1111/j.0081-1750.2006.00164.x>
- Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *The Review of Economics and Statistics* 88(3):389–432. <https://doi.org/10.1162/rest.88.3.389>
- Henderson, Daniel J., Daniel L. Millimet, Christopher F. Parmeter, and Le Wang. 2008. "Fertility and the Health of Children: A Nonparametric Investigation." In: *Modelling and Evaluating Treatment Effects in Econometrics*, edited by Tom Fomby, R. Carter Hill, Daniel L. Millimet, Jeffrey A. Smith, and Edward J. Vytlacil, 167–95. Bingley: Emerald Group Publishing Limited.
- Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–60. <https://doi.org/10.1080/01621459.1986.10478354>
- Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.
- Jacobsen, Joyce P., James Wishart Pearce III, and Joshua L. Rosenbloom. 1994. "The Effects of Childbearing on Married Women's Labor Supply and Earnings: Using Twin Births as a Natural Experiment." *Journal of Human Resources* 34(3):449–74. <https://doi.org/10.2307/146376>
- Jayachandran, Seema, and Rohini Pande. 2017. "Why Are Indian Children So Short? The Role of Birth Order and Son Preference." *American Economic Review* 107(9):2600–29. <https://doi.org/10.1257/aer.20151282>

- Kozuki, Naoko, Anne C. C. Lee, Mariangela F. Silveira, Cesar G. Victora, Linda Adair, Jean Humphrey, Robert Ntozini, Robert E. Black, and Joanne Katz. 2013. "The Associations of Birth Intervals with Small-for-Gestational-Age, Preterm, and Neonatal and Infant Mortality: A Meta-Analysis." *BMC Public Health* 13(3):S3. <https://doi.org/10.1186/1471-2458-13-S3-S3>
- Mascarenhas, Maya N., Seth R. Flaxman, Ties Boerma, Sheryl Vanderpoel, and Gretchen A. Stevens. 2012. "National, Regional, and Global Trends in Infertility Prevalence Since 1990: A Systematic Analysis of 277 Health Surveys." *PLOS Medicine* 9(12):e1001356. <http://doi.org/10.1371/journal.pmed.1001356>
- Mogstad, Magne, and Matthew Wiswall. 2016. "Testing the Quantity–Quality Model of Fertility: Estimation Using Unrestricted Family Size Models." *Quantitative Economics* 7(1):157–92. <https://doi.org/10.3982/QE322>
- Myrskylä, Mikko, Karri Silventoinen, Aline Jelenkovic, Per Tynelius, and Finn Rasmussen. 2013. "The Association between Height and Birth Order: Evidence from 652 518 Swedish Men." *Journal of Epidemiology and Community Health* 67(7):571–77. <https://doi.org/10.1136/jech-2012-202296>
- Oliveira, Jaqueline. 2016a. "The Value of Children: Inter-Generational Support, Fertility, and Human Capital." *Journal of Development Economics* 120(Supplement C):1–16. <https://doi.org/10.1016/j.jdeveco.2015.12.002>
- Oliveira, Jaqueline. 2016b. "Fertility, Migration, and Maternal Wages: Evidence from Brazil." *Journal of Human Capital* 10(3):377–98. <https://doi.org/10.1086/687416>
- Pearl, J., M. Glymour, and N. P. Jewell. (2016), *Causal inference in statistics: A primer*. Chichester: Wiley Blackwell.
- Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. London, UK: Allen Lane.
- Philipov, Dimitar, Aart C. Liefbroer, and Jane E. Klobas (eds.). 2015. *Reproductive Decision-Making in a Macro-Micro Perspective*. Dordrecht: Springer Netherlands. <https://doi.org/10.1007/978-94-017-9401-5>
- Roberts, Evan, and John Robert Warren. 2017. "Family Structure and Childhood Anthropometry in Saint Paul, Minnesota in 1918." *The History of the Family* 22(2–3):258–90. <https://doi.org/10.1080/1081602X.2016.1224729>

- Rosenzweig, Mark R., and Kenneth I. Wolpin. 1980a. "Life-Cycle Labor Supply and Fertility: Causal Inferences from Household Models." *Journal of Political Economy* 88(2):328–48. <https://doi.org/10.1086/260868>
- Rosenzweig, Mark R., and Kenneth I. Wolpin. 1980b. "Testing the Quantity–Quality Fertility Model: The Use of Twins as a Natural Experiment." *Econometrica* 48(1):227–40. <https://doi.org/10.2307/1912026>
- Rosenzweig, Mark R., and Kenneth I. Wolpin. 2000. "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature* 38(4):827–74. <https://doi.org/10.1257/jel.38.4.827>
- Rosenzweig, Mark R., and Junsen Zhang. 2009. "Do Population Control Policies Induce More Human Capital Investment? Twins, Birth Weight and China's 'One-Child' Policy." *The Review of Economic Studies* 76(3):1149–74. <https://doi.org/10.1111/j.1467-937X.2009.00563.x>
- Rubin, Donald B. 1990. "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science* 5(4):472–80. <http://www.jstor.org/stable/2245383>
- Schoen, Robert, Nan Marie Astone, Young J. Kim, Constance A. Nathanson, and Jason M. Fields. 1999. "Do Fertility Intentions Affect Fertility Behavior?" *Journal of Marriage and Family* 61(3):790–99. <https://doi.org/10.2307/353578>
- Singh, Susheela, Gilda Sedge, and Rubina Hussain. 2010. "Unintended Pregnancy: Worldwide Levels, Trends, and Outcomes." *Studies in Family Planning* 41(4):241–50. <https://doi.org/10.1111/j.1728-4465.2010.00250.x>
- Small, Dylan S., Zhiqiang Tan, Roland R. Ramsahai, Scott A. Lorch, and M. Alan Brookhart. 2017. "Instrumental Variable Estimation with a Stochastic Monotonicity Assumption." *Statistical Science* 32(4):561–79. <https://doi.org/10.1214/17-STS623>
- Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65(3):557–86. <https://doi.org/10.2307/2171753>
- Stradford, Laura, Frans W. A. Van Poppel, and L. H. Lumey. 2017. "Can Resource Dilution Explain Differences in Height by Birth Order and Family Size? A Study of 389,287 Male Recruits in Twentieth-Century Netherlands." *The History of the Family* 22(2–3):214–35. <https://doi.org/10.1080/1081602X.2016.1230510>

Swanson, Sonja A., and Miguel A. Hernán. 2018. "The Challenging Interpretation of Instrumental Variable Estimates under Monotonicity." *International Journal of Epidemiology* 47(4):1289–97. <https://doi.org/10.1093/ije/dyx038>

ⁱ We are interested in testing if the instrument(s) add sufficient variation in the instrumented variable. The F -test should therefore be conducted of the instruments, i.e. not including any potential variables we adjust the first-stage model for.

ⁱⁱ Judea Pearl argues for the usefulness of graphical representations of models in the form of directed acyclical graphs (DAGs, see Pearl, Glymour, and Jewell 2016, or Pearl and Mackenzie 2018, for introductions). He and many others have also developed tools for estimating effects and testing the implications of the graphical models. I will not apply any of these tools here but, rather, evaluate twin birth IVs in the framework of potential outcomes (primarily following Angrist, Imbens, and Rubin 1996, and Imbens and Rubin 2015) because that is the framework that has been used in a number of influential publications for the twin birth IV literature (especially Angrist and Pischke 2008, and Angrist, Lavy and Schlosser 2010).