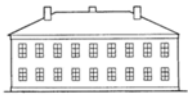
Göteborg Papers in Economic History



No. 23. April 2018 ISSN: 1653-1000

Instrumental variables based on twin births are by definition not valid

Stefan Öberg

DEPARTMENT OF ECONOMY AND SOCIETY

UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

UNIT FOR ECONOMIC HISTORY

Instrumental variables based on twin births are by definition not valid*

Stefan Öberg

stefan.oberg@econhist.gu.se

Abstract:

Instrumental variables based on twin births are a well-known and widespread method to find exogenous variation in the number of children when studying the effect on siblings or parents. This paper argues that there are serious problems with all versions of these instruments. Many of these problems have arisen because insufficient care has been given to defining the estimated causal effect. This paper discusses this definition and then applies the potential outcomes framework to reveal that instrumental variables based on twin birth violate the exclusion restriction, the independence assumption and one part of the stable unit treatment value assumption. These violations as well as the characteristics of the populations studied have contributed to hiding any true effect of the number of children. It is time to stop using these instrumental variables and to return to these important questions using other methods.

JEL: C21, C26, J13

Keywords: causal inference, natural experiments, local average treatment effect, complier average causal effect, Rubin's causal model, quantity–quality trade-off, family size

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

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

* The author gratefully acknowledges financial support from the Jan Wallanders and Tom Hedelius Foundation in the form of a Wallander Postdoc (W2014-0396:1). The author also thanks Niklas Vahlne for helpful comments and discussions.

1. Introduction

The number of children that parents desire is linked to other characteristics, such as their level of ambition in their careers, their lifestyle and the balance in their preferences for child "quality" and quantity. 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, characteristics of the parents that affect the life chances of both the children and the parents themselves. Therefore, 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. Different methods have been used to find exogenous variation in the number of children, with instrumental variables (IVs) based on twin births being a well-known and widespread solution (see Clarke (2017) for an overview of the literature).

IVs based on twin births have been considered a solution to problems of endogeneity because twin births occur at random and can therefore increase the number of children in the family exogenously, thereby creating a so-called natural experiment.[1] The randomness allows us to assume that parents who do and do not experience a twin birth have similar characteristics (in large samples). Importantly, we can assume that parents who do and do not experience a twin birth, on average, desire the same number of children.

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 studies since (Table 1). Rosenzweig and Wolpin 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. Their two 1980 publications used different specifications 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 good instruments.

[1] Or even a natural natural experiment because it is human biology that creates the situation (Rosenzweig and Wolpin (2000), p. 829).

Because the likelihood of experiencing any twin birth, as noted above, 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), Shen, Zou, and Liu (2017), Nguyen and Tran (2017), and de Jong, Smits, and Longwe (2017)) is clearly not a valid instrument. The positive association between the instrument and the 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)), will also not be valid.

Some studies have also 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.[2] If parents want even more children, then they will also proceed to have another birth after the twin birth. As shown below, the *net* association between the 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 challenge.

[2] Consequently, ending with a twin birth is approximately 1.8-1.9 times as likely as it is to experience a twin birth at a specific birth (see the results in the Appendix).

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. This issue will be discussed further below.

A number of studies have clarified and refined the method for disassociating the instrument from the desired number of children. 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 mostly excluded from the analysis because of their special characteristics (e.g., Silventoinen et al. (2013)). In these cases, families that had a twin as the first birth are therefore also excluded from the analysis.

The specification using parity-specific twin births was intimated in Rosenzweig and Wolpin (1980b), but it was further elaborated in Angrist and Evans (1998). The form of the specification was then found in Black, Devereux and Salvanes (2005), Angrist, Lavy and Schlosser ((2005), (2010)), and Cáceras-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 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 violates several necessary assumptions.

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., Moffitt (2005), Heckman (2010)). The strength of this method is its supposed high degree of internal validity (e.g., Imbens ((2010), (2014))). However, this internal validity has recently been called into question in a number of different studies (Bhalotra and Clarke (2016), Braakmann and Wildman (2016), Farbmacher, Guber, and Vikström (2018)). These studies show that violations of the necessary assumptions that are both plausible and mild lead to substantively important biases of the results. My paper contributes to this mounting critique and argues that there are even more serious issues related to the internal validity of IVs based on twin births. I argue that these IVs are by definition not valid and will produce ill-defined and biased results.

Some of these biases can be predicted to work against finding any influence of the number of children on the outcome. These biases can therefore be contributing 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, Lavy, and Schlosser (2010), Åslund and Grönqvist (2010), Marteleto and de Souza (2012), Ponczek and Souza (2012); Baranowska-Rataj, Barclay, and Kolk (2017).).

My discussion and critique of twin birth IVs are mostly based on arguments and definitions. I do not see any way to derive a formal proof that I am correct. Most likely, the reason is, in part, my lack of training: However, another reason why I make my argument mostly in verbal form is that this method does not work because of conceptual definitions. I think that previous applications have overlooked some of the problems with this method precisely 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 this method. 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. Below, I try to explain the frameworks and methods that I use. Therefore, the text will come across as basic to some readers. Please feel free to skim or skip ahead if you are one of those readers.

TABLE 1. A METHODOLOGICAL SUMMARY OF RECENT USES OF 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
Cáceras-Delpiano (2006)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Glick, Marani, 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
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
Dayioğlu, Kirdar, and Tansel (2009)	Share of twin births	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
Vere (2011)	Parity-specific twin births and *n*+ samples	Not only complete families	Mothers
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
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
Marteleto and de Souza (2012)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Cáceres-Delpiano and Simonsen (2012)	Parity-specific twin births and *n*+ samples	Not only complete families	Mothers

Ponczek and Souza (2012)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Holmlund, Rainer, and Siedler (2013)	Parity-specific twin births and *n*+ samples	Only complete families	Parents and children
Marteleto and de Souza (2013)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Kruk and Reinhold (2014)	Parity-specific twin births and *n*+ samples	Only complete families	Parents
Kolk (2015)	Parity-specific twin births and *n*+ samples	Only complete families	Children
Abdul-Razak, Abd Karim, and Abdul-Hakim (2015)	Parity-specific twin births and *n*+ samples	Not only complete families	Children
Braakmann and Wildman (2016)	Any twin birth	Not only complete families	Mothers
Baranowska-Rataj, de Luna, and Ivarsson (2016)	Parity-specific twin births and *n*+ samples	Only (or mostly) complete families	Children
Silles (2016)	Parity-specific twin births and *n*+ samples	Only complete families	Mothers
Mogstad and Wiswall (2016)	Parity-specific twin births and *n*+ samples	Only complete families	Children
Oliveira (2016a)	Twin as first birth	Only complete families	Parents and children
Oliveira (2016b)	Twin as first birth[2]	Not only complete families	Mothers
Baranowska-Rataj and Matysiak (2016)	Twin as first birth[3]	Not only complete families	Mothers
He and Zhu (2016)	Twin as first birth[4]	Not only complete families	Mothers
Shen, Zou, and Liu (2017)	Any twin birth	Only (or mostly) complete families	Children
Nguyen and Tran (2017)	Any twin birth	Not only complete families	Mothers
de Jong, Smits, and Longwe (2017)	Any twin birth[5]	Not only complete families	Mothers
Bonner and Sarkar (2017)	Being part of a twin birth	Not only complete families	Children
Brinch, Mogstad, and Wiswall (2017)	Parity-specific twin births and *n*+ samples	Only complete families	Children
Baranowska-Rataj, Barclay, and Kolk (2017)	Parity-specific twin births and *n*+ samples	Only (or mostly) complete families	Children
Zhang (2017)	Parity-specific twin births and *n*+1 samples	Not only complete families	Mothers
Arouri, Ben-Youssef, and Nguyen Viet (2017)	Twin as first birth	Not only complete families	Parents
Chen (2017)	Twin birth in first two parities	Not only complete families	Children
Dasgupta and Solomon (2018)	Twins among the younger siblings	Not only complete families	Children

Note: This summary is not an exhaustive overview of the literature.
[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.

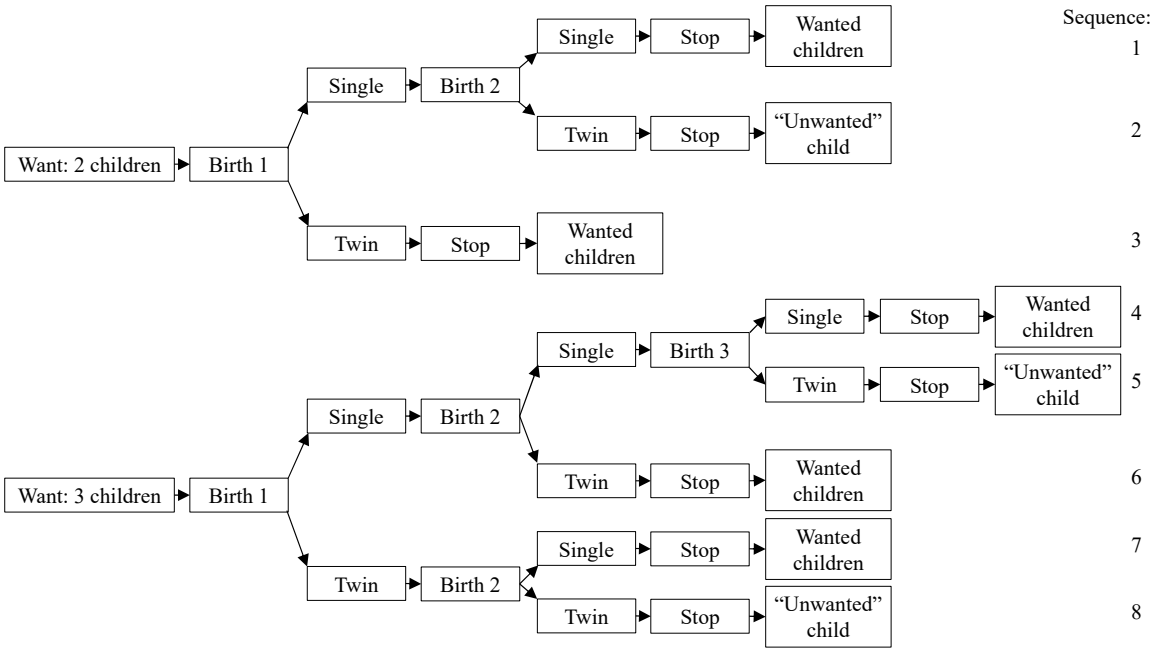
2. A definition of what we are studying

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, as noted above, 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. My aim here is to formulate a simplistic but logical 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 the 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 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.

FIGURE 1. A FLOWCHART OF THE PROCESS OF HAVING CHILDREN



I created a simulated population of parents with different birth sequences. To do so, I extended the model in Figure 1 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.

Each sequence has a probability of occurring. Sequence 1, for example, will have the probability $p_1 = \big(1 - p(twin)\big)\big(1 - p(twin)\big)$, with $p(twin)$ 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 family desires.

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), p. 232; see also, e.g., Black, Devereux, and Salvanes (2005), p. 681). This assumption might be unrealistic in real life, making it even more difficult to observe.

I used the complete set of different possible combinations of twin and single births as well as their respective probabilities to create the simulated population 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.3 I used these different distributions of realized numbers of children as proxies for hypothetical distributions of the desired numbers of children. This method 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.

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

[3] Sweden, people born 1972-1979: Åslund and Grönqvist (2010), the distribution is presented in Table 1; Norway, average year of birth 1962: Black, Devereux, and Salvanes (2005), Table II; Saint Paul, MN, USA, children aged 0-17 years in 1920: Roberts and Warren (2017), Table 3; The Netherlands, men born 1944-1947: Stradford, van Poppel, and Lumey (2017), Table 1.

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, Liefbroer, and Klobas (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 (Bhalotra and Clarke (2016); Braakmann and Wildman (2016); 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)).

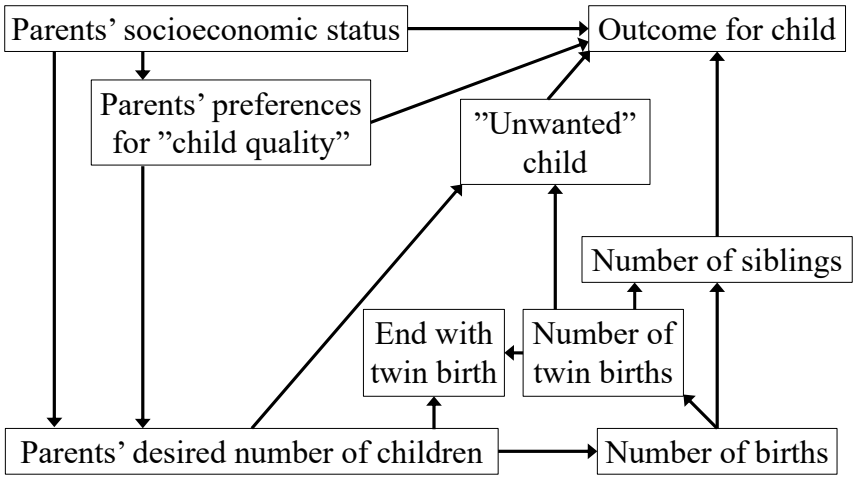
I still make these assumptions, despite knowing that they are not actually accurate, to preserve the clarity of the model and to avoid tangential issues. Realistic processes for determining the fertility preferences of parents and the birth of a child would change my deterministic, probability-weighted 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., p. 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 Bhalotra and Clarke (2016), Braakmann and Wildman (2016), and Farbmacher, Guber, and Vikström (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)).

I discuss twin birth IVs as applied to studies of how children are affected by their number of siblings. IVs based on twin births have been used as a solution to problems of endogeneity in a large number of studies investigating this topic (Table 1). Many of these studies 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 2 provides a graphical summary of an example of a model for investigating this issue.4 It shows how the estimated effect of the number of siblings will be biased through confounding from, in this case, the parents' socioeconomic

[4] 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) for an introduction). 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)).

status, the presence of "unwanted" children, and the parents' preferences for "child quality". Most often, we can adjust our estimates for the parents' socioeconomic status. However, we seldom 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 2 therefore shows one example of why and how the number of siblings becomes an endogenous variable in the model.

FIGURE 2. A GRAPHICAL SUMMARY OF A MODEL FOR INVESTIGATING HOW THE NUMBER OF SIBLINGS AFFECTS THE OUTCOME OF A CHILD



I use a simplified version of the model in Figure 2 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, Yi, and Zhang (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 want 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 and to include the possible presence of an "unwanted" child as a separate influence on 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 adjust 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". There will therefore be a measurement error in the empirical variable, which, in turn, will lead to residual confounding from the parents' socioeconomic status.[5]

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:

$$\boldsymbol{Y = \alpha_1 + \beta N + \mu_1 U + \delta_1 S + \rho_1 Q^*} \qquad (1)$$

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

[5] 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.

children, N^*, because this is one of the best predictors for the realized number of children (e.g., Schoen et al. (1999), Philipov, Liefbroer, and Klobas (2015), e.g., p. 168).[6] The following, then, is the model that is estimated in the first-stage regression:

$$\boldsymbol{N = \alpha_2 + \beta_{FS} Z + \delta_2 S + \eta_2 N^*} \quad (2a)$$

The preceding paragraphs have presented the definition of the research question, the model for investigating it and the data to be used. In the next section, I discuss the conceptual model that allows us to claim that the model estimates causal effects.

3. The counterfactual or potential outcomes framework for causal analyses

To estimate a causal effect, we must define it conceptually. There is a sometimes unrecognized distinction between defining and estimating the causal effect in which one is interested

(e.g., Holland (1986), Heckman (2005), p. 50, Imbens and Rubin (2015), chap. 1). There are different possible ways to discuss and define the causal effect that we want to estimate. I use the potential outcomes, or counterfactual, framework (Imbens and Rubin (2015), Morgan and Winship (2015)).

There is no single framework that is suitable for answering all types of scientific questions (e.g., Heckman (2005, 2010), Imbens (2010), Krieger and Davey Smith (2016)). How children and parents are affected by the number of children in the family is a substantive policy question with relevance for scientific theories that are of the type "effect of causes". They can therefore be successfully analyzed using the potential outcomes framework (Holland (1986), see also Heckman (2010), p. 361).

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

[6] I remind the reader that I use a highly stylized model of fertility behaviors allowing parents only to reach or surpass their desired number of children. Twin births can lead them to surpass their desired number of children, thus causing the birth of an "unwanted" child.

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 compare the outcome for the child after varying the treatment, that is, the number of children. 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 of interest (Imbens and Rubin (2015), p. 513). In the words of Angrist and Pischke ((2015), p. 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, 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 $Y_i\big(Z_i, W_i(Z_i)\big)$, 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, $W(Z)$.

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), p. 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 units that we choose to compare are truly comparable.

4. Estimating the causal effect using an instrumental variable

To be able to estimate any causal effect, we must make a number of assumptions (e.g., Holland (1986), Heckman (2005, 2010), Imbens and Rubin (2015)); to estimate causal effects using IVs, we must make assumptions regarding both the instrument(s) and the effect of interest.

What assumptions are needed depends on what we can assume about the model that we estimate. Angrist and Evans ((1998), p. 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).[7] The advantage of the LATE is that, given a set of assumptions and requirements, we can estimate a causal effect even if there is systematic sorting, i.e., self-selection, into treatment. (Heckman, Urzua, and Vytlacil (2006), p. 391, call this phenomenon "essential heterogeneity".) After Angrist and Pischke ((2008), p. 160–161) and Angrist, Lavy and Schlosser ((2005), (2010)) also suggested this interpretation in the twin birth IV case, it has been adopted by some other studies.[8]

Interpreting the estimated effect as a LATE enables us to allow for both heterogeneous treatment effects and variation in the effect of the instrument on the treatment. More importantly, we can also allow for systematic sorting into treatment. Doing so comes at the cost of requiring an additional assumption of the instrument, so-called monotonicity, and a stricter version of the exclusion restriction (see, e.g., Heckman, Urzua, and Vytlacil (2006), p. 391).

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

- relevant, i.e., have a substantial influence on the instrumented variable.[9]
- affecting the level of the instrumented variable in only one direction, i.e., monotonicity.[10]

---

[7] For introductions, see Imbens and Rubin (2015), chap. 23–24; Morgan and Winship (2015), chap. 9.

[8] Other studies interpreting the estimated effect as a LATE include Cáceres-Delpiano (2006, 2012b), Åslund and Grönqvist (2010), Cáceres-Delpiano and Simonsen (2012), Baranowska-Rataj, de Luna, and Ivarsson (2016), Baranowska-Rataj and Matysiak (2016), Braakmann and Wildman (2016), Silles (2016), and Brinch, Mogstad, and Wiswall (2017). The other studies in the literature have implicitly assumed homogenous treatment effects.

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

- randomly assigned, which is also called the assumption of independence.[11]
- affecting the outcome only through its effect on the treatment, the exclusion restriction.[12]

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.[13]

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

- the two parts of the stable unit treatment value assumption (SUTVAs I and II).[14]

When both the IV and the model fulfill all these necessary assumptions, we can estimate the causal effect, the LATE. The most common method of doing so is a two-stage least

---

[10] There is more agreement on the term monotonicity (e.g., Angrist, Imbens and Rubin (1996), p. 447, Angrist and Pischke (2008), p. 154, Henderson et al. (2008), p. 172, Imbens and Rubin (2015), p. 551, Swanson and Hernán (2017)), even if Heckman, Urzua and Vytlacil ((2006), pp. 391–392) suggest the term "uniformity".

[11] "Random assignment" (Angrist, Imbens and Rubin (1996), p. 446), "independence" (Angrist and Pischke (2008), p. 152), and "unconfounded type" (Henderson et al. (2008), p. 171).

[12] "Exclusion restriction" (Angrist, Imbens and Rubin (1996), p. 447, Angrist and Pischke (2008), p. 153) or "Mean independence within subpopulations" (Henderson et al. (2008), p. 171).

[13] As noted by Henderson et al. ((2008), p. 172).

[14] See, e.g., Cox ((1958), pp. 17–21), Rubin ((1990), p. 475), and Imbens and Rubin ((2015), pp. 9–12), see also Heckman ((2005), pp. 11–12, 35–36, 43). Small et al. ((2017), p. 562) also write about this issue such that there should be "no unrepresented versions of the IV". The first part of the SUTVA is sometimes discussed such that there should be no "equilibrium effects" (Heckman (2005), p. 11).

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 treatment, β_{FS}, and two, the second stage, in which we 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 also estimate the causal effect using the Wald estimator (Angrist and Pischke (2008), chap. 4, see also Imbens and Rubin (2015), chap. 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, $\hat{\beta}_{Wald} = \beta'_{\mathrm{SS}}/\beta_{\mathrm{FS}}$. 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, $\widehat{N}$, 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.

$$Y = \alpha_3 + \beta_{\mathrm{SS}}\widehat{N} + \mu_3 U + \delta_3 S + \rho_3 Q^* \qquad (3a)$$

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

quantity and "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:

$$N_i = \alpha_2 + \beta_{\mathrm{FS}} \mathrm{Z}_i + \delta_2 \mathrm{S}_i + \varepsilon_{2,i} \quad (2b)$$

$$\varepsilon_{2,i} = \mathrm{a}_2 + \eta_2 N_i^* + \mathrm{e}_{2,i} \quad (4)$$

$$Y_i = \alpha_3 + \beta_{\mathrm{SS}} \widehat{N}_i + \delta_3 S_i + \varepsilon_{3,i} \quad (3b)$$

$$\varepsilon_{3,i} = \mathrm{a}_3 + \rho_3 Q_i^* + \mu_3 U_i + \mathrm{e}_{3,i} \quad (5)$$

The error terms, $\varepsilon_{2,i}$ and $\varepsilon_{2,i}$, consist of both the unobserved factors and stochastic error terms, $\mathrm{e}_{2,i}$ and $\mathrm{e}_{3,i}$. 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 will be difficult to achieve if twin births actually increase the number of children exogenously in some families.

5. Evaluating IVs based on twin births

5.1. 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. 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)) or the children (Rosenzweig and Wolpin (1980b)). For this estimation to occur, 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. To use twin births as a source of exogenous variation in the number of children, it is therefore not enough that the twin births lead to an increase in the number of children born at a specific parity in families that experience a twin birth.

5.1.1. 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), p. 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".

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), p. 1143, Jacobsen, Pearce III, and Rosenbloom (1999), p. 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), p. 749–751fn13). 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.

5.1.2. First-stage coefficients

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 2 illustrate these influences on the first-stage coefficient.

The coefficients in Table 2 were estimated on my simulated population of families. As noted above, these families, are assumed to desire different numbers of children, with a relative distribution following the distributions in four different empirical populations. They then experience different combinations of twin and single births until they reach or surpass their desired number of children. The regressions are estimated across the 230 or 141

different combinations that are possible when using different versions of the IV. These 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 reducing the realized number of children for families with many children (that therefore also desired many children).[15]

The results in Table 2 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 *t*-values 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".[16] 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.[i] The increase in the size of the coefficient is exemplified in Table 2 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.

---

[15] I tried to simulate cutting some fertility histories short by reducing the realized number of children by more for families having (and therefore desiring) a larger number of children. I made the following changes in the realized number of children: 5→4; 6→5; 7→5; 8→6; 9→6; 10→7. Families with four children or fewer were left unchanged.

[16] For a definition of weak IVs and the problems they create, see, for example, Staiger and Stock (1997), Stock, Wright and Yogo (2002) and Murray (2006).

TABLE 2. 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}	t_{complete}	$R^2_{complete}$	$b_{\text{incomplete}}$	$t_{\text{incomplete}}$	bias (%)	
Black, Devereux, and Salvanes (2005)	Twin as first birth	0.080	0.1243	0.00007	0.172	0.3453	+53.5	230
	Twin as second birth	0.379	0.5015	0.00181	0.493	0.9354	+23.1	141
	Twin as last birth	0.543	1.1226	0.00550	0.625	1.6784	+13.1	230
	Share twin births	1.217	1.0226	0.00457	1.438	1.5704	+15.4	230
	Any twin birth	0.797	1.9050	0.01567	0.814	2.5341	+2.1	230
Åslund and Grönqvist (2010)	Twin as first birth	0.018	0.0284	0.000004	0.108	0.2287	+83.3	230
	Twin as second birth	0.402	0.5366	0.00207	0.506	0.9667	+20.6	141
	Twin as last birth	0.480	1.0404	0.00473	0.561	1.6205	+14.4	230
	Share twin births	1.053	0.9213	0.00371	1.275	1.4841	+17.4	230
	Any twin birth	0.735	1.8449	0.01471	0.746	2.4990	+1.5	230
Stradford, van Poppel, and Lumey (2017)	Twin as first birth	–0.109	–0.0951	0.00004	0.048	0.0675	——	230
	Twin as second birth	0.009	0.0066	0.0000003	0.177	0.2136	+94.9	141
	Twin as last birth	0.354	0.4250	0.00079	0.426	0.8303	+16.9	230
	Share twin births	0.866	0.3596	0.00057	1.239	0.8356	+30.1	230
	Any twin birth	1.226	2.1662	0.02017	0.881	2.5354	–39.2	230
Roberts and Warren (2017)	Twin as first birth	0.017	0.0177	0.000001	0.150	0.2245	+88.7	230
	Twin as second birth	0.167	0.1435	0.00015	0.327	0.4450	+48.9	141
	Twin as last birth	0.513	0.7054	0.00218	0.589	1.1810	+12.9	230
	Share twin births	1.314	0.6924	0.00210	1.575	1.2096	+16.6	230
	Any twin birth	1.214	2.1751	0.02033	1.010	2.6447	+20.2	230

Note: The bias was calculated as $100 \times \left(b_{incomplete} - b_{complete}\right)/b_{complete}$. This measure is not meaningful for the one case where the coefficient changes sign.

As shown in Table 2, 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), p. 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, 2015) have discussed the results in Angrist and Evans (1998) and Angrist, Lavy and Schlosser (2010). Their interpretation is the same as that which I have just presented.[ii]

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 (2016), p. 35). 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.

5.1.3. Monotonicity, only in one direction

The number of children will almost always be larger in families indicated by the twin birth IVs compared to those not indicated (Table 2). The average net effect will therefore be positive. 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), chap. 9). However, 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. In conclusion, it is not unlikely 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.

5.1.4. A binary number of children

We 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.[17] Therefore, the only variation in the number of

[17] Again, this is assuming that there are no other factors that we must include to adjust our first-stage regression. If we include other factors, then the predicted values for the

children that is used for the analyses is that families that experience a (parity-specific) twin birth, on average, have a larger number of children than families that, instead, have a single birth at the studied parity.

The 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 proposal 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 were such systematic differences, then it would be relevant to include an indicator for these families in the analytical model.

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.

Previous studies have discussed this type of violations of the exclusion restriction. Angrist, Lavy and Schlosser (2010), for example, 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" (p. 776, italics added). They proceed to discuss the reasons why (parity-specific) twin birth IVs are also associated with a larger number of children at higher parities (Angrist, Lacy and Schlosser (2010), p. 788fn15). Rosenzweig and Wolpin ((1980b), p. 234) also discuss how families that experience a twin birth are also affected in ways other than having an "extra" child being born at the studied parity (see also Angrist and Evans (1998), p. 473). These discussions are explanations of how

---

number of children are, as noted above, a linear combination of these factors plus the unique variation added by the instrument. This unique variation is still binary.

their IVs violate the exclusion restriction. Rosenzweig and Wolpin ((2000), p. 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". This assumption is also necessary when using twin birth IVs to study how children are affected by their number of siblings (Rosenzweig and Zhang (2009)).

If twin birth IVs are valid, then they should therefore make the treatment—the number of children—a binary variable in the empirical model. We therefore have a situation with a binary instrument and a binary treatment. Angrist, Imbens, and Rubin (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. I discuss twin birth IVs further below, assuming that both the instrument and the treatment are binary. 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.

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), p. 434). In the twin IV case, these are the parents who, for example, wanted two children but had three because of the twin birth. These are the only people whose treatment status, i.e.,

number of children, is affected by the instrument, i.e., the parity-specific twin birth. The treatment when using twin birth IVs is therefore essentially binary.

5.2. The definition of the treatment for twin birth IVs

The estimation of any causal effect is carried out by both conceptually defining it and empirically estimating it (Holland (1986), Heckman (2005), Imbens and Rubin (2015)). 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.[18] 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), p. 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), pp. 828–829).[19]

[18] The aim has always been to estimate the causal effect of exogenous variation in the number of children on the outcome (e.g., Bronars and Grogger (1994), p. 1142, Angrist and Evans (1998), p. 474, Jacobsen, Pearce III, and Rosenbloom (1999), p. 457, and Black, Devereux and Salvanes (2005), p. 670).

[19] "[T]he absence of models in the natural natural experiment literature [such as that using twin births as IVs] does not mean that there are no important and implausible

The ambiguity in the previous literature on the definition of the treatment 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), p. 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), p. 458). These statements correspond to treatments with subtle but important differences that have been overlooked in the literature. The first—and original—interpretation of the treatment is that twin births lead to an exogenous increase in the number of children in *some* families. In the words of Rosenzweig and Wolpin ((1980), p. 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 use any explicit potential outcomes framework 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), p. 338, (1980b), pp. 232–233).[20] 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, p. 452) is having "another" child, for example, a third child if we use a twin birth as second birth as

---

assumptions being implicitly used by the authors in interpreting the estimates they have obtained" (Rosenzweig and Wolpin (2000), p. 829).

[20] In practice—because of data limitations—Rosenzweig and Wolpin (1980b) use the share of twins among all births as the instrument. I do not understand the conceptual treatment or its causal effect in this specification, and it is not discussed in their paper.

the IV. This interpretation is different from having one more child at any parity but should be a generalizable effect at the studied parity.

The last alternative interpretation—as argued by, for example, Angrist and Pischke (2008) and Angrist, Lavy, and Schlosser (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, Lavy, and Schlosser (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, Lavy, and Schlosser (2010), p. 788).[21] 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 quite specific, and the causal effect of it is unlikely to be applicable to the number of children in general.

5.2.1. 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

[21] 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), p. 785). Bringing the desired number of children into the definition moves the interpretation in the direction of "unwanted" children.

1994; Angrist and Evans 1998; Jacobsen, Pearce III and Rosenbloom 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), p. 1142, Angrist and Evans (1998), p. 452, Jacobsen, Pearce III and Rosenbloom (1999), p. 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 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, Devereux and Salvanes (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, Devereux and Salvanes (2005), p. 681; see also, e.g., Black, Devereux and Salvanes (2010), p. 37, Cáceres-Delpiano and Simonsen (2012), p. 754). Angrist, Lavy, and Schlosser (2010) also investigate the effect on children while nonetheless arguing that the "extra" child born is the treatment. Even if the treatment and causal effect are undefined in most cases, there are therefore three different interpretations in the literature. I argue that neither of these interpretations leads to valid estimates of any causal effect. To explain why, I discuss the conceptual definition of the estimated effects and each of the assumptions necessary for estimating them.

5.2.2. A unit-level definition of the causal effect

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 situation 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. I examine the three definitions discussed 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. 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:

$$\beta_i = Y_i^{obs/mis}(Z_i = 1, W_i(Z_i = 1) = 1) - Y_i^{mis/obs}(Z_i = 0, W_i(Z_i = 0) = 0) \quad (6)$$

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.
- The family did not experience a twin birth at the studied parity and did not have an "unwanted" child.

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 did not experience a twin birth at the studied parity and did not have "another" child, e.g., only had two children.

By comparing these two situations, we estimate the effect of voluntarily or involuntarily having "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.

To move from this unit-specific definition to the population, we must ensure that the units we compare across are truly comparable. The framework presented by Angrist, Imbens, and Rubin (1996) provides tools to aid us in this evaluation. The framework presents the four types that arise from the (binary) assignment to treatment and the (binary) receipt of the assigned treatments. In the twin birth IV case, the implication is that 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. 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 3).

TABLE 3. 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 expect 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 likely 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 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 4).

The LATE is estimated by comparing the (conditional) expected values of the number of children and the outcome studied across the two columns of Table 4, 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.

TABLE 4. 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

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), p. 308fn26) provide a nice summary of the assumption allowing us to ignore them:

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

The observed differences across the columns will correspond to the differences between compliers provided that there are no systematic differences between the always-takers and never-takers who 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 definitions and formal statements of these assumptions are not always intuitive or helpful for thinking about whether the assumptions

are met, for example, when using twin birth IVs. The definitions of the assumptions that are based on the types are, at least to me, easier to comprehend and evaluate conceptually or empirically. To proceed with these evaluations, we must again discuss the definition of the treatment.

5.2.3. The perfect compliance interpretation, an "extra" child

Angrist, Lavy, and Schlosser ((2010), e.g., pp. 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), pp. 160–161) and has been adopted by some other researchers (e.g., Cáceres-Delpiano (2012a), p. 156, (2012b), p. 8, Cáceres-Delpiano and Simonsen (2012), p. 754, Baranowska-Rataj, de Luna and Ivarsson (2016), p. 1267, and Baranowska-Rataj and Matysiak (2016), p. 350).[22]

As noted above, it is somewhat unclear what Angrist, Lavy, and Schlosser (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. 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.[23] Therefore, we can also exclude the always-takers. We are left with only compliers, i.e., perfect compliance

[22] Åslund and Grönqvist (2010) also discussed compliance with the twin birth IV and, in contrast, found that "[t]he fact that the compliance rates are high is encouraging" (p. 134, italics added).

[23] It is indeed only through a twin birth at the studied parity that a family can have an "extra" child being born at the studied parity because of a twin birth.

(Table 5).[24] 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, Lavy and Schlosser (2010), p. 788). This attractive feature comes at the cost of us estimating a quite specific causal effect that is nonetheless not very generalizable.

TABLE 5. THE FOUR POSSIBLE COMBINATIONS OF ASSIGNMENT AND RECEIPT OF TREATMENT OF THE "EXTRA" CHILD BORN AT THE STUDIED PARITY BECAUSE OF THE TWIN BIRTH

		Twin birth at the studied parity?	
		No	Yes
"Extra" child born at the studied parity because of the twin birth?	No	Compliers	———
	Yes	———	Compliers

5.2.4. 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 6). In the (parity-specific) twin birth IV case, the overwhelming majority are not assigned to treatment, that is, do not experience a parity-specific twin birth. Nonetheless, there are many families that have, for example, a third child without having a twin as the second birth, that is, without being assigned to and complying with the treatment. Families that have "another" child without being assigned to have one through a twin birth are therefore not compliant with their assignment and are always-takers. In

[24] With this interpretation, the first-stage coefficient should therefore always be equal to one.

addition, because there are always-takers, compliance is not perfect. Instead, we have a situation with one-sided noncompliance (Imbens and Rubin (2015), chap. 23). With this definition of the treatment, the first-stage coefficient is the share of families that are compliers, as indicated by the instrument. This share will correspond to the share of families that have a third child because of experiencing a twin birth.

TABLE 6. THE FOUR POSSIBLE COMBINATIONS OF ASSIGNMENT AND RECEIPT OF THE TREATMENT OF HAVING "ANOTHER" CHILD

		Twin birth at the studied parity?	
		No	Yes
"Another" child?	No	Compliers	———
	Yes	Always-takers	Compliers and Always-takers

5.2.5. 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 7). The estimated LATE will therefore be no more generalizable than usual, and 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.

The compliers and never-takers (Table 7, Cell A) 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.

TABLE 7. THE FOUR POSSIBLE COMBINATIONS OF ASSIGNMENT AND RECEIPT OF THE TREATMENT OF AN "UNWANTED" CHILD

		Twin birth at the studied parity?	
		No	Yes
Having an "unwanted" child?	No	Cell A Compliers and Never-takers	Cell B Never-takers
	Yes	Cell C Always-takers	Cell D Compliers and Always-takers

The group in Cell B of Table 7—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.

The group in Cell C of Table 7—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, Sedge, and Hussain (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. 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 (Angrist, Imbens, and Rubin (1996), p. 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 C of Table 7 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.

The groups in Cell D of Table 7—compliers and always-takers—consist 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 D 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, their number of children, changed by the assignment to the treatment, here, the twin birth. These families are the families for which we estimate the LATE when using twin births as IVs for the number of children.

5.3. 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 indeed 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).

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), p. 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),
p. 160). However, recently, there have been some studies challenging this assumption by showing that even weak systematic influences on the likelihood of a twin birth have substantively important consequences for the results (Bhalotra and Clarke (2016), Braakmann and Wildman (2016), Farbmacher, Guber, and Vikström (2018)). While I think that these challenges are extremely important to consider for the literature applying this method, I ignore this problem in this paper and assume that twin births occur at random. The 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.

Twin births are not automatically unrelated to all relevant characteristics of the parents merely because they occur at random. Each birth is a possibility of a twin birth. There is therefore a positive association between the chance of experiencing a twin birth and the desired number of children. Such associations pose serious challenges to our analyses because the desired number of children, in turn, is associated with, for example, the parents' preferences for child "quality". Provided that we cannot adjust for these preferences, the fact that the IV is associated with the desired number of children is a serious problem.

Table 8 presents how different specifications of the twin birth IV are associated with the families' 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 correlations were calculated based on the simulated population presented above. 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 observation 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, Yi, and Zhang (2017)).

If twin births occur at random, then there should be no association between the type and the chance of experiencing a twin birth (Morgan and Winship (2015), pp. 311–312). 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), p. 171) call this phenomenon an "unconfounded type".

This is clearly not the case with twin birth IVs (Tables 9 and 10). The probabilities of the types vary between different specifications of the treatment and between populations. In some instances, the probabilities are almost balanced in the populations based on Black, Devereux, and Salvanes (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 8. CORRELATIONS BETWEEN THE TWIN BIRTH INSTRUMENT AND FAMILIES' 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	+0.008	–0.096
Twin as second birth	+0.000	+0.042	–0.067
Twin as last birth	+0.022	+0.074	–0.048
Share twin births	+0.031	+0.068	–0.074
Any twin birth	+0.093	+0.125	–0.014
Panel B. Distribution based on: Åslund and Grönqvist (2010)			
Twin as first birth	+0.000	+0.002	–0.106
Twin as second birth	+0.000	+0.045	–0.065
Twin as last birth	+0.015	+0.069	–0.062
Share twin births	+0.025	+0.061	–0.090
Any twin birth	+0.089	+0.121	–0.027
Panel C. Distribution based on: Stradford, van Poppel and Lumey (2017)			
Twin as first birth	+0.000	–0.006	–0.055
Twin as second birth	+0.000	+0.001	–0.049
Twin as last birth	+0.007	+0.028	–0.032
Share twin births	+0.019	+0.024	–0.063
Any twin birth	+0.139	+0.142	+0.048
Panel D. Distribution based on: Roberts and Warren (2017)			
Twin as first birth	+0.000	+0.001	–0.060
Twin as second birth	+0.000	+0.012	–0.051
Twin as last birth	+0.018	+0.047	–0.026
Share twin births	+0.029	+0.046	–0.047
Any twin birth	+0.133	+0.143	+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.

TABLE 9. THE PROBABILITY OF BELONGING TO A TYPE, TWIN AS SECOND BIRTH INSTRUMENT WITH AN "UNWANTED" CHILD AS THE TREATMENT

	The probability of belonging to a type within levels of the instrument					
	Indicated by instrument			Not indicated by instrument		
	Compliers	Never-takers	Always-takers	Compliers	Never-takers	Always-takers
Black, Devereux and Salvanes (2005)	0.11	0.11	0.77	0.34	0.34	0.32
Åslund and Grönqvist (2010)	0.11	0.10	0.79	0.34	0.34	0.32
Stradford, van Poppel and Lumey (2017)	0.10	0.43	0.46	0.34	0.33	0.33
Roberts and Warren (2017)	0.13	0.25	0.61	0.34	0.33	0.33

TABLE 10. THE PROBABILITY OF BELONGING TO A TYPE, TWIN AS SECOND BIRTH INSTRUMENT WITH "ANOTHER" CHILD, HERE, A THIRD CHILD, AS THE TREATMENT

	The probability of belonging to a type within levels of the instrument			
	Indicated by instrument		Not indicated by instrument	
	Compliers	Noncompliers	Compliers	Noncompliers
Black, Devereux and Salvanes (2005)	0.50	0.50	0.50	0.50
Åslund and Grönqvist (2010)	0.49	0.51	0.50	0.50
Stradford, van Poppel and Lumey (2017)	0.81	0.19	0.48	0.52
Roberts and Warren (2017)	0.66	0.34	0.49	0.51

5.4. 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), p. 468, Angrist and Pischke (2008), p. 153). The exclusion restriction requires that assignment to treatment not be associated with the outcome beyond changing the treatment received. The implication 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 (Equations 7 and 8). We need this implication to be true to be able to ignore the always-takers and never-takers.

Always-takers:

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

Never-takers:

$$Y_i(Z_i = 1, W_i(Z_i = 1) = 0) = Y_i(Z_i = 0, W_i(Z_i = 0) = 0) \quad (8)$$

The exclusion restriction can also be expressed as a stochastic restriction so that the distribution of potential outcomes is balanced between the two groups within each type. Below, I present the averages of the desired and realized number of children by type and value of the instrument based on my simulated population, as presented above. Again, the definition of the types depends on the definition of the treatment. I present the results for the definition of the treatment as "another", e.g., a third, child and as an "unwanted" child.

First, we analyze the exclusion restriction for when we use a twin as second birth instrument and define the treatment as having a third child. Observably, among the compliers, the families that are not indicated by the instrument both desired and have exactly two children in all four populations. The reason for this result is that the only way a family that can be a complier when it did not have a twin birth as the second birth is to desire and have exactly two children. Then, such families are not assigned to treatment, i.e., no twin birth, and are not treated, i.e., do not have a third child. The families that are indicated by the instrument have a larger realized number of children than the families that are not indicated. This difference in the realized number of children is therefore in the

direction that we want it to be. Disconcertingly, however, the differences are too large, with all being larger than one. The explanation for this result is that there are also substantive differences in the desired number of children. These differences constitute a clear indication of a violation of the exclusion restriction.

The pattern is the opposite among the always-takers. Among them, families that do not have a twin birth as the second birth both have and desire a larger number of children than do families that have twins. The size of the differences varies between the populations but is always substantial. Again, these differences indicate clear violations of the exclusion restriction.

From the first-stage coefficients in Table 2, families that experienced a twin birth, and that are therefore indicated by the instrument, have, on average, a larger number of children. This difference is a weighted average of the positive and negative differences for the compliers and always-takers, respectively.

TABLE 11. THE AVERAGE DESIRED AND REALIZED NUMBER OF CHILDREN AMONG COMPLIERS AND ALWAYS-TAKERS FOR A TWIN AS SECOND BIRTH INSTRUMENT WITH HAVING A THIRD CHILD AS THE TREATMENT

Compliers	The average desired number of children			The average realized number of children		
		Indicated by instrument?			Indicated by instrument?	
	Overall	Yes	No	Overall	Yes	No
Black, Devereux and Salvanes (2005)	2.03	2.77	2.00	2.04	3.27	2.00
Åslund and Grönqvist (2010)	2.02	2.74	2.00	2.04	3.26	2.00
Stradford, van Poppel and Lumey (2017)	2.23	4.61	2.00	2.25	4.80	2.00
Roberts and Warren (2017)	2.08	3.60	2.00	2.10	3.94	2.00
Noncompliers / Always-takers	The average desired number of children			The average realized number of children		
		Indicated by instrument?			Indicated by instrument?	
	Overall	Yes	No	Overall	Yes	No
Black, Devereux and Salvanes (2005)	3.50	2.77	3.53	3.76	3.27	3.77
Åslund and Grönqvist (2010)	3.49	2.74	3.52	3.73	3.26	3.75
Stradford, van Poppel and Lumey (2017)	5.17	4.61	5.18	5.39	4.80	5.41
Roberts and Warren (2017)	4.37	3.60	4.39	4.63	3.94	4.65

The pattern of results is very similar if we instead define the treatment as having an "unwanted" child (still using a twin as second birth instrument as our instrument). In this case, in addition to compliers, we have both always-takers and never-takers. The families indicated by the instrument have a larger number of children among the compliers and never-takers but not among the always-takers. More disturbingly, for all types we observe clear differences in the desired number of children among the families indicated by the instrument and not (Table 12). Again, these differences indicate clear violations of the exclusion restriction.[25]

TABLE 12. THE AVERAGE DESIRED AND REALIZED NUMBER OF CHILDREN AMONG COMPLIERS AND ALWAYS-TAKERS FOR A TWIN AS SECOND BIRTH INSTRUMENT WITH HAVING AN "UNWANTED" CHILD AS THE TREATMENT

	Compliers		Never-takers		Always-takers	
The average desired number of children	Indicated by instrument?		Indicated by instrument?		Indicated by instrument?	
	Yes	No	Yes	No	Yes	No
Black, Devereux and Salvanes (2005)	2.02	2.60	3.52	2.60	2.02	3.97
Åslund and Grönqvist (2010)	2.01	2.59	3.51	2.59	2.01	3.96
Stradford, van Poppel and Lumey (2017)	2.22	4.73	5.18	4.73	2.22	4.09
Roberts and Warren (2017)	2.07	3.51	4.38	3.51	2.07	4.03
	Compliers		Never-takers		Always-takers	
The average realized number of children	Indicated by instrument?		Indicated by instrument?		Indicated by instrument?	
	Yes	No	Yes	No	Yes	No
Black, Devereux and Salvanes (2005)	3.02	2.60	3.52	2.60	3.02	4.97
Åslund and Grönqvist (2010)	3.01	2.59	3.51	2.59	3.01	4.96
Stradford, van Poppel and Lumey (2017)	3.22	4.73	5.18	4.73	3.22	5.09
Roberts and Warren (2017)	3.07	3.51	4.38	3.51	3.07	5.03

[25] Because the twin birth IVs violate the exclusion restriction and, as we will observe, the SUTVA, "reduced form" estimates of the causal effect of a twin birth on the outcome (as in, for example, Baranowska-Rataj, Barclay and Kolk (2017)) will not produce valid results either.

The definition of the estimated effect for a population (following the notation of Imbens and Rubin (2015), p. 525) is as follows: (9)

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

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

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 noncompliers—always-takers—that are indicated by the instrument and not. The weights are defined by the shares of compliers, π_{co}, and noncompliers, π_{nc}, in the population. If the exclusion restriction holds, then we can assume that there are no differences between noncompliers that are or are not indicated by the instrument. This part of Equation 8 can therefore be ignored. However, if the exclusion restriction does not hold, then we can no longer estimate the LATE.

If we define having "another", i.e., a third child as the treatment, then the noncompliers that are not indicated by the instrument both desire and have a larger number of children. If there is a negative effect on child outcomes from either or both, then it will paradoxically lead to a positive bias in the estimated effect. The reason is that we estimate the effect that a twin birth has above and beyond the effect of a larger number of desired and realized children. If there is a positive effect of either or both the desired and realized number of children on child outcomes, then the bias will be negative. We can therefore know that the effect we estimate will be biased toward zero. If we instead define an "unwanted" child as the treatment, then we must expand the equation to include all four types (Imbens and Rubin (2015), p. 553). The idea is exactly the same, but predicting the direction of the bias becomes less straightforward because the patterns are opposite for always- and never-takers.

5.5. 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), pp. 9–12, see also Angrist, Imbens, and Rubin (1996), p. 446). The SUTVA includes two parts, SUTVAs I and II (Cox (1958), pp. 17–21, Rubin (1990), p. 475,

Heckman (2005), pp. 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), pp. 48–52).[26] 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), p. 11).[27] 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., $Y_i\big(Z_i, W_i(Z_i)\big) = Y_i\big(Z_i, W_i'(Z_i)\big)$.

If the effect of the treatment is not constant, then the estimated effect is "an artificial quantity" (Cox (1958), pp. 15–19). It does not necessarily correspond to the average treatment effect across the groups with the different treatment effects. Cox provides simple numerical examples illustrating that in the presence of varying treatment effects, the estimated average effect does not have to be in the range of the effects within subgroups of the population. An important reason why the effect of the treatment is not constant across units is when it depends on some (observable or unobservable) differences in the characteristics of the units (Cox (1958), p. 18). If we do not or cannot adjust our analyses

[26] The SUTVA is not mentioned in any of the papers applying twin birth IVs, and it is never evaluated. Angrist and Evans (1998), Angrist, Lavy and Achlosser (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 SUTVA (before Imbens and Rubin 2015).

[27] Small et al. ((2017), p. 567) write about this issue such that there should be "no unrepresented versions of the IV". This point leads to a very similar objection to twin birth IVs. The instrument—a parity-specific twin birth—will or will not lead to the birth of an unintended, i.e., "unwanted", child.

for these differences in characteristics, then the estimated effect cannot tell us anything about the true effect.[28]

This situation is what occurs in the twin birth IV case if we consider the extra child born at the studied parity because of a twin birth as the treatment. With this definition 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. 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 to the definition of the potential outcomes (Rubin (1990), p. 475):

$$Y_i(Z_i = 1, W_i(Z_i = 1) = 1, U_i = 0) \neq Y_i(Z_i = 1, W_i(Z_i = 1) = 1, U_i = 1) \quad (10)$$

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. Therefore, we cannot expect these potential outcomes to be the same. This part of the SUTVA is therefore not satisfied. One way to fix this problem would be to instead interpret the "unwanted" child as the treatment. However, as we have observed, doing so does not work well either because the exclusion restriction is violated.

6. Conclusions

IVs based on twin births are both a well-known and widespread solution for finding exogenous variation in the number of children in families. They have been used in many studies, with at least 17 new studies having been published in 2016 and 2017 alone. A worrying trend is that there is also an increase in the number of published studies that apply invalid versions of the twin birth instrument, such as an indicator of any twin birth. There is a need for researchers applying this method or evaluating others' applications of this method to raise the methodological standards. I think that one way of achieving these raised standards is to also interpret the methods that we use and the assumptions that we

[28] Small et al. (2017) have recently proposed a new type of causal effect that relaxes both the monotonicity assumption and the "no hidden variations of treatments" part of the SUTVA.

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.

Summarizing my findings, I think that the problems with twin birth IVs are serious enough that we should stop using this method. Useful interpretations of twin birth IVs lead to violations of the independence assumption, the exclusion restriction, and the SUTVA. Studies using twin births as IVs have estimated a casual rather than a causal effect of the number of children in the family.

The biases of this method are likely to have contributed to the current state of knowledge on these important questions. Both the common inclusion of incomplete families and the violation of the exclusion restriction will work to bias the estimated effect toward zero. The biases are therefore likely to contribute to the common pattern of results of a negative association but no causal effect when using a twin birth IV. It is time to stop using these IVs and to return to these important questions using other methods.

7. References

Abdul-Razak, N. A., M. Z. Abd Karim, and R. Abdul-Hakim. (2015), "Does Trade-Off Between Child Quantity and Child Quality Exist in Malaysia?" *Singapore Economic Review*, 60 (4). doi:10.1142/S0217590815500216

Alkema, L., K. Vladimira, M. Clare, and A. 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), 1642–1652. doi:10.1016/S0140-6736(12)62204-1

Angrist, J. D. (2004) "Treatment Effect Heterogeneity in Theory and Practice." *The Economic Journal*, 114 (494), C52–C83. doi:10.1111/j.0013-0133.2003.00195.x

Angrist, J. D. and W. N. Evans. (1998). "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *The American Economic Review*, 88 (3), 450–477. http://www.jstor.org/stable/116844

Angrist, J. D., G. W. Imbens, and D. B. Rubin. (1996), "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, 91 (434), 444–455. doi:10.2307/2291629

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

Angrist, J. D., V. Lavy, and A. Schlosser. (2010), "Multiple Experiments for the Causal Link between the Quantity and Quality of Children." *Journal of Labor Economics*, 28 (4), 773–824. doi:10.1086/653830

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

Angrist, J. D. and J.-S. Pischke. (2015), *Mastering 'Metrics: The Path from Cause to Effect*. Princeton, NJ: Princeton University Press.

Arouri, M., A. Ben Youssef, and C. Nguyen-Viet. (2017), "The More Children You Have the More Likely You Are to Smoke? Evidence from Vietnam." *Oxford Development Studies*, 45 (3), 260–275. doi:10.1080/13600818.2016.1193129

Åslund, O. and H. Grönqvist. (2010), "Family Size and Child Outcomes: Is There Really No Trade-off?" *Labour Economics*, 17 (1), 130–139. doi:10.1016/j.labeco.2009.05.003

Bachrach, C. A. and S. P. Morgan. (2013), "A Cognitive–Social Model of Fertility Intentions." *Population and Development Review*, 39 (3), 459–485. doi:10.1111/j.1728-4457.2013.00612.x

Baranowska-Rataj, A., K. J. Barclay, and M. Kolk. (2017), "The Effect of the Number of Siblings on Adult Mortality: Evidence from Swedish Registers." *Population Studies*, 71 (1), 43–63. doi:10.1080/00324728.2016.1260755

Baranowska-Rataj, A., X. de Luna, and A. 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. doi:10.4054/DemRes.2016.35.43

Baranowska-Rataj, A. and A. 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), 343–367. doi:10.1007/s12122-016-9231-6

Becker, G. S. and H. G. Lewis. (1973), "On the Interaction between the Quantity and Quality of Children." *Journal of Political Economy*, 81 (2, Part 2), S279–288. doi:10.1086/260166

Becker, G. S. and N. Tomes. (1976), "Child Endowments and the Quantity and Quality of Children." *Journal of Political Economy*, 84 (4, Part 2), S143–162. doi:10.1086/260536

Bhalotra, S. and D. Clarke (2016), The Twin Instrument. IZA DP No. 10405. http://ftp.iza.org/dp10405.pdf

Black, S. E., P. J. Devereux, and K. 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. doi:10.1093/qje/120.2.669

Black, S. E., P. J. Devereux, and K. G. Salvanes. (2010), "Small Family, Smart Family? Family Size and the IQ Scores of Young Men." *Journal of Human Resources*, 45 (1), 33–58. doi:10.3368/jhr.45.1.33

Bonner, S. and D. Sarkar. (2017), "The Quality-Quantity Trade-off among Australian Children." *Economic Modelling*, doi:10.1016/j.econmod.2017.08.010

Braakmann, N. and J. 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. doi:10.1111/rssa.12160

Brinch, C. N., M. Mogstad, and M. Wiswall. (2017), "Beyond LATE with a Discrete Instrument." *Journal of Political Economy*, 125 (4), 985–1039. doi:10.1086/692712.

Bronars, S. G. and J. Grogger. (1994), "The Economic Consequences of Unwed Motherhood: Using Twin Births as a Natural Experiment." *The American Economic Review*, 84 (5), 1141–1156. http://www.jstor.org/stable/2117765

Cáceres-Delpiano, J. (2006), "The Impacts of Family Size on Investment in Child Quality." *Journal of Human Resources*, 41 (4), 738–754. doi:10.2307/40057289

Cáceres-Delpiano, J. (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). doi:10.1515/1935-1682.2850

Cáceres-Delpiano, J. (2012b), "Can We Still Learn Something from the Relationship Between Fertility and Mother's Employment? Evidence from Developing Countries." *Demography*, 49 (1), 151–174. doi:10.1007/s13524-011-0076-6

Cáceres-Delpiano, J. and M. Simonsen. (2012), "The Toll of Fertility on Mothers' Wellbeing." *Journal of Health Economics*, 31 (5), 752–766. doi:10.1016/j.jhealeco.2012.05.006

Chen, Q. (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. doi:10.1016/j.ijedudev.2017.03.009

Clarke, D. (2017). "Children and Their Parents: A Review of Fertility and Causality." *Journal of Economic Surveys*. doi:10.1111/joes.12202

Conde-Agudelo, A., A. Rosas-Bermudez, F. Castaño, and M. 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. doi:10.1111/j.1728-4465.2012.00308.x

Cox, D. R. (1958), *Planning of Experiments*. New York: John Wiley.

Dasgupta, K. and K. 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. doi:10.1016/j.ehb.2018.01.004

Dayioğlu, M., M. G. Kirdar, and A. 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. doi:10.1111/j.1468-0084.2008.00540.x

de Chaisemartin, C. (2017), "Tolerating Defiance? Local Average Treatment Effects without Monotonicity." *Quantitative Economics*, 8 (2), 367–396. doi:10.3982/QE601

de Haan, M. (2010), "Birth Order, Family Size and Educational Attainment." *Economics of Education Review*, 29 (4), 576–588. doi:10.1016/j.econedurev.2009.10.012

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

Farbmacher, H., R. Guber, and J. Vikström. (2018), "Increasing the Credibility of the Twin Birth Instrument." *Journal of Applied Econometrics*. doi:10.1002/jae.2616

Fitzsimons, E. and B. Malde. (2014), "Empirically Probing the Quantity–Quality Model." *Journal of Population Economics*, 27 (1), 33–68. doi:10.1007/s00148-013-0474-8

Frenette, M. (2011), "How Does the Stork Delegate Work? Childbearing and the Gender Division of Paid and Unpaid Labour." *Journal of Population Economics*, 24 (3), 895–910. doi:10.1007/s00148-010-0307-y

Frenette, M. (2011), "Why Do Larger Families Reduce Parental Investments in Child Quality, but Not Child Quality per Se?" *Review of Economics of the Household*, 9 (4), 523–537. doi:10.1007/s11150-010-9115-0

Gipson, J. D., M. A. Koenig, and M. 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. doi:10.1111/j.1728-4465.2008.00148.x

Glick, P. J., A. Marini, and D. 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), 667–691. doi:10.1111/j.1468-0084.2007.00476.x

Guo, R., J. Yi, and J. Zhang. (2017), "Family Size, Birth Order, and Tests of the Quantity–Quality Model." *Journal of Comparative Economics*, 45 (2), 219–224. doi: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), 575–588. doi:10.1093/humupd/dmr015

Hall, J. A., L. Benton, A. Copas, and J. Stephenson. (2017), "Pregnancy Intention and Pregnancy Outcome: Systematic Review and Meta-Analysis." *Maternal and Child Health Journal*, 21 (3), 670–704. doi:10.1007/s10995-016-2237-0

Hatton, T. J. and R. M. Martin. (2010), "The Effects on Stature of Poverty, Family Size, and Birth Order: British Children in the 1930s." *Oxford Economic Papers*, 62 (1), 157–184. doi:10.1093/Oep/Gpp034

He, X. and R. Zhu. (2016), "Fertility and Female Labour Force Participation: Causal Evidence from Urban China." *The Manchester School*, 84 (5), 664–674. doi:10.1111/manc.12128

Heckman, J. J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." *The Journal of Human Resources*, 32 (3), 441–462. doi:10.2307/146178

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

Heckman, J. J., S. Urzua, and E. Vytlacil. (2006), "Understanding Instrumental Variables in Models with Essential Heterogeneity." *The Review of Economics and Statistics*, 88 (3), 389–432. doi:10.1162/rest.88.3.389

Heckman, J. J. (2010), "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy." *Journal of Economic Literature*, 48 (2), 356–398. doi:10.1257/jel.48.2.356

Henderson, D. J., D. L. Millimet, C. F. Parmeter, and L. Wang. (2008), "Fertility and the Health of Children: A Nonparametric Investigation." In: *Modelling and Evaluating Treatment Effects in Econometrics*, edited by T. Fomby, R. C. Hill, D. L. Millimet, J. A. Smith, and E. J. Vytlacil, 167–195. Bingley: Emerald Group Publishing Limited.

Holmlund, H., H. Rainer, and T. Siedler. "Meet the Parents? Family Size and the Geographic Proximity Between Adult Children and Older Mothers in Sweden." *Demography*, 50 (3), 903–931. doi:10.1007/s13524-012-0181-1

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

Imbens, G. W. (2014), "Instrumental Variables: An Econometrician's Perspective." *Statistical Science*, 29 (3): 323–358. doi:10.1214/14-STS480

Imbens, G. W. and J. D. Angrist. (1994), "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62 (2), 467–475. doi:10.2307/2951620

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

Jacobsen, J. P., J. W. Pearce III, and J. L. Rosenbloom. (1994), "The Effects of Childbearing on Married Women's Labor Supply and Earnings: Using Twin Births as a Natural Experiment." *The Journal of Human Resources*, 34 (3), 449–474. doi:10.2307/146376

Jayachandran, S. and R. Pande. (2017), "Why Are Indian Children So Short? The Role of Birth Order and Son Preference." *American Economic Review*, 107 (9), 2600–2629. doi:10.1257/aer.20151282

Kolk, M. (2015), "The Causal Effect of an Additional Sibling on Completed Fertility: An Estimation of Intergenerational Fertility Correlations by Looking at Siblings of Twins." *Demographic Research*, 32 (51), 1409–1420. doi:10.4054/DemRes.2015.32.51

Kozuki, N., A. C.C. Lee, M. F. Silveira, C. G. Victora, L. Adair, J. Humphrey, R. Ntozini, R. E. Black, and J. 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. doi:10.1186/1471-2458-13-S3-S3

Krieger, N. and G. D. 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. doi:10.1093/ije/dyw114

Kruk, K. E. and S. Reinhold. (2014), "The Effect of Children on Depression in Old Age." *Social Science & Medicine*, 100 (Supplement C), 1–11. doi:10.1016/j.socscimed.2013.09.003

Li, H., J. Zhang, and Y. Zhu. (2008), "The Quantity–Quality Trade-off of Children in a Developing Country: Identification Using Chinese Twins." *Demography*, 45 (1), 223–243. doi:10.1353/dem.2008.0006

Lu, Y. (2009), "Sibship Size and Education in South Africa: Black–White Variations." *Research in Social Stratification and Mobility*, 27 (2), 110–125. doi:10.1016/j.rssm.2009.04.002

Marteleto, L. J. and L. 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), 1453–1477. doi:10.1007/s13524-012-0118-8

Marteleto, L. J. and L. 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. doi:10.1093/sf/sot069

Mascarenhas, M. N., S. R. Flaxman, T. Boerma, S. Vanderpoel, and G. 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. doi:10.1371/journal.pmed.1001356

Moffitt, R. (2005), "Remarks on the Analysis of Causal Relationships in Population Research." *Demography*, 42 (1), 91–108. doi:10.1353/dem.2005.0006

Mogstad, M. and M. W. Wiswall. (2016), "Testing the Quantity–Quality Model of Fertility: Estimation Using Unrestricted Family Size Models." *Quantitative Economics*, 7 (1), 157–192. doi:10.3982/QE322

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

Murray, M. P. (2006), "Avoiding Invalid Instruments and Coping with Weak Instruments." *Journal of Economic Perspectives*, 20 (4), 111–132. doi:10.1257/089533006780387373

Myrskylä, M., K. Silventoinen, A. Jelenkovic, P. Tynelius, and F. Rasmussen. (2013), "The Association between Height and Birth Order: Evidence from 652 518 Swedish Men." *Journal of Epidemiology and Community Health*, 67 (7), 571–577. doi:10.1136/jech-2012-202296

Nguyen, C. and A. Tran. (2017), "The Effect of Having Children on Women's Marital Status: Evidence from Vietnam." *The Journal of Development Studies*, 53 (12), 2102–2117. doi:10.1080/00220388.2016.1269887

Oliveira, J. (2016), "The Value of Children: Inter-Generational Support, Fertility, and Human Capital." *Journal of Development Economics*, 120 (Supplement C), 1–16. doi:10.1016/j.jdeveco.2015.12.002

Oliveira, J. (2016), "Fertility, Migration, and Maternal Wages: Evidence from Brazil." *Journal of Human Capital*, 10 (3), 377–398. doi:10.1086/687416

Pearl, J., M. Glymour, and N. P. Jewell. (2016), *Causal inference in statistics: A primer*. Chichester: Wiley Blackwell.

Philipov, D., A. C. Liefbroer, and J. E. Klobas (eds.). (2015) *Reproductive Decision-Making in a Macro-Micro Perspective*. Dordrecht: Springer Netherlands. doi:10.1007/978-94-017-9401-5

Ponczek, V. and A. P. 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. doi:10.1353/jhr.2012.0006

Roberts, E. and J. R. Warren. (2017), "Family Structure and Childhood Anthropometry in Saint Paul, Minnesota in 1918." *The History of the Family*, 22 (2–3), 258–290. doi:10.1080/1081602X.2016.1224729

Rosenzweig, M. R. and K. I. Wolpin. (1980a), "Life-Cycle Labor Supply and Fertility: Causal Inferences from Household Models." *Journal of Political Economy*, 88 (2), 328–348. doi:10.1086/260868

Rosenzweig, M. R. and K. I. Wolpin. (1980b). "Testing the Quantity–Quality Fertility Model: The Use of Twins as a Natural Experiment." *Econometrica*, 48 (1), 227–240. doi:10.2307/1912026

Rosenzweig, M. R. and K. I. Wolpin. (2000), "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature*, 38 (4), 827–874. doi:10.1257/jel.38.4.827

Rosenzweig, M. R. and J. 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), 1149–1174. doi:10.1111/j.1467-937X.2009.00563.x

Schoen, R., N. M. Astone, Y. J. Kim, C. A. Nathanson, and J. M. Fields. (1999), "Do Fertility Intentions Affect Fertility Behavior?" *Journal of Marriage and Family*, 61 (3), 790–799. doi:10.2307/353578

Shen, G., J. Zou, and X. Liu. (2017), "Economies of Scale, Resource Dilution and Education Choice in Developing Countries: Evidence from Chinese Households." *China Economic Review*, 44, 138–153. doi:10.1016/j.chieco.2017.03.003

Silles, M. 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. doi:10.1093/oep/gpv055

Silventoinen, K., M. Myrskylä, P. Tynelius, Y. Yokoyama, and F. 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), 380–387. doi:10.1111/ppe.12054

Singh, S., G. Sedge, and R. Hussain. (2010), "Unintended Pregnancy: Worldwide Levels, Trends, and Outcomes." *Studies in Family Planning*, 41 (4), 241–250. doi:10.1111/j.1728-4465.2010.00250.x

Staiger, D. and J. H. Stock. (1997), "Instrumental Variables Regression with Weak Instruments." *Econometrica*, 65 (3), 557–586. doi:10.2307/2171753

Stock, J. H., J. H. Wright, and M. Yogo. (2002), "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments." *Journal of Business & Economic Statistics*, 20 (4), 518–529. doi:10.1198/073500102288618658

Stradford, L., F. 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), 214–235. doi:10.1080/1081602X.2016.1230510

Swanson, S. A. and M. A. Hernán. (2017), "The Challenging Interpretation of Instrumental Variable Estimates under Monotonicity." *International Journal of Epidemiology*. doi:10.1093/ije/dyx038

Vere, J. P. (2011), "Fertility and Parents' Labour Supply: New Evidence from US Census Data." *Oxford Economic Papers*, 63 (2), 211–231. doi:10.1093/oep/gpr003

Zhang, J. (2017), "A Dilemma of Fertility and Female Labor Supply: Identification Using Taiwanese Twins." *China Economic Review*, 43, 47–63. doi:10.1016/j.chieco.2016.12.005

[i] 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" (p. 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. Oliveira (2016a) studies the effect of the number of children on families in China. She uses data from a survey with retrospective fertility histories, and therefore, the data include only complete families. Nonetheless, she finds a surprisingly large and significant difference 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. This policy was 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 third child. All other studies using twins as the first birth use data on incomplete families (Bronars and Grogger 1994; Jacobsen, Pearce III and Rosenbloom 1999; Glick, Marani and Shan 2007; Oliveira 2016b; Baranowska Rataj and Matysiak 2016; He and Zhu 2016; Arouri, Ben-Youssef and Nguyen Viet 2017). The timing effect will therefore contribute to artificially strengthening the twin as first birth instrument in these applications.

[ii] See Angrist and Pischke ((2008), p. 132) on the results in Angrist and Evans (1998):

> "The twins instrument in Angrist and Evans (1998) is a dummy for a multiple second birth in a sample of mothers with at least two children. The twins first-stage is .625, ... This means that 37.5 percent of the mothers with two or more children would have had a third birth anyway; a multiple third birth increases this proportion to 1".

See Angrist and Pischke ((2015), p. 128) on the results in Angrist, Lavy, and Schlosser (2010):

> "A second twin birth, however, increases average family size by .32, that is, by about one-third of a child. Why do twin births increase family size by a Salomonic fractional child? Many Israeli parents would like three or four children; their family size is largely unaffected by the occurrence of a multiple twin birth [sic], since they were going to have more than two children either way. On the other hand, some families are happy with only two children. The latter group is forced to increase family size from two to three when the stork delivers twins. The one-third-of-a-child twins differential in family size reflects a difference in probabilities: the likelihood of having a third child increases from about .7 with a singleton second birth to a certainty when the second birth is multiple".

Göteborg Papers in Economic History

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

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

18. Tobias Karlsson and Christer Lundh: The Gothenburg Population Panel 1915–1943. GOPP Version 6.0. 2015
19. Tobias Karlsson: Pushed into unemployment, pulled into retirement: Facing old age in Gothenburg, 1923–1943. 2015
20. Stefan Öberg and Klas Rönnbäck: Mortality among European settlers in pre-colonial West Africa: The "White Man's Grave" revisited. 2016
21. Dimitrios Theodoridis: The ecological footprint of early-modern commodities. Coefficients of land use per unit of product. 2017
22. Stefan Öberg: An introduction to using twin births as instrumental variables for sibship size. 2017
23. Stefan Öberg: Instrumental variables based on twin births are by definition not valid. 2018