Essays on Fertility and Family Size

A DISSERTATION

Presented to the Faculty of Economics of the University of Oxford.

Damian Clarke
July, 2015
58,127 words

Abstract

In these papers I discuss the causal estimation of the effects of fertility and fertility planning developments on mother and child outcomes. A number of concerns are raised with existing identification techniques, and alternative methodologies to consistently estimate the effect of interest are proposed. These concerns and new techniques are illustrated using microdata on slightly more than 43,000,000 births ocurring between 1972 and 2013.

In the first substantive chapter (written with Sonia Bhalotra), we discuss the validity of the use of twin births in fertility research. We demonstrate that twin births are not random. Successfully taking twins to term depends upon positive maternal health behaviours and investments in the periods preceding birth. We show that this is of considerable concern for estimation techniques which rely on twin births being (conditionally) randomly assigned to identify causal effects. To illustrate, we consider the estimation of the child quantity–quality (QQ) trade-off, and show that existing instrumental variable estimates are inconsistent in the contexts examined. Upon partially correcting for the fact that twin births are not random, a statistically significant QQ trade-off begins to emerge. We close by examining a number of partial identification techniques to bound the true effect of fertility on child outcomes.

In the second substantive chapter, I examine the effect of fertility control policies on the fertility decisions and outcomes of women. I consider the case of the emergency contraceptive pill in Chile. The staggered arrival of this technology to Chile over the last decade has resulted in the availability of the first safe and legal post-coital birth control policies. In a context of high teenage pregnancy rates, difference-in-difference (DD) style estimates suggest that this policy has accounted for reductions in short-term teen childbearing by as much as 7%, an effect similar to the arrival of abortion in the USA. This policy is also shown to reduce fetal deaths reported in early gestation with no similar reduction in late gestation: suggestive evidence that an alternative fertility control policy may reduce costly and dangerous illegal abortions.

Finally, I turn to the use of DD estimators as a policy-analysis tool. I discuss how such estimators perform in the case of reforms which may not be sharply demarcated to treatment and control clusters, but rather subject to local spillovers or externalities. I propose an extension of the typical DD estimator: a spillover-robust DD estimator. This methodology is applied to estimate the effect of two localised fertility control reforms in Mexico and Chile, where women close to treatment clusters who were not themselves subject to the reform may nonetheless travel to access treatment.

Contents

A Note on Authorship

Chapters 1, 3 and 4 were written alone. Chapter 2 was coauthored with Professor Sonia Bhalotra, to whom I am indebeted for her willingness to share knowledge. The work is indeed a co-authored project, which has evolved over a number of years. Both authors have made substantive contributions to the ideas, derivations and empirical results presented in the paper.

A Note on Reproducability

In the interests of openness, code to replicate the entirety of this thesis is publicly available for inspection at http://github.com/damiancclarke/PhD. This code allows any and all interested parties to fully generate any result from the thesis. All code is released under the GNU General Public License, implying that it is free software. The thesis can be replicated from source (including the download of publicly available data) on any machine, provided that a number of languages or compilers are installed. These requirements are (the free) Python, R, Octave, Fortran, git and LaTeX, and the (non-free) Stata. Replicating the entire analysis is relatively easy. On a Unix machine:

```
$ git clone https://github.com/damiancclarke/PhD.git
$ make
```

runs the entirety of the analysis from source. Full details are provided at:
https://github.com/damiancclarke/PhD/wiki.

Introduction

Fertility decisions are fundamental to human welfare. As animals we are driven to reproduce, while a deep desire for the well-being of our offspring promotes investments to maximise their chances of success. These behaviours play out continually over our lives: first as the subjects of fertility choices, and later as their architects.

This thesis examines fertility choices and their effects on human well-being. It joins an enormous range of literature, both academic and non-academic, and from economics as well as other fields. Particularly, this thesis is a collection of papers which discuss how to estimate the causal effects of fertility decisions on the individuals and families which make them. The estimation of causal effects in an area of life (child-bearing) that is related to and affects many other areas of life is not trivial. However, in the existing economic literature a number of techniques have been developed to circumnavegate the difficulties involved in the estimation of causal links between child-bearing and other human outcomes. In the chapters contained in this thesis I provide extensions and caveats to these techniques.

The body of this work consists of four chapters. Three of these are central to the thesis and contain new empirical and theoretical considerations and results. The fourth provides a survey of the state of the art, along with a much more precise definition of the terminology and field in which this thesis is located. This overview chapter is the first chapter, and a useful starting point for a reader interested in quickly gaining an understanding of the methodologies which are commonly employed to estimate the causal effects of fertility, and the empirical results in the literature.

The research contribution of this thesis begins with chapter 2. This chapter focuses on twins, and their use as an instrumental variable. The prevailing logic is that twin births are (at least conditionally) randomly assigned to women who conceive. If this is the case, twins are the perfect example of an instrumental variable: they have strong effects on completed fertility, while being

unrelated to all other parental and family characteristics. This has given rise to a number of influential papers using twin births to estimate the effect of fertility at the extensive margin on women's and men's labour supply, as well as sibling quality.

Chapter 2 documents that the assumption of conditional exogeneity of twin birth does not hold in the many cases examined. It is argued (and demonstrated) that healthier women are more likely to give birth to twins. It is shown that conditional on twin conception (which may or may not be random), healthier mothers are more likely to take twins to term. This is shown to be true prior to the introduction of assisted reproductive technologies like IVF, and to hold in a large number of contexts, including administrative data from the USA, UK, and Spain, and survey data from Africa, Asia and Latin America.

The nature of this violation of exogeneity is problematic. We document that healthy mothers are more likely to give live birth to twins. Maternal health is very hard (or impossible) to fully capture in data. In various contexts we show that health as proxied by height, alcohol and drug consumption prior to birth, maternal stress, smoking, and access to prenatal care all predict twinning. It seems unlikely that all such measures (and the many which we don't observe) could ever be fully observable in data. What's more, these measures are very likely to be correlated with unobserved factors which predict the very outcomes we are interested in analysing (labour force participation or child quality).

In order to illustrate the problems involved in this methodology we turn to the quantity-quality (QQ) trade-off. Recent empirical analysis has suggested that an additional birth within a family may have very little causal impact on existing children in a family. As we demonstrate, any two-stage least squares estimates which ignore the role of maternal health in twinning are likely to attenuate estimates of the QQ trade-off. We partially correct for this bias, and show that upon partial correction, a significant trade-off does begin to emerge in data from the developing world, as well as in data from the USA.

The estimates we produce are at best lower bounds of the trade-off. In closing the chapter we turn to ways to tighten the bounds on the estimate of the effect of interest. We use partial identification techniques to bound both IV as well as OLS estimates, and then use our IV and OLS estimates (which are likely to suffer different biases) to find the range in which the QQ trade-off lies.

Chapter 3 turns to the effect of contraceptive technologies on women's outcomes. I focus on the effect of the emergency contraceptive pill in Chile. This is a particularly interesting context

given the high rates of adolescent (and unplanned) pregnancy, and the absence of any other legal post-coital birth control technologies. The arrival of the emergency contraceptive pill heralded a large shift in the ability to post-coitally contracept, which previously was illegal and risky.

The few existing studies of the effects of the emergency contraceptive pill (which have been conducted in the USA) on fertility suggest that its role may be minor when considering rates of pregnancy. However, these studies are conducted in considerably different circumstances: namely, where free or cheap options to interrupt pregnancy already existed (via surgical abortion). The evidence from Chile, however, suggests that in the absence of alternative post-coital technologies, the emergency contraceptive pill can act as an important technology in reducing rates of pregnancy in the short term.

In order to identify the effect, I focus on the staggered arrival of the emergency contraceptive pill to the country. Following juridical and constitutional rulings in 2008, the decision of whether or not to freely distribute the emergency contraceptive pill in municipal health centers was left in the hands of the mayor of each municipality. This interim legislation (which existed from 2008 until 2011) resulted in a situation in which approximately half of the municipalities stated that they would prescribe the pill, while the remainder did not. In this period, over 15,000 pills were freely prescribed and issued to women.

I document the effect of this reform on rates of child-bearing using a flexible difference-in-differences (DD) methodology. Naïve DD estimates suggest a 4.5% reduction in rates of teen birth, and a 3% reduction for women aged between 20-34. I also document the effects of the reform on rates of fetal death. DD estimates suggest that the reform had important effects on rates of early term fetal deaths reported by hospitals, but no similar effect on late term fetal deaths. Although a reduction in fetal deaths in morning after pill municipalities may be an indicator of many things, the fact that this reduction *only* occurs early in gestation is suggestive of a reduction in illegal and risky abortions.

In closing the chapter I examine the possibility that women may have crossed municipal boundaries (defying their treatment status) to access the emergency contraceptive pill. I demonstrate that there is evidence of reform 'spillovers' which may travel up to 30km from the nearest treatment municipality, and which have important impacts on estimated treatment effects.

I return to the idea of spillovers much more extensively in chapter 4—the final content chapter of the thesis. Frequently in the economic literature, policy analysis is undertaken using a DD methodology. This is particularly the case where a reform arrives to certain areas or groups

within a country but not to others, allowing for the clear classification of individuals as treatment or control based on their area of residence. This methodology relies explicitly on the stable unit treatment value assumption (SUTVA), which, among other things, implies that no spillovers can occur between treatment and control areas.

This is at odds with the incentives put in place in many policy reforms. Where policies are geographically (or otherwise) bounded, and where the policy offers clear benefits to those in treatment regions, nearby individuals will have the incentive to access treatment, violating their status, and the SUTVA. I propose a methodology to estimate causal treatment effects in the absence of SUTVA. Rather, I outline a weaker set of conditions, where this assumption need only hold between *some* units. I derive a full set of conditions under which the average treatment effect on the treated (ATT) and the average treatment effect on the close to treatment (ATC) can be estimated in the presence of local spillovers.

This spillover-robust DD methodology is demonstrated by examining two contraceptive reforms. Contraceptive reform is a perfect example of such a situation given that the costs of having an undesired pregnancy are very high, the arrival of new contraceptive technology to certain areas is often slowed by legal challenges, and violation of treatment status is possible if individuals from non-covered areas can travel and convince health care providers in covered areas to prescribe treatment. This is shown with application to the 2007 abortion reform in Mexico DF, and the aforementioned emergency contraceptive pill reform in Chile in 2008.

In both cases of sub-national reform it is shown not only that the policy reduced rates of teenage pregnancy treatment areas, but also that reductions occured in areas 'close to' treatment. In the case of Mexico DF (where abortion was only available in one geographical area of the country), the existence of spillovers is limited only to areas local to the state in question. These spillovers are shown to *not* have significant effects on traditional DD estimates. In the case of Chile, however, treatment was geographically disperse. Here, like Mexico, evidence of spillovers is seen, but in this case spillovers reach a wider range of untreated areas. Unlike Mexico, once these spillovers are properly taken into account, the estimated treatment effect is *larger* than that estimated when SUTVA is assumed to hold.

Overall then, this thesis seeks to push forward the applied microeconometric literature on the estimation of fertility and fertility control. As more and more data on babies and long-term outcomes becomes available, significantly more precise effects of fertility on human outcomes can be estimated. However, as this thesis documents, even with considerable amounts of information and carefully designed estimation strategies, the assumptions underlying our methodologies

should be tested demandingly. Fertility decisions are complex, and economic estimates must make account for this complexity when purporting to report causal estimates for policy and academics.

Chapter 1

Causality and Fertility

Identification with Endogenous Decisions

Chapter Abstract

Child-bearing decisions are not made in isolation. They are taken in concert with decisions regarding work, marriage, health investments and stocks, as well as many other observable and non-observable considerations. Drawing causal inferences regarding the effect of child birth is complicated by these endogenous factors. This review chapter outlines the identification of causal estimates of fertilty, and the assumptions underlying the range of estimators and methodologies proposed in the economic, as well as the non-economic, literature.

1.1 Introduction

Human decisions regarding births, and how these decisions affect individual outcomes, are central to human welfare. They are also widely relevant to the world population. In 2014, 18.7 per every 1,000 people had a child, which is the equivalent of 4.3 births per second (CIA, 2014). Over the course of her lifetime, the average woman in 2013 will have 2.46 births, down from 4.98 in 1960 (The World Bank, 2015). The importance of choice, and control over the timing and number of children has been documented as early as $c1800$ BC, with discussions of a range of contraceptive methods included in the Kahun Gynaecological Papyrus (O'Dowd and Philipp, 1994). Estimates suggest that between then and today, contraceptive use has grown to reach global coverage of 60% of all married, fertile-aged women (Darroch, 2013). This review chapter examines the effect of childbirth and birth timing decisions on human outcomes.

I review the literature on the effect of individual fertility behaviours on individual outcomes: namely, the microeconomic effects of fertility. This chapter, and indeed the literature on which it is based, largely focuses on the effect of a woman or family's fertility decisions on the mother's life outcomes and the outcomes of her children. Some, although relatively less, focus is paid to the effect on her partner (if existing and present).

Rather than focus on correlations between fertility and other outcomes, this article is centered on the *causal* analysis of the effects of fertility. I discuss the theoretical and empirical considerations required to infer causality in a behaviour which cannot be manipulated directly in an experimental context. Given the interruptive nature of child birth on a large range of other life outcomes, any study of causal effects must isolate changes in fertility from corresponding changes in simultaneously determined, or dynamically dependent, outcomes. As a simple example, if women jointly choose to exit the labour market and have a child, any inference regarding the effect of fertility on her or her child's *other* outcomes must be independent of her labour market choice.

Since the boom in fertility related research in microeconomics in the mid-1970's,[1] a range of methodologies have been proposed to permit inference in precisely these circumstances. These include the use of instrumental variables (IV), combining difference-in-difference (DD) with IV, the use of quasi-experimental fertility shocks, or trying to artificially construct a treatment and

[1] From 1970 onwards, the frequency of the occurrence of the word "fertility" in titles of all articles published in the *Journal of Political Economy* is 46.5 per 100,000 words. Prior to the 1970s, the frequency was 1.45 per 100,000 words. Though admittedly based on a small sample, the frequency by decade is 11.17 ('60s), 72.72 ('70s), 66.33 ('80s), 29.41 ('90s), and 41.07 ('00s). It was not mentioned in titles of articles published between 1863 and the 1960s.

control group using family members or other matching methods. In the sections which follow, I describe these methodologies, the papers which propose and estimate them, and the identifying assumptions that motivate causality in each case.

Empirical considerations regarding the causal effects of fertility require the consideration of (at least) three questions: 'Effects on what?', 'Effects at what margin?', and 'Effects of what? (timing or quantity)'. Frequently, these questions are split further, and examined for specific groups of women or children.

Regarding the first question, the theoretical and empirical microeconomic literature has hypothesised that marginal births may have causal implications for many individual-level outcomes. This includes effects on children: their health indicators, cognitive and non-cognitive achievements, long-term labour market outcomes, education inputs, and social outcomes such as age at marriage and crime incidence; as well as effects on parents. Parental outcomes often considered are centered around labour market participation and returns, rates of education completion, marriage market outcomes and socioeconomic indicators such as welfare-receipt. I outline these measures in the following sections, along with the evidence that a causal link exists (or does not exist) with fertility.

Human fertility decisions exist at two (very different) margins. The choice of whether or not to have a child (the extensive margin), and, conditional on choosing to have any children, the decision regarding *how many* births to have (the intensive margin). The nature of links between fertility decisions and outcomes vary considerably when considering the intensive and the extensive margin. The consensus in the literature is that the causal effects are certainly not a linear function of births, with important non-linearities, and indeed non-monotonic relationships, described between fertility and (some) of the previously mentioned outcomes. In order to quantify effects at different margins, a range of estimation samples and methodologies need be employed.

Finally, causal effects of fertility are not independent of mother's age at birth. Often, rather than estimating the effect of a marginal birth, we will be interested in determining the effect of child birth at a particular age (such as during adolescence). Considerations of these effects have important life-cycle implications for future investment decisions.

The study of fertility is common to a huge range of fields: social sciences, physical sciences, demography and medicine, and in many subfields within disciplines. As a result, any moderate-length review of the literature must be necessarily pointed. As such, this chapter firmly focuses

on the effects of fertility on other individual-level outcomes, and not the determinants of fertility[2], the macroeconomic effects (Enke, 1966, 1971), or discussions of the broader effects of population control policies (Miller, 2010; Rosenzweig and Wolpin, 1986).

In what remains of this chapter I briefly layout a number of considerations surrounding causal analysis of endogenous fertility decisions, before turning to particular methodologies and samples used to generate empirical estimates. I consider the causal effects of fertility on child and on parental outcomes in turn, outlining main results from relevant papers in the field.

1.2 Causality and Fertility

Consider an outcome Y_i, for each member i of a sample, $i \in \{1, \ldots, N\}$, where Y denotes an outcome variable of interest, and the sample is drawn from a population of child-bearing mothers (or, as discussed later, their children). We are interested in determining the effect of manipulations of fertility, which we denote F_i, on our outcome variable of interest Y_i. It is assumed that Y_i is a function of fertility, an unobserved variable U_i, and a series of other variables which are summarised as the error term ε_Y:

$$Y_i = f_Y(F_i, U_i, \varepsilon_Y). \tag{1.1}$$

Fertility is assumed to be a function of the unobserved U_i, and stochastic ε_F:

$$F_i = f_F(U_i, \varepsilon_F), \tag{1.2}$$

and finally $U_i = f_U(\varepsilon_U)$. These error terms $\boldsymbol{\varepsilon}$ are assumed mutually independent. To fix ideas, we could consider an outcome variable Y_i as average years of education of i's children, F_i as completed fertility, and U_i as unobserved positive health behaviours of the mother. By iterative substitution of the $\boldsymbol{\varepsilon}$ terms into (1.1) and (1.2) it becomes apparent (in the defined system of equations) that changes in fertility and health are unrelated to ε_Y, but that both average years of education and fertility are related to unobserved maternal health behaviours.[3]

[2] More recently, Kearney and Levine (2012) provide an extensive discussion of teenage childbearing and its determinantes in the USA, Bailey (2013) reviews the old and new evidence on the effect of access to contraception, and Moffitt (2005) has provided a discussion and application of causal inference to the effects of teenage childbearing on child outcomes. All provide extremely useful reviews of the relevant literature. The handbook chapter of Schultz (2008) is perhaps the definitive reference for microeconomists interested in an analysis with a very broad scope. There are many papers discussing education and fertility (ie Black et al. (2008)), which won't be discussed in this chapter.

[3] *Ie* F_i and U_i are not functions of ε_Y, however Y_i and F_i are functions of ε_U.

In causal terms as per Haavelmo (1943, 1944) (and particularly, the recent exposition in Heckman and Pinto (2015)), we are interested in the change in Y resulting from the hypothetical manipulation of fertility F, while other elements of the system of equations $(U, \boldsymbol{\varepsilon})$ remain unchanged. We define b as a particular draw of F, and are thus interested in the causal effect of manipulating fertility from b to $b+1$, which throughout this paper we will call β:[4]

$$\beta \equiv \mathbb{E}_{U_i, \varepsilon_Y}[Y_i(b_i + 1) - Y_i(b_i)] \tag{1.3}$$

The hypothetical manipulation envisioned by Haavelmo is generally not feasible in real-world fertility decisions. And in observational studies using data over space or time, the presence of factors similar to U considerably hinders the estimation of causal effects. In the sections which follow we return to this system of equations, and outline the existing techniques which aim to recover causal estimates despite the lack of explicit exogenous manipulation of F.

1.3 The Effects of Family Size on Children

In 1973, the *Journal of Political Economy* released a special issue dedicated to new economic approaches to fertility. Interest in determining the causal effect of total fertility on the outcomes of children in the household blossomed from the articles it contained. A common theme in a number of articles in this issue (Becker and Lewis, 1973; De Tray, 1973; Willis, 1973) concerns a family's decisions regarding fertility (the quantity of children) and investments in child human capital (the 'quality' of children). Abstracting from intra-household variations in child quality[5], each of the aforementioned articles demonstrates the theoretical existence of a quantity–quality (QQ) trade-off.[6]

The QQ trade-off described in the above series of articles as well as in Becker and Tomes

[4] For now we make no distinction between different values of b in (1.3). Generally however, we will be interested in at least two seperate situations. The first, comparing having any children to having no children (the extensive decision), while the second refers to having $b+1$ children versus having b children for $b \in 1, \ldots, k$ (the intensive margin). We return to discussions of b for different parities when discussing empirical results in the sections which follow.

[5] An extensive literature exists which looks at *intra*-household endowment and investment decisions among siblings. Behrman et al. (1982) provides initial discussion, and Aizer and Cunha (2012) embed these considerations in a quantity–quality-type framework.

[6] As Willis (1973) succinctly describes:

> 'Thus, parents not only balance the satisfactions they receive from their children against those received from all other sources not related to children ..., but they must also decide whether to augment their satisfaction from children at the "extensive" margin by having another child or at the "intensive" margin by adding to the quality of a given number of children.'

(1976, 1986) owes to the joint entry of quality and quantity in the household budget constraint. As the number of children enters in the shadow price of quality, and the quality of desired children enters the shadow price for quantity, decisions regarding fertility and quality cannot be made in isolation. Holding all else constant, increases in fertility increase the shadow price of quality, and increases in quantity increases the shadow price of the marginal birth. This considerably complicates causal inference. What's more, as recognised in early articles by Ben-Porath and Welch (1972); Ben-Porath (1976), quality decisions may *directly* feed back to quantity via child mortality.

1.3.1 Observational Data

Given the aforementioned theoretical structure of the relationship between child quality and child quantity, it is apparent that estimating OLS on observational data will lead to consistent estimates of β only in very particular circumstances. To see this, we return to equation 1.1. If we consider standard OLS with a linear model, we re-write (1.1) as:

$$Y = \beta F + U + \varepsilon_Y,$$

where we assume that $\mathbb{E}[\varepsilon_Y] = 0$. To estimate β from the above, we can consider conditioning on two distinct values of F:

$$\begin{aligned} \mathbb{E}[\hat{\beta}] &= \mathbb{E}[Y|F = b+1] - \mathbb{E}[Y|F = b]. \\ &= \mathbb{E}[\beta(b+1) + U|F = b+1] - \mathbb{E}[\beta(b) + U|F = b] \\ &= \beta + \{\mathbb{E}[U|F = b+1] - \mathbb{E}[U|F = b]\} \end{aligned} \tag{1.4}$$

Thus, (1.4) is only identical to the causal estimate in (1.3) in a very limited set of circumstances: above this is when $\mathbb{E}[U|F = b+1] = \mathbb{E}[U|F = b]$. This is simply a specific example of the well-known OLS requirement that the independent variable of interest (F) must be uncorrelated with the omitted error term, given that $\text{plim}(\hat{\beta}) = \beta + \text{Cov}(F, U)/\text{Var}(F)$. Where variation in fertility in a cross-sectional dataset is correlated with movements of other variables related to the outcome of interest (which here we summarise as U), we will fail to identify the true causal effect of fertility given the lack of Haavelmo's hypothetical manipulation of F.

Due to the limitations laid out in the preceding paragraph, very few papers in the literature aim to infer causality by estimating linear models with cross-sectional data.[7] However many

[7] Early work, such as Desai (1995), provides across-sectional descriptive evidence to document correlations, while Hanushek (1992) estimates some cross-sectional (though value-added) models.

papers which use alternative methods to infer causality (discussed in the sections which follow) estimate OLS as a base specification, which can provide some information on the type and degree of bias in OLS. Beyond recognising that a bias is likely to exist, relatively few of these papers provide an explicit discussion of why this may be. Notable exceptions include Qian (2009), who suggests joint parental preferences for more education and fewer children as well as optimal stopping rules which depend on the quality of the first child, and Black et al. (2010), who additionally note that family size effects are confounded with birth order effects. Indeed there are a number of reasons one could use to suggest bias. These include parental education, discount rates, maternal health, or network effects driving both fertility and child quality. Generally it seems likely that these factors will induce a negative bias in OLS estimates of the effect of fertility, given that factors which lead to fewer births (contraceptive knowledge, opportunity cost of time, aspirations, and so forth) also seem likely to drive greater investments in children who are eventually born. Empirically, this overwhelmingly seems to be the case, with OLS estimates of the effect of fertility being universally lower (more negative) than more credibly causal estimates. We return to provide more details on these estimates in the sections which follow.

1.3.2 Instrumental Variables

In systems of equations of the type described in 1.3, one way to drive inference is through the use of shifters (or instrumental variables) which affect the quantity of one of the variables without affecting the other. In order to identify the effect of fertility on children's outcomes, this instrumental variable must affect only fertility, with no indirect effects on quality.[8] Returning to the nomenclature introduced in section 1.2, consider Y_i as child quality, F_i as child quantity, and the unobserved U_i, all generated as described in 1.2. However, now consider the case where a new variable from outside the system is observed, denoted Z_i. Z_i is assumed to directly affect F_i:

$$F_i = f_F(U_i, Z_i, \varepsilon_F).$$

The function 1.1 is unchanged, reflecting the fact that the only channel with which Z_i affects Y_i is through F_i (in other words, the exclusion restriction holds). Finally, to close, assume that $Z_i = f(\varepsilon_Z)$, and once again the error terms $\boldsymbol{\varepsilon}$ are assumed mutually independent.

The above situation leads to an explicit way of generating the hypothetical variation discussed in section 1.2. By taking advantage of variation in F induced by variation in Z, the effect of F on Y can be identified in the absence of any movement in U. The most simple way to consider

[8] *Ie* the exclusion restriction must hold, implying that the estimation of the structural equation which contains quality on fertilty and the instrument must result in a coefficient on the instrumental variable which is precisely equal to zero.

this is by observing the Wald estimator:

$$\hat{\beta} = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[F|Z=1] - \mathbb{E}[F|Z=0]}. \tag{1.5}$$

Here rather than explicitly being based on a movement from $F = b$ to $F = b + 1$ estimation is driven by the effect which Z has on Y, scaled by the degree to which it moves F. If the instrument increases birth by exactly one, then (1.5) collapses to an expression similar to (1.4). Fundamentally in strategies of this type, the identifying assumption shifts from concerns regarding correlations between U and F to correlations between U and Z. Consistent causal estimation now requires that $\text{Cov}(U, Z) = 0$.

The earliest discussion of these types of shifters and the corresponding exclusion restriction required for the estimation of the causal effects of fertility was in Rosenzweig and Wolpin (1980a). They point out that if multiple births are unanticipated, their occurrence will cause some families to exceed their desired fertility, shifting the total number of births in the absence of any change in parental considerations of quality investments. This has motivated estimation in a number of papers, where twin births are employed as instrumental variables. Twin instruments have been employed in a range of contexts and to examine various different 'quality' outcome variables of children. These include Black et al. (2005); Cáceres-Delpiano (2006); Li et al. (2008); Dayioğlu et al. (2009); Sanhueza (2009); Black et al. (2010); Angrist et al. (2010); Fitzsimons and Malde (2010) and Ponczek and Souza (2012), and focus on child quality measures including years of education, IQ, private school enrolment, BMI and height, college completion and age at marriage. The evidence on the existence of a QQ trade-off in these studies is mixed, although recent influential results suggest that the evidence in favour of a trade-off may be weak. In table 1.1 I lay out outcome variables, contexts, and estimates of β presented in the IV literature.

As per the above series of equations, causal estimates rely on the fact that Z truly is independent of U. This has been questioned in a number of ways. Rosenzweig and Zhang (2009) suggest that the close birth-spacing of twins, and the fact that twins have lower health stocks at birth (Almond et al., 2005) means that parents may change behaviours to reinforce or compensate for intra-household human capital differences. While this can be tested directly, it requires data on early live human capital endowments such as birthweight. Bhalotra and Clarke (2015) question the exogeneity assumption in another way. They demonstrate that healthier mothers are more likely to take twin births to term, and at the same time that healthier mothers are more likeley to have additional resources to invest in child quality later in life. This critique is laid out and estimated in chapter 2 of this thesis. At the very least however, both of these critiques will lead to predictable biases in estimates of β, resulting in bounds on the effect of fertility on child outcomes.

A frequently used alternative to twin births consists of instrumenting with the gender mix of children born in the family. Generally, it is argued that parents prefer to have offspring of both genders (Conley and Glauber, 2006; Angrist et al., 2010; Becker et al., 2010; Millimet and Wang, 2011; Fitzsimons and Malde, 2014), and so those having various children of the same sex are more likely to continue childbearing. Alternatively, in some circumstances it is argued that parents have a son preference, and so are more likely to continue after having early-birth girls (Lee, 2008; Kumar and Kugler, 2011). In both cases these are empirically shown to be important drivers of fertility. Again, like estimates driven by twin births, empirical results are mixed, although recent evidence seems to point to statistically insignificant (though nearly universally negative) estimates of the trade-off, as outlined in panel B of table 1.1.

Causality in this case requires that child sex mix has no direct effect on quality. This implies (among other things), that there are no gender-specific economies of scale which facilitate child quality investments more when children are of the same sex (Butcher and Case, 1994). While one could argue (and indeed hope) that goods which could be employed in the household for boy's education could also be employed for girl's education, generally there are other concerns. Dahl and Moretti (2008) show that gender composition affects the likelihood that parents live together. Butcher and Case (1994) provide extensive discussion of the potential that different child gender mixes may affect child costs, and demonstrate that in the USA girls with sisters are significantly less educated than girls with brothers, postulating that this may be due to a reference group effect where parents have lower aspirations for their children when all children are girls. Concerns such as these cast doubt on the validity of the exclusion restriction described earlier in this section.

A range of other instruments have been proposed, including infertility (Bougma et al., 2015), miscarriage (Hotz et al., 1997; Maralani, 2008; Miller, 2009) and distance to family planning (Dang and Rogers, 2013). The outcomes and empirical results related to these studies are displayed in table 1.1. While these instruments—all generally related to the ability to conceive or control conception—clearly drive fertility, in each case the exclusion restriction is questionable. This is explicitly treated in Hotz et al. (1997), who motivate techniques to recover bounds on the estimate of the effect of fertility. At the very least, in each case if unhealthy women are more likley than healthy women to be infertile or suffer miscarriage, this suggests a positive bias in IV estimates of β.

Beyond general threats to inference discussed in this section, instrumental variable estimates lead to the question of 'inference for whom?'. Estimates based on IV lead to a local average treatment effect (LATE), not an average treatment effect for the population in general (Imbens and Angrist, 1994). This LATE implies that any estimates of β holds for that group of the

Table 1.1: Empirical Results: Fertility and Child Outcomes (IV)

Author	Country	Outcome	Estimate (Std. Err.)
Panel A: Twins			
Black et al. (2005)	Norway	Years of Educ	-0.16(0.44)
Cáceres-Delpiano (2006)	USA	Private School	-0.000(0.005)
		Behind cohort	0.005(0.004)
Li et al. (2008)	China	Educ (categorical)	-0.027(0.014)
		Educ (enrolment)	-0.025(0.013)
Dayioğlu et al. (2009)	Turkey	Attendance	0.203(0.245)
Sanhueza (2009)	Chile	Years of Educ	-0.280(0.092)
Black et al. (2010)	Norway	IQ (standardised 1-9)	-0.170(0.052)
Angrist et al. (2010)	Israel	Years of Educ	0.167(0.117)
		Some college	0.059(0.036)
		College grad	0.052(0.032)
Fitzsimons and Malde (2010)	Mexico	Years of Educ (F)	0.096(0.063)
		Enrolment (F)	-0.019(0.014)
Ponczek and Souza (2012)	Brazil	Years of Educ (F)	-0.634(0.194)
		Years of Educ (M)	-0.060(0.164)
Panel B: Gender Mix			
Conley and Glauber (2006)	USA	Private school	-0.061(0.021)
		Grade repetition	0.007(0.004)
Lee (2008)	Taiwan	Total ln(educ spend)	0.328(0.088)
Angrist et al. (2010)	Israel	Years of Educ	-0.067(0.120)
		Some college	-0.025(0.025)
		College grad	-0.032(0.022)
Becker et al. (2010)	Prussia	Enrolment	-0.430(0.189)
Kumar and Kugler (2011)	India	Years of Educ	-0.363(0.061)
Fitzsimons and Malde (2014)	Mexico	Years of Educ (F)	-0.015(0.125)
Millimet and Wang (2011)	Indonesia	BMI for Age	0.049(0.013)
Panel C: Fertility Shock			
Bougma et al. (2015)	Burkina Faso	Years of Educ	-0.99(0.40)
Maralani (2008)	Indonesia	Years of Educ (early)	-0.167(0.117)
		Years of Educ (late)	-0.054(0.055)
Hotz et al. (1997)	USA	Complete highschool	-0.147(0.406)
Dang and Rogers (2013)	Vietnam	Years of Educ	-0.589(0.392)
		Private tutoring	-0.318(0.147)

Notes: In the case that various samples are reported in the papers, the pooled estimate for female and male children of all women from the most recent time period is reported. In the case of twins estimates, the 3+ sample (twins at third birth as instrument) is reported. Where the original studies report p-values associated with estimates rather than standard errors, these are converted into standard errors for inclusion in this table.

population who would be induced to change their behaviour (ie their fertility) by the instrument in question. Thus, all instrumental estimates (even assuming causality) should be cast in terms of the sub-population (compliers) of interest. This is a point explicitly discussed in Angrist et al. (2010), who suggest that the twin instrument is relevant for the whole population, while sex-composition instruments are relevant for only certain groups. Rosenzweig and Wolpin (1980a)'s original article, although based on a reduced form equation, suggests that twin births are relevant for a more specific group than that suggested by Angrist et al. (2010): namely, those families who have a twin birth where the twin birth causes them to exceed their desired fertility. I return to discuss LATE and external validity in the following section of this paper.

1.3.3 Natural Experiments

An alternative manner to deal with correlation between F and U consists of taking advantage of externally defined (to U) reforms. If reforms are applicable to a subgroup of a particular population and are designed to affect fertility, this suggests a natural 'treatment' and 'control' group which can be compared. Those who receive the fertility reform are considered treated, and those who don't are considered as controls. If reforms are truly put in place for reasons entirely divorced from U, causal conclusions can be drawn regarding the effect of the reform. Typically the effect of reforms is estimated using difference-in-differences (DD). This compares pre-reform differences between treated and control units with post-reform differences, inferring that any change in the level of differences is driven by the reform, or stated in another way, that *no* differential and simultaneously occurring phenomena seperate treatments from controls. This is the well known 'parallel trends assumption' and is central to this line of inference.

These studies can be broadly split into two groups: those which examine the effect of public policies or other natural experiments on fertility itself, and those which leverage the externally-defined effect on fertility to quantify the effect of fertility on some other outcome of interest. In the latter case, the differentiation between difference-in-differences and IV estimates is artificial, as the (DD estimated) effect of the policy on fertility is simply plugged in as the first stage in a 2SLS IV framework.[9] The first set of studies are of fundamental importance in analysing the *determinants* of fertility and the effect of new contraceptive methods on life-cycle childbearing, but do not directly quantify the causal effects of fertility itself. Nevertheless, given their relevance both as a first stage in causal estimates and as a reduced form estimate itself, I outline a number of these studies below, before moving on to a more comprehensive discussion of their link to causal estimates.

[9] Duflo (2001) is a well known example of this design. We discuss examples of this framework applied to fertility later in this section and in 1.4.1.

Table 1.2: The Estimated Effect of Reforms on Fertility (Selected Studies)

Author	Abortion Effect	Pill Effect	Note
Angrist and Evans (1996)	-0.012(0.004)		$a, (x = 19)$
Levine et al. (1996b)	-0.019(0.007)		b
Gruber et al. (1999)	-0.059(0.005)		c
Bailey (2006)	-0.093(0.043)	-0.074(0.057)	$a, (x = 22)$
Guldi (2008)	-0.100(0.054)	-0.085(0.041)	
Bailey (2009)	-0.012(0.007)	0.028(0.048)	$a, (x = 22)$
Ananat and Hungerman (2012)	-0.043(0.015)	-0.088(0.023)	

NOTE: All figures report the results of short term access of a fertility reform on birth rates of young women unless otherwise specified in notes.

[a] Binary model with outcome 1=first birth by age x. Bailey (2009) is an erratum for 2006.

[b] Estimate expressed as births per woman. Mean rate is 0.110

[c] Estimate for states adopting 1974-1975. Estimate for 1971-1973 is -0.021(0.005).

Of the large number of studies which use reforms of fertility-control policies[10] to examine the effect on fertility in a DD-style framework, only a relatively small number then employ this as the first stage to estimate the causal effect of fertility—the focus of this review chapter. Among those that *do* directly estimate the effect of fertilty on child outcomes are Gruber et al. (1999); Ananat et al. (2009) and Ananat and Hungerman (2012). Gruber et al. (1999) examines the effect of fertility (via 2SLS) on the likelihood that a child lives with single parents, lives in poverty, receives welfare, and on rates of infant mortality and low birth rates. Of these, it is suggested that fertility significantly increases the probability of living in poverty and having single parents, as well rates of infant mortality. Ananat et al. (2009) also examine these outcomes, and suggest that in the long-run, the marginal child is more likely to have lived with a single parent, receive welfare, and not have graduated college. Finally, Ananat and Hungerman (2012) return to these same outcomes and report a Wald ratio as in (1.5). These Wald estimates allow them to look at the characteristics of marginal child not born due to both the diffusion of the pill, and the legalisation of abortion. Their results suggest that the two fertility control policies had remarkably different effects on marginal child characteristics. In agreement with the above studies they suggest that the marginal child not born due to abortion legalisation would have been 49.2% (se=25.5) more likely to live in a welfare-receiving household. However, the marginal child not born due to pill diffusion looks remarkably different: 8.0% (se=4.4) *less* likely to belong to a welfare receiving household. These comparisons make manifestly clear the distinction between compliers for different instruments discussed at the end of section 1.3.2. Given that the group of 'compliers' in the two policies had very different characteristics, estimated effects of fertility on outcomes are very different despite being plausibly causal in both cases.

[10]For abortion: Ananat et al. (2007, 2009); Angrist and Evans (1996); Charles and Melvin (2006); Cook et al. (1999); Currie et al. (1996); Gruber et al. (1999); Guldi (2008); Kane and Staiger (1996a); Levine et al. (1996b,a, 1999); Pop-Eleches (2005, 2006), for the oral contraceptive pill: Ananat and Hungerman (2012); Bailey (2006, 2010, 2012, 2013); Christensen (2012); Goldin (2006); Goldin and Katz (2002a,b); Kearney and Levine (2009) and for the emergency contraceptive pill: Durrance (2013); Gross et al. (2014) and chapter 3 of this thesis.

Despite not directly estimating the causal effect of fertility on child outcomes, a number of other contraceptive-based natural experiment papers estimate the effect of the natural experiment directly on child outcomes. This reduced form technique provides an estimate of the numerator of the ratio in (1.5), and so can be thought of as an unscaled estimate of the effect of fertility. Papers of this type include Pop-Eleches (2006) who finds that the illegalisation of abortion in Romania worsens child education and labour market outcomes (conditional on parental characteristics) and Bailey (2013) who reports that US contraceptive pill laws had long-standing impacts on children's eventual college completion, labour force participation, and family incomes.

The validity of using policies of this type to isolate the effects of childbearing on child outcomes hinges upon the fact that the timing (or allowance) of fertility control reforms should not depend upon pre-existing differences between areas affected and those not affected by the reform. Any phenomena which will imply that 'treated' and 'untreated' areas would follow different paths *in the absence* of the reform will lead to inconsistent estimates of the effect of fertility on child outcomes. Generally, papers which propose estimation by leveraging reforms of this type run a series of tests, including event-study analysis, placebo regressions, or a regression of receipt of treatment on pre-existing characteristics.[11] However, directly testing the validity of such estimation methodologies is, of course, impossible, given that the counterfactual outcome—the world where the pill was not available—is never observed. This has lead to back-and-forth discussion, questioning the validity of the use of policy-defined reforms to drive estimation (for example, see Joyce (2013), who questions the exclusion restriction vs Bailey et al. (2013) who defend current state-of-the-art results). While concerns that reforms may be systematically correlated with other unobservable factors are of course justifiable, the best sets of studies aim to use judiciously chosen control groups (including using women of different ages subject to the same geographic factors and institutions), to minismise concerns such as these. An alternative concern in identification strategies of this type surrounds the possiblity that *local* reforms have more widely spread effects. For example, the availability of abortion in one region does not necessarily imply that nearby non-treated individuals cannot travel to treated areas, defying their quasi-experimental status to receive treatment (Levine et al., 1999). Fortunately, violations of this type will, at worst, bias downwards estimated results. Often concerns such as these are examined empirically, as is the case in Christensen (2012). This is a point returned to more extensively, and tested explicitly in chapter 4 of this thesis.

Finally, a number of other natural experiments have been used in the literature to examine

[11] For example, Bailey (2006) demonstrates that early access to the pill was unrelated to education, fertility norms, poverty rates, availability of household technologies such as washers and dryers, as well as labour market participation at a state level. The probability of early access is, however, related to the percent of Catholic residents in a state.

the effect of fertility on child outcomes.[12] Perhaps most notably among these, Qian (2009) uses the relaxation of China's one child policy to estimate the causal effect of movements from one child to two child households. This study is unique for two reasons: the low parity shift of the experiment (an expansion from one to two children), and the fact that it finds that higher fertility in this case *increases* child schooling outcomes, especially among households who have two children of the same gender. These results suggests that estimates of fertility at the intensive margin may not be linear, and indeed may not even be monotonic by parity, changing from positive to negative at higher orders.

1.3.4 Structure and Dynamics

A number of dynamic or dynamic structural papers motivate estimation of the effects of fertility (or fertility timing) on birth outcomes based on a finite horizon, rather than static, estimation framework. While these papers allow for a more extensive examination of *timing* and life-cycle decisions, estimation generally relies on an exclusion restriction similar to those discussed in section 1.3.2. Rosenzweig and Wolpin (1995) motivate the estimation of a dynamic model to examine the effect of early fertility (teen motherhood) on child birth outcomes (gestation and birthweight). By formulating an over-identified series of equations where identification (in a FE-IV framework) comes from family background variables, and idiosyncratic elements shared by siblings. Rosenzweig and Schultz (1985) take advantage of variations in fecundity, or births per attempt, which they suggest are unobserved by parents prior to contraception attempts, but observable after the fact in data, as births per period. Identification in these studies usually depends on correct functional form and distributional assumptions for the stochastic error term(s), though it is important to point out that precisely the same conditions are the case for parametric OLS (DD) and IV estimates discussed in previous subsections.

1.4 The Effects of Child Birth on Mothers

Beyond the analysis of a child's effect on his or her siblings' outcomes, a birth, at the extensive or the intensive margin, has myriad impacts on parents or other carers. The analysis of these effects has received considerable and ongoing attention in the economics literature. Much of the

[12] Similarly, Bleakley and Lange (2009) use a natural experiment: the eradication of hookworm in USA, to test the QQ hypothesis. However, the elimination of hookworm is used as a shifter for child quality, *not* child quantity. This allows them to quantify the effect of quality increases on subsequent fertility decisions of households, and they find that increases in quality do lead to fertility declines in line with the QQ model discussed earlier.

focus of this work falls on the effect of marginal births on mothers' labour market outcomes and trajectories.

Fleisher and Rhodes (1979) provides a summary of the early literature, with considerable coverage also provided in the *JPE* Fertilty issue described in section 1.3 (Willis, 1973; Gronau, 1973). As is the case with child investment and fertility decisions, choices regarding fertility, labour market participation, and (adult) human capital attainment are linked, and dynamic in nature. Total fertility, and, if child-bearing, birth timing have important impacts on labour market participation, non-labour market work, accrued experience, and wages, while participation, experience and wages also influence timing and fertility decisions. Inferring causality in systems of this type is once again challenging, relying on the use of plausible instruments, natural experiments, structural estimation, or a combination of methods.

1.4.1 Natural Experiments

Frequently, natural experiments experiments of the type discussed in section 1.3.3 are leveraged to quantify the effect of fertility on parent outcomes. Estimation is based on the fact that—at the level of the family—living in treatment or non-treatment areas is a randomly assigned variable which can be used to isolate effects on fertility in the absence of changes in other outcomes. This requires that these experiments be clearly demarcated, unexpected, and not propogate from treated to untreated areas.

One of the most common natural experiments employed in these types of analyses is the arrival of new birth control technologies to a particular geographic area. There are a very large range of microeconomic studies which discuss the effect of these types of programs on a mother's total fertility. These can be broadly split into those which examine the short-run effects of contraceptives on fertility,[13] and long-run analyses, which account for both short-run delays and long-run rearrangements in timing afforded by new technologies. Short run analyses include those examining the contraceptive pill (Bailey, 2006, 2009; Christensen, 2012), abortion (Guldi, 2008; Levine et al., 1999), the morning after pill (Gross et al.,2014; Durrance, 2013; chapter 3) and medicare access (Kearney and Levine, 2009), while those examining the long-run effects of contraceptive reform on completed fertility include (among others) Bailey (2010, 2013, 2012) for the contraceptive pill and Angrist and Evans (1996); Ananat and Hungerman (2012) for abortion.

[13]This does not imply using data over a short time frame, but rather examining the effect of a birth control method up to an age *less* than the end of the fertile life (eg Bailey (2006)'s focus on childbearing before the age of 22).

Once again however, beyond the direct relevance of this swarth of studies for policy focused on fertility control, in order to apply these results to *causal* analysis of parental outcomes, the specifications discussed above can only act as a first-stage effect. For the full system of equations, we are interested in a two-step process: first quantifying the effect of reforms on fertility, and then, from this, the flow-on effect that exogenous shifts in fertility have on mother (or carer) outcomes.

Only a subset of papers which focus on fertility reforms then go on to examine the second stage of interest here. As discussed in sections 1.3.2 and 1.3.3, consistent estimation relies on an exclusion restriction assumption, whereby the only effect of the program on outcomes is driven by its effect on fertility. Ananat and Hungerman (2012) use both the pill and abortion to examine different groups of compliers, and report Wald estimates of the effect of fertility on single parenthood: for pill-compliers, marginal fertility reductions occur in contexts with less single parenthood, while the reverse is true for abortion compliers.[14] Angrist and Evans (1996) report similar estimates for 1970 abortion reforms in the USA. They report that the effects of a particular type of child-bearing (teen and unmarried) reduces the education and employment probability, particulary of black women. Other significant outcomes discussed in this framework include Bailey et al. (2012); Bailey (2006, 2013), who show that it has effects on wages over the life cycle or female labour force participation rates, and Christensen (2012), who finds (reduced form) effects on cohabitation.

1.4.2 Instrumental Variables

The use of instrumental variables to examine the effect of fertility on *mothers'* outcomes (rather than children's outcomes as described in section 1.3.2) follows a similar logic to that outlined in equation (1.5). An external variable which has strong effects on fertility but no direct effects on the outcome of interest except via its effect on fertility can be used to drive causal estimates. Instrumental variable estimates are a popular methodology employed to determine the effect of fertility on mothers.

Outcome variables of interest are typically related to parental labour force outcomes,[15] including female labour force participation (Agüero and Marks, 2008, 2011; Chun and Oh, 2002;

[14] They also show that the long-term effect of the pill on mothers results in a higher likelihood of college completion, having a college educated spouse, and a lower likelihood of divorce.

[15] Largely, these papers focus on maternal labour market participation rates. Kim and Aassve (2006) studies both mothers' and fathers' responses to fertility, using fecundity (births per attempt) as an instrument. They find that on average mothers reduce hours of work in the short run, while paternal hours of work increase (in rural areas).

Cáceres-Delpiano, 2008; Angrist and Evans, 1998), or earnings (Cáceres-Delpiano, 2006; Hotz et al., 1997; Jacobsen et al., 1999). A range of instruments has been proposed including twins, as in section 1.3.2 (Rosenzweig and Wolpin, 1980b; Jacobsen et al., 1999; Bronars and Grogger, 1994), gender mix (Agüero and Marks, 2008, 2011; Chun and Oh, 2002), and fertility shocks (Miller, 2011; Cristia, 2008; Rosenzweig and Schultz, 1987).[16] The effect sizes of existing estimates of the effect of fertility on labour market outcomes largely point towards significant negative effects, though not universally so. A summary of point estimates and confidence intervals of estimates is presented in table 1.3.

Table 1.3: Fertility and Mother's Labour Market Outcomes

Authors	$\hat{\beta} \pm$ se$(\hat{\beta})$	β
Labour Force Participation		
Bronars and Grogger (1994)		−0.123
Angrist and Evans (1996)		−0.143
Angrist and Evans (1998)		−0.113
Jacobsen et al. (1999)		−0.016
Cáceres-Delpiano (2006)		−0.068
Cáceres-Delpiano (2008)		−0.064
Agüero and Marks (2008)		−0.004
Angrist et al. (2010)		0.032
Agüero and Marks (2011)		−0.006
	−0.4 −0.3 −0.2 −0.1 0 0.1	
Hours per Week		
Hotz et al. (1997)		−1.46
Angrist and Evans (1998)		−4.59
Jacobsen et al. (1999)		−1.14
Cáceres-Delpiano (2006)		−1.24
Angrist et al. (2010)		2.35
	−6 −4 −2 0 2 4 6	

Notes to table: Points represent coefficients, while error bars represent 95% confidence intervals. Estimates are ordered by date of publication. In the case that various samples are reported in the papers, the pooled estimate for all women from the most recent time period is reported. In the case of twins estimates, the 3+ sample (twins at third birth as instrument) is reported.

While the majority of these instruments can only be used to estimate the effect of fertility at the intensive margin, interestingly, those based on fertility shocks can also be applied to quantify the effects of extensive margin births. Cristia (2008) for example proposes using the outcome of fertility treatments (pregnant or not) as an instrument, suggesting that delays in child-bearing lead to an increase in wages and hours worked.

[16] Ribar (1994) proposes three alternative exclusion restrictions (age at first period, availability of Ob/Gyn, and local abortion rates) for use in selection models. While the identification methodology is different to IV, the requirements for inferring causality are identical.

Finally, miscarriage has been proposed as an alternative IV that can be employed to estimate the effect of child birth on maternal outcomes (Hotz et al., 2005; Fletcher, 2012). This line of argument relies on fetal deaths in utero being randomly assigned to mothers, in order to compare treated (live births) to control (no live births) women. This also relies on miscarriage not having any other effect on (propsective) mothers' outcomes of interest, beyond its direct effect on fertility. Hotz et al. (2005) suggest that following this line of argument, early (teenage) child-bearing is associated with small effects on educational attainment, and life cycle changes in labour market rates.

The use of this instrument is of course complicated if characteristics which predict miscarriage are also correlated with mother unobservables. Given that miscarriage is considerably more likely for unhealthy mothers, this seems likely, and a range of studies address these concerns. Foremost is Hotz et al. (1997) who discuss how to bound the effect of fertility where the instrumental variable is composed of a mixture of both women who randomly miscarry, and those who non-randomly miscarry. They show that tight bounds on the effect of fertility can be estimated, if the proportion of non-random and random miscarriages can be estimated. Applying these bounds estimates they suggest that teenage child-bearing significantly increases the number of hours worked during early adulthood, and (weakly) decreases the likelihood of completing a GED certificate in the USA. Fletcher and Wolfe (2009) provide additional discussion of the challenges in estimating causal effects using miscarriage. They suggest that unobserved *community*-level characteristics are likely correlated with miscarriage, and once including community fixed-effects find that teen child-bearing reduces education and wages, and increase the likelihood of welfare receipt.

1.4.3 Other Methods

A range of other methods have been employed in the economic and non-economic literature to examine the effects of fertility on mother's outcomes. These involve RCTs (DiCenso et al., 2002) between-effects using siblings (Holmlund, 2005; Geronimus and Korenman, 1992; Ribar, 1999), and other matching methods (Chevalier and Viitanen, 2003; Levine and Painter, 2003). In the case of the last two methods (between-effects and matching), the identification of casual effects relies on the comparison method fully controlling for relevant differences between those having children, and those not having children. In matching this collapses to an assumption regarding 'selection on observables' (which is to say that any characteristic predicting child-bearing is observed by the econometrician), and in siblings or relative fixed effects, that on average, those who become pregnant early in life are otherwise identical to those who become pregnant later

in life. Ribar (1999) and Rosenzweig and Schultz (1985) provide additional discussion, and examination of, the validity of these estimation techniques.

Finally, Rosenzweig and Wolpin (1980b)—in their initial proposition of twins as an exclusion restriction—return to the Beckerian (1973) simultaneous equation framework for fertility, child quality, *and* life-cycle (mother's) labour supply. They are the first to use twins to estimate the structural equation linking fertility and labour supply. They estimate that for younger women, additional births reduce labour supply, but this fades as women age. Once again—as they indeed highlight—consistent estimation relies on twins being entirely orthogonal to labour supply. This assumption is questioned in previous sections of this review chapter. Using the presumed exogeneity of twins as an identifying assumption, Rosenzweig and Wolpin (1980b) provide a very interesting series of tests casting considerable doubt on the assumption that fertility is exogenous to labour supply decisions, as maintained in the prevailing literature at the time of their work.

1.5 Conclusion

This review chapter serves to provide an overview of the causal estimation of the effect of fertility on child and parental outcomes. It surveys the wide range of methodologies employed in the existing microeconometric literature, and discusses how various techniques aim to skirt issues of endogenous fertility choices. In each case, I outline the identifying assumptions implicitly or explicitly invoked, as well as the threats to which these are subject.

The evidence discussed in this chapter is mixed. While there seems to be quite clear evidence in favour of moderate-to-large effects of marginal child births and early births on parental labour market outcomes, the existing microeconometric child-level estimates are less compelling. Despite a large body of theoretical microeconomic work which posits that such a QQ trade-off may exist, causal estimates are certainly not conclusive, and seem to suggest that the trade-off is small or non-existent. While there are a number of papers which *do* find significant effects on a number of outcomes, these are context- and complier-specific.

In some cases this lack of evidence may be due to threats to exclusion restrictions or other identifying assumptions. It is in line with this that the present thesis now moves forward. The remaining chapters are essays on the causal estimate—or more specifically, the challenges and involved in causally estimating—the effect of fertility and contraceptive programs on human outcomes, and some solutions which may be employed to recover bounds and causal estimates of

these effects.

Chapter 2

The Twin Instrument*

Estimating the Quality–Quantity Trade-off

Chapter Abstract

The incidence of twins has been used to identify the impact of changes in fertility on measures of investment in children born prior to the twins, and the emerging consensus in this literature is that there is no evidence of a quantity-quality (Q–Q) trade-off. We argue that the standard approach is flawed. Even if twin conception is random, bringing twins to term is a function of maternal health which is difficult to fully observe and which tends to be correlated with child quality, rendering the instrument invalid. The neglect of this fact in the existing literature will tend to lead to under-estimation of the Q–Q trade-off and so could contribute to explaining the negative results in the literature. Our contention that women who produce twin births are positively selected is demonstrated using data from richer and poorer countries. Using large samples of microdata from developing countries and from the USA which include indicators of maternal health characteristics and behaviours, we show that a significant trade-off emerges upon correcting for these biases. We show that this result is likely to be only a *lower* bound of the true Q–Q trade-off and discuss how to estimate the size of these bounds. These results have important implications for twin studies in all contexts examined here.

*This paper is coauthored with Professor Sonia Bhalotra. We are grateful to Paul Devereux, James Fenske, Cheti Nicoletti, Carol Propper, Atheen Venkataramani, Marcos Vera-Hernandez and Frank Windmeijer, along with various seminar audiences and discussants for helpful comments. We thank Emilia Del Bono, Climent Quintana-Domeque, Pedro Ródenas, Libertad González, Anna Aevarsdottir, Martin Foureaux Koppensteiner and Ryan Palmer who have very kindly shared data and/or source code from their work.

2.1 Introduction

Since the pioneering work of Rosenzweig and Wolpin (1980a), economists have attempted to leverage the occurrence of twin births to estimate the effect of family size on child outcomes. If twin births occur at random, their occurrence constitutes a fertility shock that is uncorrelated with family characteristics, including parental preferences, and other unobservables which may be related to child quality. This provides the exogenous variation (in quantity) required to estimate the quantity-quality model of Becker (1960); Becker and Lewis (1973); Becker and Tomes (1976). The essential idea of these studies is that the shadow price of child quality is increasing in child quality and *vice versa*. Thus, by comparing those families who unexpectedly produced additional children with those who always produced one child per birth, it is argued that the effect of an additional sibling on a child's human capital attainment can be uniquely identified.

However, the consistent estimation of this effect is based upon the untestable assumption that twinning is exogenous. This requires not only that twin conceptions are randomly assigned to families, but also that taking a twin conception to term does not depend upon a woman's behaviours during pregnancy or on her endowments prior to pregnancy. This is at odds with the evidence.

We show that endowments and behaviours affect the chances of twins being born. In data from a large sample of developing countries, we show that taller women and women with a higher body-mass index are significantly more likely to give birth to twins.[2] Maternal health is likely to be an especially significant determinant of birth outcomes in poorer countries where many women are chronically under-nourished (an indicator of which is their final stature), exhibit anemia and low-BMI, and are prone to infections. In these conditions only relatively healthy women will have the resources to support a successful twin pregnancy. However, our critique applies to richer countries too. Using administrative data from the Scotland, the USA, Sweden and Spain, and survey data from the UK and Chile, we find that women who are taller and less likely to engage in risky behaviours such as smoking, drug taking or alcohol consumption during pregnancy are significantly more likely to have a (live) twin birth.[3]

Overall, our argument is that women who give birth to twins are positively selected and that the common tendency to ignore this will result in under-estimation of the Q–Q trade-off.

[2]Height is an index of the stock of health of a woman and is a function of investments over her growth period and, especially, the early years of her life (see references in Bhalotra and Rawlings (2013)).

[3]Simiarly, it is well known that fertility treatments increase the likelihood of twinning. While we do not discuss this extensively here, we note that this is likely to be less of an identification challenge given that IVF treatment is an entirely observable behaviour, unlike maternal health and pregnancy behaviours, which are multidimensional and largely unobservable.

We focus upon maternal health because no previous work has highlighted it as a determinant of twinning, and because it is inherently impossible to fully control for. Even if we had data that included all of the indicators of health we mention above, we may not observe whether women skip breakfast (Mazumder and Seeskin, 2014), whether they are stressed (Black et al. (2014) and references therein), whether they seek and adhere to antenatal care and so on. Our contention adds a novel twist to a recent literature which suggests that mothers' health and fetal environment matter and may alter the birth weight of children, the sex ratio at birth and a range of future human capital outcomes (Almond et al., 2011; Bhalotra and Rawlings, 2013; Barker, 1995). Like birth weight and the probability of a boy relative to a girl birth, the probability of a twin birth is increasing in the health of the mother/the fetal environment.

The emerging consensus in the literature using twin births to test the Q–Q model is that there is in fact no significant or substantial trade-off between fertility and investments in children. In this paper we suggest that this result may, in principle, follow from the bias created by twin-mothers being positively selected. We re-examine the validity of these results, accounting for the innovation discussed previously. If twin births are not truly exogenous, but instead depend upon maternal health stocks and behaviours, there is an identification challenge. Specifically, if healthier mothers are both more likely to give birth to twins, and more likely to invest in child human capital in later life, existing estimates may significantly underestimate the size of the Q–Q trade-off. When ignoring these considerations we find that the effect of an additional birth on human capital attainment (education) is minor, or even weakly positive. However, when taking into account this innovation we find that a significant trade-off does exist, and that an additional birth reduces standardised schooling behaviour by at least 4% of a standard deviation.

The estimated effect of $\sim 4\%$ of an s.d. is at most the lower bound for the true size of the Q–Q trade-off. Given that we suggest that maternal health predicts twinning, and given that maternal health is multidimensional in nature and difficult to observe fully, we will only ever be able to include a partial set of controls to account for the inconsistency in IV estimates. As such, we examine a number of methods to estimate plausible bounds on the Q–Q trade-off. First, we argue that the true estimate is bounded by the OLS and IV estimtes. We also follow Conley et al. (2012) in conceptualizing twins as a plausibly exogenous event, and derive estimates by assuming that the traditional exclusion restriction is 'close to' holding.

Empirically, work which further probes the Q–Q trade-off is important, especially in the developing country setting on which we place considerable focus in this paper. In order to conduct empirical tests, we construct a large microdata set from 68 developing countries with observations on more than 2.5 million children and nearly 1 million mothers. The macro level trends in this data suggest that educational attainment has risen considerably while completed

and desired fertility has fallen sharply over the past 50 years (see figures 2.1a and 2.1b). Similar effects have been described extensively in the economic and demographic literature (*eg* Hanushek (1992)). It is of considerable relevance to researchers and to policy makers to determine whether such a trend is (at least partially) causal. In practical terms, a significant number of government bodies report that family planning is considered an important concern,[4] and in some cases these concerns have resulted in aggressive, and at times group-specific, fertility control policies.

This paper unfolds as follows. In the next section we discuss the existing literature which estimates the Q–Q model using twins. We then discuss the methodology that we will use to examine twinning, and to bound the Q–Q trade-off. Section 2.4 discusses our data sources and estimation samples from both low- and high-income countries, and section 2.5 presents results. We briefly conclude in the final section.

2.2 Twins

The ocurrence of twins has fascinated people, not least of all social scientists, for as long as recorded history exists. Stories of Romulus and Remus, the mythological founders of Rome, date from at least the fourth century BC. However, the first use of twins as an exogenous increase in family size in the economic literature came much later, in Rosenzweig and Wolpin (1980a).

Rosenzweig and Wolpin (1980a) proposed to incorporate twins into an estimation strategy in order to circumnavigate problems of the joint determination of child quantity and quality first raised by Becker (1960); Becker and Lewis (1973); Willis (1973); De Tray (1973); Becker and Tomes (1976). Under a number of assumptions, they show that twins as a shock to family size will be sufficient to directly identify the interaction between quality and quantity. As well as the occurrence of twins pushing at least some families above their desired total fertility, this requires that twin births are random. Empirically then, Rosenzweig and Wolpin (1980a) estimate the Q–Q trade-off, accounting for the increase in rates of twinning by parity (a biological relationship)[5], and by total number of births (a mechanical relationship). Beyond these variables, twinning must be random to produce consistent estimates.[6]

[4] A recent survey of national governments suggests that fertility was perceived as too high in 50% of developing countries, with this figure rising to 86% among the least developed countries United Nations (2010).

[5] Such a relationship is an empirical regularity in all data examined. Rosenzweig and Wolpin (1980a) report rates which increase by parity in USA. In DHS data a similar pattern is observed, as presented in figure 2.2.

[6] It is important to note that in their original paper this *is* tested formally. The authors report that a joint test of the effect of farm size, non-farm earnings and parental education on the rate of twins per births did *not* allow for the rejection of no statistical effect.

The twin instrument has been widely used since Rosenzweig and Wolpin's (1980a) initial work. As well as its use in the estimation of the quantity-quality trade-off (Black et al. (2005); Cáceres-Delpiano (2006); Li et al. (2008); Angrist et al. (2010), among others), it has been used to estimate the effect of childbearing on female labour force participation (Rosenzweig and Wolpin, 1980b; Jacobsen et al., 1999; Angrist and Evans, 1998), and the effect of unwed childbearing on marriage market outcomes, poverty and welfare receipt (Bronars and Grogger, 1994). Typically, estimation is based on two-stage least squares, with the set of controls included in the first- and second-stage varying slightly over time.[7]

Table 2.1 documents the principal studies in the Q–Q trade-off literature where twins are employed. Along with the data sample and time period under study, we list the set of controls included in each case. The more recent wave of these studies make a similar conditional randomness assumption, however refurbish the controls in IV estimates to include variables such as a mother's race and her educational attainment. In some cases the validity of such assumptions is probed by regressing twinning on observable family outcomes, or testing for the equality of means of certain characteristics between twin and non-twin families. Black et al. (2005), Li et al. (2008) and Sanhueza (2009) report joint F-tests suggesting that twinning is not related to parental education in their data samples, while Rosenzweig and Zhang (2009) report t-tests showing equality of means across twin and non-twin groups. However, as is well known and acknowledged in each case, any such tests are at best partial evidence in support of instrumental validity. While twins can be shown to be unrelated to observable or measured characteristics, similar tests cannot be run for variables which are either unobservable, or not recorded in survey data. We return to this point in the following sections.

The most comprehensive controls considered in the economic literature (often by necessity due to data restrictions), include maternal age, parental education and measures of income and/or goods. However, recent evidence from the medical literature points to the fact that twinning may depend more deeply upon a mother's health behaviours or endowments. Hall (2003) for example suggests that follicle-stimulating hormone (FSH) is associated with an increased likelihood of twinning, and is found in higher concentrations in older, heavier and taller mothers. Further, she suggests "that adequate maternal folic acid consumption could affect the number of twins coming to term" (see p. 741, and further discussion in Li et al. (2003)).

Unrelated to health measures *per se*, recent studies seek to control for the fact that multiple births are correlated with fertility treatments. Typically, such an analysis requires either fo-

[7] Twins have been widely used in the economic, medical, biology and psychology literature in a number of ways. In this paper we focus only on the use of twin births as an instrument for total fertility, and not on the so called 'twin studies', which base inference on between-twin comparisons using maternal fixed effects.

cusing on offspring born before the introduction of fertility treatments (Cáceres-Delpiano, 2006; Angrist et al., 2010), or, in the case of sufficiently rich data, removing families undergoing fertility treatment from estimation samples (Braakmann and Wildman, 2014). Once again, consistent estimation in this case is based on the assumption that beyond fertility treatment and family controls listed above, twin births are as good as random.

Finally, the twin instrument is not without critique for other reasons. Existing critiques of the twin instrument have focused on parental behaviours in response to twins, rather than on the likelihood that parental behaviours (or endowments) may affect the likelihood of twinning. Rosenzweig and Zhang (2009) question the effect that close (or indeed no) birth spacing and an endowment effect—where parental behaviours respond to the lower health at birth of twins compared to single births[8]—has on investments in pre-twin siblings. They demonstrate that if parents behave in such a manner, bounds for the Q–Q trade-off can be calculated. This hypothesis is tested in Angrist et al. (2010), and applied in Fitzsimons and Malde (2014). We turn to bounds estimation in section 2.5.4 of this paper.

2.3 Methodology

Empirical analyses of the quality-quantity trade-off focus on producing consistent estimates of β_1 in the following equation:

$$educ_{ij} = \beta_0 + \beta_1 fert_j + \boldsymbol{X}\boldsymbol{\beta} + u_{ij}. \tag{2.1}$$

Here, quality is proxied by the educational attainment of child i in family j, ($educ$) and fertility ($fert$) is measured as the total births in a child's family. A vector of family and child controls is included, denoted $\boldsymbol{X}$. As has been extensively discussed in prior literature, estimation of β_1 using OLS with cross-sectional data will result in biased coefficients given that child quality and quantity are jointly determined (Becker and Lewis, 1973; Becker and Tomes, 1976), and given that unobservable parental behaviours and attributes influence both fertility decisions, and investments in children's education (Qian, 2009).

[8]Using data from the United States, Almond et al. (2005) document that twins have substantially lower birth weight, lower APGAR scores, higher use of assisted ventilation at birth and lower gestion period than singletons. In our data samples similar endowment differences are observed. For example, appendix figure 2.10 documents the much larger average reported birth size of twins versus singletons in DHS data. Birth weight figures show similar patterns.

2.3.1 Quantity-Quality with Twins

One proposed solution has been to employ 2SLS estimation, where fertility is is instrumented using twin births.[9] The corresponding first stage is:

$$fert_j = \pi_0 + \pi_1 twins_j + \boldsymbol{X}\boldsymbol{\pi} + \nu_j, \tag{2.2}$$

where $twins_j$ is an indicator for whether the n^{th} birth in a family is a twin birth. As described further in section 2.4.2, the sample in each case is the so-called $n+$ group, consisting of children born before birth n in families with at least n births. As such the twins themselves are excluded from the estimation sample.[10] The logic, in quasi-experimental terms, is that existing children (the subjects) are randomly assigned either one (control group) or two (treatment group) siblings at the n^{th} birth.

Consistent estimation of β_1 can thus proceed provided (among other things) that instrumental validity holds:

$$\text{plim}_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} twin_j u_{ij} = 0 \tag{2.3}$$

The typical challenge in IV estimates arises when considering (2.3); as the error term u_{ij} consists, by nature, of unobservable components, whether or not the equality holds cannot be tested formally. There is however nothing which stops us from partially testing (2.3) by removing a subset of observable components from u_{ij} and testing whether their (conditional) correlation with $twin_j$ is significantly different from zero. The error term u in (2.1) is a function of a large number of elements:

$$\begin{aligned} u = f(&\text{maternal health stocks, fertility behaviour, positive pregnancy investments,} \\ &\text{parental education, fetal environment, } \ldots) \end{aligned} \tag{2.4}$$

While many of the relevant elements are either completely or partially unobservable, some of these variables, such as maternal education and incomplete measures of health stocks and behaviours, can be observed. Thus a partial test of the twin methodology consists of estimating the following regression:

$$twin_j = \alpha_0 + \boldsymbol{X}\boldsymbol{\alpha}_1 + \boldsymbol{S}\boldsymbol{\alpha}_2 + \boldsymbol{H}\boldsymbol{\alpha}_3 + \varepsilon_j. \tag{2.5}$$

Here $\boldsymbol{X}$ refers to the initial vector of family and child controls, $\boldsymbol{S}$ to additional family socioeco-

[9] Other instruments and methodologies are also used including gender mix of children (Conley and Glauber, 2006), policy experiments such as the arrival and loosening of the one child policy in China (Qian, 2009; Argys and Averett, 2015), historical time series variation in schooling (Bleakley and Lange, 2009), or fertility transitions during industrialisation (Dalgaard and Strulik, 2015).

[10] Typically, the argument is made that twins are different to single births, and hence should not be compared in analysis.

nomic variables such as income and parental education, and $\boldsymbol{H}$ to maternal health variables.

If twin birth is indeed an event which is as good as random, the coefficients on maternal health and family socieoeconomic variables in the above regression should not be significantly different to zero. We thus test the following hypothesis:

$$H_0 : \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_3 = 0. \tag{2.6}$$

Rejection of the null would raise difficulties in proceeding with IV estimation using the twin instrument. Of course, if the rejection of the null were only due to one or a number of *observable* element(s) which predicted twinning, these variables could simply be included in the first and second stages above, much like occurs with maternal age and race in the existing literature. However, more generally it would be difficult to be argue for instrumental validity if twinning is shown to depend upon (a limited set of) measurable family characteristic or choice variables, while many similar variables are not observed.

Given the biological demands placed on a mother pregnant with twins, we may expect that healthier mothers, or mothers with more resources to invest in their pregnancy, are more likely to take twin conceptions to term. Similarly, we may suspect that mothers more able to invest in their children during pregnancy will also be more able to invest in their child's human capital after birth. If this is the case, we would see that (at the very least) $\boldsymbol{\alpha}_2 > 0$.

An alternative test of whether twins appear to be as good as random consists of comparing women who give birth to twins with those who give birth to singletons *before* these children are born. If twin births occur randomly in the population, the two groups of mothers should appear identical before these births occur. In order to compare health stocks before twins, we run tests comparing the rate of infant mortality—a completely predetermined variable—of children in each of the $n+$ groups described above. If, as we contend, healthier mothers are more likely to give birth to twins, this should be captured in lower infant mortality rates (IMR) among twin mothers in early births.

If, following the tests described above, twins are found to be positively selected among healthy mothers, traditional IV estimates of β_1 will be inconsistent. We turn to discuss this bias now. Assuming additive separability of the elements in the omitted error term[11], we can re-write u_{ij}

[11]This assumption can be loosened with little implication on the analysis which follows. If we do not assume additive seperability, covariance terms between each error component must be included. However, given that the covariance between elements of $\boldsymbol{S}$ and $\boldsymbol{H}$ is likely to be positive, and given that the covariance between each of these and u^*_{ij} is likely to be of the same sign as the covariance between *twin* and u^*_{ij}, the omission of the covariance terms does not effect the inequalities in equation 2.9. This is something which we test empirically later in this paper.

from (2.2) and (2.4) as:

$$u_{ij} = u_{ij}^S + u_{ij}^H + u_{ij}^*.$$

Here u_{ij}^S and u_{ij}^H correspond to the (observable) elements included as $\boldsymbol{S}$ and $\boldsymbol{H}$ in (2.5), while u_{ij}^* represents the remaining (unobserved) components. We can thus re-write our IV estimate for β_1 as:

$$\hat{\beta}_1^{IV} = \beta_1 + \text{plim}_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} twin_j(u_{ij}^S + u_{ij}^H + u_{ij}^*) \tag{2.7}$$

Typically, this is the coefficient estimated in the existing twin literature which assumes that twinning is a conditionally exogenous event. If, however, the likelihood of taking twin conceptions to term increases for healthier and/or wealthier mothers, we should include $\boldsymbol{S}$ and $\boldsymbol{H}$ in the first and second stages, giving

$$\hat{\beta}_1^{IV,S+H} = \beta_1 + \text{plim}_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} twin_j u_{ij}^*, \tag{2.8}$$

where the superscript $S+H$ signifies that socioeconomic and health variables have been included as additional controls and, correspondingly, have been removed from the stochastic error term. What's more, if both the likelihood that a woman takes twins to term, and a family's subsequent investment in child human capital, are positively correlated with positive health behaviours and other positive socioeconomic variables such as parental education, we would expect[12] that:

$$\hat{\beta}_1^{IV} > \hat{\beta}_1^{IV,H} > \hat{\beta}_1^{IV,S+H} > \beta_1. \tag{2.9}$$

It should be noted in the above series of inequalities that even conditional upon socioeconomic and health variables, IV estimation will *not* result in a consistent estimate of β_1 if twinning is correlated with unobservable elements in u_{ij}^*. We return to this point, and how to bound β_1 in the following sub-section.

2.3.2 Bounding the Q–Q Trade-off

In the previous subsection, Q–Q estimation using twins is motivated in equations (2.1) and (2.2). Consistent IV estimation imposes the (strong) prior belief that twin births can be excluded from the second stage equation, or that the sign of γ in the following is equal to zero:

$$educ_{ij} = \beta_1 fert_j + \gamma twin_j + \boldsymbol{X}\boldsymbol{\beta} + u_{ij}. \tag{2.10}$$

[12] This follows from the remainder terms in (2.7) and (2.8). If twinning is positively correlated with the omitted error term in the second stage equation, $twin_j u_{ij}^S > 0$, $twin_j u_{ij}^H > 0$ and $twin_j u_{ij}^* > 0$, and resultingly, any IV estimates of β_1 will be biased upwards.

As we discuss above, this will not be the case if maternal health controls omitted from (2.1) are correlated with both the likelihood of taking twin conceptions to term, and with eventual measures of child quality. Specifically, γ will reflect the effect of unobserved maternal health on child quality, interacted with the degree to which twin mothers are healthier than non-twin mothers.[13]

However, even in cases such as this where we are not confident that $\gamma = 0$, we can still estimate bounds on the Q–Q tradeoff if we are confident in making some statement of prior belief about the distribution from which γ is drawn. Conley et al. (2012) describe such a process, which they refer to as *plausible exogeneity*.[14] We invoke this terminology here, and refer to twins as a plausibly exogenous event, implying that we have reason to believe that γ may be close to, but not necessarily precisely equal to, zero. Specifically, we are concerned that healthier mothers are more likely to give birth to twins, and, all else constant, healthier mothers are more likely to be able to invest more in their children post-pregnancy. Thus, γ, the coefficient on twins in (2.10), reflects the interaction between the partial correlation of a mother's health and her likelihood of giving birth to twins, which we denote ϕ_t, and the partial correlation between her health and child quality, denoted ϕ_q.

In this paper we estimate β_1 under a range of assumptions regarding the true nature of γ. Firstly we estimate β_1 by simply assuming a support assumption for γ: namely that γ falls between zero (implying instrumental validity) and some (positive) number δ:

$$\gamma \in [0, \delta]. \tag{2.11}$$

This is a relatively weak assumption, however, as Conley et al. (2012) show, it allows for us to recover a 'union of confidence intervals' (hereafter UCI) for estimates of β_1 over the entire support of γ. This UCI, then, provides bounds for β_1 even in the case that twin exogeneity does not strictly hold. We also estimate by imposing a stronger prior: specifically we fully specify the distribution of γ as:

$$\gamma \sim \mathcal{N}(\mu_\delta, \sigma_\delta^2). \tag{2.12}$$

This stronger assumption allows for a tighter estimate of the bounds on β_1. Conley et al. (2012) provide a full derivation of this result, and we follow them in referring to this as a local-to-zero

[13] This is a useful way to think about the IV bias. If one or other of these conditional correlations is equal to zero, IV estimates will not be biased. If twin mothers are not more healthy than non-twin mothers, then our instrumental validity concerns (due to health) can be dismissed. Similarly, if twin mothers *are* healthier than non-twin mothers, but maternal health has no effect on child quality, instrumental validity concerns will once again be irrelevant.

[14] Nevo and Rosen (2012) propose a bounds estimate for IV of a similar nature. Inference in this case depends on assumptions about the direction of correlation between the instrument and the error term. They denote this as ρ_{zju}, which is analogous to the sign of γ in Conley et al.'s framework which we apply here.

(LTZ) approximation.

Assumptions (2.11) and (2.12) depend upon the values of δ (and $\mu_\delta, \sigma^2_\delta$) that we believe hold in the case of twinning and the Q–Q equation. In order to form a prior for γ and its distribution, we turn to a specific case which allows us to causally estimate each of ϕ_t and ϕ_q, and hence the coefficient γ. By using a well documented shock to mother's health—the arrival of sulfonamide antibiotics in the USA in 1937 (Bhalotra and Venkataramani, 2014; Jayachandran et al., 2010)—we can observe, firstly, whether exogenous improvements in maternal health do lead to improvements in child quality. Similarly, using the same data we can estimate the degree to which the health of twin mothers exceeds the health of non-twin mothers, scaling the direct effect of maternal health on quality to estimate γ. We provide a comprehensive discussion and derivation in of this method in appendix 2.B, and return to estimate of γ and bounds on our estimand of interest β_1 in section 2.5.4.

2.4 Data and Estimation Samples

We consult two data sets for our main IV and OLS analysis, and a large number of auxiliary datasets when only considering the link between twinning and maternal health. In order to estimate the (health augmented) specification 2.1, we require information on each child's siblings, his or her mother's characteristics, as well as the child's eventual 'quality' measure. We focus our empirical tests on two principal datasets which contain all of these variables. Firstly, the Demographic and Health Surveys (DHS), which have been applied over 20 years in a range of developing countries, and secondly, the United States National Health Interview Surveys (NHIS), which has been applied in an identical way from 2004-2014. In what follows, we describe the characteristics of these datasets. A comprehensive description of all data, including that which is used only to illustrate twinning and its relation to maternal characteristics, is provided in data appendix 2.A.

2.4.1 Data

The DHS

The DHS are a set of nationally representative surveys which have been administered in low- and middle-income countries between 1985 and the present. Women aged between 15–49 in surveyed

households respond to an in-depth series of questions reporting their full fertility history (listing all surviving and non-surviving children), their actual and desired contraceptive use and number of births, education level, marital status, plus the measurement of a number of health endowments such as height and body mass index. For all other members living in the household, a shorter series of responses are recorded, including the individual's educational attainment.

This results in two distinct sets of data to be merged. One database contains one line for each birth reported by every 15–49 year-old woman surveyed with a limited number of child-level covariates such as the child's date of birth, type of birth (single or multiple), and the child's survival status. The other database contains one line for each member currently living in the survey household. This database includes each member's educational status. We merge these two databases where all children who live in the same household as their mother merge without loss. We are thus able to generate data for the educational attainment of each of a woman's children currently residing in the household as well as their mother's health and educational status. This database is selected in two ways: firstly it only contains children who have survived up until the survey date, and secondly it only contains children who have remained living in the same household as their mother. We drop from our sample children aged 18 and over, due to concerns that these will *not* be representative of the general population.

We pool all publicly available DHS data resulting in microdata on 3,297,318 children ever-born to women who responded fully to any DHS survey. A full list of the DHS countries and years of surveys which make up this sample is provided in an online appendix (table 2.24). Of the 3,297,318 offspring reported in survey data, 2,033,510 remain living in the same household as their mother. The majority of these 2,033,510 children are aged 18 and under (92.96%) and hence make up our principal estimation sample (in future we will refer to this as the 'household sample'). The remaining 1,263,808 offspring were not recorded as living in the same household as their mother. Of these children not in the household, and hence for whom education is not recorded, the majority (53.9%) were aged over 18 or had died prior to the date of survey.[15]

The NHIS

The National Health Interview Survey (NHIS) is a yearly survey, conducted from 1957 and ongoing as at 2015, with participants drawn from each of the 50 US States as well as the District

[15]Children aged under 18 who are alive but not living in the same household as their mother are statistically quite different to those children who do remain in the household. In our data sample, they are on average 2.7 years older, born to less educated and younger mothers, and are slightly more likely to be males.

of Columbia each year.[16] We pool all survey data from 2004 until 2014, resulting in data on 119,111 mothers and 227,213 children. We focus on this period given that prior to 2004, changes in a number of key variables make it difficult to compare between years, and post-1996 the survey was considerably revised.

Each set of surveys is collected at the level of the household. For our analysis we use all households which consist of a biological mother and her children, whether or not any father is present. For all children who remain in the household, the survey records total fertility. We infer twin status by assuming that all children who share a birth month, birth year and biological mother must be twins. For each child and mother, we have a number of measures of usage of health care along with a self-reported measure for health status, whether or not the mother smokes, and the level of completed education (at the time of the survey) of mothers and children. Once again, we subset to children aged below 18, and for education measures, children who are aged above 6 years old, and hence who are able to be enrolled in school. Descriptive statistics of this and DHS data are provided in section 2.4.3.

2.4.2 Estimation Samples

The quasi-experimental variation exploited in twin studies is to leverage the effect of an unexpected additional child on siblings who were born before the extra birth. Thus, all first-born children in families of at least two births are split into two groups: the treatment group, which consists of first-born children in a family whose second birth results in two children (twins), and a control group consisting of first-born children in families where the second birth results in just one child.

For our main IV specification, we follow the existing literature in defining birth-order specific estimation samples, as laid out in the preceding paragraph. These samples are referred to as the 2+, 3+, and 4+ samples. These samples are defined $\forall\ n \in \{2, 3, 4\}$ such that they include first-born to $n-1$ born children in families with at least n births.[17] As an example, the 2+ sample (described in the previous paragraph) consists of first-borns in families with at least two births, and the 3+ sample consists of first- and second-borns in families with at least 3 births. Such a sample decision is important when estimating the Q–Q trade–off using twinning as an instrument. Given that family size is endogenously chosen by parents and rates of twin birth are

[16] The NHIS has a survey design to oversample Hispanic and African American people. We use NHIS-specific probability weights in all analyses.

[17] Existing studies such as Angrist et al. (2010) focus mainly on the 2+ and 3+ samples. Given the higher fertility in the DHS data, we also include higher a birth-order group, 4+.

not constant by birth-order, twin-births will occur more frequently in families that have a higher fertility preference (see figure 2.2). This point is addressed by (among others) Rosenzweig and Wolpin (1980a) and Black et al. (2005) who first suggested combining $n+$ groups with twinning at birth order n as a way to ensure that twin and non-twin families in the sample would have similar fertility preferences.

2.4.3 Descriptive Statistics

Table 2.2 provides summary statistics for DHS data, and table 2.3 describes NHIS data. Fertility and maternal characteristics are described at the level of the mother, while child education and survival are described at the level of the child. The number of observations at each level is provided at the bottom of the table.

For DHS data, survey countries are classified according to country income level in order to allow for a disaggregation of Q–Q results by income group.[18] We present summary statistics by birth type (singleton or twin), and by country income status. Twin births make up 1.85% of all births. A simple comparison of means suggests that healthy mothers (as proxied by height, BMI and probability of being underweight) are more likely to give birth to twins, and that twin births are more likely to occur in low-income countries. This apparent contradiction can be explained given that twins are (both mechanically and biologically) a positive function of fertility, and fertility is higher in the low-income sample. Figure 2.2 describes this positive relationship: while twins account for less than 1% of all first-borns, they account for greater than 4% of all tenth-born children. As expected, twin families are larger than non-twin families. Figure 2.3 describes total fertility in twin and non-twin families. The distribution of family size in families where at least one twin birth has occurred dominates the corresponding distribution for all-singleton families. This is expected given imperfect fertility control and—even were fertility perfectly controlled by families—given that some twins will occur on a family's final desired birth. Such a result is required for instrumental relevance when using twining to estimate a Q–Q trade-off.

Similar patterns are observed when turning to NHIS data. Twin mothers have (unconditionally) higher education, and greater health stocks as measured by BMI, percent underweight (BMI<18.5) and self-reported health. This is despite having higher total fertility, and being somewhat older. The proportion of twins in the USA sample is slightly higher (at 2.57%). This

[18] This classification is obtained from the World Bank, with DHS surveyed countries falling into two broad groups based on their GNI per capita at the moment of the DHS survey. These groups consist of countries classed as low-income economies, and countries classed as middle-income economies (either lower-middle or upper-middle). Details regarding this classification can be found in Appendix Table 2.24.

value is very similar to the value reported in the universe of all births in the United States from birth certificate data (see for example figure 2.5).

Child 'quality' is measured using each child's educational attainment. Our principal outcome variable in each case is a standarised score for schooling (Z-Score). This Z-Score is calculated by comparing each child's total years of completed education to his or her cohort or reference. In the case of DHS data, this cohort is made up of all children in the same country and birth cohort, while in NHIS data, it is made up of children with the same month and year of birth. The use of a standardised score rather than just total years of education allows us to express all effect-sizes in terms of a one standard deviation increase in total educational attainment.

2.5 Results

2.5.1 Twinning

In table 2.4 we present results of a regression of a child's twin status (one if a twin, zero if a singleton) on their mother's health, education, and a range of other demographic and family characteristics.[19] These results suggest that twin births are not random, even after conditioning on maternal age and child birth order as is typical in the recent twin literature summarised in table 2.1. The inclusion of a full set of country and year-of-birth dummies (not displayed in table 2.4) will capture any systematic trend in the frequency of twin births across time or regions, and country dummies will absorb all time invariant differences in the probability of a twin birth across countries. The estimated coefficients and signs support the idea discussed in section 2.3 that higher investments (for example in maternal health) required to maintain multiple healthy fetuses in utero may result in non-random twin births. We return to the mechanism by which twin selection may occur in the following section.

Initially, results from the pooled DHS data are presented as this provides a particularly large sample with which to test the hypothesis of twin exogeneity. This is presented in table 2.4 column (1) and provides considerable evidence that live twin births are related to family choice variables such as education (tests for the joint significance of socioeconomic variables and health variables are rejected with p-values of <0.00). Regressions displayed here are estimated by OLS, however

[19] In our principal specification, the full set of controls are country, child year of birth, and age dummies; a cubic function of mother's age at time of birth; mother's age at time of first birth; mother's education and education squared; and mother's height and BMI. We cluster standard errors at the level of the mother.

are robust to alternative functional forms and estimation methods.[20]

Columns (2)–(5) suggest that these results are not due only to the most low income countries, or the post-IVF time period, especially when considering maternal health. Given that the frequency of multiple births increases in cases where the mother undergoes fertility treatment, column (5) presents regression results for births in a period not potentially affected by IVF.[21] Pre- and post-1990 results are qualitatively similar although education is no longer significant prior to 1990 (in the smaller sample). When examining coefficients on mother's height and BMI, measures of health stocks are positively correlated with twinning regardless of the sample. Similarly, this result is not driven by a particular country or region. Figure 2.4 provides evidence that healthy women (as proxied by height) are significantly more likely to have twins in nearly all of the 68 countries included in DHS surveys. Along with higher average rates of twinning in countries with taller women, a positive within-country gradient exists, with taller women in a given country more likely to have twins than their shorter counterparts. The size of DHS estimates are considerable. Increasing a woman's height by 1 standard deviation increases the probability of twinning by 0.44% (as compared to a mean rate of twins of 1.85%).

Results for the infant mortality test described in section 2.3.1 are presented in table 2.5. In row 1, we regress IMR for first-borns on the twin status of second-borns.[22] Mothers who have second-born twins have much lower rates of infant mortality *before* the twins than women who had second-born singletons. This alternative test supports the regression results from table 2.4, providing additional evidence that mothers of twins are healthier when considering entirely pre-determined measures of health. Similarly, the rates of infant mortality among first- and second-borns are much lower in families of women who have third-born twins than in those who have third-born singletons. This holds for all parity levels examined.

Much of the existing twin literature focuses on the USA, or other developed countries. In table 2.6, we provide similar regressions for women in the USA based on the full set of NHIS surveys. These results show that twins are not as good as random, even in the context of a country with a more developed healthcare system and social safety nets. Taller mothers, heavier mothers, and mothers who don't smoke prior to conception (a positive health behaviour) are

[20] Significant and quantitatively similar results are found if a logit model is estimated rather than a linear probability model, and when running separate models for twinning at each birth order. Similarly, if we run the regression at the level of the mother or include any combination of fertility measures, similar patterns are observed. Alternatively, rather than running a regression we can run (unconditional) balance of characteristics tests by twin status. These are available in the online appendix (table 2.25). The findings are similar.

[21] In order to be conservative, we estimate for the period preceeding 1990, the date which coincides with the first reported successful use of IVF in South Africa, an early-adopter among DHS countries.

[22] IMR is defined as 1 for children who die before their first birthday. We remove from the sample any children who were not yet 1 at the time of the following birth, as these children have not yet been entirely exposed to the risk of infant mortality.

significantly more likely to have twins.

The dependence of twinning on positive maternal health stocks and behaviours is a consistent and quantitatively important phenomenon in all data sets we have examined. We have compiled data and run similar regressions using vital statistics data from the USA (NVSS), Spain, Scotland and Sweden, and additional survey data from Chile and the United Kingdom (see tables 2.13–2.16 for results). In each case, the probability of twinning increases as mothers become more healthy and are less likely to engage in risky health behaviours before and during pregnancy. Along with the results described in tables 2.4 and 2.6, these additional sources of data show that mothers who consume alcohol, tobacco or other drugs, who suffer from chronic disease, stress during the second or third trimester of pregnancy, or who have less access to prenatal care are significantly less likely to give birth to twins. Similarly, using the methodology and data described in Quintana-Domeque and Ródenas-Serrano (2014), we find that women exposed to exogenous increases in stress during the second and third trimester of pregnancy (from ETA bombs in Spain) are significantly less likely to have live twin births (table 2.17).

Finally, if twinning is related to positive health stocks and behaviours of prospective mothers and families, we can examine how rates of twinning respond to time-series variations in (female) health outcomes. While only suggestive, as many other environmental events or changes in the composition of mothers may explain changes in twinning, time-series evidence from the USA leads to similar conclusions. Figure 2.5 plots the rate of twinning from vital statistics data since birth type (single or multiple) was first recorded. Interestingly, the rate of twins has increased steadily over time, even before the advent of IVF and other fertilisation treatments. This is in line with increasing trends in female health over this period, which is proxied by female life expectancy and plotted in the same figure.

2.5.2 Selection Into Twinning: Mechanisms

In a wide variety of contexts, healthier women are more likely to give birth to twins. There are a number of competing hypotheses which may explain why this is the case. Firstly, it may simply be the case that healthier mothers are more likely to conceive twins. This may reflect some underlying biological process, such as that mediated by follicle stimulating hormone as discussed in Hall (2003). Secondly, conditional on conceiving twins, healthier mothers may be more likely to take both fetuses to term. Finally, it may be the case that (conditional on conceiving twins and taking them to term), healthier mothers may be more likely to survive the birth, and hence appear in survey or vital statistics data. In broad terms we will refer to these as the conception

mechanism, the gestation mechanism and the birth (survival) mechanism.

When considering IV estimates with twins, any of these processes is sufficient to invalidate causal inference insofar as observing twins depends upon hard-to-measure maternal behaviours and characteristics. Nonetheless, we may be interested in determing which of these are the relevant channels in explaining the results from the previous section. Particularly, the mechanism may be relevant when considering the use of the instrument. For example, if twins are less likely *only* due to selective maternal death, then as mothers become more likely to survive childbirth (ie moving from high maternal mortality countries to low maternal mortality countries), threats to instrumental validity become considerably less relevant.

We test these mechanisms below. In order to determine whether twin selection could be entirely explained by selective maternal survival, we follow Alderman et al. (2011) in simulating estimates under the counterfactual scenario that unhealthy women—who are more likely to die in childbirth—were all carrying twins. Using DHS data described in section 2.4.1, we observe a woman's height, BMI, pregnancy outcomes, and the maternal mortality status of all her sisters. As we do not observe health stocks of women who died in childbirth, we assume that her sister's health (height and BMI) is a reasonable proxy for the health of the woman who died within 42 days of giving birth (which is classed as a maternal death). Appendix figure 2.11 shows that maternal mortality is much higher among more unhealthy women. Women shorter than the mean height of 155.5 cm are considerably more likely to suffer maternal death, with this being particularly so below heights of 145cm.

To test the potential importance of maternal survival in explaining twin selection, we simulate observations for the number of women who, according to DHS data, would exist in the sample if it were not for the fact that they died in childbirth. We then examine the coefficients of interest in our twin regression (2.5), if all unhealthy women who died were pregnant with twins, while all healthy women who died were not. As this relies on a binary 'healthy vs unhealthy' distinction, we define this in various ways, based on height and BMI. These results are presented in table 2.7. The first column shows the estimated coefficients on height and BMI in the unaltered sample of women from DHS countries where maternal mortality data is available. In this sample, a BMI increase of 1 point is associated with a 0.046% increase in the probability of twinning. The remaining columns add in observations based on maternal mortality rates among sisters of surveyed women. For example, in the second column, we examine the effect of adding to the sample unhealthy and healthy women based on the maternal mortality rate in each group, and then assuming that all unhealthy women would give birth to twins, and all healthy mothers would not. As expected, this reduces the importance of positive maternal health in predicting twinning, with the coefficient on BMI falling from 0.0460 to 0.0437. The other columns continue

in this manner, however using continually less conservative assumptions in assigning members to the unhealthy group who are defined as giving birth to twins. Even in the final column, where the entire bottom half of the anthropometric distribution is assumed as being unhealthy, the coefficient on both height and BMI remains positive and significant.[23]

These results suggest that selective maternal death alone is not enough to explain why healthier mothers are more likely to have twins. Turning to the gestation mechanism, we are able to test whether less healthy women who are pregnant with twins are more likely to miscarry than healthier women who are also pregnant with twins. In order to do so, we require data which records miscarriages, mother's health, as well as the type of fetus (single or multiple) which the mother was pregnant with. Such data exists in administrative datasets for the USA (the National Vital Statistics System) and Spain. For each of these datasets, we thus run a series of regressions where miscarriage is the dependent variable, and the independent variables are measures of poor maternal health, whether the pregnancy is a twin pregnancy, and interactions between pregnancy type and poor maternal health. We would expect that both poor maternal health and a non-singleton pregnancy increase the likelihood of miscarriage, however we are interested in determing if more unhealthy mothers are *more* likely to miscarry twins than healthy mothers carrying twins. Thus, we are interested in testing if the coefficients on the interaction terms are significantly larger than zero.

These regression results are reported in table 2.8 (for the USA), and table 2.19 for Spain. As expected, less healthy mothers and mothers exhibiting risky behaviour such as smoking are more likely to miscarry. For example, in the USA, mothers who smoke prior to birth are 3% more likey to miscarry than non-smokers. Similarly, twins are approximately 10% more likely to miscarry than singleton births. In order to test the gestation mechanism, the coefficients of interest are interactions between twin pregnancies and maternal health. For nearly all health variables examined (smoking, alcohol consumption, and hypertension), less healthy women pregnant with twins are observed to be additionally more likley to miscarry. These results provide evidence in favour of the gestation mechanism, as conditional on conceiving twins, it is found that unhealthy women are more likely to miscarry these fetuses before taking them to term, and hence less likely to be subject to the fertility shock than mothers who engage in healthier behaviours or have higher health stocks.

[23]Examining selection in this way (as per Alderman et al. (2011)) is only one way to examine the effect of selection on estimated coefficients. An alternative measure as proposed by Lee (2009) involves trimming the control and treatment group (in our case unhealthy and healthy mothers), to account for differential selection by treatment status. This results in bounds estimates of the effect of treatment (good health) on the outcome variable (twinning). We report Lee bounds in appendix table 2.18, however note that these bounds are based on the assumption that treatment is random, which here it is not. Nonetheless, Lee (2009) bounds agree with the simulated estimates in table 2.7, providing further evidence that selective maternal survival is not enough to explain the correlation between maternal health and twinning.

2.5.3 The Twin Instrument and the Q–Q Trade-off

Results from section 2.5.1 show that twinning is not as good as random, even when conditioning on race, maternal age, and parental education. It is demonstrated in a large range of countries, that healthier mothers are more likely to give birth to twins. If these mothers also invest more after birth, the Q–Q trade-off will be under-estimated, and hence *less* negative than the true trade-off (we can see this by considering the term $twin_j u_{ij}^H$ in equation (2.7)). However, progressively including additional health controls in our first and second stage equations should drive IV estimates in the direction of identifying the true Q–Q trade-off, as we partially correct for the bias introduced by the positive correlation between twinning and maternal health.

Conversely, OLS estimates are typically thought to over-estimate the true magnitude of the Q–Q trade-off. If unobserved parental behaviours favour both lower fertility and higher child human capital,[24] OLS estimates of β_1 will be negatively biased, and hence *more* negative than the true trade-off. As items are removed from the error term and included in the principal equation to be estimated by OLS, we thus expect that these estimates should approach the true parameter from above.

We examine this intuition by estimating β_1 from (2.1) by OLS and IV. Our principal data is the large DHS sample, where quality is measured by school Z-score, a child's educational attainment compared to his or her country and year birth cohort. After considering DHS estimates we turn to NHIS data from USA.

Table 2.9 reports pooled OLS estimates from all DHS data, and in low and middle-income country groups. OLS is estimated separately by the 2+, 3+, and 4+ fertility groups, and included in appendix table 2.20. As is typically found in empirical studies of the Q–Q tradeoff, conditional correlations between family size and child outcome variables are negative, and strongly significant. The results in table 2.9 suggest that an additional sibling is associated with approximatelty a 0.1 s.d. decrease in standardised schooling outcomes. The magnitude of these estimates decreases as additional controls for maternal health and family socioeconomic variables are included in the regressions. This is in line with the hypothesised effect of these variables. To the degree that components from the error term are removed which are positively related to high desired quality and to low desired family size, the bias in OLS estimates should be reduced, and the estimated magnitude of the trade-off should move in the direction of zero. Of course, as long as any such variables remain unobserved and as part of the stochastic error term, OLS estimates *will not*

[24] As a (highly stylised) example, consider a prospective mother's eventual labour market plans. A mother who plans to join the labour market may prefer fewer children, facilitating more immediate labour force participation, but have more resources to invest in child quality.

converge to unbiased values. The following sections examine alternative estimation methods to bound the Q–Q trade-off.

The Twin Instrument: Estimates of the Q–Q Trade-off

In table 2.10, we turn to IV estimates using twins. As we outline in section 2.3.1, the assumption of 'as good as random' twin births is unlikely to hold, even when conditioning on the augmented set of controls proposed in (2.5). If this is the case, we will also be unable to consistently estimate β_1 using twin births.

However, it is likely that the $\hat{\beta}_1$ which we estimate using twin births will provide us with a strict lower bound of the magnitude of the Q–Q trade-off as outlined in (2.8). We expect that the bias in this estimate is due to those mothers who invest more in their children in utero, or who have greater initial health endowments, being more likely to give birth to twins, thus resulting in larger family sizes. At the same time, we expect healthier mothers to invest more in their children after birth, and hence have higher quality children. By relegating health variables to the error term, these two positive correlations will result in a positive bias on the fertility coefficient estimated via IV. In order to determine the effect that these omitted variables have on estimates of the Q–Q trade-off, we turn to results for equation (2.1), both first omitting, and the including, maternal health and socioeconomic variables.

The main specification is displayed in the top row of table 2.10, with separate columns for the 2+, 3+ and 4+ sample groups. For each parity group, the base case (controlling for maternal and child age, country, and year of birth) results in insignificant, and at times weakly positive, estimates of the effect of an additional birth on a child's educational attainment. These results suggest that the inclusion of maternal health and socioeconomic controls may be of considerable importance. Despite the lack of results when using the 'typical' set of twin controls from the twin Q–Q literature, including health (columns 2, 4 and 6) reduces point estimates on fertility from an effect of approximately 0% of a standard deviation, to -3 or -4% of a standard deviation in standardised educational attainment. Further, conditioning on maternal education results in slightly more precise estimates, suggesting a statistically significant (or close to statistically significant in the case of the 2+ sample) Q–Q trade-off of at least 3 or 4%.

Thus, in low- and middle-income country data, the inclusion of health indicators in the twin instrument does have an important effect on IV estimates, moving as hypothesised in (2.9). In table 2.11, we present identical estimates based upon NHIS data from the USA. As this survey

focuses on health, as well as standardised educational attainment, we examine the effect of additional siblings on the reported health of children. Results for both variables show similar patterns to those reported based on DHS data. Focusing on the 3+ group, the inclusion of health and socioeconomic controls results in OLS estimates moving closer to zero, and IV results further away from zero. In nearly all cases examined, at all parity levels in both DHS and NHIS, the movements of point estimates is precisely as hypothesised in the series of inequalities described in (2.9), suggesting considerable improvements in IV estimates from accounting for twin endogeneity. In the case of self reported health status, the inclusion of additional twin predictors is sufficient to result in statistically significant evidence in favour of the Q–Q trade-off for twins at birth three, while for education results, a significant Q–Q trade-off emerges for the 2+ group despite the reasonably imprecise standard errors of estimated coefficients. While for the other NHIS subgroups, the effect of a a twin does not always results in statistically significant results, it is important to note that the point estimates become more negative, moving in favour of a negative β_1, consistent with all other NHIS and DHS results.

More generally, we should not be surprised that confidence intervals on these point estimates are wide. It is well recognised that IV estimates are much less precise than their OLS counterparts, and this is more true with the case of the twin instrument, where a small ($\sim$2%) of the population are twins. The precision of twin estimates is discussed further in Angrist et al. (2010). While in the DHS data we are able to produce tighter confidence intervals given the large population of births, in USA NHIS data, despite the fact that point estimates fully support the hypothesis that a Q–Q trade-off does emerge upon correcting for non-random twin births, the relatively small sample size makes these estimates less precise than their DHS counterparts.

Heterogeneity

Theoretical derivations of the Q–Q model are based on the assumption that all children in a family are of the same quality. More recent work (for example the theoretical work of Aizer and Cunha (2012)) has loosened this assumption. Among other things, this allows for reinforcing behaviours by parents in child human capital investment decisions.[25] If this is the case, the coefficient β_1 may vary by children in the family. More generally, β_1 may be context specific, depending upon the returns to human capital in a given time-period or economy.

Empirically, we find that estimates of the Q–Q trade-off are heterogeneous across birth orders, country income level, and the gender of the child affected by the additional birth. The

[25] An empirical review of early life human capital and reinforcing versus compensating behaviour, (however not explicitly related to the Q–Q hypothesis), is provided by Almond and Mazumder (2013).

magnitude and significance of the results is lowest when considering the effect on the first-born child of moving from two to three births (the 2+ group), and higher when considering moving from three to four births or four to five births. However, in lower fertility environments the effect is, as expected, concentrated on lower birth orders. The third row of table 2.10 suggests that in middle-income countries the effect is largest on first borns, and progressively smaller, but still considerable, at higher birth orders.[26]

Estimates of the magnitude of the Q–Q trade-off by country income level suggest that the trade-off is considerably larger in middle- rather than low-income countries. In low-income countries point estimates on fertility suggest (insignificant) trade-offs centred around 2-3% of a standard deviation, while in middle-income countries results are significant, and considerably larger, reaching as much as 9% of a standard deviation: only slightly lower than OLS estimates for this group.

Similarly, effects of the Q–Q trade-off vary considerably depending upon a child's gender. In appendix table 2.22 we present regression results estimated separately by the gender of the index child. These results suggest that females may bear the brunt of additional births, with estimates being negative and significant for girls, while insignificant for male children. Interestingly, recent empirical of the Q–Q trade-off from other (middle income) contexts finds similar gender-biased results (Ponczek and Souza, 2012), also suggesting that girl children pay a steeper price.

2.5.4 Bounding the Q–Q Trade-off

The results from the previous subsection provide consistent estimates of β_1 via 2SLS *if* the full set of controls in the first and second stage equations completely account for those characteristics and behaviours which predict giving live birth to twins. However, given that we have shown that twinning is predicted by a wide range of health behaviours, and given that maternal health variables in these datasets do not exhaustively capture all aspects of health stocks and behaviours, it seems unlikely that all relevant variables are included in these specifications. As such, we turn to Conley et al.'s 2012 methodology to estimate bounds for the Q–Q trade-off.

As outlined in section 2.3.2 (and further in appendix 2.B), this involves the definition of some prior belief over the sign and magnitude that the coefficient on twinning would take in the

[26] In this paper we focus nearly exclusively on the internal validity of twins estimates (IV consistency). In recent work, Deheija et al. (2015) examine the external validity of an alternative fertility natural experiment, when examining various contexts using many countries' censuses. Like them, our results in this section suggest that homogeneity between countries and contexts (and even birth orders) is not seen when using twins to estimate the Q–Q trade-off.

structural equation 2.10. Results for DHS are displayed in figures 2.6 and 2.7, for 2+ and 3+ groups respectively. The corresponding figure for the 4+ group is provided as appendix figure 2.12. Similarly, figures 2.8 and 2.9 provide estimates using NHIS (USA) data. At each point on the horizontal axis of these figures, the bounds for β_1 are displayed, along with the corresponding point estimate under the assumption that γ is distributed $\sim U(0, \delta)$. Dashed lines present the 95% confidence interval, while the solid line represents the point estimate.

While this technique allows us to agnostically estimate bounds over a range of values for γ, we have thus far made no restriction over the true magnitude of δ. While in figures 2.6 and 2.7 we have assumed that this is less than 0.1 standard deviations, we can can form far more precise bounds by estimating γ directly. As discussed more extensively in appendix 2.B, this requires us to (causally) estimate the effect of maternal health shocks on twinning, and the degree to which twin mothers are healthier than non-twin mothers. By taking the product of these, we can calculate the partial correlation between twinning (which occurs with higher frequency for healthier mothers), and child quality, and then we can plug our estimate of γ (where a non-zero value for γ implies instrumental invalidity) in Conley et al.'s bound estimator.

We follow the specification outlined in Bhalotra and Venkataramani (2014) (and appendix 2.B of this paper) to estimate the effect of the positive health shock associated with the mother's receipt of Sulfanide drugs on child quality (school Z-score), as well as the health differential between twin and non-twin mothers. These estimates are presented in appendix table 2.23. Based on these estimates, we find that γ (the direct effect of coming from a twin family on child quality) is around 1% of a standard deviation (0.0091 s.d.). For what remains of this section, we use this estimate of the violation of the exclusion restriction to calculate the true bounds on the Q–Q trade-off.

In table 2.12 we provide estimates of Conley bounds using our more precisely estimated γ. These are our preferred bounds estimates for β_1. As described in appendix 2.B, for the UCI approach this implies an assumption that $\gamma \in [0, 2\hat{\gamma}]$ (or $\gamma \in [0, 2 \times 0.0091]$). For the LTZ approach, we assume that $\gamma \sim N(\mu_{\hat{\gamma}}, \sigma_{\hat{\gamma}})$. The assumption of normality is driven by resampling (bootstrap) estimates of $\hat{\gamma}$, which allows us to construct a distribution for γ. This is then tested for equality against a normal distribution, and the null hypothesis of different distributions is not rejected with a p-value of 0.1805. The empirical and analytical distributions for γ that are applied in the LTZ method are displayed in appendix figure 2.13b-2.13a.

In all cases, our preferred bounds estimates are those in the right-hand columns of table 2.12, as these allow us to produce more efficient bounds based on the estimated bootstrap distribution.

For DHS results, these bounds estimates are informative for the three-plus and four-plus groups, and very close to being informative for the 2+ group also. For these groups, both the upper and lower bound fall entirely below zero. In all cases for DHS, the average trade-off for the pooled group (the middle of these bounds) falls between 3% and 5% of a standard deviation. An additional sibling thus *does* appear to affect a child's educational attainment, and this is of the order of magnitude of 3-5% of a standard deviation compared to his or her peers. Similarly, in the USA, despite having much wider bounds, these are informative for 2+ (education) and 3+ and 4+ (health). Once again, we generally find a mid-point average effect of approximately 5% of a standard deviation from an additional birth, although these values are slightly higher (and less precise) when considering education rather than health.

2.6 Conclusion

Twin births are not random. Rather, they appear to be far from it, in a wide variety of environments, time periods and contexts. Based on a considerable body of evidence compiled from vital statistics and survey data from low- and high-income countries, we demonstrate that mothers with greater health stocks, those who engage in positive health-related behaviours, and those living in healther environments are much more likely to take twins to term. It is demonstrated that these mothers are healthier *prior* to twinning, and this results in a greater likelihood of taking twins to term, conditional upon conceiving two fetuses.

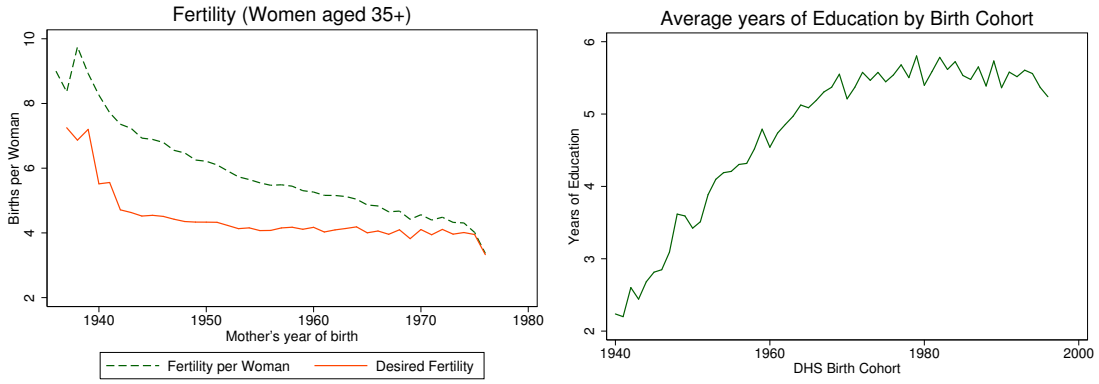
These results have important implications for empirical work which aims to identify the causal effect of child quantity (more siblings) on child quality (higher human capial). The existing evidence from the Q–Q literature is mixed. A range of studies which estimate the effect of child quantity on child quality within a family find that the effect is small, or frequently not statistically different from zero. By assembling large datasets linking a child's human capital outcomes with her mother's health—in both the developing world and the USA—we show that partially correcting for twin endogeneity is sufficient to push estimates of the trade-off up by about 3%-4% of a standard deviation, potentially explaining the lack of significant results in the existing literature. Using partial identification to bound the effect of child quantity on child quality suggests that the *true* effect size, once accounting for the entire health differential in favour of twin families, may even be as high as 8% of a standard deviation, though is typically centered around 5% of a standard deviation.

We are able to conclude that additional unexpected births do have quantitatively important effects on their siblings' educational outcomes. A 5% of a standard deviation increase is equivalent

to an additional 0.2 years in the classroom. While the true effect of these 0.2 years of course depend on the quality of education imparted, time in school is at the very least a necessary condition for more general increases in learning outcomes. The implications of these findings are wide-reaching, both in terms of the vindication of Beckerian theory, and particularly, in an applied sense, when considering human well-being in developing countries which are yet to fully pass through the demographic transition.

Figures

Figure 2.1: Education and Fertility



(a) Trends in Fertility

(b) Trend in Education

Note to figure 2.1: Cohorts are made up of all individuals from the DHS who are over 35 years (for fertility), and over 15 years (for education). In each case the sample is restricted to those who have approximately completed fertility and education respectively.

Figure 2.2: Proportion of Twins by Birth Order

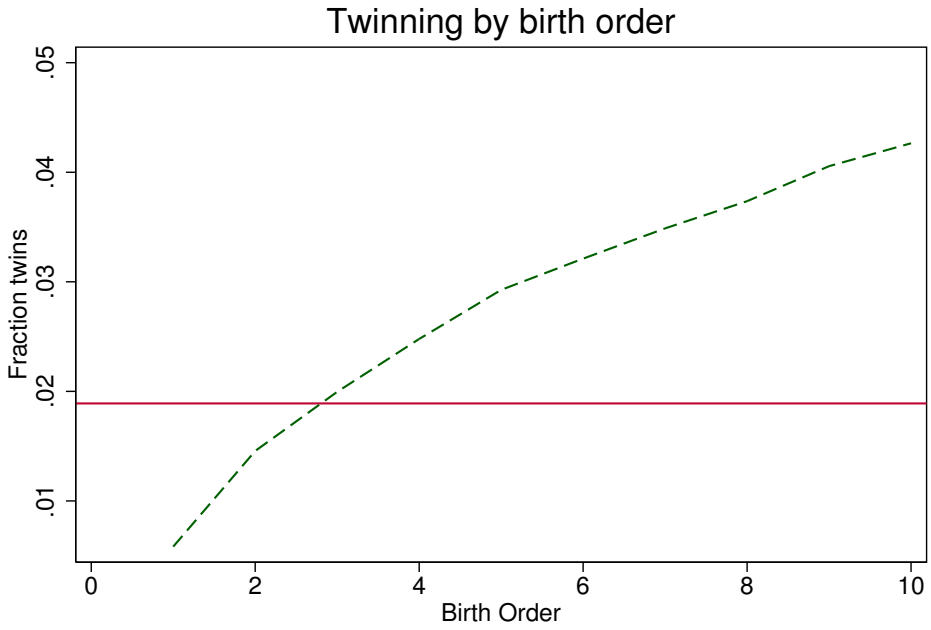


Figure 2.3: Twin Births and Total Fertility

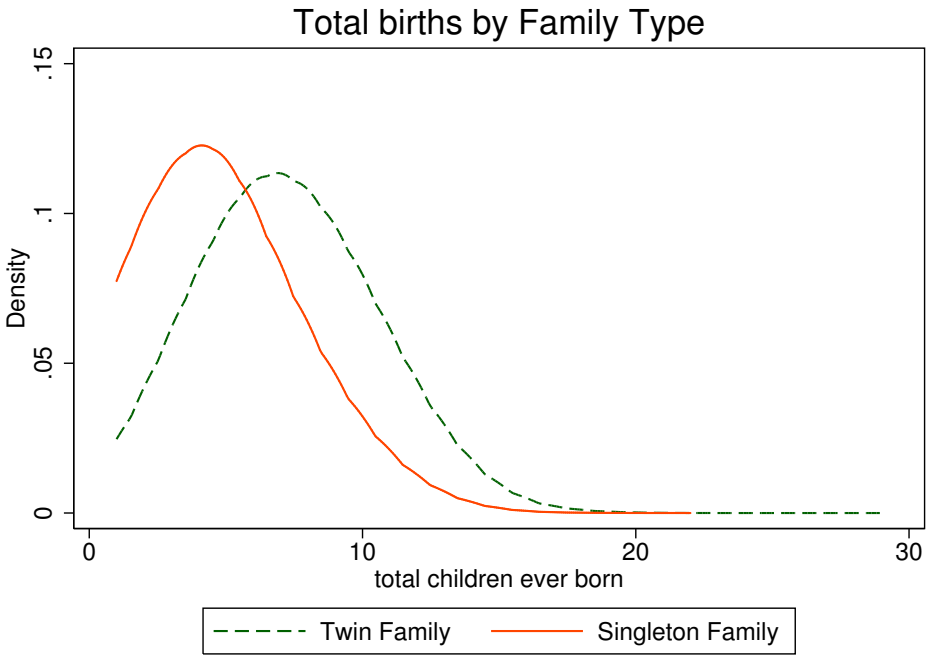
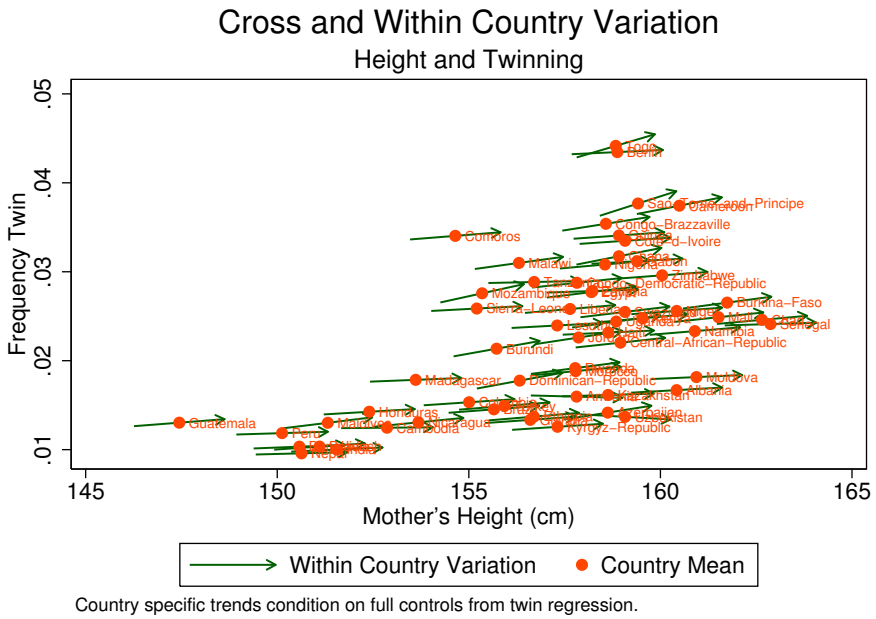
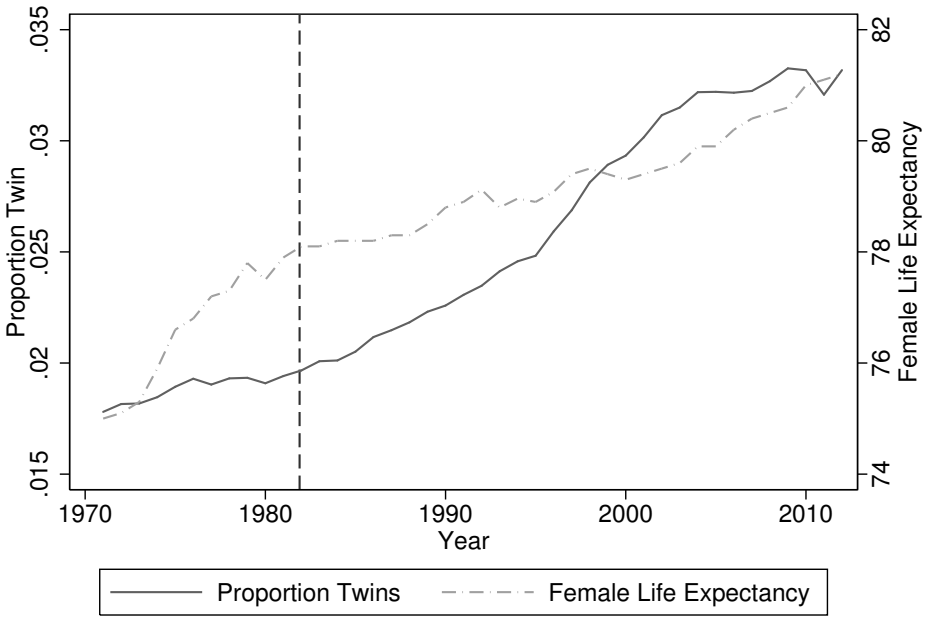


Figure 2.4: Intra- and Inter-country trends: height and twinning



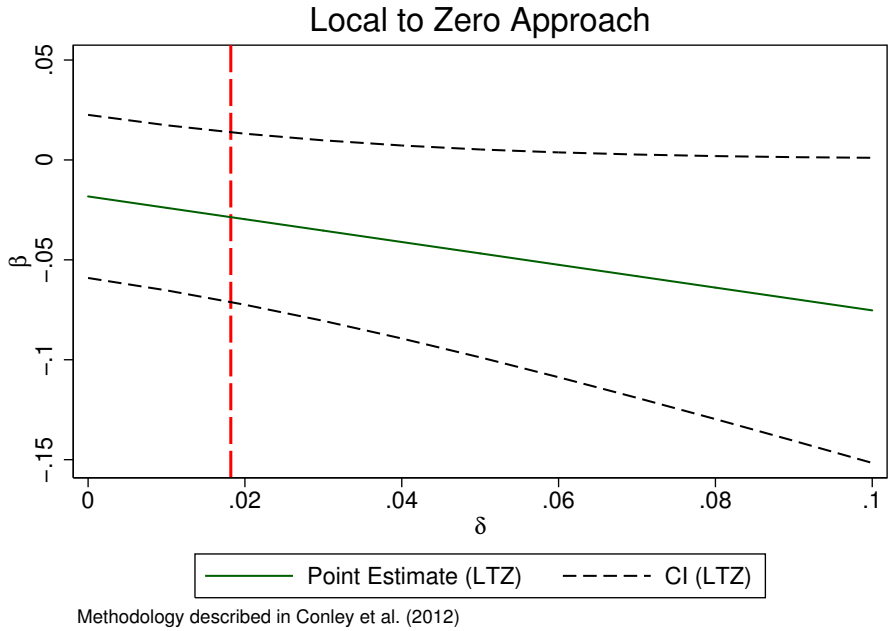
Country specific trends condition on full controls from twin regression.

Figure 2.5: Proportion of Twins of All Births (USA)



Twin data from NVSS Birth Certificate Data. Female Life Expectancy data from World Bank.
Dotted line represents first ever IVF birth in USA.

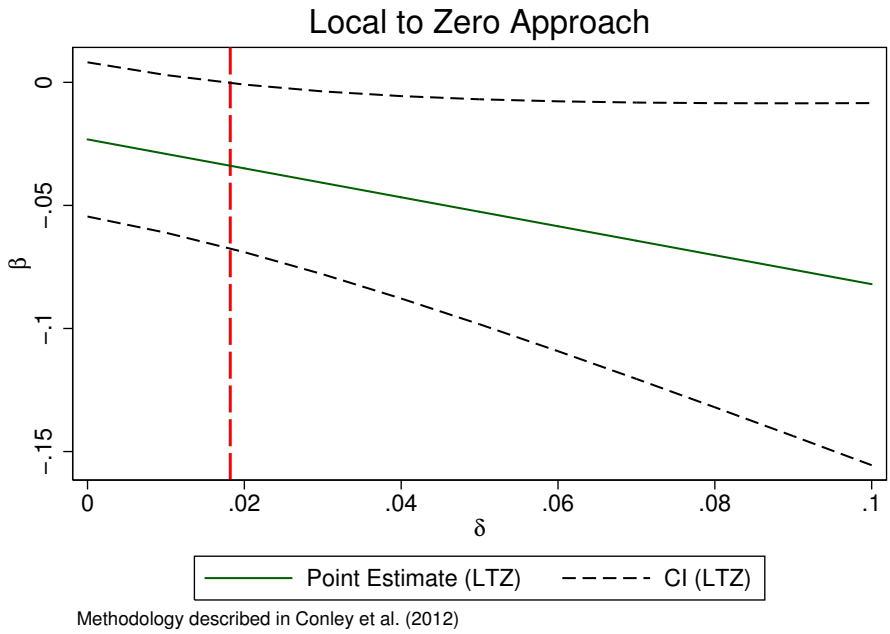
Figure 2.6: Relaxing Strict Exogeneity (two plus)



Methodology described in Conley et al. (2012)

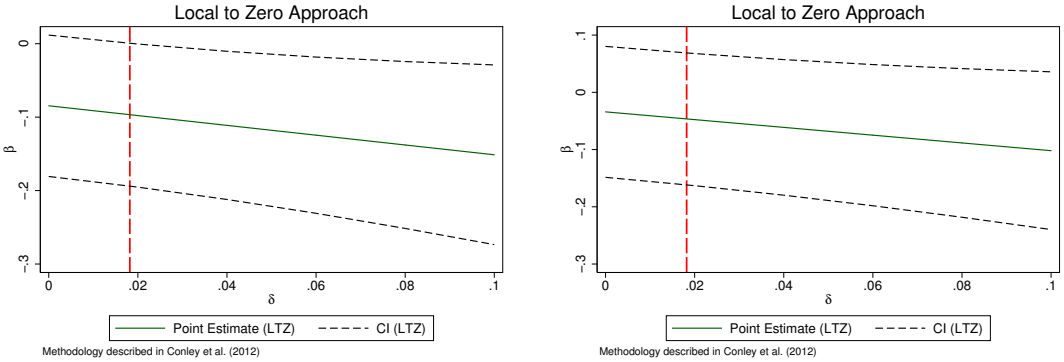
Note to figure 2.6: See note to Figure 2.7

Figure 2.7: Relaxing Strict Exogeneity (three plus)



Note to figure 2.7: Confidence intervals and point estimates are calculated according to Conley et al. (2012). Estimates reflect a range of priors regarding the validity of the exclusion restriction required to consistently estimate $\hat{\beta}_{fert}$ using twinning in a 2SLS framework. The local to zero (LTZ) approach applied here assumes that γ, the sign on the instrument when included in the first stage, is distributed $\gamma \sim U(0, \delta)$. The vertical dashed line indicates $2\times\hat{\gamma}$, the point at which the estimate for γ lies precisely halfway between [0,δ]. Further discussion is provided in appendix 2.B and table 2.12.

Figure 2.8: Plausibly Exogenous Bounds: School Z-Score (USA)

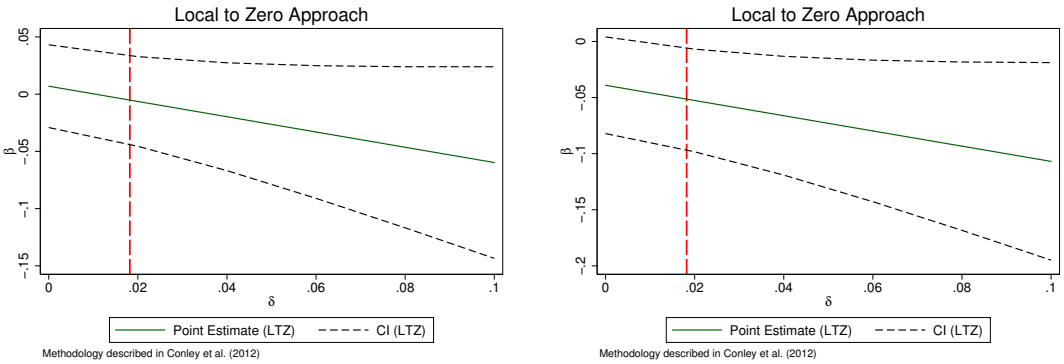


(a) Two Plus

(b) Three Plus

Notes to figure 2.8: See notes to figure 2.7.

Figure 2.9: Plausibly Exogenous Bounds: Excellent Health (USA)



(a) Two Plus

(b) Three Plus

Notes to figure 2.9: See notes to figure 2.7.

Tables

Begin overleaf.

Table 2.1: Fertility and the Twin Instrument: Literature

Author	Data, Period	Controls Included	Sample	Estimates	
				OLS	IV
(1) Black et al. (2005)	Norway matched administrative files of individuals aged 16-74 during 1986-2000, (children > 25 years). Outcome is completed years of education.	Age, parents' age, parents' education, sex.	Two Plus	-0.060	-0.038
				(0.003)	(0.047)
			Three Plus	-0.076	-0.016
				(0.004)	(0.044)
			Four Plus	-0.059	-0.024
				(0.006)	(0.059)
(2) Cáceres-Delpiano (2006)	USA 1980 Census Five-Percent Public Use Micro Sample. Children aged 6-16 years. Outcome (reported here) is an indicator of whether the child is behind his or her cohort.	Age, state of residence, mother's education, race, mother's age, sex.	Two Plus	0.011	0.002
				(0.000)	(0.003)
			Three Plus	0.017	0.010
				(0.001)	(0.006)
(3) Angrist et al. (2010)	Israel 20% public-use microdata samples from 1995 and 1983 censuses, 18-60 year old respondents. Outcome (reported here) is highest grade completed.	Age, missing month of birth, mother's age, age at first birth and age at immigration, mother's and father's place of birth, and census year.	Two Plus	-0.145	0.174
				(0.005)	(0.166)
			Three Plus	-0.143	0.167
				(0.005)	(0.117)
(4) Li et al. (2008)	The 1 percent sample of the 1990 Chinese Population Census. Subjects are 6-17 year olds with mothers who are 35 years of age or younger. Outcome (reported here) is years of schooling.	Child age, gender, ethnic group, birth order, and place of residence. Parental age and educational level.	Two Plus	-0.031	0.002
				(-29.6)†	(0.18)†
			Three Plus	-0.038	-0.024
				(-21.4)†	(-1.70)†
(5) Fitzsimons and Malde (2014)	Mexican Survey data (ENCASEH) from 1996-1999. Subjects are 12-17 year olds. Outcome (reported here) is years of schooling.	Parent's age, parents' years of schooling and schooling dummies, birth spacing, household goods (rooms, land, water, etc).	Two Plus	-0.020	-0.019
				(0.001)	(0.015)
			Three Plus	-0.020	0.007
				(0.001)	(0.025)
			Four Plus	-0.018	-0.032
				(0.002)	(0.036)

Author	Data, Period	Controls Included	Sample	Estimates	
				OLS	IV
(6) Rosenzweig and Zhang (2009)	The Chinese Child Twins Survey (CCTS), 2002-2003. Individuals selected from twins' (aged 7-18) and non-twin households. Outcome (reported here) is years of schooling	Mother's age at time of birth, child gender and age.	Reduced Form		-0.307 (1.92)†
			Reduced Form + Bwt		-0.225 (1.31)†
(7) Ponczek and Souza (2012)	1991 Brazilian Census microdata, 10 and 20% sample. Children of 10-15 years, and 18-20 years old. Outcome reported here is years of school completed.	Child's gender, age and race controls,; mother and family head's years of schooling, and age.	Two Plus (M)	-0.233 (0.010)	-0.137 (0.146)
			Two Plus (F)	-0.277 (0.015)	-0.372 (0.198)
			Three Plus (M)	-0.230 (0.010)	-0.060 (0.164)
			Three Plus (F)	-0.283 (0.015)	-0.634 (0.194)

Notes: Individual sources discussed further in the body of the text. Estimates reported in each study are presented along with their standard errors in parenthesis. Parentheses marked as † contain the t-statistic rather than the standard error.

Table 2.2: Summary Statistics

	Low Income		Middle Income		
	Single	Twins	Single	Twins	All
FERTILITY					
Fertility	3.670	6.093	3.348	5.425	3.609
	(2.365)	(2.582)	(2.272)	(2.609)	(2.372)
Desired Family Size	4.182	5.296	3.340	4.128	3.892
	(2.500)	(2.832)	(2.083)	(2.498)	(2.403)
Fraction Twin	0.0194		0.0173		0.0185
	(0.1379)		(0.1306)		(0.1348)
Birth Order Twin	4.680		4.017		4.420
	(2.463)		(2.370)		(2.448)
MOTHER'S CHARACTERISTICS					
Age	30.92	34.18	31.99	34.94	31.40
	(7.980)	(7.166)	(8.105)	(7.156)	(8.038)
Education	3.984	3.337	6.817	6.115	5.030
	(4.370)	(4.033)	(4.794)	(5.048)	(4.735)
Height	155.6	157.7	155.6	157.2	155.7
	(7.084)	(6.987)	(6.956)	(6.957)	(7.042)
BMI	21.89	22.47	25.83	26.50	23.38
	(3.983)	(4.098)	(5.066)	(5.437)	(4.822)
Pr(BMI)<18.5	0.172	0.123	0.0344	0.0276	0.119
	(0.377)	(0.328)	(0.182)	(0.164)	(0.324)
Actual Births>Desired	0.297	0.513	0.319	0.567	0.311
	(0.457)	(0.500)	(0.466)	(0.495)	(0.463)
CHILDREN'S OUTCOMES					
Education (Years)	3.695	3.212	5.438	4.999	4.465
	(3.581)	(3.270)	(3.859)	(3.734)	(3.805)
Education (Z-Score)	-0.00869	-0.0130	0.0121	-0.0366	0.000177
	(1.001)	(0.961)	(0.998)	(0.987)	(1.000)
No Education (Percent)	0.200	0.213	0.0634	0.0766	0.140
	(0.400)	(0.409)	(0.244)	(0.266)	(0.346)
Infant Mortality	0.0860	0.165	0.0489	0.113	0.0758
	(0.280)	(0.371)	(0.216)	(0.316)	(0.265)
Child Mortality	0.122	0.208	0.0616	0.131	0.102
	(0.327)	(0.406)	(0.240)	(0.338)	(0.303)
Number of Countries	42	42	34	34	68
Number of Mothers	491,905	7,457	297,413	4,317	850,032
Number of Children (Education)	1,176,513	25,003	714,751	14,333	1,930,600
Number of Children (Ever Born)	1,716,247	43,866	940,204	21,302	2,721,619

NOTES: Summary statistics are presented for the full estimation sample consisting of all children 18 years of age and under born to the 850,032 mothers responding to any publicly available DHS survey. Group means are presented with standard deviation below in parenthesis. Education is reported as total years attained, and Z-score presents educational attainment relative to country and cohort (mean 0, std deviation 1). Infant mortality refers to the proportion of children who die before 1 year of age, while child mortality refers to the proportion who die before 5 years. Maternal height is reported in centimetres, and BMI is weight in kilograms over height in metres squared. For a full list of country and years of survey, see appendix table 2.24.

Table 2.3: Summary Statistics (NHIS)

	Single	Twins	All
FERTILITY			
Fertility	1.925	3.107	1.955
	(1.001)	(1.176)	(1.022)
Fraction Twin		0.0257	
		(0.0158)	
Birth Order Twin		2.196	
		(1.064)	
MOTHER'S CHARACTERISTICS			
Age	36.05	36.88	36.07
	(8.396)	(7.997)	(8.387)
Education	12.54	12.70	12.54
	(2.326)	(2.232)	(2.323)
BMI	27.45	28.01	27.47
	(6.628)	(7.247)	(6.645)
Pr(BMI)<18.5	0.0206	0.0168	0.0205
	(0.142)	(0.128)	(0.142)
Excellent Health	0.320	0.325	0.320
	(0.466)	(0.468)	(0.466)
CHILDREN'S OUTCOMES			
Education (Years)	5.135	4.633	5.123
	(3.835)	(3.736)	(3.833)
Education (Z-Score)	0.00236	-0.0980	0.0000
	(1.001)	(0.949)	(1.000)
Pr(Excellent Health)	0.526	0.536	0.526
	(0.499)	(0.499)	(0.499)
Number of Children	221,381	5,832	227,213

NOTES: Summary statistics are presented for the full estimation sample consisting of all children 18 years of age and under included in NHIS surveys from 2000-2013. Group means are presented with standard deviation below in parenthesis. Education is reported as total years attained, and Z-score presents educational attainment relative to month-of-birth cohort (mean 0, std deviation 1). Excellent health is self reported by the mother of the child, and BMI is weight in kilograms over height in metres squared.

Table 2.4: Probability of Giving Birth to Twins

Twin*100	(1) All	(2) Income	(3)	(4) Time	(5)	(6) Prenatal
		Low inc	Middle inc	1990-2013	1972-1989	
Age	0.594***	0.613***	0.554***	0.646***	0.314***	0.632***
	(0.029)	(0.036)	(0.050)	(0.033)	(0.075)	(0.040)
Age Squared	-0.008***	-0.008***	-0.007***	-0.009***	-0.003*	-0.009***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Age First Birth	-0.053***	-0.093***	0.005	-0.052***	-0.055***	-0.041***
	(0.009)	(0.012)	(0.014)	(0.010)	(0.019)	(0.013)
Education (years)	0.040**	0.086***	-0.005	0.046**	0.021	-0.070**
	(0.017)	(0.022)	(0.029)	(0.020)	(0.034)	(0.028)
Education squared	-0.002	-0.006***	0.001	-0.002	0.001	0.003
	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)
Height	0.058***	0.057***	0.059***	0.062***	0.043***	0.059***
	(0.004)	(0.005)	(0.007)	(0.005)	(0.008)	(0.007)
BMI	0.048***	0.063***	0.039***	0.045***	0.054***	0.045***
	(0.006)	(0.009)	(0.009)	(0.007)	(0.011)	(0.011)
Prenatal (Doctor)						0.913***
						(0.128)
Prenatal (Nurse)						0.073
						(0.108)
Prenatal (None)						-0.484***
						(0.132)
R-squared	0.01	0.01	0.01	0.01	0.01	0.01
Observations	1930600	1201516	729084	1524894	405706	615908

NOTES: All specifications include a full set of year of birth and country dummies, and are estimated as linear probability models. Twin is multiplied by 100 for presentation. Height is measured in cm and BMI is weight in kg divided by height in metres squared. l Prenatal care variables are only recoreded for recent births. As such, column (6) is estimated only for that subset of births where these observations are made. *p<0.1; **p<0.05; ***p<0.01

Table 2.5: Test of hypothesis that women who bear twins have better prior health

Infant Mortality (per 100 births)	Base	+S&H	Observations
Treated (2+)	-2.065***	-2.110***	503785
	(0.212)	(0.213)	
Treated (3+)	-4.619***	-4.632***	686931
	(0.201)	(0.201)	
Treated (4+)	-4.257***	-4.243***	676303
	(0.183)	(0.183)	
Treated (5+)	-3.353***	-3.324***	587919
	(0.183)	(0.183)	

Notes: The sample for these regressions consist of all children who have been entirely exposed to the risk of infant mortality (ie those over 1 year of age). Subsamples 2+, 3+, 4+ and 5+ are generated to allow comparison of children born at similar birth orders. For a full description of these groups see the the body of the paper or notes to table 2.10. Treated=1 refers to children who are born before a twin while Treated=0 refers to children of similar birth orders not born before a twin. Base and S+H controls are described in table 2.10.*p<0.1; **p<0.05; ***p<0.01

Table 2.6: Probability of Giving Birth to Twins USA (NHIS)

Twin×100	All	Time	
		1982-1989	1990-2013
Age	0.0198	-0.569**	0.0306
	(0.0432)	(0.225)	(0.0462)
Age Squared	-0.000903	0.00513**	-0.000883
	(0.000601)	(0.00248)	(0.000649)
Age First Birth	0.153***	0.200*	0.143***
	(0.0131)	(0.106)	(0.0139)
Education (years)	0.0157	0.0793*	0.0107
	(0.0154)	(0.0453)	(0.0163)
Height	0.0341*	-0.0163	0.0386*
	(0.0201)	(0.0597)	(0.0213)
BMI	0.00852***	0.0158*	0.00770**
	(0.00304)	(0.00849)	(0.00324)
Smokes (pre-birth)	-0.186*	0.171	-0.206*
	(0.112)	(0.312)	(0.119)
Observations	114,037	10,114	103,923
R^2	0.003	0.006	0.003

All specifications include a full set of survey year, region of birth, and mother's race dummies and are estimated as linear probability models. Twin is multiplied by 100 for presentation. Height is measured in cm and BMI is weight in kg divided by height in metres squared. Standard errors clustered by mother are included in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 2.7: Can Selective Maternal Survival Explain Twinning Rates?

Twin×100	MMR Sample	<140cm or BMI <16	<145cm or BMI <16.5	<150cm or BMI <17	<155cm or BMI <17.5
Height	0.0657***	0.0635***	0.0590***	0.0514***	0.0417***
	(0.00414)	(0.00414)	(0.00418)	(0.00420)	(0.00425)
BMI	0.0460***	0.0437***	0.0427***	0.0409***	0.0405***
	(0.00637)	(0.00636)	(0.00637)	(0.00643)	(0.00650)
Observations	844,638	848,642	848,686	848,557	848,667
R^2	0.024	0.024	0.024	0.023	0.022

Each column represents a separate regression of maternal characteristics on twinning. For a full list of variables included see table 2.4. Only health variable are included in regression output. Column 1 includes the full sample of women surveyed in countries where the DHS maternal mortality module is applied. Columns 2-5 inflate samples in line with maternal mortality rates, where 'unhealthy', is defined as described in the column title. Full details are available in the body of the text. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 2.8: Are Twins More Likely to Miscarry to Unhealthy Mothers? (USA Vital Stats)

Miscarriage× 100	All (1)	Alcohol (2)	Health (3)
African American	2.545***	6.205***	4.692***
	(0.0253)	(0.0469)	(0.0430)
Primary Education	1.413***	0.173**	1.000***
	(0.0359)	(0.0794)	(0.0726)
Secondary Education	0.719***	1.220***	1.042***
	(0.0227)	(0.0398)	(0.0364)
Consumed tobacco (pre-birth)	1.426***	3.702***	3.040***
	(0.0353)	(0.0620)	(0.0567)
Twin	8.334***	11.02***	15.07***
	(0.677)	(1.159)	(1.074)
Twin × Tobacco	1.176***	1.362***	1.590***
	(0.209)	(0.365)	(0.334)
Twin × No Education	4.578***	5.144**	3.865*
	(1.703)	(2.363)	(2.160)
Twin × Primary Education	-2.593***	2.108***	0.728
	(0.220)	(0.527)	(0.482)
Twin × Secondary	-0.681***	1.071***	0.858***
	(0.122)	(0.216)	(0.198)
Twin × African American	-0.0244	-0.402	0.961***
	(0.134)	(0.248)	(0.226)
Twin × Alcohol		4.630***	5.435***
		(1.339)	(1.224)
Pregnancy related hypertension			0.179**
			(0.0716)
Eclampsia			6.885***
			(0.0728)
Twin × Hypertension			3.690***
			(0.279)
Twin × Eclampsia			-5.543***
			(0.289)
Consumed alcohol (pre-birth)		3.788***	4.046***
		(0.215)	(0.196)
Constant	87.97***	78.70***	53.14***
	(0.103)	(0.171)	(0.162)
Observations	3,955,099	1,753,396	1,753,396
R^2	0.464	0.519	0.598

Data is a 10% sample of NVSS birth and fetal death data from 2003–2012. Each regression includes month and year of birth fixed effects, and mother age fixed effects (interacted with a binary variable for twins) and is estimated by OLS. Miscarriage is a binary variable, and is multiplied by 100 for presentation. The omitted education variable is tertiary. Standard errors are included in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 2.9: OLS Estimates of the Q-Q Trade-off

	Base Controls	+ Health	+Health &Socioec	Bord Controls	Desired	Altonji Ratio 1	Altonji Ratio 2
Panel A: All Countries							
Fertility	-0.117***	-0.109***	-0.0774***	-0.0846***	-0.0740***	13.625	1.955
	(0.000879)	(0.000870)	(0.000838)	(0.000998)	(0.000895)		
Fertility×desire					-0.00616***		
					(0.000522)		
Observations	1,132,227	1,132,227	1,132,227	1,132,227	1,132,227		
R^2	0.092	0.107	0.159	0.160	0.159		
Panel B: Low Income							
Fertility	-0.114***	-0.106***	-0.0759***	-0.0832***	-0.0715***	13.25	1.992
	(0.00114)	(0.00111)	(0.00106)	(0.00125)	(0.00114)		
Fertility×desire					-0.00722***		
					(0.000639)		
Observations	704,812	704,812	704,812	704,812	704,812		
R^2	0.092	0.117	0.175	0.176	0.175		
Panel C: Middle Income							
Fertility	-0.124***	-0.118***	-0.0846***	-0.0912***	-0.0830***	19.667	2.147
	(0.00139)	(0.00139)	(0.00136)	(0.00166)	(0.00143)		
Fertility×desire					-0.00321***		
					(0.000899)		
Observations	427,415	427,415	427,415	427,415	427,415		
R^2	0.098	0.104	0.144	0.147	0.144		

Notes: Base controls consist of child gender, mother's age and age squared mother's age at first birth, child age, country, and year of birth dummies. Socioeconomic augments 'Base' to include mother's education and education squared, and Health includes mother's height and BMI. "Desire" takes 1 if the child is born before the family reaches it's desired size, and 0 if the child is born after the desired size is reached. The Altonji et al. (2005) ratio determines how important unobservable factors must be compared with included observables to imply that the true effect of fertilty on educational attainment is equal to zero. Ratio 1 compares no controls to socioeconomic controls, while ratio 2 compares no controls to socioeconomic and health controls. Standard errors are clustered at the level of the mother. *p<0.1; **p<0.05; ***p<0.01

Table 2.10: Principal IV Results

	2+			3+			4+		
School Z-Score	Base	+H	+S&H	Base	+H	+S&H	Base	+H	+S&H
All									
Fertility	0.024	-0.011	-0.011	-0.004	-0.036*	-0.040*	-0.013	-0.034	-0.034*
	(0.029)	(0.027)	(0.026)	(0.023)	(0.022)	(0.021)	(0.024)	(0.022)	(0.020)
Observations	260704	260704	260704	396619	396619	396619	410515	410515	410515
Low-Income									
Fertility	0.052	0.020	0.024	0.012	-0.020	-0.029	-0.008	-0.030	-0.024
	(0.034)	(0.031)	(0.030)	(0.029)	(0.027)	(0.025)	(0.028)	(0.026)	(0.024)
Observations	158139	158139	158139	247949	247949	247949	265419	265419	265419
Middle-Income									
Fertility	-0.045	-0.069	-0.074	-0.037	-0.072**	-0.065*	-0.022	-0.045	-0.051
	(0.055)	(0.051)	(0.049)	(0.040)	(0.036)	(0.036)	(0.043)	(0.039)	(0.036)
Observations	102565	102565	102565	148670	148670	148670	145096	145096	145096

Notes: The two plus subsample refers to all first born children in families with at least two births. Three plus refers to first- and second-borns in families with at least three births, and four plus refers to first- to third-borns in families with at least four births. Each cell presents the coefficient of a 2SLS regression where fertility is instrumented by twinning at birth order two, three or four (for 2+, 3+ and 4+ respectively). Different rows of the table correspond to different sub-groups or specifications. In order these correspond to: all children, grouped by country income status, adjusting fertility to correct exclude children who did not survive to one year, and including both pre-twins *and* twins in the regression. Base controls include child age, mother's age, and mother's age at birth fixed effects plus country and year-of-birth FEs. In each case the sample is made up of all children aged between 6-18 years from families in the DHS who fulfill 2+ to 4+ requirements. First-stage results in the final panel correspond to the second stage in row 1. Full first stage results for each row are available in table 2.21. Standard errors are clustered by mother.*p<0.1; **p<0.05; ***p<0.01

Table 2.11: NHIS Estimates: Education and Health

	2+			3+			4+		
	Base	+H	+S&H	Base	+H	+S&H	Base	+H	+S&H
OLS									
School Z-Score	-0.043***	-0.031***	-0.025***	-0.040***	-0.031***	-0.024***	-0.024	-0.018	-0.013
	(0.006)	(0.006)	(0.006)	(0.009)	(0.009)	(0.009)	(0.019)	(0.019)	(0.019)
Excellent Health	-0.011***	-0.005**	-0.004*	-0.016***	-0.009***	-0.007**	-0.028***	-0.018***	-0.016***
	(0.003)	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)	(0.007)	(0.005)	(0.005)
IV									
School Z-Score	-0.085	-0.097	-0.102*	-0.005	-0.013	-0.013	-0.135	-0.144	-0.158
	(0.066)	(0.061)	(0.060)	(0.068)	(0.067)	(0.067)	(0.153)	(0.157)	(0.152)
Excellent Heatlth	0.027	0.030	0.027	-0.035	-0.058*	-0.059*	0.029	-0.024	-0.034
	(0.027)	(0.021)	(0.021)	(0.039)	(0.032)	(0.032)	(0.059)	(0.052)	(0.050)
First Stage									
School Z-Score	0.650***	0.699***	0.703***	0.735***	0.740***	0.740***	0.804***	0.805***	0.837***
	(0.026)	(0.026)	(0.026)	(0.048)	(0.047)	(0.047)	(0.082)	(0.081)	(0.080)
Excellent Health	0.689***	0.738***	0.743***	0.751***	0.754***	0.756***	0.801***	0.806***	0.837***
	(0.025)	(0.025)	(0.025)	(0.045)	(0.044)	(0.044)	(0.077)	(0.076)	(0.076)
Observations	69695	69695	69695	52949	52949	52949	24105	24105	24105
Joint F-test Educ (IV)		164.5	64.7		101.3	39.6		38.0	7.7
Joint F-test Health (IV)		34469.6	163.9		15335.6	28.4		5276.4	17.1

NOTES: Each cell presents the coefficient of interest from a regression using NHIS survey data (2004-2014). Base controls include child age FE (in months), mother's age, and mother's age at first birth plus race dummies for child and mother. In each case the sample is made up of all children aged between 6-18 years from families in the NHIS who fulfill 2+ to 4+ requirements. Descriptive statistics for each variable can be found in table 2.3. Standard errors are clustered by mother.

Table 2.12: 'Plausibly Exogenous' Bounds

	UCI: $\gamma \in [0, 2\hat{\gamma}]$		LTZ: $\gamma \sim \mathcal{N}(\mu_{\hat{\gamma}}, \sigma_{\hat{\gamma}})$	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
Panel A: DHS				
Two Plus	-0.0734	0.0155	-0.0634	0.0060
Three Plus	-0.0700	0.0022	-0.0608	-0.0071
Four Plus	-0.0818	-0.0167	-0.0714	-0.0250
Panel B: USA (Education)				
Two Plus	-0.2277	-0.0028	-0.1780	-0.0155
Three Plus	-0.1335	0.0835	-0.1430	0.0500
Four Plus	-0.4429	0.1054	-0.2517	0.0084
Panel B: USA (Health)				
Two Plus	-0.0322	0.0618	-0.0362	0.0257
Three Plus	-0.1237	-0.0185	-0.0882	-0.0147
Four Plus	-0.1138	0.0237	-0.1046	-0.0071

NOTES: This table presents upper and lower bounds of a 95% confidence interval for the effects of family size on (standardised) children's education attainment. These are estimated by the methodology of Conley et al. (2012) under various priors about the direct effect that being from a twin family has on educational outcomes (γ). In the UCI (union of confidence interval) approach, it is assumed the true $\gamma \in [0, 2\hat{\gamma}]$, while in the LTZ (local to zero) approach it is assumed that $\gamma \sim \mathcal{N}(\mu_{\hat{\gamma}}, \sigma_{\hat{\gamma}})$. The consistent estimation of $\hat{\gamma}$ and its entire distribution is discussed in appendix 2.B.

Appendices

2.A Data Appendix

Main IV and OLS results for this paper are based on DHS and NHIS data described in section 2.4. These data are downloaded directly off the web and merged to form the estimation samples of interest. For DHS data, we use two surveys: the Individual (woman) Recode (IR), and the Household Recode (HR) providing education for each household member. For NHIS data, we merge three of the datafiles made available by the CDC: familyxx, household, and person. In each case, full generating code for this process is made available on the authors' websites. This code downloads, merges and cleans DHS and NHIS data to produce the datasets (one line per child) used in analysis.

In auxiliary regressions examining the characterstics of mothers and the relationship these characteristics and twin births and miscarriage, we consult a large number of other datasets. These are the following:

- United States National Vital Statistics Birth Data
- United States National Vital Statistics Fetal Death Data
- Spanish Vital Statistics (INE)
- The Swedish Medical Birth Register
- Scottish Vital Statistics
- Longitudinal Early Life Survey, Chile (ELPI)

In the case of the first 5 datasets (administrative records of births and/or fetal deaths), we use all recorded instances, focusing on twins as our outcome variable of interest. Depending upon the data source, we use all available measures of pre-determined maternal health stocks or family socioeconomic indicators. The ELPI survey from Chile focuses on child early life, and records mother's behaviours before, during and after pregnancy, along with child birth outcomes. We use all children from the first wave of this survey to run the twin regression included in the appendix tables. Further notes regarding each dataset and the particular years and number of births can be found in the notes to each table.

2.B Plausibly Exogenous Bounds and Estimating γ

From (2.10), we are interested in forming a consistent estimate of $\gamma = \frac{\partial educ}{\partial twin}|_X$. From Bhalotra and Venkataramani (2014) we have:

$$Y_{stc} = \alpha + \phi(Post_t \times basePneumonia_s) + \theta_{rs} + \eta_{rt} + \varphi \mathbf{X}_{st} + \lambda_{rc} + \theta_s \times \eta_t + \varepsilon_{stc} \tag{2.13}$$

where ϕ is the effect of access to sulfanide drugs on the outcome variable Y_{stc}.[27] To begin, we estimate the effect of this exogenous shock to maternal health on child quality:

$$educ_{stc} = \alpha^q + \phi^q(Post_t \times basePneumonia_s) + \cdots + \varepsilon^q_{stc} \tag{2.14}$$

where superscript q refers to coefficients in the quality regression, and the remainder of (2.14) follows specification (2.13). Under typical difference-in-difference assumptions, we can thus causally estimate $\phi^q = \frac{\partial educ}{\partial bP}|_X$ by OLS, where bP is the $Post_t \times basePneumonia_s$ variable indicating the extent to which sulfanide drugs result in maternal health improvements.

From (2.14) we isolate the effect of a positive maternal health shock (the reduction in rates of pneumonia) on child quality. However, for γ—the violation of the exclusion restriction—we need to estimate the twin-mediated effect of maternal health on child quality. Thus, to estimate γ, we must know both the effect of a particular health shock on quality (ϕ^q), as well as the relative levels of health of twin and non-twin mothers.[28] We refer to this as quantity as ϕ^t, and this captures the difference in average rates of (state) pneumonia mortality between twin and non-twin families:

$$\phi^t = \overline{bP}_{twin=1} - \overline{bP}_{twin=0} = \left.\frac{\partial bP}{\partial twin}\right|_X, \tag{2.15}$$

where we condition on the controls from (2.14). Finally, with these two quantities in hand, we can estimate γ by taking their product:

$$\phi^q \times \phi^t = \frac{\partial educ}{\partial bP} \times \left.\frac{\partial bP}{\partial twin}\right|_X = \left.\frac{\partial educ}{\partial twin}\right|_X = \gamma. \tag{2.16}$$

This is our estimand of interest, and we can plug it into our estimates of the bounds on β_1 using Conley et al.'s method.

In order to estimate (2.16) we turn to US census data described in Bhalotra and Venkatara-

[27] This is an unbiased effect of sulfanide drugs on health if parallel trends are satisfied between high- and low-intensity states. Evidence of this is provided by Bhalotra and Venkataramani (2014).

[28] As a simple example, if the effect of a maternal health variable is to increase child quality by 0.1 standard deviations, and twin mothers have 20% higher stocks of that maternal health variable than non-twin mothers, this suggests an estimate of γ (the twin-mediated effect of maternal health on child quality) of 0.1*0.2=0.02 s.d.

mani (2014), firstly estimate ϕ^t and ϕ^q, and then calculate product of these values to form an estimate of γ, which we denote $\hat{\gamma}$. In the case of the UCI approach, this is sufficient to estimate the bounds of β_1, assuming that: $\gamma \in [0, 2\hat{\gamma}]$.[29] In the case of the more precise LTZ approach (our preferred bounds estimates), the logic is similar, however now we must form a prior over the entire distribution of γ. Calculating the variance of γ is not as straightforward as using the variance-covariance matrix corresponding to each of the estimates $\hat{\phi}^t$ and $\hat{\phi}^q$. In this case however we can use bootstrapping to calculate J replications of $\hat{\phi}^t \times \hat{\phi}^q$, and from these estimates construct an estimated distribution of $\hat{\gamma}$, which allows us to determine our prior for the distribution of γ. From this empirical distribution, we observe the estimated mean and standard deviation, and finally test whether the distribution is normal using a Shapiro Wilk test for normality.[30]

Estimates of ϕ^q, ϕ^t and γ are presented in table 2.23. The first column suggests that the effect of improved maternal health from sulfa drugs on the next generation is an increase in 0.0497 (or 4% of a standard deviation) in school z-score. This is the quantity ϕ^q. In the second column (conditional on health and fertility controls, we see that twin mothers come from, on average, areas with 18.4% *lower* rates of pneumonia (or in other words, lower rates of pneumonia mortality are associated with higher rates of twinning, in line with the findings in the body of this paper). We interact $\hat{\phi}^q$ and the inverse of $\hat{\phi}^t$ (given that twin mothers have lower rates of pneumonia and hence are *more* healthy) to form γ. From the third column, this results in an estimate for γ of 0.0091 s.d. (0.91% of a standard deviation). This our estimate of the direct effect of coming from a twin family (having a healthier mother) on school Z-scores, and the value we use to estimate Conley et al. (2012)'s bounds in section 2.5.4.

Finally, estimates of the full distribution are presented in figures 2.13a and 2.13b. These are the estimated $\hat{\gamma}_j$ from $j \in \{1, \ldots, 100\}$ bootstrap replications using the census data described above. The first figure overlays the empirical (bootstrap) distribution with the analytical distribution for a normally distributed random variable with the same mean and variance, while the second figure is overlaid with a log normal distribution. In all cases tested (uniform, normal, log normal, χ^2), the normal distribution provides the best fit of the analytical with the empirical distribution. We cannot reject that γ is normal with a p-value of 0.2184. Although we can't reject that γ is log normal, the p-value is much lower, at 0.1019.

[29] We scale $\hat{\gamma}$ by the factor of 2 in order for this value to fall precisely in the middle of the range. Conley et al. (2012) provide a similar example to calculate the returns to education using the UCI approach.

[30] We also use Kolmogorov-Smirnov tests for equality of distributions to test whether the distribution is more likely to be log normal, uniform, and a number of other analytical distributions. In order to do this, we first estimate the empirical distribution as described previously. We then observe the mean $\hat{\mu}$ and the standard deviation $\hat{\sigma}$, and run a one-sample test do determine whether the observed empirical distribution is is significantly different to each analytical distribution $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$, $U(\hat{\mu}, \hat{\sigma}^2)$ or $\ln\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$.

2.C Appendix Tables

Table 2.13: Probability of Giving Birth to Twins USA (Vital Statistics)

	Time		Alcohol	Health
Twin×100	1980-1989	2003-2012	Controls	Controls
African American	0.629***	0.585***	0.477***	0.373***
	(0.0228)	(0.0288)	(0.0451)	(0.0455)
Other Race	-0.379***	0.0185	0.140***	-0.578***
	(0.0481)	(0.0280)	(0.0470)	(0.0634)
Secondary Education	0.0608***	-0.0126	0.0190	0.734***
	(0.0193)	(0.0278)	(0.0504)	(0.0645)
Tertiary Education	-0.0470**	1.115***	1.100***	1.963***
	(0.0217)	(0.0294)	(0.0499)	(0.0662)
Consumed tobacco (pre-birth)		-0.334***	-0.445***	-0.440***
		(0.0397)	(0.0576)	(0.0580)
Consumed alcohol (pre-birth)			-1.394***	-1.332***
			(0.204)	(0.205)
Mother Anemic				-1.362***
				(0.106)
Mother Cardiac Disease				-0.550**
				(0.244)
Mother Chronic Lung Disease				-0.639***
				(0.142)
Mother Diabetic				-0.341***
				(0.0886)
Mother Chronic Hypertension				-0.795***
				(0.167)
Pregnancy Associated Hypertension				-4.393***
				(0.0846)
Eclampsia				-5.959***
				(0.280)
Married	-0.173***			
	(0.0215)			
Observations	3,566,621	3,891,882	1,660,669	1,222,212
R^2	0.001	0.008	0.008	0.012

Each regression includes full maternal age and child birth year and birth order fixed effects and is estimated as a linear probability model. The outcome variable (a twin birth) is multiplied by 100 for presentation. Heteroscedasticity robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 2.14: Probability of Giving Birth to Twins (Chile)

	(1)		
Twin×100			
PRE-PREGNANCY		PREGNANCY	
Income p.c.	-0.006	Smoked	-0.573
	(0.011)		(0.416)
Income p.c. squared	0.000	Drugs (infrequent)	-0.119
	(0.000)		(1.646)
Secondary Education	0.142	Drugs (frequent)	-1.872***
	(0.300)		(0.344)
Tertiary Education	1.507***	Alcohol (infrequent)	-0.002
	(0.583)		(0.570)
Low Weight	-0.589	Alcohol (frequent)	-1.891***
	(0.471)		(0.290)
Obese	-1.997***	No Check-ups	-1.031
	(0.766)		(0.966)
Mother's Age	0.410***	Hospital Birth	0.939***
	(0.133)		(0.344)
Mother's Age Squared	-0.007***	Diabetes	-0.255
	(0.002)		(0.505)
Indigenous	-1.027***	Depression	0.031
	(0.395)		(0.416)
Observations	14268	R^2	0.01

NOTES: Data comes from the Encuesta Longitudinal de Primera Infancia (ELPI) from Chile. Education at each level are dummy variables, primary education is the omitted base. Regional controls and child age fixed effects are omitted for clarity. Heteroscedasticity robust standard errors are presented in parenthesis.*p<0.1; **p<0.05; ***p<0.01

Table 2.15: Probability of Giving Birth to Twins (Scotland)

	(1)		
Twin×100			
PRE-PREGNANCY		PREGNANCY	
Deprivation Index (Quintile 2)	-1.628**	Smoker	0.001
	(0.958)		(0.669)
Deprivation Index (Quintile 3)	-0.188	Previous Smoker	1.717**
	(0.967)		(0.877)
Deprivation Index (Quintile 4)	-0.421	Alcohol (1-2 per week)	-4.498*
	(0.934)		(1.935)
Deprivation Index (Quintile 5)	-1.132	Alcohol (3+ per week)	-3.030*
	(0.920)		(1.543)
Height	0.306***	Overweight	-0.092
	(0.044)		(0.643)
Married	3.272***	Obese	1.350**
	(0.878)		(0.746)
Age	-0.337	Diabetes	-0.188
	(0.400)		(0.967)
Age Squared	0.020***		
	(0.007)		
Observations	193,254	R-squared	0.01

NOTES: Data comes from Scottish birth records. All births occurring after 24 weeks for the period 1997-2012 are included. Twin births account for 1.51% of the sample. *p<0.1; **p<0.05; ***p<0.01

Table 2.16: Probability of Giving Birth to Twins (Sweden Medical Birth Registry)

Dep Var: Twin×100	(1)	(2)	(3)	(4)	(5)	(6)
Mother's Age	-0.058**	-0.053**	-0.053**	-0.053**		
	(0.026)	(0.026)	(0.027)	(0.027)		
Mother's Age Square	0.002***	0.001***	0.001***	0.001***		
	(0.0004)	(0.0004)	(0.0005)	(0.0005)		
Mother's Height	0.052***	0.052***	-0.151**	0.052***	-0.197***	-0.197***
	(0.002)	(0.002)	(0.076)	(0.002)	(0.074)	(0.074)
Mother's Height Square			0.001***		0.001***	0.001***
			(0.0002)		(0.0002)	(0.0002)
Mother's Weight		0.002	0.002	0.002		
		(0.001)	(0.001)	(0.001)		
Native	0.351***	0.372***	0.384***	0.371***	0.367***	0.367***
	(0.036)	(0.037)	(0.037)	(0.037)	(0.036)	(0.036)
Smoking 1st Trimester	0.115	0.079	0.080	0.079	0.116	0.117
	(0.072)	(0.075)	(0.075)	(0.075)	(0.072)	(0.072)
Smoking 3rd Trimester	-0.985***	-0.974***	-0.974***	-0.973***	-0.985***	-0.985***
	(0.074)	(0.078)	(0.078)	(0.078)	(0.074)	(0.074)
Diabetes				-0.311		-0.400**
				(0.204)		(0.196)
Kidney Disease				-0.424**		-0.397**
				(0.187)		(0.180)
Hypertension				-0.139		-0.136
				(0.232)		(0.221)
Observations	1,326,615	1,240,627	1,240,627	1,240,627	1,326,615	1,326,615
R-squared	0.009	0.009	0.009	0.009	0.009	0.009
Mother's Age FE					Y	Y

Notes: Data includes all births from the Swedish Medical Birth Registry from 1990 to 2011. The mean of the dependent variable is 2.385 (2-385% of all births are twins). All regressions are conditional on birth cohort and birth order FEs. Heteroscedasticity robust standard errors are reported. *** p<0.01, ** p<0.05, * p<0.1

Table 2.17: Twinning and Stress *in Utero* (Quintana-Domeque and Ródenas-Serrano, 2014)

	Twins (per 1,000)	Twins (per 1,000)	Twins (per 1,000)
ETA Bomb casualities 1st trimester of pregnancy	0.023	-0.015	-0.015
	(0.058)	(0.057)	(0.043)
ETA Bomb casualities 2nd trimester of pregnancy	-0.103***	-0.096***	-0.099***
	(0.036)	(0.037)	(0.036)
ETA Bomb casualities 3rd trimester of pregnancy	-0.124*	-0.130*	-0.129**
	(0.065)	(0.075)	(0.058)
Observations	6,793,890	6,759,120	6,759,120
Year×month and province FE	Y	Y	Y
Socio-demographic controls		Y	Y
Provice-specific linear year-month trends			Y

NOTES: Data consists of live births conceived between January 1980 and February 2003. Treatment is defined as number of ETA bomb casualties in the province of conception. Full details are provide in Quintana-Domeque and Ródenas-Serrano (2014). Standard errors are clustered at the level of the province (50 provinces). *p<0.1; **p<0.05; ***p<0.01.

Table 2.18: Can Selective Maternal Survival Explain Twinning Rates? (Lee Bounds)

Twin×100	>140cm & BMI >16	>145cm & BMI >16.5	>150cm & BMI >17	>155cm & BMI >17.5
Upper Bound	0.371	2.339	0.723	-0.253
	(2.172)	(0.977)	(0.663)	(0.517)
Lower Bound	6.566	6.652	6.680	6.786
	(0.199)	(0.203)	(0.211)	(0.233)

Estimates of Lee (2009) bounds of the effect of treatment (positive health) on twinning. Selection is due to death during pregnancy, proxied by any sister of the index women suffering a maternal death. In each column, "healthy" is defined in the column title.

Table 2.19: Twins, Miscarriage and Maternal Health (Spain Vital Statistics)

Miscarriage×100	(1) All		
Primary	-0.602***	Twin×Primary	-0.662***
	(0.014)		(0.209)
Secondary	-0.720*	Twin×Secondary	-0.559***
	(0.015)		(0.209)
Tertiary	-0.800***	Twin×Tertiary	-0.651***
	(0.015)		(0.209)
Immigration	-0.072***	Twin×Immigration	0.228
	(0.0171)		(0.296)
Married	-0.074***	Twin×Married	-0.090
	(0.008)		(0.119)
No Father	0.686***	Twin×No Father	3.252
	(0.238)		(4.093)
Observations	2,869,329	R^2	0.01

Note: Spanish vital statistics data, 2007-2012. Outcome is a binary variable for miscarriage (late term fetal death) multiplied by 100 for presentation. Standard errors are presented in parenthesis.*p<0.1; **p<0.05; ***p<0.01

Table 2.20: OLS Estimates of the Q-Q Trade-off (n+ groups)

	Base Controls	+ Health	+Health &Socioec	Bord Controls
Two Plus				
Fertility	-0.152***	-0.0993***	-0.0990***	-0.0989***
	(0.00181)	(0.00178)	(0.00178)	(0.00177)
Observations	261,434	261,434	261,434	261,434
R^2	0.109	0.177	0.177	0.179
Three Plus				
Fertility	-0.142***	-0.0942***	-0.0939***	-0.0936***
	(0.00168)	(0.00162)	(0.00162)	(0.00161)
Observations	397,750	397,750	397,750	397,750
$R^2$0.093	0.167	0.168	0.170	
Four Plus				
Fertility	-0.123***	-0.0831***	-0.0829***	-0.0824***
	(0.00180)	(0.00170)	(0.00170)	(0.00170)
Observations	411,690	411,690	411,690	411,690
R^2	0.081	0.158	0.159	0.161

Notes: Two plus refers to all first borns in families with at least two births. Three plus refers to all first and second borns in families with at least three births. Four plus refers to first to third borns in families with at least four births. For additional notes, refer to table 2.9.

Table 2.21: First Stage Results

	2+			3+			4+		
FERTILITY	Base	+H	+S&H	Base	+H	+S&H	Base	+H	+S&H
All									
Twin	0.776***	0.821***	0.822***	0.794***	0.827***	0.826***	0.840***	0.859***	0.861***
	(0.031)	(0.029)	(0.028)	(0.027)	(0.027)	(0.026)	(0.027)	(0.027)	(0.026)
Observations	249,536	249,536	249,536	375,987	375,987	375,987	385,389	385,389	385,389
Low-Income									
Twin	0.826***	0.853***	0.848***	0.810***	0.828***	0.834***	0.867***	0.873***	0.869***
	(0.038)	(0.038)	(0.037)	(0.033)	(0.033)	(0.032)	(0.033)	(0.033)	(0.033)
Observations	149,602	149,602	149,602	232,371	232,371	232,371	246,622	246,622	246,622
Middle-Income									
Twin	0.718***	0.774***	0.784***	0.757***	0.817***	0.801***	0.783***	0.831***	0.839***
	(0.050)	(0.045)	(0.043)	(0.046)	(0.045)	(0.043)	(0.047)	(0.044)	(0.042)
Observations	99,934	99,934	99,934	143,616	143,616	143,616	138,767	138,767	138,767

NOTES: Each cell represents the coefficient from the first-stage of a two-stage regression. The first-stage represents the effect of twinning at parity N on total fertility where N is 2, 3 or 4 for 2+, 3+ and 4+ groups respectively. The 2+ group includes all first borns in families with at least 2 births, the 3+ group includes first and second borns in families with at least 3 births, and the 4+ group includes all first to third borns in families with at least four births. In each regressions the sample is made up of all children aged between 6-18 years from families in the DHS who fulfill these birth order conditions. Controls in each case are identical to those described in table 2.10. Standard errors are clustered at the level of the mother.*p<0.1; **p<0.05; ***p<0.01

Table 2.22: Q-Q IV Estimates by Gender

	Females				Males			
	Base	Socioec	Health	Obs.	Base	Socioec	Health	Obs.
Two Plus	0.005	-0.039	-0.037	122,414	0.010	-0.010	-0.015	127,122
	(0.043)	(0.039)	(0.038)		(0.040)	(0.038)	(0.036)	
Three Plus	-0.024	-0.056*	-0.052*	187,098	0.016	-0.015	-0.022	188,889
	(0.033)	(0.030)	(0.029)		(0.030)	(0.028)	(0.027)	
Four Plus	-0.029	-0.052*	-0.053**	192,714	-0.005	-0.020	-0.018	192,675
	(0.032)	(0.029)	(0.027)		(0.030)	(0.028)	(0.027)	

NOTES: Female or male refers to the gender of the index child of the regression. All regressions include full controls including socioeconomic and maternal health variables. The full list of controls are available in the notes to table 2.10. Standard errors are clustered by mother. *p<0.1; **p<0.05; ***p<0.01

Table 2.23: Consistent Estimates of γ Using a Maternal Health Shock

	$\frac{\partial Educ}{\partial Health}$	$\frac{\partial Health}{\partial Twin}$	$\gamma = -\frac{\partial Educ}{\partial Twin}$	γ (bootstrap)
Estimate	0.0497***	-0.184***	0.0091	0.0091
	(0.0181)	(0.0181)		(0.0027)
Observations	943,038	897,112		
R-squared	0.011	0.110		

NOTES: Regression results for (2.14) and (2.15) use the 5% sample of 1980 census data. Specifications and samples are identical to those described in Bhalotra and Venkataramani (2014). The estimate of γ is formed by taking the product of panel A and panel B estimates. A full description of this process, along with the non-pivotal bootstrap process to estimate the standard error of γ is provided in appendix 2.B, and figure 2.13a

Table 2.24: Full Survey Countries and Years

		Survey Year						
COUNTRY	INCOME	1	2	3	4	5	6	7
Albania	Middle	2008						
Armenia	Low	2000	2005	2010				
Azerbaijan	Middle	2006						
Bangladesh	Low	1994	1997	2000	2004	2007	2011	
Benin	Low	1996	2001	2006				
Bolivia	Middle	1994	1998	2003	2008			
Brazil	Middle	1991	1996					
Burkina Faso	Low	1993	1999	2003	2010			

Burundi	Low	2010						
Cambodia	Low	2000	2005	2010				
Cameroon	Middle	1991	1998	2004	2011			
Central African Republic	Low	1994						
Chad	Low	1997	2004					
Colombia	Middle	1990	1995	2000	2005	2010		
Comoros	Low	1996						
Congo Brazzaville	Middle	2005	2011					
Congo Democratic Republic	Low	2007						
Cote d Ivoire	Low	1994	1998	2005	2012			
Dominican Republic	Middle	1991	1996	1999	2002	2007		
Egypt	Low	1992	1995	2000	2005	2008		
Ethiopia	Low	2000	2005	2011				
Gabon	Middle	2000	2012					
Ghana	Low	1993	1998	2003	2008			
Guatemala	Middle	1995						
Guinea	Low	1999	2005					
Guyana	Middle	2005	2009					
Haiti	Low	1994	2000	2006	2012			
Honduras	Middle	2005	2011					
India	Low	1993	1999	2006				
Indonesia	Low	1991	1994	1997	2003	2007	2012	
Jordan	Middle	1990	1997	2002	2007			
Kazakhstan	Middle	1995	1999					
Kenya	Low	1993	1998	2003	2008			
Kyrgyz Republic	Low	1997						
Lesotho	Low	2004	2009					
Liberia	Low	2007						
Madagascar	Low	1992	1997	2004	2008			
Malawi	Low	1992	2000	2004	2010			
Maldives	Middle	2009						
Mali	Low	1996	2001	2006				
Moldova	Middle	2005						
Morocco	Middle	1992	2003					
Mozambique	Low	1997	2003	2011				
Namibia	Middle	1992	2000	2006				
Nepal	Low	1996	2001	2006	2011			
Nicaragua	Low	1998	2001					
Niger	Low	1992	1998	2006				
Nigeria	Low	1990	1999	2003	2008			
Pakistan	Low	1991	2006					
Paraguay	Middle	1990						
Peru	Middle	1992	1996	2000				
Philippines	Middle	1993	1998	2003	2008			
Rwanda	Low	1992	2000	2005	2010			
Sao Tome and Principe	Middle	2008						
Senegal	Middle	1993	1997	2005	2010			
Sierra Leone	Low	2008						
South Africa	Middle	1998						
Swaziland	Middle	2006						
Tanzania	Low	1992	1996	1999	2004	2007	2010	2012
Togo	Low	1998						
Turkey	Middle	1993	1998	2003				

Uganda	Low	1995	2000	2006	2011
Ukraine	Middle	2007			
Uzbekistan	Middle	1996			
Vietnam	Low	1997	2002		
Yemen	Low	1991			
Zambia	Low	1992	1996	2002	2007
Zimbabwe	Low	1994	1999	2005	2010

NOTES: Each year listed represents a DHS survey. Country income status is based upon World Bank classifications described at http://data.worldbank.org/about/country-classifications and available for download at http://siteresources.worldbank.org/DATASTATISTICS/Resources/OGHIST.xls (consulted 1 April, 2014). Income status varies by country and time. Where a country's status changed between DHS waves only the most recent status is listed above. Middle refers to both lower-middle and upper-middle income countries, while low refers just to those considered to be low-income economies.

Table 2.25: Test of Balance of Observables: Twins versus Non-twins

	Non-Twin Family	Twin Family	Diff. (Diff. SE)
Panel A: Two Plus			
Mother's Education	5.082	5.114	-0.0316
			(0.0963)
Mother's Height (cm)	155.6	157.2	-1.561***
			(0.142)
Prenatal care available	0.942	0.949	-0.00707
			(0.00469)
Mother's Age in Years	20.27	20.77	-0.492***
			(0.0774)
Age First Birth	19.85	20.34	-0.489***
			(0.0769)
Total Fertilty	3.627	4.524	-0.897***
			(0.0307)
Infant Mortality (pre-twin)	0.00370	0.00358	0.000120
			(0.00122)
School Z-score (pre-twin)	0.0995	0.107	-0.00790
			(0.0193)
Percent male child (pre-twin)	0.510	0.508	0.00268
			(0.0100)
Panel B: Three Plus			
Mother's Education	4.057	4.174	-0.116
			(0.0767)
Mother's Height (cm)	155.7	157.1	-1.436***
			(0.125)
Prenatal care available	0.935	0.949	-0.0133**
			(0.00428)
Mother's Age in Years	21.35	21.82	-0.471***
			(0.0669)
Age First Birth	19.29	19.68	-0.388***
			(0.0620)
Total Fertilty	4.421	5.270	-0.848***
			(0.0263)
Infant Mortality (pre-twin)	0.00622	0.00508	0.00114
			(0.00131)
School Z-score (pre-twin)	0.00496	0.0167	-0.0117
			(0.0162)
Percent male child (pre-twin)	0.504	0.512	-0.00765
			(0.00752)

Notes: All variables are at the level of the mother from full DHS data described in table 2.2. Panel A contains mothers of children who have had at least two births, where twin is defined as a twin at the second birth. Panel B contains mothers of children who have had at least three births, where twin is defined as a twin at the third birth. Education is measured in years, mother's height in centimetres, and prenatal care is binary, taking 1 if available in the mother's region. Diff. SE is calculated using a two-tailed t-test. *p<0.1; **p<0.05; ***p<0.01

2.D Appendix Figures

Figure 2.10: Birth Size of Twins versus Singeltons

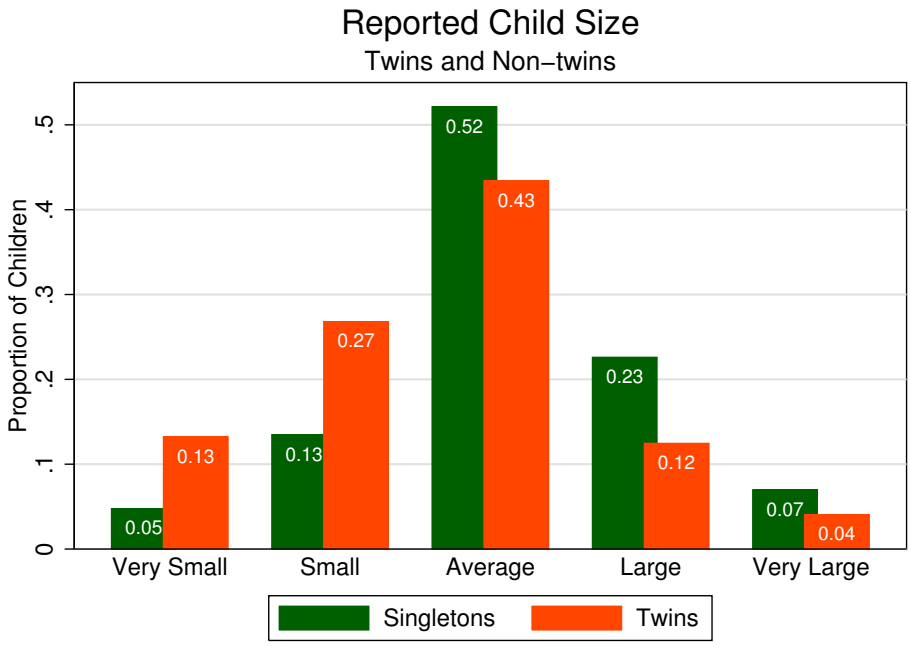


Figure 2.11: Height and Selective Survival

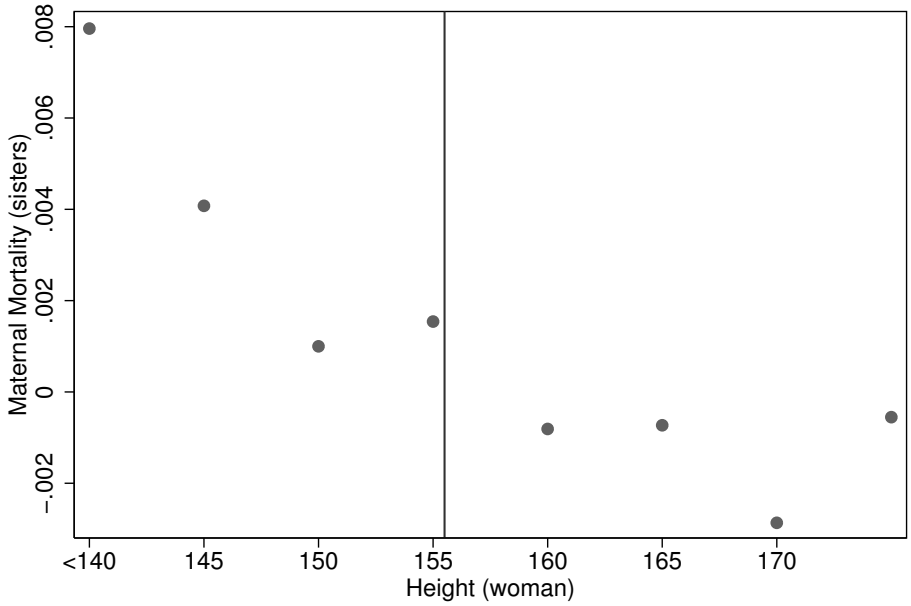
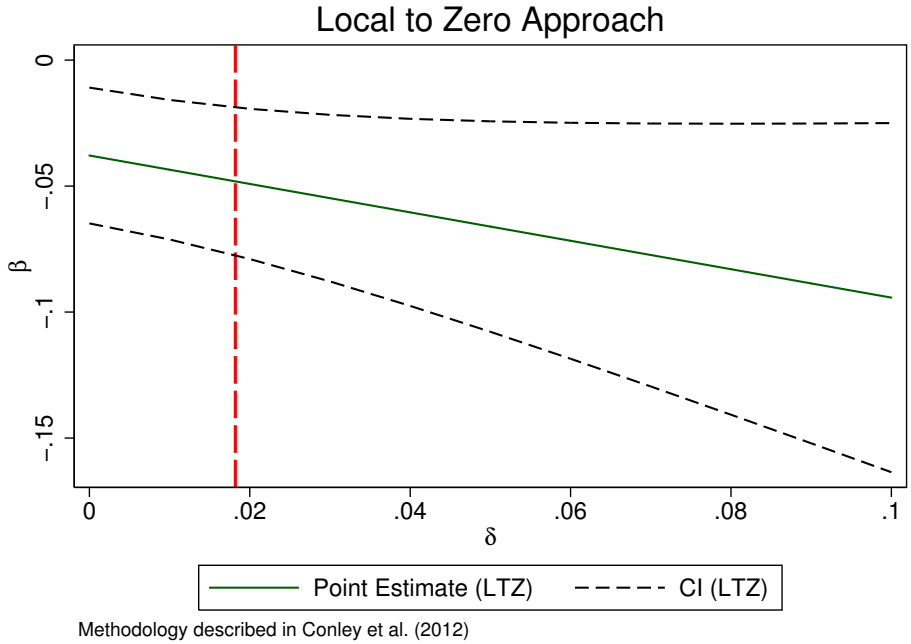


Figure 2.12: Relaxing the Exclusion Restriction (four plus)



Note to figure 2.12: See notes to figure 2.7

Figure 2.13a: Bootstrap Estimates of $\hat{\gamma}$ (Normal)

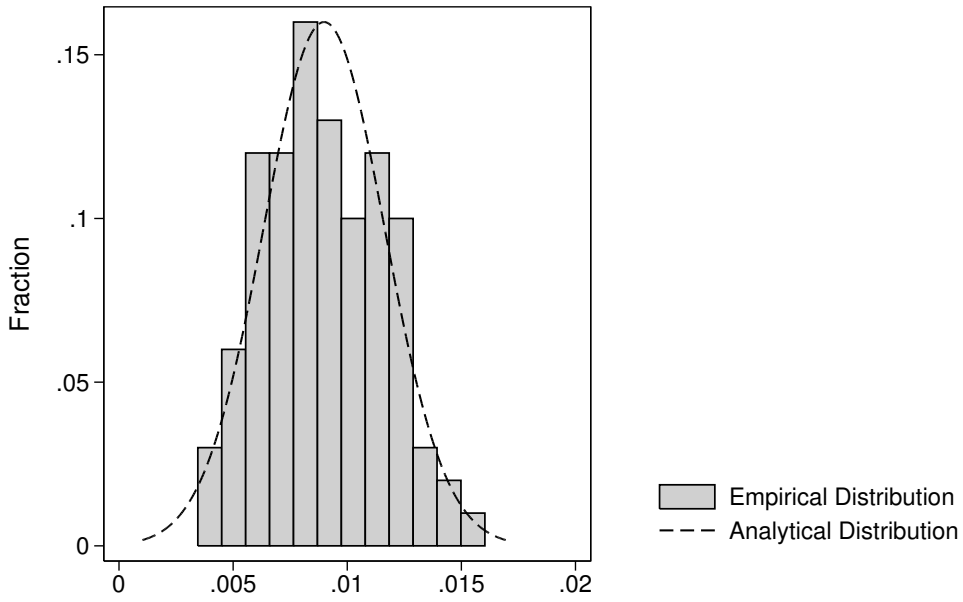
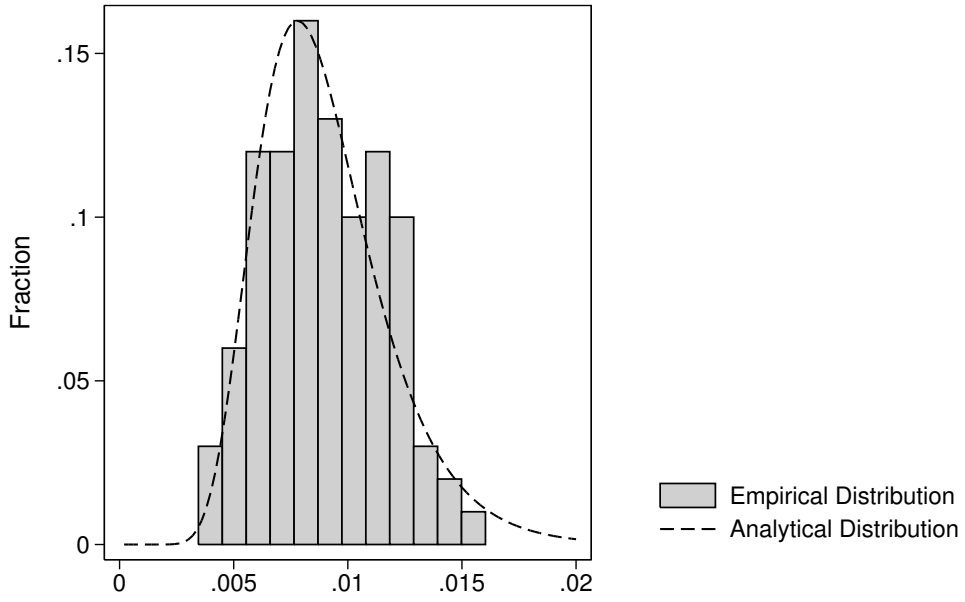


Figure 2.13b: Bootstrap Estimates of $\hat{\gamma}$ (log Normal)



Note to figures 2.13b-2.13a: The empirical distributionis generated by performing J=100 bootstrap replications to estimate ϕ^t and ϕ^q (see discussion in appendix 2.B). The overlaid analytical distribution in figure A is normal $\sim N(\mu_{\hat{\gamma}}, \sigma_{\hat{\gamma}})$, and in figure B is log Normal with the same mean and standard deviation. The first and second moments estimated in the bootstrap replications are $\mu_{\hat{\gamma}} = 0.00898$ and $\sigma_{\hat{\gamma}} = 0.00265$.

Chapter 3

Assessing Plan B*

The Effect of the Morning After Pill in Chile

Chapter Abstract

I examine the effect of quasi-experimental variation in the availability of the emergency contraceptive ("morning after") pill in Chile. Using censal data on all births and fetal deaths over the period 2005-2011 I show that the availability of the pill reduces pregnancy and early gestation fetal death, which I argue proxies for illegal abortion. These effects are particularly pronounced among teenagers and young women: point estimates suggest a 6.9% reduction in teenage pregnancy and 4.2% reduction for 20-34 year olds. I suggest that diffusion of the morning after pill between quasi treatment and control areas played an important role, and suggest a way to estimate unbiased treatment effects where the stable unit treatment value assumption does not hold locally. This paper is the first to provide censal evidence of the emergency contraceptive's effect, and the first to examine the technology in a country where no other (legal) post-coital fertility control options exist.

*I acknowledge the excellent support and advice of a number of members of the Government of Chile who provided extremely useful access to, and advice regarding, national databases. Principally, I thank Rodrigo Alarcón S., Andrés Álvarez A., Carlos Arce Mártinez, Ximena Carrasco and Nico Muñóz of the Ministries of Health, Social Development and Education. Much care was taken by all parties to respect all necessary privacy clauses, and data analysis was undertaken in line with Law 19.628 of Protection of Private Life (*Ley 19.628 de Protección de la Vida Privada*).

3.1 Introduction

Undesired pregnancy—particularly among young and adolescent women—is a considerable contributor to poor maternal and child outcomes, and to a lack of intergenerational mobility. The last half-century has seen a remarkable increase in contraceptive technology, with considerable impacts on rates of such undesired pregnancy and with far-reaching consequences for the social and productive structure of modern society. The widespread introduction of the oral contraceptive pill has brought with it lower birth rates, delays in childbearing and marriage, higher rates of human capital attainment and labour market participation for women (Angrist and Evans, 1996; Bailey, 2006; Goldin and Katz, 2002a,b), reductions in the gender wage differential (Bailey et al., 2012), and, theoretically at least, more empowered women (Chiappori and Oreffice, 2008). In the long-run, these outcomes have led to generations of children less likely to have divorced parents, and more likely to live with college educated mothers (Ananat and Hungerman, 2012).

While the contraceptive pill has had a remarkable impact on a woman's capability to control the timing of her fertility decisions, these treatments require an expensive ongoing investment, which is difficult or impractical for certain groups of women. In contrast to the rich literature on the effects of the contraceptive pill, very little evidence is available regarding the effects of post-coital (non-abortive) birth control. In this paper I examine the effect of fully-subsidised provision of the emergency contraceptive pill. This so called "morning after pill" offers an alternative form of contraception in cases where other forms were not used or failed during intercourse, or in the case of rape.

The scarce existing literature on this topic suggests that the emergency contraceptive (EC) pill may have had surprisingly little effect on both pregnancy and abortion (Gross et al., 2013; Durrance, 2013). Here I present considerable evidence that, at least in the case of Chile, access to emergency contraception does have significant effects on births and abortions, and that these effects are concentrated on teenagers and young women. I identify a plausibly exogenous policy decision in Chile which affects a woman's access to the (fully subsidised) emergency contraceptive pill. Using censal data on each woman's pregnancy status in each year, and the outcomes of each pregnancy in Chile, I demonstrate that the availability of the morning after pill reduced the likelihood of pregnancy and illegal abortion, and that this effect was transversal rather than being enjoyed overwhelmingly by one social class.

The reform under examination comes from a series of constitutional challenges between 2005-2008, which meant that the introduction of the emergency contraceptive pill in Chile was entirely

controlled by the Supreme court and constitutional Tribunals. Legal challenges resulted in the 2008 finding that it would be illegal for all nationally run health centres and hospitals to prescribe the emergency contraceptive pill, however that in each of the 346 municipalities of Chile health centres were at liberty to do so. This resulted in a situation in which a woman's access to the pill entirely depended upon the decisions taken by her mayor. Due to this reform it is shown that around half the municipalities in Chile made the pill available, while the other half did not.

Using this reform, I estimate the effect that the staggered arrival of the emergency contraceptive had on women and children, including its effect on births and abortions. The arrival of this new technology is associated with significant reductions in these outcomes. Further, the effects identified are of considerable magnitude. It is estimated that among teenage girls, the widespread availability of emergency contraception reduces births by around 7%, and may more than halve rates of illegal abortion. Among older women the reductions in births and illegal abortion are more moderate, however still quantitatively important. For example, among 20-34 year olds, the emergency contraceptive pill reduces births by an estimated 4.2% and appears to reduce illegal abortion by around 20%.

Naive estimates of the effect of the morning after pill on pregnancies, abortions, and other outcomes, are based on the assumption that the arrival of the emergency contraceptive to approximately half of the women in the country had no effect on those women who did not live in areas where the pill was available. I examine the validity of this assumption by comparing women who live 'close' to areas where the pill was available to those who live considerably further away. It is shown that significant treatment spillovers may occur, and so suggested that naive estimates off the effect of the emergency contraceptive pill may significantly underestimate the true effect of the expansion of availability in Chile.

Given the spatial nature of these spillovers, a methodology is proposed which allows for the recovery of a consistent treatment effect even in the presence of local spillovers. It is shown that under certain assumptions regarding the nature of the stable unit treatment value assumption, the estimated treatment effect will be significantly attenuated if this consideration is not made. I then propose a number of ways to determine which control clusters should be considered 'close to' treatment clusters. It is shown that for the morning after pill, treatment spillover is a quantitatively important consideration, and that in some groups, diffusion may exist for anywhere up to 30km from a treatment location.

This study makes a number of contributions. Foremost, it is one of the first—if not the first—study of the effects of emergency contraception using censal microdata on a national scale.

It is also the first large scale study of which the author is aware which addresses these questions in a country other than the United States. This is of considerable importance given that Chile, the country under study here, does not offer legal abortion, and so the emergency contraceptive pill is the first legal mechanism for post-coital fertility control.

The results of this study add to the nascent literature on the emergency contraceptive pill. Recent studies such as Gross et al. (2013); Durrance (2013) which have been the first to address this question in the economic literature have provided evidence to suggest that the effects of this technology may be minor. Here I offer considerable evidence to the contrary, suggesting that the expansion in the availability of emergency contraceptives may offer important effects in certain countries, with large impacts on pregnancy and abortion rates, especially among young women.

Finally, I raise a number of novel points regarding the estimation of treatment effects in the presence of spillovers between treatment and control clusters. This is fundamentally different to, despite sharing some characteristics with, the literature on estimating treatment effects in the presence of spillovers between treated and non-treated individuals within treatment clusters. I suggest that even in the presence of such spillovers, unbiased treatment effects can be estimated if these spillovers occur in a predictable way, as is likely to be the case when distance to treatment varies in an exogenous manner. I propose a number of flexible ways to consider estimating treatment effects in these circumstances.

In what remains of this paper I first provide background regarding the emergency contraceptive pill, and the reform under study in Chile in section 3.2. Section 3.3 discusses the censal data sets I will use to assess the effects of the reform, while 3.4 discusses identification and methodology. In section 3.5 I present results on the contraceptive's effect on births, abortions and aggregate human capital endowments. Finally, section 3.6 concludes.

3.2 The History of the Emergency Contraceptive Pill

The emergency contraceptive pill is a hormonal treatment which can be used within 5 days of an unprotected sexual relationship to reduce the probability of conception. There are a number of alternative types of emergency contraceptive pills, however principally these are composed of doses of the progestin levonogestrel, or a combined dose of estrogen and progestin. Typically these are taken as a single pill or two pills in a 12 hour period (von Hertzen et al., 2002), however similar doses of hormones can be obtained by combining normal birth control pills (Ellerson

et al., 1998).

This form of contraception has been shown to be relatively effective at avoiding undesired pregnancy. Estimates of around 75%-85% effectiveness based on typical usage are common, depending upon the method of emergency contraception used.[2] The success of these treatments is dependent upon the delay between intercourse and taking the drug, so widespread—or at least quickly available—access is important in reducing undesired pregnancies. While most effective when taken within 12 hours after intercourse, effectiveness can continue when taken within as much as 120 hours (von Hertzen et al., 2002).

The emergency contraceptive pill is not an abortive agent, but rather is a 'postcoital contraceptive' which acts to prevent ovulation (Novikova et al., 2007; Noé et al., 2011). This contraceptive method has been of clinical interest since at least the late 1960s (Demers, 1971), however access to these methods, either by prescription or over the counter, is still not universal. The fact that emergency contraception is non-abortive has meant that it is available in many countries in which abortion is absolutely prohibited, or prohibited in all cases except where concerns for maternal survival exist. Some countries have made the EC pill available as early as the mid-1980s (UK Family Planning Association, 1996), while many more countries have legalised this method of contraception during the last decade.

3.2.1 The Morning After Pill in Chile

The introduction of the emergency contraceptive pill in Chile has followed a complicated path, with early legislation frequently blocked by conservative groups in office and in civil society.[3] While initial discussions and administrative inquiries took place in 2001, it was not until 2005 that significant advances in legislature were made. In December of this year the Chilean supreme Court determined that the Institute of Public Health—the pharmaceutical regularity body of Chile—was *not* acting unconstitutionally by approving the provision of an emergency contraceptive drug on the pharmaceutical register. However, this finding was quickly challenged by detractors, with cases presented before ordinary and Constitutional tribunals (Casas Becerra, 2008).

These tribunals were followed by a number of years' worth of legislations and litigations,

[2]The WHO's Task Force on Postovulatory Methods of Fertility Regulation (1998), for example suggests that a levonogestrel routine reduces pregnancy rates by 85%, with a 95% confidence interval of 74-93%.

[3]The Chilean political framework is marked by a strong conservative axis, and a constitution which favours the maintenance of the status quo in economic and valoric policies. This has been the case since the return to democracy in 1990, with an alliance of right wing parties (and some members of the presiding left wing coalition) who have "resisted more liberal changes in the poorly named value judgements" (Casas Becerra (2008), p.6, author's translation.)

which resulted in sporadic availability of the morning after pill, occasionaly freely available from state clinics or by purchase in private pharmacies. However, these were generally short-lived and emergency contraception was not consistently stocked, with both political and economic ramifications for groups providing access to the pill.[4] Details regarding this intervening process and laws passed by parliament theoretically requiring the provision of emergency contraception are discussed more fully in appendix 3.A.

The period of interest for this study follows a decision taken by the Chilean Constitional Tribunal in 2008. This finding, responding to a demand placed by 36 parlamentary deputies in 2006, made it expressly illegal for the centralised health system to distribute the emergency contraceptive. This requirement held for all centres under direct administration of the national Ministry of Health, but, fundamentally, provided all municipal-level centres and hospitals the freedom to distribute the pill. Given that these centres are administered by the mayor of each municipality (or *comuna*), the availability in each municipality was entirely under the control of the mayor (Dides C. et al., 2011, 2010; Dides et al., 2009).[5] This resulted in a situation in which around half of the municipalities in Chile distributed the morning after pill freely, while the remaining half refused to distribute it, or to distribute it only in a very restrictive set of circumstances. At the level of the woman, her municpality's treatment status was essentially exogenously determined, being based on the whim of the mayor or representative public health bodies in her area of residence. This strange policy environment endured for approximately four years, until a law was passed mandating that the emergency after pill must be available to all women who request it. This new law became operational in May of 2013.

The Chilean context is one in which emergency contraception may be expected to have particularly important effects on pregnancy and maternal health. Abortion is entirely illegal in Chile, meaning that in the absence of emergency contraception, undesired or accidental pregnancies must either be taken to term, or a woman must risk undertaking a dangerous and illegal clandestine abortion (Shepard and Casas Becerra, 2007). Figures on the frequency and method of clandestine abortion are unclear, however Shepard and Casas Becerra (2007) suggest that the primary method is by taking the abortive drug misoprostol, which can be legally prescribed for treatment of ulcers. However, the cost of accessing this drug without prescription is high. Dated (2007) figures suggest prices of 38,000-50,000 Chilean pesos, or around one third of the minimum monthly wage at this time.

[4]For example, the subsecretary of health was removed from cabinet due to his announcement in 2005 that emergency contraception would be available to all women who sought it.

[5]Of the 346 municipalities in Chile, 320 have their own health systems, while the remaining 26 depend entirely upon the Ministry of Health. These 320 municipalities make up 94% of the population of Chile. Municipal health centres make up the majority of health centres in Chile. Of the 2501 registered health centres and hospitals, 2049 are under the control of municipalities (DEIS Ministerio de Educación, Gobierno de Chile, 2013).

3.3 Data

The data of interest for this study comes from matched administrative data files recording all live births and fetal deaths in Chile. I use birth outcomes for all women aged between 15 and 49 (inclusive) at some point during 2006-2012. This is crossed with data recording all women in Chile and their municipality of residence, resulting in a record of each woman, and her pregnancy status in each period (live birth, fetal death or no pregnancy). Along with a woman's birth status, we observe her baby's birth weight and gestational length in the case that a birth or fetal death was recorded.[6]

This results in a total sample of 1,391,565 births and 11,387 fetal deaths. The number of births per year in Chile has remained relatively stable over the last decade. Figure 3.1 displays total births, along with total fetal deaths during the period under study (a similar plot for births per women is available in figure 3.3). Total births vary between around 220,000-250,000 per year, while total fetal deaths recorded in the Ministry of Health data (all fetal deaths in any hospitals or clinics in Chile), vary between 1700 and 2100.

Our measure for the pill is a binary variable which records whether the emergency contraceptive was freely available to a woman upon request at her municipal health centre in the year before her birth outcome is observed. We consult two sources to collect data on pill availability. First, in each of 2009, 2010 and 2011 an independent survey was conducted, asking health care workers from each municipality whether they were able to prescribe the morning after pill (Dides et al., 2009; Dides C. et al., 2010, 2011). This should directly reflect the decision by each mayor regarding whether his or her municipality could prescribe the pill after the 2008 Constitutional ruling. In each case, survey respondents were also asked to list the circumstances in which they could prescribe the pill. All municipalities which reported that they could prescribe the pill freely to women were recorded as treated, while all others were recorded as untreated.[7] Secondly, the Ministry of Health has made available administrative data on all pill requests and dispursements at municipality clinics and hospitals. This allows us to determine the veracity of the survey data discussed above, while also providing concrete numbers regarding the use of the emergency contraceptive pill following the reform of interest. However, I do not use pill disbursements as the main measure of treatment. I focus on reported availability, given that disbursements are

[6] Since 1999, each baby born in Chile (in public and private hospitals and in private homes with midwives) is weighed and measured, and their gestational time is recorded. This data is collated by the Ministry of Health, and is available as a public birth weight census. As well as the baby's characteristics, the mother's education, age, labour market status and municipality of residence is collected. Similar details are collected in the case that a woman enters hospital and suffers a miscarriage.

[7] A small number of municipalities reported that they could prescribe the emergency contraceptive, however that this was only following cases of rape. These municipalities were classed as *un*treated given the lack of widespread availability.

endogenous, and jointly determined by demand as well as supply.

In total, 224 of Chile's 346 municipalities report being able to prescribe the pill in at least one year after the 2008 Tribunal result (see table 3.1). Figure 3.2 displays the quantity of municipalities reporting pill disbursements over time. Here, the number of prescribers increases over time in line with greater awareness of the legality of distributing the emergency contraceptive pill. While less than half of all municipalities report pill availability in 2009, this has increased to around two thirds by 2011. Official records of pill prescriptions suggest reasonably large fluctuations over time. While nearly 8000 women were reported as requesting the pill in 2009, this fell to slightly under 4000 the following year. Recent figures suggest that this number has been stable at around 6000-7000 requests in 2010-2013 (the most recent two years have been omitted from this study, and from graphical output, given that official birth records for 2012 and 2013 are not yet finalised). Figure 3.4 displays the geographic variation of pill availabilitly. This suggests that the pill is available in all parts of the country. With the exception of the large and very sparsely populated southern region of the country (the 10^{th} region) which has no municipal health centres, no obvious spatial patterns exist.

I examine fetal deaths as a manner to proxy illegal abortion. While it is certainly not the case that all (or even the majority) of fetal deaths observed in administrative data are results of abortive drugs, there is some evidence that these are the result of abortion in some cases, although they are recorded in a number of different ways in official figures to avoid criminal charges against women (Shepard and Casas Becerra, 2007). To avoid concerns that reductions in fetal deaths may be simply due to greater investments in public health, I examine a number of subgroups of interest. Firstly I focus on deaths occurring between 1-20 weeks of gestation, as this is the period in which nearly all abortions are conducted. Secondly, I remove deaths which, based on their ICD code,[8] are clearly not related to abortion, such as those due to congenital malformations, deformations and chromosomal abnormalities. By using this methodology, a clear validity check exists by comparing reductions in fetal deaths during 1-20 weeks (which may represent abortions and should respond to the morning after pill), to those occurring from week 21 and above, which should be largely or entirely unaffected by emergency contraceptive availability.

Full summary statistics are provided in Table 3.1. These statistics are subdivided by whether or not the municipality has the pill in a given year. We see that there are some differences between pill and non-pill municipalities, such as higher education and health spending in pill municipalities. However, this is largely due to the fact that all years in which the pill was observed occur after 2008 while non-pill status is observed over the entire time period under

[8]The ICD refers to the International Classification of Disease, and refers to a set of standardised codes by which deaths can be classed. All deaths on the birth register report this code, (the ICD-10).

study. Surprisingly, we see that municipalities in both groups are approximately balanced in terms of the 'conservativeness' of the party of the mayor, however we do see that female mayors are more likely to be associated with pill municipalities. Later in this study I describe pre-treatment trends in pill and non-pill municipalities.

3.4 Methodology

I take advantage of the quasi-random nature of the expansion of the availability of the morning after pill to women in Chile. A woman i, living in municipality j in time t is considered as treated if public health centres report that the pill is available upon request. A woman's child bearing status $birth_{ijt}$ is regressed on the availability of the pill ($pill_{jt}$) in the preceding year:

$$birth_{ijt} = \alpha + \delta \cdot \mathbb{I}\{Pill_{jt-1}\} + \phi_t + \eta_j + \eta_j \cdot t + X_{jt-1}\gamma + \varepsilon_{ijt}. \tag{3.1}$$

In (3.1), full municipality and year fixed effects are included, and municipality-specific time trends are allowed. Standard errors are estimated which allow for auto-correlation by municipality. The identifying variation in availability of the pill is by municipality and year. Prior to the legal reform the pill was unavailable to all women, while posterior to the reform the pill was available to those women living in municipalities where the mayor did not restrict access. This provides a flexible differences-in-differences (hereafter diff-in-diff) framework, and allows us to causally estimate the effect of the morning after pill if we believe that typical diff-in-diff assumptions hold. Namely, we require that unobserved components ε_{ijt} in the above specification evolve similarly over time in the treated and untreated municipalities.

Given the geographically disperse, and, as discussed in previous sections, plausibly exogenous nature of the arrival of the morning after pill, we may be willing to accept that this assumption is valid. However, to minimise the potential that spurrious events confound the arrival of the pill, we progressively include higher order time trends and other factors that vary non-linearly over time across municipalities. These factors, X_{jt-1}, include controls for political and social outcomes such as the mayor's party (and implicitly the conservativeness of views), the degree of voter support for the mayor, the mayor's gender, health and education inputs including staffing and training investments, and time varying measures of female empowerment by municipality.

We are interested in determining whether the morning after pill affects fertility at both the extensive and the intensive margins. We thus measure pregnancy in a number of ways: firstly, at the intensive margin by examining whether a woman gives birth at any parity level, and secondly,

only at the extensive margin by examining whether she moves from 0 to 1 births. Similarly, we are interested in determining the degree of heterogeneity of access by age groups, and look at teenagers (15-19 year olds), 20-34 year olds, and 35-49 year olds.

Similar estimations are run replacing $birth_{ijt}$ with $fetal\ death_{ijt}$, which—for certain subsets—we believe proxy illegal abortion (as discussed in section 3.3). After assessing the pill's impact on pregnancy, and abortion I estimate reduced form effects of the pill's arrival on various measures of mother and child outcomes. These include maternal education, employment status and marital status, and child birthweight and gestational length. While I don't believe that these regressions are demonstrating causality in the case of mother's outcomes, these regressions are a useful test to determine whether certain groups are more likely to access the pill leading to aggregate compositional change in the cohorts of women who give birth.

3.4.1 Identifying Spillovers Between Municipalities

Our diff-in-diff estimates in the previous section potentially underestimate the true effect of the morning after pill. Principally, we may be concerned that there exist spillovers between treatment and control clusters due to the porous nature of municipal boundaries. Given that a woman can access municipal health centres in neighbouring comunas, if she is denied access to the pill in her comuna she may travel to obtain it elsewhere, or otherwise rely on the close geographic distance between her municipality and a treatment municipality to gain access to the morning after pill[9]. This motivates the following specification:

$$y_{ijct} = \alpha + \delta \cdot \mathbb{I}\{Pill_{jt-1}\} + \sum_{c=0}^{C} \zeta_c \cdot close_{cdjt-1} + \phi_t + \eta_j + \eta_j \cdot t + X_{jt-1}\gamma + \varepsilon_{ijct}. \tag{3.2}$$

where

$$close_{cdjt} = \begin{cases} 1 & \text{if } dist_{jt} > c \wedge dist_{jt} \leq c + d \\ 0 & \text{if } dist_{jt} \leq c \vee dist_{jt} > c + d. \end{cases}$$

where $dist_{jt}$ is the distance (in km) to the nearest treatment municipality (ie a municipality which reports prescribing the pill). By definition, this takes 0 for all treatment municipalities.

This specification is identical to that in (3.1), however here we include a number of *close* controls (indexed by c). These *close* variables are designed to capture spillovers between the pill treatment areas and surrounding areas which may also be affected by this treatment status, but

[9]This may be the case for example if women rely on friends or contacts in neighbouring municipalities to gain access.

which were not themselves treated. As defined above, at most one of these *close* dummies can be switched on for a given (non-treated) municipality. By judiciously selecting an appropriate series of *close* dummy variables, the true effect of the morning after pill can be recovered even in the case that spillovers occur between certain treatment and control clusters.[10]

To see this, define y_{1ijt} as the potential outcome for a woman in the presence of treatment. Likewise, y_{0ijt} is the potential outcome for a woman in the absence of treatment. As is well known in the treatment effects literature[11], difference-in-differences will allow us to estimate the causal effect of treatment if we are willing to make a common trends assumption about treament and control municipalities. Implicitly, this common trends assumption nests an assumption about spillovers between treatment and control muncipalities: that no such spillovers may exist, as these will affect the pre-existing trend in the control state.[12] This is analogous to the Stable Unit Treatment Value Assumption (SUTVA) of the Rubin Causal Model.

Now, rather than dealing with the two potential outcomes statuses above, we define new outcomes. First, we define y_{0ijtc}, the potential outcome for woman i in untreated municipality j in time t and with close status c. For simplicity, in what follows we will consider c as binary, indicating whether j is 'close' or 'not close' to a treatment municipality, although the results for a categorical variable follow logically. Similarly, we define y_{1ijtc} as the potential outcome in a treated municipality with close status c.[13]

As is typical in a double-differences framework, an additive structure for y_{ijtc} is assumed which consists of a municipality effect, a time effect, an indicator for treatment (D_{jt}), and in our case, an indicator for being 'close' to treatment ($close_{jtc}$):

$$y_{ijtc} = \eta_j + \phi_t + \delta D_{jt} + \zeta close_{jtc} + \varepsilon_{ijtc} \tag{3.3}$$

Now, if we consider the single differences which make up a double-differences estimate, we have, for the treatment group:

$$E[y_{ijtc}|j = Pill, t = 2, c = 1] - E[y_{ijtc}|j = Pill, t = 1, c = 1] = \phi_2 - \phi_1 + \delta. \tag{3.4}$$

[10] Thus, these controls are determined by c, the minimum distance to a treatment cluster, and $c+d$, the maximum distance to the treatment cluster. For example, $close_{0,15,jt}$ will take the value of 1 for any municipality which does not itself prescribe the EC pill, but is within (0,15] km of a treatment municipality. Similarly $close_{15,30,jt} \Rightarrow (15, 30]$.

[11] See for example Card and Krueger (1994).

[12] Fortunately for the naive estimates of treatment effects in this case, any estimates will be attenuated rather than overstated, given that the mixture of treated units with the control group will cause outcomes in control group to look more like those in the treatment group.

[13] However, in the case of treated municipalities c will always take the value of 0 given that these municipalities are themselves treated rather than simply close to a treated municipality.

This is the traditional single difference which forms one half of a typical double-differences estimator. However, in the case of the control group, the single difference is no longer simple. It will now be made up of two components: the difference over time in control municipalities who are 'close' to treatment municipalities:

$$E[y_{ijtc}|j = NoPill, t = 2, c = 1] - E[y_{ijtc}|j = NoPill, t = 1, c = 1] = \phi_2 - \phi_1 + \zeta \tag{5a}$$

and the difference over time for 'non-close' control municipalities:

$$E[y_{ijtc}|j = NoPill, t = 2, c = 0] - E[y_{ijtc}|j = NoPill, t = 1, c = 0] = \phi_2 - \phi_1. \tag{5b}$$

If we were to naively combine close and non-close control municipalities to make one large control group, we would have that our second difference consists of the weighted sum of (5a) and (5b). Were we then to combine the first difference (3.4) and the second difference (the weighted average of 5a and 5b) to form our double-differences estimator, this would give:

$$\begin{aligned} &\{E[y_{ijtc}|j = Pill, t = 2, c = 1] - E[y_{ijtc}|j = Pill, t = 1, c = 1]\} - \\ &\Bigg(\frac{N_c}{N_c + N_{nc}}\{E[y_{ijtc}|j = NoPill, t = 2, c = 1] - E[y_{ijtc}|j = NoPill, t = 1, c = 1]\} + \\ &\frac{N_{nc}}{N_c + N_{nc}}\{E[y_{ijtc}|j = NoPill, t = 2, c = 0] - E[y_{ijtc}|j = NoPill, t = 1, c = 0]\}\Bigg) = \\ &\delta - \frac{N_c}{N_c + N_{nc}}\zeta. \end{aligned} \tag{6}$$

Here we clearly see that our naive estimator fails to recover the true parameter of interest δ.[14] Generally, we would suspect that if spillovers exist, then they are likely to be of the same direction as the effect of treatment, meaning that δ and ζ will have the same sign. If this is the case, the inclusion of $\frac{N_c}{N_{nc}+N_c}\zeta$ in the naive estimate attenuates the treatment effect.

However, in the above discussion, we are concerned that SUTVA has been violated in a specific manner. We are concerned that the treatment status of women in treatment municipalities spillover to those of control muncipalities 'close' to these treatment municipalities. Conversely then, we assume that women in control muncipalities 'far enough away' from those in treatment municipalities are not affected by their treatment status, and so SUTVA still holds in these cases. Specifically, imagine that our double-difference estimator is now only based upon those control

[14] This estimator includes as a limiting case the typical diff-in-diff estimator, as in this case we assume that no spillovers are present and SUTVA holds, meaning that $\zeta = 0$. Similarly, if no municipalities were close enough to experience spillovers, we would have that $N_c = 0$, and once again δ would be recovered.

municipalities which are not classed as belonging to *close*. In this case:

$$\begin{aligned}\{E[y_{ijtc}|j = Pill, t = 2, c = 1] - E[y_{ijtc}|j = Pill, t = 1, c = 1]\}- \\ \{E[y_{ijtc}|j = NoPill, t = 2, c = 0] - E[y_{ijtc}|j = NoPill, t = 1, c = 0]\} = \delta,\end{aligned} \tag{7}$$

and using the sample analogue of these population parameters we are able to correctly recover the true effect of interest. Discussion of how to precisely determine which municipalities are and are not part of the *close* group is delayed until section 3.5.3.

Estimating (3.2) provides a flexible regression-based framework for (7). Both $Pill_{jt}$ and $close_{cjdt}$ switch on only in those municipalities who are affected by the pill (either directly or via spillover) in the date after the morning after pill has become available. In this case both δ and ζ (from 3.2) identify the effect of living in a pill or close-to-pill municipality by comparing them to treatment municipalities which are sufficiently far from the morning after pill that we can plausibly assume SUTVA. In the case of the coefficient on pill this is simply estimating our effect of interest (7), while the coefficient on *close* identifies the marginal effect of being close to the pill.[15] Given that we are assuming geographic dependence in these estimates, we use Conley's (1999) spatial standard errors. This involves defining a reasonably flexible covariance matrix which inversely weights observations to allow for dependence across individuals based on distance.

3.5 Results

3.5.1 The Effect of Emergency Contraception on Births

Table 3.2 provides estimates for specification (3.1). This has been estimated using a logistic regression, and all coefficients are cast as log odds. Here, I examine two fertility outcomes: the probability that a woman gives birth to any child (columns 1-4), and the probability that a woman gives birth to her first child (columns 5-8). The latter outcome captures just the effects of the emergency contraceptive pill at the extensive margin, while the prior outcome captures both extensive (first birth), and intensive (more births) effects.

[15] In the framework above, this can be viewed as:

$$\begin{aligned}\{E[y_{ijtc}|j = NoPill, t = 2, c = 1] - E[y_{ijtc}|j = NoPill, t = 1, c = 1]\}- \\ \{E[y_{ijtc}|j = NoPill, t = 2, c = 0] - E[y_{ijtc}|j = NoPill, t = 1, c = 0]\} = \zeta.\end{aligned}$$

In each case we estimate first the simple diff-in-diff specification without time-varying controls, and then gradually add time varying controls which may confound results of the original specification. Initial results suggest that the effect on pregnancies may be large, particularly so for teenagers. Point estimates on "All Births" for the 15-19 year old group suggest that the pill is associated with a highly significant 6.2% reduction in pregnancy (1-exp(−0.064)), or a 4.0% reduction when including potentially confounding time-varying controls. The coefficients on these time varying controls are omitted from table 3.2 for the sake of clarity; however a full output for column (4) of each panel is provided in appendix table 3.10.

When we examine the effect only on first births, we see a somewhat smaller, but still important 3.5% reduction in births (or an imprecisely estimated 2.1% reduction when including controls for condom availability). This "First Births" column must necessarily be less than or equal to the effect of the morning after pill on all births, given that all births include first births, along with higher order births. The difference between the results in these two columns allows for a rough examination of the magnitude of extensive versus intensive effects on fertility. Were the entire effect of the pill working at the extensive (first birth) margin, we would expect that the coefficients for "First Births" should equal those on "All Births". As is, for the teenage group, we see that the coefficient on first births is 51% of that on all births, suggesting that while the extensive margin is important, the morning after pill also has important effects at the intensive margin.

The effects on older age groups are more moderate than the effect on teenagers, consistent with the fact that a greater proportion of teenage births are undesired. However, for 20-34 year olds we still see that access to the emergency contraceptive reduces pregnancy, by 3.0% for all births, and 2.1% for first births. In contrast to younger women, there appears to be no effect of the morning after pill on women aged 35 and above. All estimates for the 35-49 year old group are not significantly different to zero.

3.5.2 The Effect of Emergency Contraception on Abortions

In table 3.3, difference-in-difference estimates of the effect of the emergency contraceptive pill on fetal deaths are presented. Once again these are estimated using a logit model. In this case the denominator (or 0 in the outcome variable) is assigned to each live birth, while a fetal death is assigned a 1. All effects are thus interpreted as fetal deaths per live births. As discussed in section 3.3, by using certain subsets of fetal deaths we aim to proxy for illegal abortion. We expect that if the emergency contraceptive pill affects abortion, this should turn up in fetal deaths occurring

from 0-20 weeks, however should not turn up in deaths occuring later in the gestational period, given that abortions rarely take place beyond the 20^{th} week.

Column (1) of table 3.3 presents the effect of the pill on *all* fetal deaths. We are, however, most intersted in columns (2) and (3), which present results for early (0-20 weeks), and late (21-39 weeks) respectively. In these columns we have removed from the sample any fetal deaths which have been classified in ICD class Q (a minority of fetal deaths), as these represent causes such as congenital defects, which are very unlikely to proxy abortion.

For the 15-19 year old group, significant evidence is found to suggest that the morning after pill may reduce the prevalence of (illegal) abortion. Some effect is found when examining the effect on all fetal deaths, however when this is examined by subgroups, the effect is entirely driven by early gestation deaths. The size of the coefficient is empirically very important: it suggests a reduction in early gestation deaths by 55%, which I interpret as strong evidence in favour of reductions of illegal abortion. When compared to the null effect on late-term deaths, this seems to provide more support to this claim.

A similar pattern is observed for the 20-34 year old group of women, however effects are smaller and somewhat less significant. While no significant effect is found when examining all births, there is evidence (at the 10% significance level), that the arrival of emergency contraceptive reduces early gestation deaths by 17%. Once again, no significant effect is found in late gestation fetal death.

The group of women aged 35 years and above is somewhat less clear, and, while the effect sizes of the coefficients follows the pattern outlined above, the significance on late gestation fetal deaths is somewhat surprising. Given that fetal death is much more common as maternal age increases, it is perhaps unsurprising that we find some effect for this group. One possible explanation for this finding is that the morning after pill allows less healthy women to select out of child bearing, although given the lack of covariates recording mother's health at the time of childbirth, this cannot be explored fully.

3.5.3 Municipality Spillover and Imperfect 'Compliance'

We augment our naive estimates from sections 3.5.1 and 3.5.2 to account for between-cluster spillovers in table 3.5. These results are estimated according to equation (3.2) using a logit regression. The discussion is section 3.4.1 proposes including controls for areas which are 'close

enough' to treatment municipalities that they are likely to be affected by spillovers. However, discussion regarding how to determine the threshold has been putoff until this point. In this section I examine two related ways which this can be done. Both of these ways rely on the data and specific context of the treatment in question to determine the range over which municipalities should be considered as close.

The first method involves a series of consecutive regressions and tests on the coefficient $\widehat{\delta}$. First, a regression is run including no close controls and $\hat{\delta}^0$ is observed along with its standard errors (where superscript 0 refers to the number of close controls included). Then, a single close control is included for municipalities within d km of the treatment municipality (where d can be some small number). From this regression, we observe $\hat{\delta}^1$, and test for the equality of δ^0 and δ^1 using a t-test. If this test is rejected, we add another close control, this time indicating muncipalities located within d and $2d$ km of the treatment municipality. Again we run a t-test for the equality of δ^1 and δ^2. This iterative process is continued until the point that we cannot reject the test that $\delta^{C-1} = \delta^C$. At this point we accept that we have saturated our model with sufficient 'close' controls to recover a consistent estimate of δ, and assume:[16]

$$\left|\delta - \frac{c}{c+nc}\zeta\right| \simeq |\hat{\delta}^0| < |\hat{\delta}^1| < \cdots < |\hat{\delta}^{C-1}| = |\hat{\delta}^C| \simeq |\delta|. \tag{8}$$

The second method follows a similar iterative process, but rather than testing each δ^c against its predecesor, we run a t-test with the null: $\zeta = 0$. The logic in this case is that rather than assuming that we have a consistent estimate of δ once this coefficient is stable, we assume that we have included enough 'close' municipalities once spillover effects are no longer found in the marginal municipality.

In order for these methodologies to uncover a consistent estimate of δ, all we require is that there actually are at least *some* control municipalities far enough away from treatment municipalities in which no spillover effects are felt. As described in (7), these 'non-close' municipalities act as the control group for our diff-in-diff estimator, so if no non-close municipalities exist, no consistent estimator can be formed. Given the relatively large distance between some non-treated municipalities and their nearest treated counterpart in the Chilean context, this seems unlikely in this case, although we cannot reject this formally.

Panel A of table 3.5 estimates using this methodology. In this case, using either of the above methods results in an identical number of close controls. For both 15-19 year olds and

[16] One situation in which this will not provide a consistent estimate of δ is the case in which spillovers converge, but do not converge to zero. If for example beyond a certain distance C the effect of spillovers reach some fixed constant, then the null that $\delta^{C-1} = \delta^C$ will not be rejected, even though the marginal ζ term is not equal to zero.

20-34 year olds, it appears that living within 30 km of a treatment municipality results in a spillover effect, while for the case of 35-49 year olds no spillover is observed. Now, based on these updated estimates it appears that the true effect of the morning after pill may be significantly higher than that estimated in section 3.5.1. Compared to the 4.1% reduction in teenage births estimated from specification (3.1), here we estimate a 6.9% reduction for women living in treated municipalities ($1 - \exp(-0.071)$), with sizeable effects also found for those living in close, non-treated municipalities. Similar patterns are observed for the 20-39 year old group, however in this case estimates are increased in magnitude from 3.0% to 4.2%. Figures 3.5 and 3.6 provide graphical support of this methodology. Here, estimates are calculated based on a wide range of *close* controls, with a step size d of 2.5 km. In each case, the estimated $\hat{\delta}$ appears to converge when controlling for spillovers of up to 30km.

In Panel B, similar tests are run for fetal deaths. In this case it appears that spillovers act only over a shorter range or not at all, however this may owe partially to the fact that fetal death is a relatively uncommon event, so estimates are more imprecise. Once again however, the inclusion of close municipality controls act to increase the magnitude of treatment effects estimated. For 15-19 year olds, point estimates move from a 51.7% reduction in early-gestation fetal deaths, to a 56.6% reduction, and for 20-39 year olds these point estimates move from 13.0% to 16.1%, however it is worth noting that these changes are not statistically significant.

3.5.4 Emergency Contraception and Aggregate Human Capital at Birth

Table 3.4 examines the effect of emergency contraception on aggregate human capital indicators of pregnant women and newborn babies. While it is not suggested that the morning after pill itself will affect a woman's human capital attainment over such a short time frame, if certain subgroups of the population are more likely to access the contraceptive, it is likely that aggregate compositional changes will be seen in both maternal and child human capital outcomes. There is considerable evidence to this effect when considering access to the oral contraceptive pill (Bailey et al., 2012; Ananat and Hungerman, 2012; Chiappori and Oreffice, 2008), and the arrival of legal abortion (Whitaker, 2011; Ananat et al., 2009).

I examine three outcome variables for mothers: years of education, employment status, and a binary variable for marriage, and three outcome variables for newborns: weight at birth, weeks of gestation, and length (in cm) at birth.[17] Each model is estimated as outlined in (3.1) using OLS. Surprisingly, we find that the morning after pill has had no, or very little, effect on aggregate

[17]These outcomes, particularly birthweight, have been shown to improve outcomes including educational attainment and income throughout life (Behrman and Rosenzweig, 2004)

human capital indicators. This is the case among mothers, and consequently among newborn babies.

Panel A of table 3.4 presents estimates by age group. For both teenagers and 20-34 year olds, no effect is seen on any of the variables examined. In general, these results seem to suggest that access to the morning after pill is transversal, and is not centred on highly educated or employed women. Moving to the 35-49 year old group, some evidence exists to suggest that the aggregate education of women giving birth is slightly higher where the morning after pill is available. This would be consistent with less educated (and perhaps less healthy) women selecting out of child bearing in this age group, which is consistent with the results found for this age group in table 3.3.

Panel B provides estimates for all children born over the period under study. Once again, very little evidence is found to suggest that the emergency contraceptive pill has created large-scale compositional effects to birth cohorts. Given the lack of effect found in mothers, it is not surprising that similar results are found in babies. In each case, no effect is observed on birthweight, gestational period, or length at birth (with the exception of a very small reduction in gestational length for babies born to 20-34 year olds). Each of the reported significance levels is based on a two-tailed t-test. If I were to correct for multiple hypothesis testing using a Bonferroni correction, finding a significant result would be even less likely.

3.5.5 Placebo Tests

Robustness of the main estimates to the addition of time-varying controls and municipal-specific time trends provides some confidence in the results, however does not directly examine the differential trends assumption underlying diff-in-diff estimation. In order to examine this assumption more closely, I run a number of placebo tests. These placebo tests allow us to examine whether the results are driven by pre-existing differences or trends in treatment and control municipalities.

I thus run analogous tests to (3.1) and (3.2), however rather than looking at births following the introduction of the pill, I examine births *preceding* the introduction of the pill. The logic underlying these tests is that if is the arrival of the EC pill which reduces undesired births, then there should be no difference between trends in births in pill and non-pill municipalities in years preceding the reform. If however, the effects are due to general differences in trends in non-pill and pill municipalities, we may expect that an effect would be seen even in the absence of the

EC pill. We this run the following series of tests:

$$birth_{ijt-l} = \alpha + \delta \cdot \mathbb{I}\{Pill_{jt}\} + \phi_t + \eta_j + \eta_j \cdot t + \varepsilon_{ijt}, \tag{9}$$

where l refers to a series of lags $l \in 3, 4, 5$ years. I choose lags of at least 3 years so that all births observed will occur entirely before the arrival of the EC pill in 2008.

These results are presented in table 3.6, both for specification (9) and an analogous specification where placebo close municipalities are defined. These placebo tests support the diff-in-diff specification estimated. In all but 3 of 30 coefficients, small and statistically insignificant results are observed on placebo treatments. In 3 of 30 cases, significant effects are found, although these are always on placebo 'close' treatments, and not on the main treatment itself. Generally this is quite strong evidence in favour of an absence of pre-treatment differential trends, as at 10% significance levels, it is expected that approximately 3 in 30 coefficients should be falsely accepted (ie a type I error should occur).

Along with these formal placebo tests, we can examine trends by eye based on full data on all births occurring in Chile in the past decade and a half. I present graphical results as appendix figures 3.9 and 3.10. These figures suggest that indeed, the sharp discontinuity in births occurs precisely following the arrival of the EC pill to Chile: further evidence in favour of these results owing to the morning after pill, rather than to alternative actions taken in pill and non-pill municipalities.

3.6 Conclusions

This study provides the first censal estimates of effect of the emergency contraceptive pill. In contrast to existing studies based on data from the United States, this study focuses on a reform in Chile, a country with high rates of teenage pregnancy and undesired childbearing, and where abortion is entirely outlawed. The lack of abortion or other post-coital birth control technologies means that the arrival of the emergency contraceptive pill heralded the first opportunity for women to control fertility in cases where alternative forms of birth control were not used or failed during intercourse.

By taking advantage of a legal finding which left decisions regarding the availability of the morning after pill in the hands of the mayor of each of Chile's 346 municipalities, I estimate the effect of this technology on fertility, abortion and aggregate human capital outcomes. In contrast

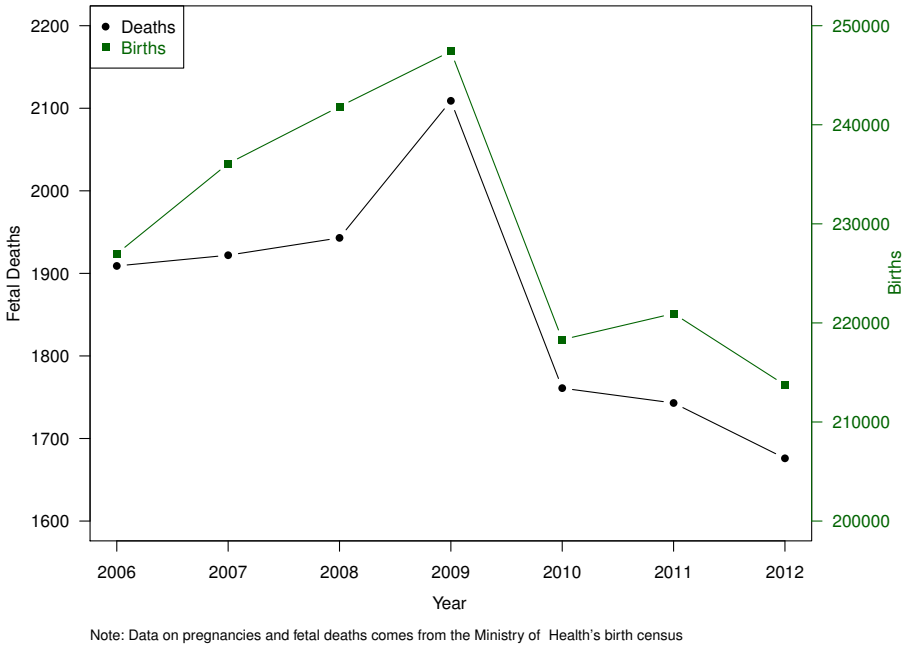
to the literature currently available, I find the emergency contraception has large and significant effect on births and early gestation fetal deaths. For teenagers, this effect is estimated to be a reduction of 6.9% and a remarkable 55% in births and early-gestation fetal deaths respectively, while for 20-34 year old women these figures are a smaller, but still significant 4.1% and 16.0%. It is argued that these early-gestation deaths proxy for illegal abortion, and comparisons with late term deaths add support to this claim.

Given the permeable nature of municipal boundaries within a country, I examine the possibility that the arrival of the pill to a given municipality is not restricted only to women who live within its boundaries. Results suggest that this may be the case, and that treatment spillovers may endure for as much as 30km. I propose an identification strategy which flexibly allows for such spillover effects to be accounted for, while simulatenously recovering consistent estimate of the effect of the treatment in the presence of contaminated (local) control groups.

All told, this paper provides considerable evidence that emergency contraception may play an important role in a woman's contraceptive behaviour. This finding is of particular importance to the country under study given that only recently has law been implemented making the morning after pill available to all. This also suggests that despite evidence to the contrary in the United States, the emergency contraceptive pill may be an important interim technology in the many countries which currently do not allow alternative forms of post-coital contraception.

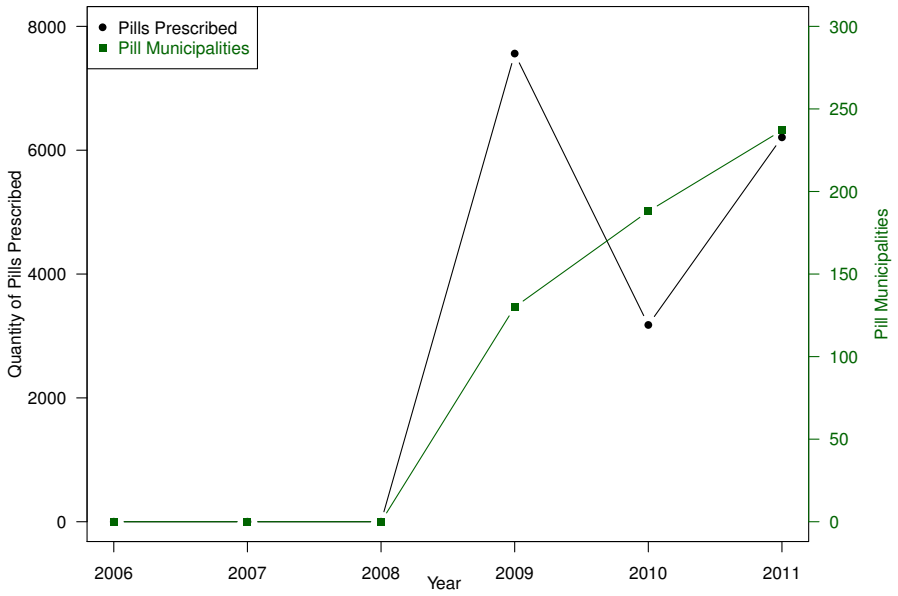
Figures

Figure 3.1: Total Recorded Births and Fetal Deaths, 2006-2011



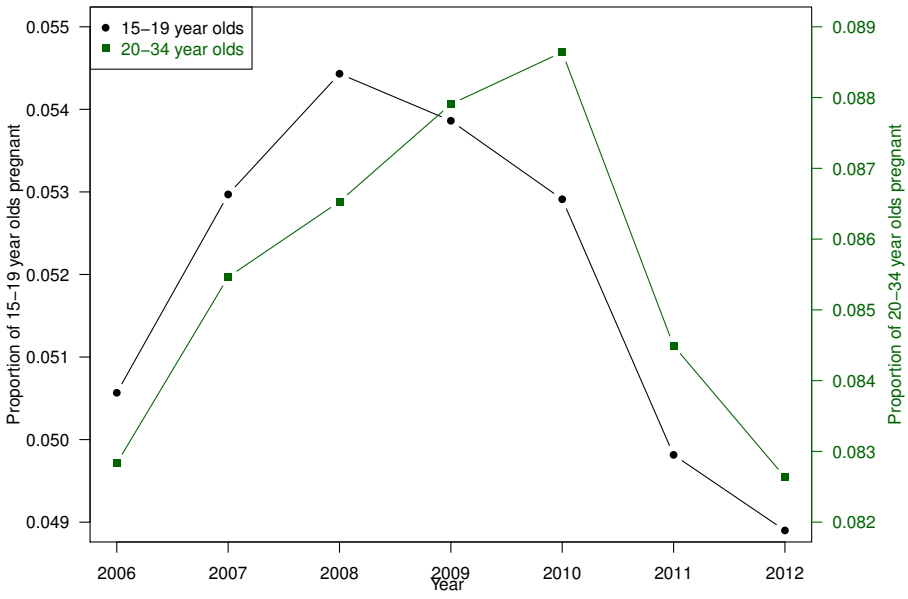
Note: Data on pregnancies and fetal deaths comes from the Ministry of Health's birth census

Figure 3.2: Pill Prescriptions and Availability by Time



Note: Prescription data is from the Ministry of Health's administrative data on medications and medical attention. Municipality data is from an independent survey conducted by Dides et al, (2010;2011;2012).

Figure 3.3: Pregancies by Age Group and Time



Note: Data on pregnancies comes from the Ministry of Health's birth census

Figure 3.4: The Availability of the Pill by Geographic Region

NOTE: THIS FIGURES HAS BEEN REMOVED TO REDUCE FILE SIZE BUT WILL BE ADDED IN TO THE PRINT VERSION.

Figure 3.5: Estimates of $\hat{\delta}^c$ for Pregnancy (15-19)

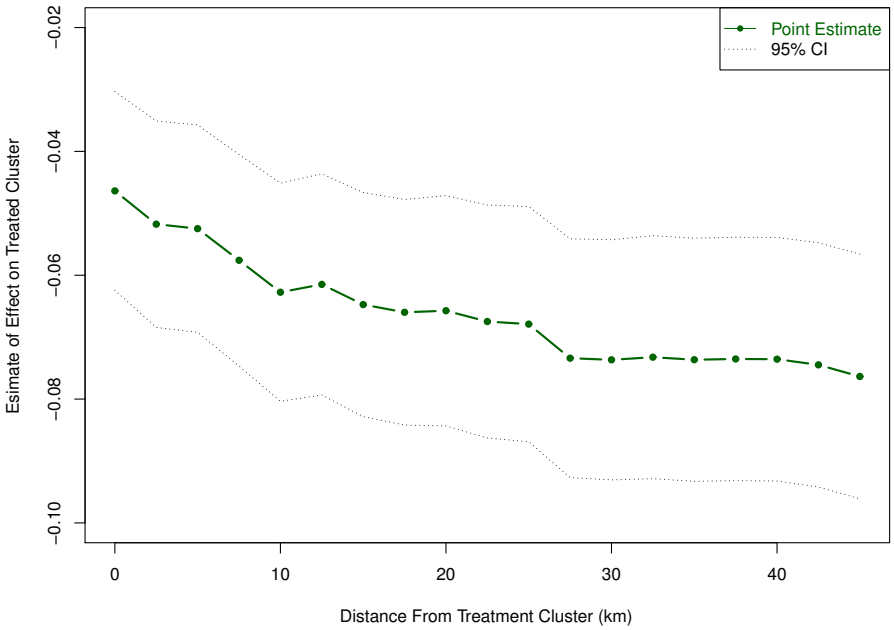


Figure 3.6: Estimates of $\hat{\delta}^c$ for Pregnancy (20-34)

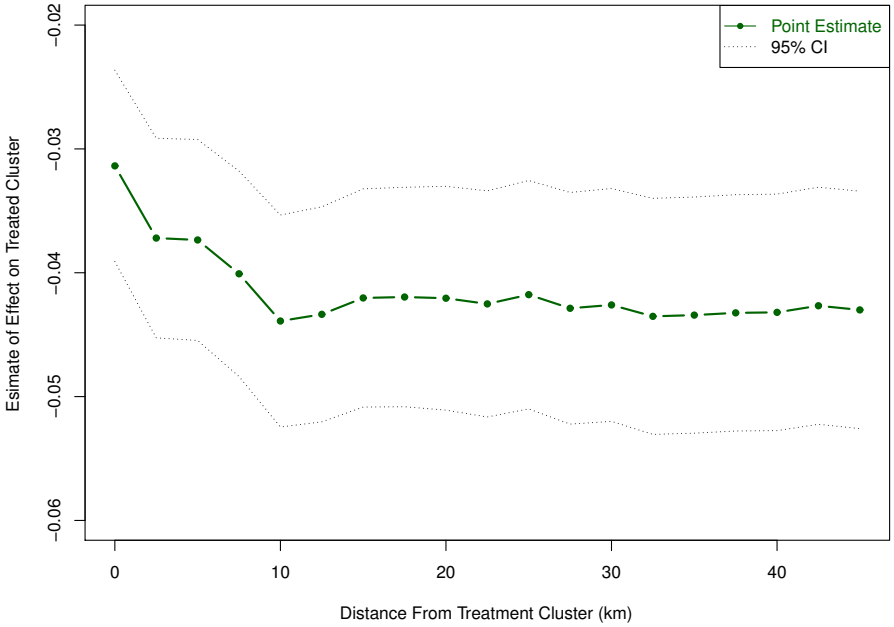


Figure 3.7: Event Study: 15-19 Years

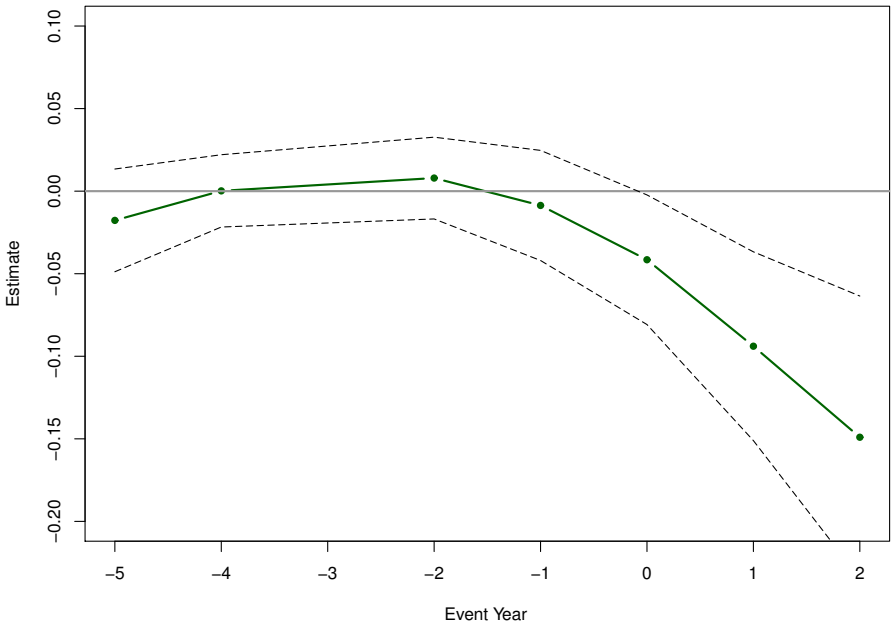
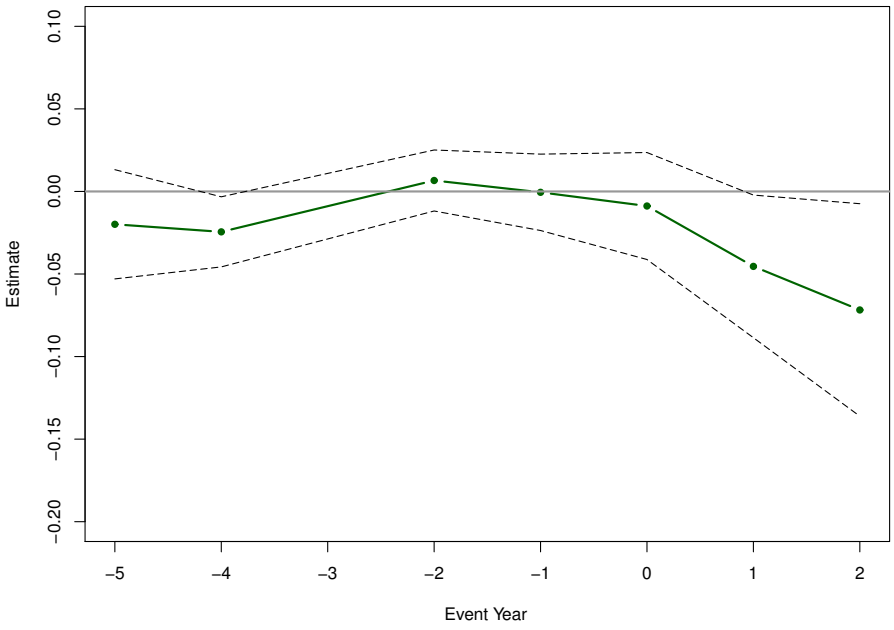


Figure 3.8: Event Study: 20-34 Years



Note to figures 3.7-3.8: Points and confidence intervals represent estimates for a full event study. Each point represents an indicator for the treatment group (pill municipality) interacted with n years prior or posterior to the reform. On the x-axis, 0 represents the first year the reform arrived, 1 represents the reform having run for one year, and so forth. The (arbitrarily chosen) omitted base in each case is 3 years prior to the reform.

Tables

Table 3.1: Summary Statistics

	No Pill Available	Pill Available	Total
Municipality Characteristics			
Poverty	16.4	17.0	16.6
	(7.47)	(7.56)	(7.49)
Conservative	0.286	0.267	0.281
	(0.452)	(0.443)	(0.45)
Education Spending	4,817	5,980	5,108
	(5,649)	(6,216)	(5,818)
Health Spending	1,866	2,788	2,096
	(2,635)	(3,381)	(2,867)
Out of School	4.07	3.98	4.05
	(3.16)	(3.06)	(3.13)
Female Mayor	0.120	0.134	0.123
	(0.325)	(0.341)	(0.329)
Female Poverty	60.5	62.0	60.8
	(10.64)	(9.48)	(10.4)
Pill Distance	5.94	0.00	4.46
	(18.4)	(0.0)	(16.1)
Individual Characteristics			
Live Births	0.054	0.053	0.054
	(0.226)	(0.224)	(0.226)
Fetal Deaths	0.0558	0.0513	0.0547
	(0.269)	(0.256)	(0.266)
Birthweight	3322.7	3334.3	3324.7
	(540.0)	(542.3)	(540.4)
Maternal education	11.92	12.03	11.94
	(2.967)	(2.894)	(2.955)
Percent working	0.295	0.395	0.312
	(0.456)	(0.489)	(0.463)
Married	0.340	0.309	0.335
	(0.474)	(0.462)	(0.472)
Age at Birth	27.05	27.15	27.07
	(6.777)	(6.790)	(6.779)
N Comunas	346	280	346
N Fetal Deaths	9,999	3,064	13,063
N Births	1,214,088	391,212	1,605,300

Notes: Group means are presented with standard deviations below in parentheses. Poverty refers to the % of the municipality below the poverty line, conservative is a binary variable indicating if the mayor comes from a politically conservative party health and education spending are measured in thousands of Chilean pesos, and pill distance measures the distance (in km) to the nearest municipality which reports prescribing emergency contraceptives. Pregnancies are reported as % of all women giving live birth, while fetal deaths are reported per live birth. All summary statistics are for the period 2006-2012.

Table 3.2: The Effect of the Morning After Pill on Pregnancy

	All Births				First Births			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
15-19 YEAR OLDS								
Morning After Pill	−0.060***	−0.061***	−0.056***	−0.046***	−0.029***	−0.034***	−0.031***	−0.021*
	(0.010)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)
Observations	4,775,811	4,775,811	4,775,811	4,775,811	4,745,189	4,745,189	4,745,189	4,745,189
McFadden's R^2	0.669	0.669	0.670	0.671	0.632	0.633	0.635	0.636
20-34 YEAR OLDS								
Morning After Pill	−0.038***	−0.037***	−0.035***	−0.031***	−0.022**	−0.025***	−0.025**	−0.021**
	(0.007)	(0.007)	(0.007)	(0.007)	(0.009)	(0.009)	(0.010)	(0.010)
Observations	12,815,945	12,815,945	12,815,945	12,815,945	12,166,333	12,166,333	12,166,333	12,166,333
McFadden's R^2	0.785	0.785	0.786	0.786	0.707	0.708	0.709	0.710
35-49 YEAR OLDS								
Morning After Pill	−0.014	−0.012	−0.008	−0.010	0.037	0.039	0.038	0.033
	(0.010)	(0.010)	(0.010)	(0.010)	(0.024)	(0.026)	(0.026)	(0.025)
Observations	12,210,243	12,210,243	12,210,243	12,210,243	11,985,343	11,985,343	11,985,343	11,985,343
McFadden's R^2	0.559	0.560	0.560	0.560	0.656	0.657	0.657	0.657
Trends & FEs	Y	Y	Y	Y	Y	Y	Y	Y
Political Controls		Y	Y	Y		Y	Y	Y
Health, Educ, Gender Controls			Y	Y			Y	Y
Condom Availability				Y				Y

NOTES: All Births and First Births are binary variables taking the value of 1 in the case that a women gives live birth and that this occurs at any birth order, or is her first birth (respectively). All models are estimated using logistic regression and include comuna and year fixed. Standard errors are clustered at the level of the comuna. All coefficients are reported as log odds and in each case Pill is a binary variable referring to the availability of the morning after pill in the woman's comuna and (lagged) year. Political controls include party dummies for the mayor in power, the mayor's gender, and the vote margin of the mayor. Health and education controls include the percent of girls out of highschool, education spending by both the municipality and the Ministry of Education and total health spending and health spending on staff and training. Gender controls are the percent of female heads of households living below the poverty line, and the percent of female workers in professional positions in the Municipality. *p<0.1; **p<0.05; ***p<0.01

Table 3.3: The Effect of the Morning After Pill on Fetal Deaths

	All Deaths	Early Gestation	Late Gestation
15-19 YEAR OLDS			
Morning After Pill	−0.063	−0.468**	−0.017
	(0.105)	(0.218)	(0.148)
Mean (deaths/live birth)	0.008	0.002	0.005
Observations	250,331	248,941	249,539
McFadden's R^2	0.199	0.345	0.219
20-34 YEAR OLDS			
Morning After Pill	−0.042	−0.090	−0.067
	(0.055)	(0.105)	(0.065)
Mean (deaths/live birth)	0.007	0.002	0.004
Observations	1,103,721	1,098,062	1,100,414
McFadden's R^2	0.187	0.370	0.149
35-49 YEAR OLDS			
Morning After Pill	−0.370***	−0.727***	−0.412***
	(0.084)	(0.174)	(0.149)
Mean (deaths/live birth)	0.011	0.003	0.006
Observations	264,311	262,155	225,329
McFadden's R^2	0.239	0.394	0.219

NOTES: Total fetal deaths for each group are 1,991, 8,076, and 2,996 for 15-19, 20-34 and 35-49 year olds respectively. All regressions include year and comuna fixed-effects, and comuna-specific trends. Each regression also includes the full set of time varying controls described in table 3.2. Standard errors are clustered by comuna. *p<0.1; **p<0.05; ***p<0.01;

Table 3.4: Emergency Contraception and Aggregate Human Capital

	15-19 year olds			20-34 year olds			35-49 year olds		
Panel A: Mother Characteristics	(1) Yrs Educ	(2) Working	(3) Married	(4) Yrs Educ	(5) Working	(6) Married	(7) Yrs Educ	(8) Working	(9) Married
Morning After Pill	0.022	-0.002	0.000	0.001	-0.004*	-0.003	0.061**	-0.004	-0.001
	(0.021)	(0.002)	(0.001)	(0.014)	(0.002)	(0.005)	(0.028)	(0.005)	(0.007)
Observations	131,605	131,746	131,614	896,230	897,363	896,318	198,885	199,472	198,906
R^2	0.02	0.01	0.01	0.14	0.04	0.17	0.21	0.03	0.247
Panel B: Child Characteristics	(1) Weight	(2) Gestation	(3) Length	(4) Weight	(5) Gestation	(6) Length	(7) Weight	(8) Gestation	(9) Length
Morning After Pill	-1.377	-0.020	0.039	-0.636	-0.023***	0.02	-4.923	-0.016	0.030
	(5.944)	(0.019)	(0.028)	(2.532)	(0.008)	(0.016)	(5.602)	(0.016)	(0.024)
Observations	131,746	131,471	129,880	897,363	895,671	885,932	199,472	198,745	195,863
R^2	0.01	0.01	0.03	0.01	0.01	0.03	0.09	0.01	0.03

Notes: Each column represents an OLS regression, and full controls listed in table 3.2 are included. Working and Married are binary variables, Weight is measured in grams, Gestation in weeks, and Length in centimetres. Summary statistics for these variables are available in table 3.1. Standard errors are clustered at the level of the municipality. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 3.5: The Morning After Pill and Treatment Spillovers

	15-19 Year olds	20-34 Year olds	35-49 Year olds
Panel A: Births			
Morning After Pill	-0.076^{***}	-0.044^{***}	-0.001
	(0.015)	(0.011)	(0.013)
Close $<$ 15 km	-0.056^{***}	-0.033^{***}	0.015
	(0.015)	(0.012)	(0.014)
Close 15-30 km	-0.049^{**}	-0.001	
	(0.019)	(0.012)	
Close 30-45 km	-0.022	0.000	
	(0.022)	(0.019)	
Observations	4,775,811	12,815,945	8,201,384
McFadden's R^2	0.672	0.787	0.549
Panel B: Fetal Deaths			
Morning After Pill	-0.510^{*}	-0.093	-0.750^{**}
	(0.281)	(0.159)	(0.305)
Close $<$ 15 km	-0.078	0.038	-0.175
	(0.312)	(0.205)	(0.282)
Observations	248,941	1,098,062	224,627
McFadden's R^2	0.345	0.370	0.394

Notes: All models are estimated using logistic regressions, and coefficients are reported as log odds. Each regression includes comuna and year fixed effects and comuna-specific trends, and the full set of time-varying controls described in table 3.2. Conley (1999) standard errors are reported. *p<0.1; **p<0.05; ***p<0.01;

Table 3.6: Placebo Tests

	Lag = 3 years		Lag = 4 years		Lag = 5 years	
	(1)	(2)	(3)	(4)	(5)	(6)
PANEL A: 15-19 YEAR-OLDS						
Morning After Pill	−0.002	−0.003	0.009	0.016	−0.002	−0.003
	(0.011)	(0.016)	(0.010)	(0.016)	(0.011)	(0.016)
Close < 15 km		−0.005		0.022		0.018
		(0.020)		(0.021)		(0.024)
Close 15-30 km		0.004		−0.010		0.020
		(0.022)		(0.020)		(0.027)
Close 30-45 km		0.007		−0.011		0.051
		(0.025)		(0.019)		(0.033)
Observations	4,771,813	4,771,813	4,722,210	4,722,210	4,661,289	4,661,289
McFadden's R^2	0.230	0.230	0.228	0.228	0.230	0.230
PANEL A: 20-34 YEAR-OLDS						
Morning After Pill	0.009*	0.007	−0.001	0.006	0.009*	0.007
	(0.005)	(0.012)	(0.007)	(0.015)	(0.005)	(0.012)
Close < 15 km		−0.001		0.018		−0.017
		(0.012)		(0.016)		(0.012)
Close 15-30 km		−0.004		−0.007		0.004
		(0.014)		(0.016)		(0.012)
Close 30-45 km		−0.019		−0.021		0.012
		(0.014)		(0.021)		(0.014)
Observations	12,491,789	12,491,789	12,398,246	12,398,246	12,318,983	12,318,983
McFadden's R^2	0.235	0.235	0.221	0.221	0.214	0.214

NOTES: All specifications are identical to those estimated in tables 3.2 and 3.5. However, instead of using births 1 year subsequent to the reform the outcome variable in each case is births and preceeding the reform by lag= $l \in 3, 4, 5$ years, and hence entirely unaffected in both treatment and control municipalities. *p<0.1; **p<0.05; ***p<0.01.

Appendices

3.A The Chilean Legislative Environment and the Adoption of Emergency Contraception

Discussions surrounding the introduction of emergency contraception in Chile have taken place since at least 1996, when the Chilean Institute of Reproductive Medicine (ICMER for its initials in Spanish) proposed the use of this method to avoid undesired pregnancies in a country where abortion was entirely outlawed (Dides Castillo, 2006). However, the first legislative attention given to this matter occurred when the aforementioned (see section 3.2.1) Institute of Public Health emitted a resolution allowing for the production and sale of 'Postinol', a drug containing levonogestrel by a Chilean laboratory in 2001. The Constitionality of this was quickly challenged, and the drug was prohibited by the Supreme Court.

The emergency contraceptive pill again entered legislative attention in 2004, following the Ministry of Health's publication of a guide suggesting that emergency contraception be used following cases of rape. Following this in 2005, the Subsecretary of Health Dr. Antonio Infante announced that emergency contraception would be freely available to *all* women who requested it, however the President of Chile and the Ministry of Health later declared that this was not the case, leading to removal of the Subsecretary from office.

In November of 2005, the Supreme Court of Chile provided the first constitutional support for the emergency contraceptive pill, voting 5-0 to reverse the decision taken in 2001, allowing emergency contraception to be provided in the case that the mother's life was in danger. Once again however, this finding was challenged shortly thereafter. The same non-governmental institution which had earlier raised a case against ICMER, now challenged the private commercial laboratory in charge of producing and distributing the drug. However, before this case could reach court, this laboratory voluntarily gave up their license to produce the drug, in a three line statement issued by the General Director of the company on February 14, 2006 (Casas Becerra, 2008).

In the same year, a group of 36 parliamentary deputies from conservative parties raised a case with the Constitional Tribunal, claiming that the provision of the emergency contraceptive pill contravened the "National Laws for the Regulation of Fertility", a set of rules issued by the Ministry of Health. This case was only resolved in 2008, with the Constitional Tribunal's finding in favour of this group, hence making illegal any provision by hospitals or health centres controlled by the Ministry of Health (and hence under the jurisdiction of the National Fertility Laws). Fundamentally however, this left the door open for Municipal health centres to distribute the pill freely to women. These Municipal Health Centres are run under the directive of the elected mayor of each Municipality, leaving all remaining legislation regarding the distribution

of the pill up to the 346 mayors in Chile.

In this study I study the period surrounding this 2008 legislation as the cutoff of interest. However, even after this finding the emergency contraceptive pill has not been far from legislative action, with a number of other cases raised. These cases never entirely threatened the continuity of supply of the morning after pill by municipalities, however did cause some confusion for mayors and municipal health bodies in determining whether or not they were legally allowed to prescribe the contraceptive. These cases also resulted in the passing of a number of laws and standards. Most importantly, they resulted in national Law 20.418 which "creates standards for information, guidance and regulatory services in fertility" (author's translation), and the passing of a decree on March 3, 2013, which makes obligatory the provision of the morning after pill to women of any age in any health centre in Chile. This became operative on May 28, 2013, meaning that—at least officially—there are no longer any restrictions in place in the country.

3.B Data Appendix

With the exception of raw birth and fetal death data which requires that the user agree to a number of privacy clauses, all raw and processed data used in this paper is made available online at: https://github.com/damiancclarke/morning-after-pill/tree/master/Data. Birth and fetal death data can be downloaded online at: http://www.deis.cl/descargar-bases-de-datos/ and I make available full processing scripts which convert this into the final dataset used here. In the remainder of this appendix, I provide further details regarding each data source used.

3.B.1 Main Data on Births and Fetal Deaths

Data on all births and deaths in Chile is publicly available for download at http://www.deis.cl/descargar-bases-de-datos/. This contains microdata registers of every birth and fetal death occurring Chile between 1999 and 2012. This is censal data, and is unlikely to miss any births given the importance of registering every child born with authorities in order to receive a national identity number used in all contact with public and private organisations including hospitals and schools. The main analysis in this paper is based on births and fetal deaths occurring between 2005 and 2012 (see table 3.1), however in placebo tests earlier birth data is also used.

3.B.2 Population Data

In order to link the number of births to the number of women of fertile age in each municipality and time period, I consult data from the National Institute of Statistics of Chile (INE). This is made available at http://www.ine.cl/canales/chile_estadistico/familias/demograficas_vitales.php and provides full demographics by age, municipality, and gender.

3.B.3 Time-Varying Municipality and Region Controls

Time-varying municipal controls such as education and health spending, and the number of females working in public government is downloaded from the National System of Municipal Information (SINIM). This provides data as far back as 2005, and is freely available for download online at http://www.sinim.gov.cl/indicadores/busq_serie.php.

Data on municipal elections, mayor's gender, party and vote share is accessed from the Electoral Service of Chile (SERVEL). This provides all electoral results from municipal elections for the full time period of this study. Raw data is available online at http://www.servel.cl/ss/site/mobile/padron_electoral_comunal_por_ano_informe_comunal_anual.html or processed as one line per municipality at the data page of the author's website linked to above.

Finally, I calculate data for alternative contraceptive use based on a series of regionally representative surveys collected every 3 years beginning in 1994. The National Survey of Youth asks respondents whether they use any method of contraception in both their first and most recent sexual activity. In the case that they did not use a condom, they are asked whether this is because they did not have access. Based on this survey, access to condom is calculated as an additional time-varying control. However, it should be noted that this variable can only be calculated at the level of the region (one level above the municipality), given that this survey is not representative at the level of the municipality. Once again, processed data and processing scripts are made available at the data section of the author's site, and, if desired, raw data is available on the web: http://extranet.injuv.gob.cl/Encuesta_Nacional_de_la_Juventud/contenido/index.php.

3.C A Back of The Envelope Consistency Check of Effect Sizes

Using the official Ministry of Health data on the number of pills distributed in each year, I am able to determine whether the effect sizes identified in this study seem to be of reasonable magnitude. These calculations should of course be taken as illustrative only, given that we do not know if all pills distributed were taken by the recipient, nor the rates of pregnancy avoidance conditional upon taking the pill.

According to the adminstrative medications data, 16,857 emergency contraceptive pills were prescribed (in total) in 2009, 2010 and 2011. Of these, 5,736 were prescribed to women 18 years of age and younger, while the remaining 11,121 we prescribed to women over that age of 18. In order to have a rough idea of whether the estimates we find are reasonable, we can compare the approximate reduction in pregnancy estimated from our preferred specification, with the number of pills given out over the period of interest.

Given that the Ministry of Health's administrative data on prescriptions only records the ages of women accessing the pill as 18 and under and 19 and over, I estimate our specification for these

two subgroups. I also calculate the total number of pregnancies in treated (and close to treated) municipalities during the period in which the pill was available. These figures are displayed in table 3.12. In order to determine the reduction in pregnancies which these estimates imply, I compare the theoretical number of pregnancies without the pill, to the number recorded with the pill. For example, in the case of the 18 and under group, the pill acts to reduce pregnancies by $1 - \exp(-0.069) = 0.067$, or 6.7%. So, we inflate the total number of pregnancies for this group (which was 20,713), suggesting that the total number of pregnancies without the pill would have been 22,612 (which we calculate as $\frac{20,713}{1-0.067}$). Thus, the approximate effect of the pill for this group is estaimated as a reduction of 22,200-20,713=1,487 pregnancies. Similar calculations can be run for each subset, to calculate the total estimated effect in each age group.

Based on this methodology, our estimates suggest that the pill accounted for 3,212 fewer pregnancies in the 18 and under age group, and 11,742 fewer pregnancies in the 18 and over age group.[18] Comparing these to total pill disbursements of 5,736 and 11,121, the estimated effects seem to be quite close to actual data on pills acquired. Although the estimates are slightly higher than expected for the 19 and over group (implying -0.86 births per pill dispursed) and perhaps slightly lower than expected for the 18 and under group (-0.45 births per pill dispursed), this back of the envelope consistency check performs remarkably well, and when considering the standard errors on our estimates, certainly falls within the margins that we would expect given the number of pill requests.[19]

If I instead compare the total pill disbursements over the period to the total estimated reduction in pregnancies,[20] this implies an efficiency rate of 71.9% (or that 71.9% of pills should result in an avoided pregnancy to account for the reduction in births. For reference, the United States FDA reports an effectiveness rate of 89% based on typical usage.

[18] The full calculation for the 18 and under group is:

$$\left(\frac{20,713}{1-0.067} - 20,713\right) + \left(\frac{10,370}{1-0.072} - 10,370\right) + \left(\frac{6,141}{1-0.048} - 6,141\right) = 2,596$$

and a similar calculation for the 19 and over group gives

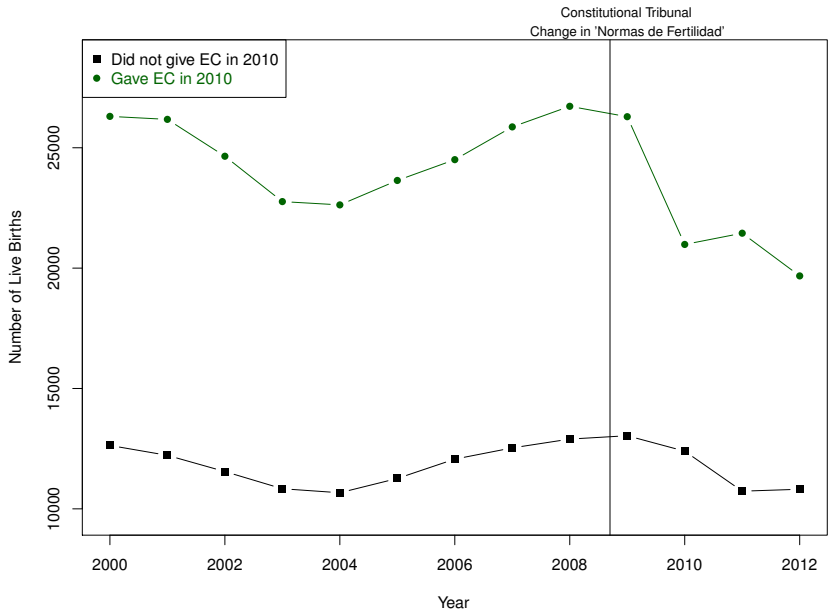
$$\left(\frac{172,557}{1-0.032} - 172,557\right) + \left(\frac{100,749}{1-0.032} - 100,749\right) + \left(\frac{48,756}{1-0.013} - 48,756\right) = 9,525.$$

[19] Further, when considering that pills may have be transferred between women who received the prescription and women who ultimately took the pill, we may be more interested in overall rates for both age groups.

[20] It seems reasonable to make such a comparison given the the spillover effects estimated in this paper suggest that the person accessing the pill may not be the same as the person using the pill.

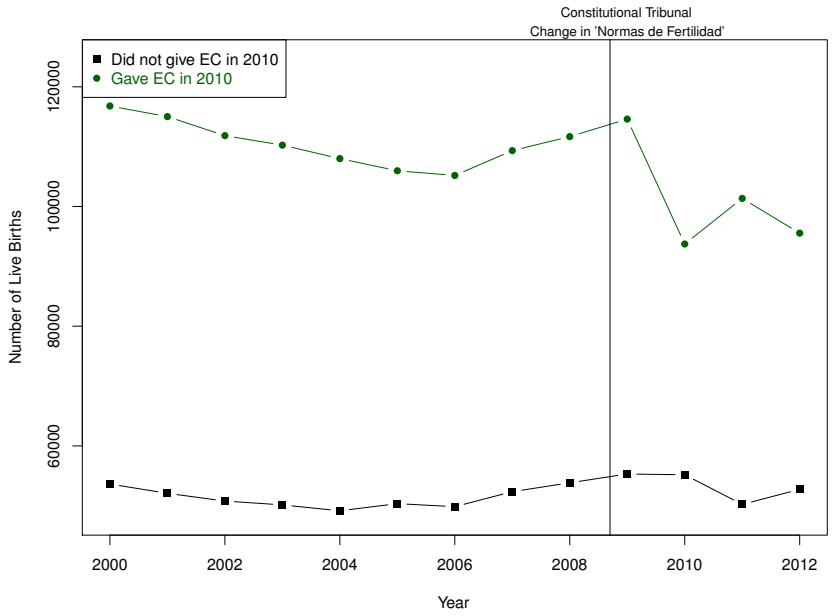
3.D Appendix Figures

Figure 3.9: Birth Trends 2000-2011: 15-19 Years



Note: Some municipalities which did not give the EC pill in 2010 did give the EC pill in 2011 (and vice versa).

Figure 3.10: Birth Trends 2000-2011: 20-34 Years



Note: Some municipalities which did not give the EC pill in 2010 did give the EC pill in 2011 (and vice versa).

3.E Appendix Tables

Table 3.7: OLS Estimates Based on Aggregate Comunal Data

	N Births (1)	N Births (2)	N Births (3)
Pill (15-19 years)	-8.600^{***}	-5.668^{***}	-8.049^{***}
	(1.535)	(1.746)	(2.611)
Observations	2,157	2,157	2,157
R-squared	0.995	0.995	0.995
Pill (20-34 years)	-16.751^{***}	-13.073^{***}	-18.899^{***}
	(3.649)	(4.328)	(7.188)
Observations	2,184	2,184	2,184
R-squared	0.999	0.999	0.999
Pill (35-49 years)	-2.421^{*}	-1.839	0.109
	(1.318)	(1.394)	(1.986)
Observations	2,159	2,159	2,159
R-squared	0.997	0.997	0.997
Pill (All Groups)	-27.742^{***}	-20.567^{***}	-26.821^{***}
	(5.279)	(6.040)	(9.984)
Observations	2,189	2,189	2,189
R-squared	0.999	0.999	0.999
Year & Comuna FEs	Y	Y	Y
Trend, Controls		Y	Y
Spillovers			Y

NOTES: Each panel presents difference-in-difference results for a regression of the total count of pregnancies for the age group in each municipality. All models are estimated by OLS and standard errors are clustered at the level of the comuna. Controls are described in table 3.2. *p<0.1; **p<0.05; ***p<0.01

Table 3.8: OLS Estimates Based on Aggregate Age-Specific Fertility Rate

	ASFR (1)	ASFR (2)	ASFR (3)
Pill (15-19 years)	−4.269***	−2.595**	−3.389**
	(1.206)	(1.236)	(1.638)
Observations	2,157	2,157	2,157
R-squared	0.787	0.793	0.793
Pill (20-34 years)	−3.141***	−1.987**	−1.980*
	(0.881)	(0.909)	(1.149)
Observations	2,184	2,184	2,184
R-squared	0.844	0.848	0.849
Pill (35-49 years)	−0.524	−0.383	0.024
	(0.399)	(0.429)	(0.557)
Observations	2,159	2,159	2,159
R-squared	0.719	0.724	0.725
Pill (All Groups)	−2.200***	−1.437***	−1.529**
	(0.489)	(0.511)	(0.659)
Observations	2,189	2,189	2,189
R-squared	0.878	0.882	0.882
Year & Comuna FEs	Y	Y	Y
Trend, Controls		Y	Y
Spillovers			Y

NOTES: Each panel presents difference-in-difference results for a regression of the age-specific fertility rate (ASFR) for the age group in each municipality. ASFR is defined as the number of births per 1,000 women. In the case of all women, this is called the General Fertility Rate (GFR). All models are estimated by OLS and standard errors are clustered at the level of the comuna. Controls are described in table 3.2. *p<0.1; **p<0.05; ***p<0.01

Table 3.9: OLS Estimates: Fetal Deaths/Live Birth

	All Deaths	Early Gestation	Late Gestation
15-19 YEAR OLDS			
Morning After Pill	0.994	−1.344*	2.175
	(1.674)	(0.730)	(1.467)
Observations	2,157	2,157	2,157
R-squared	0.304	0.331	0.295
20-34 YEAR OLDS			
Morning After Pill	0.378	0.394	0.071
	(0.691)	(0.350)	(0.630)
Observations	2,184	2,184	2,184
R-squared	0.383	0.443	0.349
35-49 YEAR OLDS			
Morning After Pill	−0.540	−0.227	−0.512
	(2.032)	(1.512)	(1.400)
Observations	2,159	2,159	2,153
R-squared	0.317	0.327	0.327

NOTES: Each regression uses as its dependent variable fetal deaths divided by live births in the comuna and age group and is estimated by OLS. All regressions include year and comuna fixed-effects, and comuna-specific trends. Each regression also includes the full set of time varying controls described in table 3.2. Standard errors are clustered by comuna. *p<0.1; **p<0.05; ***p<0.01;

Table 3.10: The Morning After Pill and Pregnancy: Full Covariates

	Pregnancy		
	15-19 year olds	20-34 year olds	35-49 year olds
	(1)	(2)	(3)
Morning After Pill	-0.041***	-0.030***	0.006
	(0.010)	(0.005)	(0.010)
Female Mayor	0.016	-0.005	-0.007
	(0.026)	(0.013)	(0.026)
Mayor's Support	0.054	0.017	-0.129
	(0.084)	(0.042)	(0.085)
Out of School	-0.004	-0.001	-0.001
	(0.003)	(0.001)	(0.003)
Total Education Spending	0.001*	-0.00004	0.001**
	(0.0003)	(0.0002)	(0.0003)
Municipal Education Spending	-0.004***	-0.001**	-0.001**
	(0.001)	(0.0003)	(0.001)
Health Spending	-0.0002	0.0001	-0.0005
	(0.001)	(0.0003)	(0.001)
Health Training	-0.080***	-0.036***	0.005
	(0.024)	(0.012)	(0.025)
Health Staff	0.001	0.001***	-0.0004
	(0.001)	(0.0005)	(0.001)
Female Poverty	-0.0004	-0.0001	0.002***
	(0.001)	(0.0003)	(0.001)
Female Workers	-0.001	-0.0005	0.001
	(0.001)	(0.0004)	(0.001)
Years × Municipality	1,929	1,934	1,934

NOTES: Each model is identical to column (4) of table 3.2. A description of each variable is also provided in table 3.2. Municipality dummies and trends and political party dummies have been omitted for clarity. *p<0.1; **p<0.05; ***p<0.01

Table 3.11: The Morning After Pill and Fetal Death: Full Covariates

	Fetal Death (0-20 Weeks)		
	15-19 year olds	20-34 year olds	35-49 year olds
	(1)	(2)	(3)
Morning After Pill	-0.815***	-0.189*	-0.776***
	(0.237)	(0.113)	(0.217)
Female Mayor	0.987*	0.096	-0.270
	(0.593)	(0.293)	(0.528)
Mayor's Support	1.861	1.168	-0.416
	(1.886)	(0.989)	(1.783)
Out of School	-0.005	-0.003	0.074
	(0.083)	(0.032)	(0.064)
Total Education Spending	0.0001	-0.00003	-0.00003
	(0.0001)	(0.00004)	(0.0001)
Municipal Education Spending	0.0004*	0.0001	0.0001
	(0.0002)	(0.0001)	(0.0001)
Health Spending	-0.0001	0.0002**	0.0001
	(0.0002)	(0.0001)	(0.0001)
Health Training	0.005	-0.003	0.005
	(0.004)	(0.002)	(0.004)
Health Staff	0.00004	0.00004	0.0001
	(0.0002)	(0.0001)	(0.0002)
Female Poverty	0.017	0.002	0.002
	(0.016)	(0.007)	(0.013)
Female Workers	0.008	-0.006	0.018
	(0.019)	(0.008)	(0.014)
Years × Municipality	1,887	1,912	1,891
Akaike Inf. Crit.	2,594.943	4,244.940	2,811.065

NOTES: Each model is identical to column (2) of table 3.3. A description of each variable is also provided in table 3.2. Municipality dummies and trends and political party dummies have been omitted for clarity. *p<0.1; **p<0.05; ***p<0.01

Table 3.12: Back of the Envelope Calculation of Effect Sizes

	18 & Under (1)	19 & Over (2)
Morning After Pill	-0.068^{***}	-0.035^{***}
	(0.017)	(0.009)
Close $<$ 15 km	-0.057^{***}	-0.026^{**}
	(0.021)	(0.010)
Close 15-30 km	-0.042^{*}	0.001
	(0.023)	(0.011)
N Preg (pill)	20,713	172,557
N Preg (close 15)	10,370	100,749
N Preg (close 30)	6,141	48,756
Pills Disbursed	5,736	11,121

NOTES: Regression coefficients and standard errors are calculated in line with specification (3.2). The number of pills disbursed is calculated from administrative data described in figure 3.2, and number of avoided pregnancy is based on regression estimates and total births in administrative data. Further details are provided in appendix 3.C. *p<0.1; **p<0.05; ***p<0.01

Chapter 4

Difference-in-Differences in the Presence of Spillovers*

Chapter Abstract

I propose a method for difference-in-differences (DD) estimation in situations where the stable unit treatment value assumption is violated locally. This is relevant for a wide variety of cases where spillovers may occur between quasi-treatment and quasi-control areas in a (natural) experiment. A flexible methodology is described to test for such spillovers, and to consistently estimate treatment effects in their presence. This methodology is illustrated using two recent examples of contraceptive reform. It is shown that with both the arrival of abortion to Mexico DF, as well as the arrival of the emergency contraceptive pill to certain areas of Chile, reductions in teenage pregnancy occurred in both reform neighbourhoods as well as nearby (but theoretically untreated) neighbourhoods. Where reforms are geographically disperse, I demonstrate that spillovers can cause considerable concern regarding the unbiasedness of the traditional DD estimates widely employed in the economic literature.

*I thank participants in the Impact Evaluation Meeting at the Inter-American Development Bank for useful comments on this draft. Source code, including the Stata module `cdifdif` is available at github.com/damiancclarke/spillovers.

4.1 Introduction

Natural experiments often rely on territorial borders to estimate treatment effects. These borders separate quasi-treatment from quasi-control groups with individuals in one area having access to a program or treatment while those in another do not. In cases such as these where geographic location is used to motivate identification, the stable unit treatment value assumption (SUTVA) is, either explicitly or implicitly, invoked.[2]

However, often territorial borders are porous. Generally state, regional, municipal, and village boundaries can be easily, if not costlessly, crossed. Given this, researchers interested in using natural experiments in this way may be concerned that the effects of a program in a treatment cluster may spillover into non-treatment clusters—at least locally.

Such a situation is in clear violation of the SUTVA's requirement that the treatment status of any one unit must not affect the outcomes of any other unit. In this paper I propose a methodology to deal with such spillover effects. I discuss how to test for local spillovers, and if such spillovers exist, how to estimate unbiased treatment effects in their presence. It is shown that this estimation requires a weaker condition than SUTVA: namely that SUTVA holds between *some* units, as determined by their distance from the treatment cluster. I show how to estimate treatment and spillover effects, and then propose a method to generalise the proposed estimator to a higher dimensional case where spillovers may depend in a flexible way on an arbitrary number of factors.

It is shown that this methodology recovers unbiased treatment estimates under quite general violations of SUTVA. While it is assumed that the distance of an individual to the nearest treatment cluster determines whether stable unit treatment type assumptions hold for that individual, 'distance' is defined very broadly. It is envisioned that this will allow for phenomena such as information flowing from treated to untreated areas, or of untreated individuals violating their treatment status by travelling from untreated to treated areas. In each case distance plays a clear role in the propogation of treatment; either information must travel out, or beneficiaries must travel in. Similarly, this framework allows for local general equilibrium-type spillovers, where a tightly applied program may have an economic effect on nearby markets, but where this effect disipates as distance to treatment increases.

[2]The SUTVA has a long and interesting history, under various guises. Cox (1958) refers to "no interference between different units", before Rubin (1978) introduced the concept of SUTVA (the name SUTVA did not appear until Rubin (1980)). Recent work of Manski (2013), refers to this assumption as Individualistic Treatment Response (ITR).

Turning to empirics, this methodology is illustrated with two examples. I examine how spillovers of reforms across municipal boundaries may contaminate 'traditional' difference-in-differences (DD) estimators. This is applied to two contraceptive reforms where individuals from contiguous or nearby areas can travel to a treatment region to access the reform. It is shown that both the arrival of the morning after pill to certain municipalities of Chile, and abortion to certain districts in Mexico, results in a reduction of births in the given area, as well as in close-by quasi-control areas. As a result, the spillover-robust DD estimator propsed here flexibly captures this effect, correcting for any (local) spillover bias that traditional DD fails to identify.

Although in both these examples distance is a geographic measure, calculated variously as Euclidean distance, shortest distance over roads, and shortest travel times between areas, this methodology should not be considered as limited to spatial spillovers. Univariate measures of distance including propogation through nodes in a network, ethnic distance, ideological distance, or other quantifiable measures of difference between units can be used in precisely the same manner. I also show how multivariate measures of distance, or interactions between distance and other variables, can be similarly employed. This is particularly useful for cases where the effects of spillovers may be expected to vary by individual characteristics such as age, socioeconomic status, access to transport or access to information.

This paper joins recent literature which aims to loosen the strong structure imposed by the SUTVA. Perhaps most notably, it is (in broad terms) an application of Manski's (2013) social interactions framework, focusing on the case where spillovers are restricted to areas local to treatment clusters. However, unlike recent developments focusing on spillovers between treated and control units *within* a treatment cluster (notable examples include McIntosh (2008); Baird et al. (2014); Angelucci and Maro (2010)), this paper focuses on situations where entire clusters are treated, and the status of the *cluster* may affect nearby non-treated clusters. This is likely the case for quasi-experimental studies, where 'experiments' are defined based on geographic boundaries, such as administrative political regions which set different policies.[3]

4.2 Methodology

Define $Y(i,t)$ as the outcome for individual i and time t. The population of interest is observed at two time periods, $t \in \{0,1\}$. Assume that between $t = 0$ and $t = 1$, some fraction of the

[3] A very different case is that of (for example) PROGRESA/Oportunidades, where treatment clusters (ie localities or *localidades*) contained both treatment and control individuals, and the literature is concerned with spillovers between treatment and control individuals within this treatment cluster.

population is exposed to a quasi-experimental treatment. As per Abadie (2005), I will denote treatment for individual i in time t as $D(i,t)$, where $D(i,1) = 1$ implies that the individual was treated, and $D(i,1) = 0$ implies that the individual was not directly treated. Given that treatment only exists between periods 0 and 1, $D(i,0) = 0 \; \forall \; i$.

It is shown by Ashenfelter and Card (1985) that if the outcome is generated by a component of variance process:

$$Y(i,t) = \delta(t) + \alpha D(i,t) + \eta(i) + \nu(i,t) \tag{4.1}$$

where $\delta(t)$ refers to a time-specific component, α as the impact of treatment, $\eta(i)$ a component specific to each individual, and $\nu(i,t)$ as a time-varying individual (mean zero) shock, then a sufficient condition for identification (a complete derivation is provided by Abadie (2005)) is:

$$P(D(i,1) = 1|\nu(i,t)) = P(D(i,1) = 1) \; \forall \; t \in \{0,1\}. \tag{4.2}$$

In other words, identification requires that selection into treatment does not rely on the unobserved time-varying component $\nu(i,t)$. If this condition holds, then the classical DD estimator provides an unbiased estimate of the treatment effect:

$$\begin{aligned} \alpha = &\{E[Y(i,1)|D(i,1) = 1] - E[Y(i,1)|D(i,1) = 0]\} \\ &- \{E[Y(i,0)|D(i,1) = 1] - E[Y(i,0)|D(i,1) = 0]\} \end{aligned} \tag{4.3}$$

where E is the expectations operator.

Assume now, however, that treatment is not precisely geographically bounded. Specifically, I propose that those living in control areas 'close to' treatment areas are able to access treatment, either partially or completely. Such a case allows for a situation where individuals 'defy' their treatment status, by travelling or moving to treated areas, or where spillovers from treatment areas are diffused through general equilibrium processes. Define $R(i,t)$ where:

$$R(i,t) = \begin{cases} 1 & \text{if an individual resides close to, but not in, a treatment area} \\ 0 & \text{otherwise} \end{cases}$$

As treatment occurs only in period 1, $R(i,0) = 0$ for all i. Similarly, as living in a treatment area itself excludes individuals from living 'close to' the same treament area, $R(i,t) = 0$ for all i such that $D(i,t) = 1$.

Generalising from (4.1), now I assume that $Y(i,t)$ is generated by:

$$Y(i,t) = \delta(t) + \alpha D(i,t) + \beta R(i,t) + \eta(i) + \nu(i,t) \tag{4.4}$$

If we observe only $Y(i,t)$, $D(i,t)$ and $R(i,t)$, a sufficient condition for estimation now consists of (4.2) and the following assumption:

$$P(R(i,1) = 1|\nu(i,t)) = P(R(i,1) = 1) \ \forall \ t \in \{0,1\}. \tag{4.5}$$

This requires that both treatment, and being close to treatment cannot depend upon individual-specific time-variant components. To see this, write (4.4), adding and subtracting the individual-specific component $E[\eta(i)|D(i,1), R(i,1)]$:

$$Y(i,t) = \delta(t) + \alpha D(i,t) + \beta R(i,t) + E[\eta(i)|D(i,1), R(i,1)] + \varepsilon(i,t) \tag{4.6}$$

where, following Abadie (2005), $\varepsilon(i,t) = \eta(i) - E[\eta(i)|D(i,1), R(i,1)] + \nu(i,t)$. We can write $\delta(t) = \delta(0) + [\delta(1) - \delta(0)]t$, and write $E[\eta(i)|D(i,1), R(i,1)]$ as the sum of the expectation of the individual-specific component $\eta(i)$ over treatment status and 'close' status[4]. Finally define μ (the intercept at time 0) as:

$$\mu = E[\eta(i)|D(i,1) = 0, R(i,1) = 0] + \delta_0,$$

τ, a fixed effect for treated individuals, as

$$\tau = E[\eta(i)|D(i,1) = 1, R(i,1) = 0] - E[\eta(i)|D(i,1) = 0, R(i,1) = 0],$$

γ, a similar fixed effect for individuals close to treatment, as

$$\gamma = E[\eta(i)|D(i,1) = 0, R(i,1) = 1] - E[\eta(i)|D(i,1) = 0, R(i,1) = 0]$$

and δ, a time trend, as $\delta = \delta(1) - \delta(0)$. Then from the above and (4.6) we have:

$$Y(i,t) = \mu + \tau D(i,1) + \gamma R(i,1) + \delta t + \alpha D(i,t) + \beta R(i,t) + \varepsilon(i,t). \tag{4.7}$$

Notice that this (estimable) equation now includes the typical DD fixed effects τ and δ and the double difference term α. However it also includes 'close' analogues γ (an initial fixed effect), and β: the effect of being 'close to' a treatment area. This equation partitions observations into three mutually exclusive binary groups: treated, close to treated, and un-treated, and, as such, $R(i,t) \cdot D(i,t) = 0$ for all individuals.

[4] $E[\eta(i)|D(i,1), R(i,1)] = E[\eta(i)|D(i,1) = 0, R(i,1) = 0] + (E[\eta(i)|D(i,1) = 1, R(i,1) = 0] - E[\eta(i)|D(i,1) = 0, R(i,1) = 0]) \cdot D(i,1) + (E[\eta(i)|D(i,1) = 0, R(i,1) = 1] - E[\eta(i)|D(i,1) = 0, R(i,1) = 0]) \cdot R(i,1)$.

From the assumptions in (4.2) and (4.5) it holds that $E[(1, D(i,1), R(i,1), D(i,t),\ R(i,t)) \cdot \varepsilon(i,t)] = 0$, which implies that all parameters from (4.7) are consistently estimable by OLS. Importantly, this includes consistent estimates of α and β: the effect of the program treatment and spillover effects on outcome variable $Y(i,t)$. Then, from (4.7), a our coefficients of interest α and β are:

$$\begin{aligned}\alpha = &\{E[Y(i,1)|D(i,1)=1, R(i,1)=0] - E[Y(i,1)|D(i,1)=0, R(i,1)=0]\} \\ &- \{E[Y(i,0)|D(i,1)=1, R(i,1)=0] - E[Y(i,0)|D(i,1)=0, R(i,1)=0]\},\end{aligned}$$

and

$$\begin{aligned}\beta = &\{E[Y(i,1)|D(i,1)=0, R(i,1)=1] - E[Y(i,1)|D(i,1)=0, R(i,1)=0]\} \\ &- \{E[Y(i,0)|D(i,1)=0, R(i,1)=1] - E[Y(i,0)|D(i,1)=0, R(i,1)=0]\}.\end{aligned}$$

where the sample estimate of each parameter is generated by a least squares regression of (4.7) using a random sample of $\{Y(i,t), D(i,t), R(i,t) : i = 1, \ldots, N, t = 0, 1\}$.

4.3 A Spillover-Robust Double Differences Estimator

We are interested in estimating difference-in-difference parameters α and β from (4.7). I will refer to these estimators respectively as the average treatment effect on the treated (ATT), and the average treatment effect on the close to treated (ATC). Average treatment effects are cast in terms of the Rubin (1974) Causal Model.

Following a potential outcome framework, I denote $Y^1(i,t)$ as the potential outcome for some person i at time t if they were to receive treatment, and $Y^0(i,t)$ if the person were not to receive treatment. Our ATT and ATC are thus:

$$ATT = E[Y^1(i,1) - Y^0(i,1)|D(i,1)=1] \tag{4.8}$$

$$ATC = E[Y^1(i,1) - Y^0(i,1)|R(i,1)=1], \tag{4.9}$$

As is typical in the potential outcome literature, estimation is hindered by the reality that only one of $Y^1(i,t)$ or $Y^0(i,t)$ is observed for a given individual i at time t. The realised outcome can thus be expressed as $Y(i,t) = Y^0(i,t)\cdot(1-D(i,t))(1-R(i,t)) + Y^1(i,t)\cdot D(i,t) + Y^1(i,t)\cdot R(i,t)$, where, depending on an individual's time varying treatment and close status, we observe either $Y^0(i,t)$ (untreated) or $Y^1(i,t)$ (treated or close). Thus, in order to be able to estimate the quantities of interest, we rely on averages over the entire population, rather than average of

individual treatment effects. As is typical in difference-in-differences identification strategies, consistent estimation requires parallel trends assumptions. In the case of treatment *and* local spillovers, this relies on:

Assumption 1. ***Parallel trends in treatment and control:***
$E[Y^0(i,1) - Y^0(i,0)|D(i,1) = 1, R(i,1) = 0] = E[Y^0(i,1) - Y^0(i,0)|D(i,1) = 0, R(i,1) = 0]$,

Assumption 2. ***Parallel trends in close and control:***
$E[Y^0(i,1) - Y^0(i,0)|D(i,1) = 0, R(i,1) = 1] = E[Y^0(i,1) - Y^0(i,0)|D(i,1) = 0, R(i,1) = 0]$.

In other words, assumption 1 and 2 state that in the absence of treatment, the evolution of outcomes for treated units and for units close to treatment would have been parallel to the evolution of entirely untreated units. This is the fundamental DD identifying assumption of parallel trends, generalised to hold for treatment *and* close to treatment status. Note that in the above, we no longer need to make *any* assumptions regarding parallel trends between treatment and close to treatment units allowing for direct interactions between those living in treatment areas, and those living close by.

However, as a matter of course, in order to consistently estimate any treatment effect, some form of the SUTVA must be invoked. Typically, this requires that each individual's treatment status does not affect each other individual's potential outcome. Here, I loosen SUTVA. In the remainder of this article, it will be assumed that:

Assumption 3. ***SUTVA holds for some units:***
There is some subset of individuals $j \in J$ of the total population $i \in N$ for whom potential outcomes (Y_j^0, Y_j^1) are independent of the treatment status $D = \{0,1\}$ $\forall_{i \neq j} \in N$.

Fundamentally, this assumption implies that SUTVA need not hold among all units. Now, rather than identification relying on each unit not affecting each other unit, it relies on there existing at least some subset of units which are not affected by the treatment status of others.[5]

Finally, I assume that spillovers, or violations of SUTVA, do not occur randomly in the population:

Assumption 4A. ***Assignment to close to treatment depends on observable $X(i,t)$:***
There exists an assignment rule $\delta\Big(X(i,t)\Big) = \{0,1\}$ which maps individuals to close to treatment

[5]This is an identifying assumption. If all 'non-treatment' units are affected by spillovers from the treatment area, a consistent treatment effect cannot be estimated using this methodology. This is a general rule and can be couched in Heckman and Vytlacil (2005)'s terms: 'The treatment effect literature investigates a class of policies that have partial participation at a point in time so there is a "treatment" group and a "comparison" group. It is not helpful in evaluating policies that have universal participation.' (or in this case, universal participation and spillovers.

status $R(i,t)$*, where* $R(i,t) \equiv \delta\Big(X(i,t)\Big) = \mathbf{1}_{X(i,t)<d}$*,* $X(i,t)$ *is an observed covariate, and* d *is a fixed scalar cutoff.*

This restriction is quite strong, and is loosened in coming sections. In other words, it simply states that violations of SUTVA occur in an observable way. For example, if SUTVA does not hold locally to the treatment area, assumption 4A implies that we are able to define what 'local' is. While this article focuses on an X_i representing geographic distance, these derivations do not imply that this must be the case. The 'close' indicator $R(i,t)$ could depend on a range of phenomena including euclidean space, ethnic distance, edges between nodes in a network, or, as I return to discuss in section 4.3.3, multi-dimensional interactions between measures such as these and economic variables.

Proposition 1. *Under assumptions 1 to 4A, the ATT and ATC can be consistently estimated by least squares when controlling, parametrically or non-parametrically, for* $R(i,t) = \mathbf{1}_{X(i,t)\leq d}$*.*

Proof of Proposition. Provided in appendix 4.A. ■

In the following two subsections I examine these estimands in turn.

4.3.1 Estimating the Treatment Effect in the Presence of Spillovers

From proposition 1, we can consistently estimate α and β, our estimands of interest, with information on treatment status, and close to treatment status, along with outcomes $Y(i,t)$ at each point in time. In a typical DD framework, we observe $Y(i,t)$ and $D(i,t)$, however, do not fully observe $R(i,t)$, an individual's close/non-close status.

We do however, assume that $X(i,t)$, the variable measuring 'distance' to treatment is observed. From assumption 4A, we could thus map $X(i,t)$ to $R(i,t)$ using the indicator function, *if* we know the scalar value d, which represents the threshold of what is considered 'close to treatment'. *Ex ante*, in the absence some economic model, there is no reason to believe that d will be observed by researchers.[6] In the remainder of this section I discuss how to determine $R(i,t)$ based on $X(i,t)$, in the absence of a known value for d.

[6] That is not to say that economic intuition cannot play a role in suggesting what a reasonable value of d might be. For example, if treatment is the receipt of a program with a clear expected value and travel costs to access the program increase with distance, there will exist a cut-off point beyond which individuals will be unwilling to travel. Similarly, if treatment must be accessed in a fixed amount of time and propogation of treatment is not instantaneous, a limit for d may be calculable. This is a point I return to in empirical estimates where one illustration is based on access to the emergency contraceptive pill.

In order to do so, we re-write (4.7) as:

$$\tilde{Y}(i,t) = \mu + \alpha D(i,t) + \upsilon(i,t). \tag{4.10}$$

where $\upsilon(i,t) = \beta R(i,t) + \varepsilon(i,t)$, and for ease of notation the fixed effects $D(i,1), R(i,1)$ and t have been concentrated out to form $\tilde{Y}(i,t)$ in line with the Frisch–Waugh–Lovell (FWL) theorem. If we were to estimate $\hat{\alpha}$ from the above regression ignoring the potential presence of spillovers, then we have that the expectation of $\hat{\alpha}$ is:

$$\begin{aligned} E[\hat{\alpha}] &= \alpha + \beta \frac{\text{Cov}[D(i,t), R(i,t)]}{\text{Var}[D(i,t)]} + \frac{\text{Cov}[D(i,t), \varepsilon(i,t)]}{\text{Var}[D(i,t)]} \\ &= \alpha + \beta \frac{\text{Cov}[D(i,t), R(i,t)]}{\text{Var}[D(i,t)]}, \end{aligned} \tag{4.11}$$

where the second line comes from (4.2), which implies that $E[\text{Cov}(D(i,t), \varepsilon(i,t))] = 0$. So far I have attached no functional form to $R(i,t)$. Define $R(i,t)$ as:

$$R(i,t) = \sum_{k=1}^{K} R^k(i,t) \tag{4.12}$$

where:

$$R^k(i,t) = \begin{cases} 1 & \text{if } \; X_i > (k-1) \cdot h \;\; \text{and} \;\; X_i \leq k \cdot h \\ 0 & \text{otherwise} \end{cases} \qquad \forall k \in (1, 2, \ldots, K). \tag{4.13}$$

In the above expression h refers to a bandwidth type parameter, which partitions the continuous distance variable X_i into groups of distance h.[7]

From the above, we have partitioned X_i into K different groups. However, we are still unable to say anything about the distance d above which spillovers no longer occur. From assumptions 2 and 3, we do however know that $d < max(X_i)$, implying that there are at least some units for whom spillovers do not occur. From (4.12) and the preceding logic, this suggests that d can be recovered following the iterative procedure laid out below.

If we start by estimating a typical DD specification like (4.10), our estimated treatment

[7] So, if for example X_i refers to physical distance to treatment and the minimum and maximum distances are 0 and 100km respectively, h could be set as 5km, resulting in at most 19 different indicators R^k (and one omitted base for living between 95-100km from treatment), of which each individual i in time t can have at most one switched on.

effect, which I now denote $\hat{\alpha}^0$ is:

$$E[\hat{\alpha}^0] = \alpha + \beta\frac{\text{Cov}[D(i,t), R^1(i,t)]}{\text{Var}[D(i,t)]} + \beta\frac{\text{Cov}[D(i,t), R^2(i,t)]}{\text{Var}[D(i,t)]} + \\ \ldots + \beta\frac{\text{Cov}[D(i,t), R^K(i,t)]}{\text{Var}[D(i,t)]}.$$

If spillovers exist below some distance d, then $\text{Cov}[D(i,t), R^k(i,t)] > 0 \quad \forall \quad kh < d$, given that $D(i,t)$—the treatment status in a treated area—affects the close to treated status in nearby areas. If this is the case, and if spillovers work in the same direction as treatment, then $|E[\hat{\alpha}^0]| < |\alpha|$, implying that the estimated treatment effect will be attenuated by treatment spillover to the control group.

We can then re-estimate (4.10), however now *also* condition out $R^1(i,t)$ prior to estimating α. Our resulting estimate, $\hat{\alpha}^1$, will have the expectation:

$$E[\hat{\alpha}^1] = \alpha + \beta\frac{\text{Cov}[D(i,t), R^2(i,t)]}{\text{Var}[D(i,t)]} + \ldots + \beta\frac{\text{Cov}[D(i,t), R^K(i,t)]}{\text{Var}[D(i,t)]}.$$

Once again, if spillovers exist and are of the same sign as treatment, then the estimate $\hat{\alpha}^1$ will be attenuated, but not as badly as $\hat{\alpha}^0$ given that we now partially correct for spillovers up to a distance of h. In this case: $|E[\hat{\alpha}^0]| < |E[\hat{\alpha}^1]| < |\alpha|$. If, on the other hand, spillovers do not exist, then we will have that $|E[\hat{\alpha}^0]| = |E[\hat{\alpha}^1]| = |\alpha|$. This leads to the following hypothesis test, where for efficiency reasons $\hat{\alpha}^0$ and $\hat{\alpha}^1$ are estimated by seemingly unrelated regression:

$$H_0 : \alpha^0 = \alpha^1 \qquad H_1 : \alpha^0 \neq \alpha^1.$$

From Zellner (1962), the test statistic has a χ^2_1 distribution. If we reject H_0 in favour of the alternative, this indicates that partially correcting for spillovers affects the estimated coefficient α, implying that spillovers occur at least up to distance h, and that further tests are required.

Rejection of the null suggests that another iteration should be performed, this time removing $R^1(i,t)$ and $R^2(i,t)$ from the error term $\upsilon(i,t)$ in (4.10), and the corresponding parameter α^2 be estimated. If spillovers do occur at least up to distance $2h$, we expect that $|E[\hat{\alpha}^0]| < |E[\hat{\alpha}^1]| < |E[\hat{\alpha}^2]| < |\alpha|$, however if spillovers only occur up to distance h, we will have $|E[\hat{\alpha}^0]| < |E[\hat{\alpha}^1]| = |E[\hat{\alpha}^2]| = |\alpha|$. This leads to a new hypothesis test:

$$H_0 : \alpha^1 = \alpha^2 \qquad H_1 : \alpha^1 \neq \alpha^2,$$

where the test statistic is distributed as outlined above. Here, rejection of the null implies that spillovers occur at least up to distance $2h$, while failure to reject the null suggests that spillovers

only occur up to distance h.

This process should be followed iteratively up until the point that the marginal estimate $\hat{\alpha}^{k+1}$ is equal to the preceding estimate $\hat{\alpha}^{k}$. At this point, we can conclude that units at a distance of at least kh from the nearest treatment unit are not affected by spillovers, and hence a consistent estimate of α can be produced. Finally, this leads to a conclusion regarding d and the indicator function $R(i,t) = \mathbf{1}_{X(i,t)\leq d}$. When controlling for the marginal distance to treatment indicator no longer affects the estimate of the treatment effect α^{k}, we can conclude that $d = kh$, and thus correctly identify $R(i,t) = \mathbf{1}_{X(i,t)\leq kh}$ in data.

4.3.2 Estimation the Magnitude of Spillovers

In section 4.3.1, I discuss the consistent estimation of α, the effect of being in a treatment area. The extension of this methodology to consistently estimate β, the effect of being close to treatment, is reasonably straightforward. Once the scalar value d has been determined, and with data $\{Y(i,t), D(i,t), X(i,t) : i = 1, \ldots, N, t = 0, 1\}$ in hand, we can use d to map $X(i,t)$ into $R(i,t)$. Given the above we can now estimate (4.7), and form consistent estimates $\hat{\beta}$ and $\hat{\alpha}$ using OLS.

The estimate $\hat{\beta}$ will be the average treated effect on the close to treated (ATC), and will be one summary value for all areas to which spillovers occur. However, more information regarding the precise manner of propogation can be observed by estimating with the re-parametrized $R(i,t) = R^{1}(i,t) + \ldots + R^{d/h}(i,t)$ from (4.12) instead of the indicator variable $R(i,t)$. In the case of the former, the spillover coefficient can be observed in bins of various distance from treatment, rather than only one average "close" paramater where the latter binary $R(i,t)$ is included.

This suggests an alternative spillover test, in the style of that proposed in section 4.3.1. Rather than observing $\hat{\alpha}^{j}$ at each stage of the estimation process, $\hat{\beta}^{j}$ can be directly observed. If $\hat{\beta}_{j} \neq 0$, this suggests that the effect on the marginal close to treatment area is different to the effect in the (remaining) control area. If spillovers are the estimand of interest, additional $R^{j}(i,t)$ controls can be added until the hypothesis: $H_0 : \beta_j = 0$ cannot be rejected for the marginal parameter. The empirical illustrations in section 4.4 estimate both the treatment effect, as well as spillovers at varying distances from treatment.

4.3.3 Estimating with Multidimensional Spillovers

Previously it has been assumed that $R(i,t)$ is a function of a unidimensional distance measure $X(i,t)$. I now generalise this to a multidimensional case where $R(i,t)$ may depend upon an arbitrary number of variables $\mathbf{X}(i,t)$. This allows for cases where distance to treatment may interact with some other variable, such as income, ownership of a vehicle or access to information (among other things). In order to allow for spillovers to depend upon a range of observable variables, we must generalise assumption 4A. In order to do this, the following new terminology is introduced, following Zajonc (2012). An assignment rule, δ, maps units with covariates $\mathbf{X} = \mathbf{x}$ to close assignment $R(i,t)$:

$$\delta : \mathcal{X} \to \{0,1\}.$$

This leads to a close-to-treatment assignment set $\mathbb{T}$ defined as:

$$\mathbb{T} \equiv \{\mathbf{x} \in \mathcal{X} : \delta(\mathbf{x}) = 1\}$$

whose complement $\mathbb{T}^c$ is known as the control assignment set. Finally then, we can write the treatment assignment rule[8]:

$$\delta(\mathbf{x}) \equiv \mathbf{1}_{\mathbf{x}\in\mathbb{T}}. \tag{4.14}$$

With this (multidimensional) treatment assignment rule in hand, a more general version of assumption 4A can now be provided:

Assumption 4B. ***Assignment to close to treatment depends on observable*** $\mathbf{X}(i,t)$***:*** *An multidimensional assignment rule* $\delta(x) = \mathbf{1}_{\mathbf{x}\in\mathbb{T}}$ *exists which maps individuals to close to treatment status* $R(i,t)$*, where* $\mathbf{X}(i,t)$ *are observed covariates, and* $\mathbb{T}$ *is a fixed function of* $\mathbf{X}(i,t)$*.*

Proposition 2. *Under assumptions 1–3 and 4B, the ATT and ATC can be consistently estimated by least squares when controlling, parametrically or non-parametrically, for* $R(i,t) = \mathbf{1}_{\mathbf{x}\in\mathbb{T}}$*.*

Proof of Proposition. Provided in appendix 4.A. ■

Now, in the same manner, we can go about generating our estimands of interest, replacing $R(i,t) = \mathbf{1}_{X_i \leq d}$ with $R(i,t) = \mathbf{1}_{\mathbf{x}\in\mathbb{T}}$. The most compuationally demanding step in this estimation procedure is in forming a parametric or non-parametric version of the underlying function $R(i,t)$ over which to search. In a unidimensional framework it is reasonably straightforward to form local linear bins for $R(i,t)$. However, in the multidimensional framework this is no longer the

[8] The uni-dimensional case discussed up to this point is just a particular application of the treatment assignment rule where $\mathbf{X}(i,t) = X(i,t)$ and $\mathbb{T} \equiv \{x < d : \delta(x) = 1\}$

case. Additionally, as the dimensionality of $\mathbf{X}$ rises, the number of search dimensions for spillovers also rises, leading to curse of dimensionality type considerations in the estimation of α.

The particular functional form assigned to $R(i,t)$ will be context-specific, and ideally driven by economic theory. As mode of example, below we consider the case where $R(i,t) = f(X_1, X_2)$ is a function of two variables, one binary and the other continuous. Such a case would be appropriate for a situation in which spillovers depend upon distance to treatment and some indicator, such as exceeding some income threshold. Consider the case where $X_1 \in \{0,1\}$ is binary, and X_2 continuous. Then we can parametrise $R(i,t)$ as:

$$\begin{aligned} R(i,t) &= f(X_1, X_2) \\ &= X_1 \cdot [R_2^1(i,t) + \cdots + R_2^K(i,t)] \\ &+ (1 - X_1) \cdot [R_2^1(i,t) + \cdots + R_2^K(i,t)]. \end{aligned}$$

where $R_2^k(i,t)\ \forall\ k \in 1 \dots K$ is defined based on $X_2(i,t)$ as per (4.13). Estimation of α can then proceed iteratively as in section 4.3.1. First a traditional DD parameter is estimated ignoring the possiblity that spillovers exist, leading to the proposed estimate $\hat{\alpha}^0$. Then $X_1 \cdot R_2^1(i,t)$ and $(1 - X_1) \cdot R_2^1(i,t)$ are included in the regression, leading to an updated estimate $\hat{\alpha}^1$. If the hypothesis $H_0 : \alpha^0 = \alpha^1$ cannot be rejected this suggests that spillovers are not a relevant phenomenon for either group, and the estimate of $\hat{\alpha}^0$ is accepted as the ATT. Otherwise, an additional iteration is made until the inclusion of the marginal $R_2^k(i,t)$ indicators for $X_1 \in \{0,1\}$ no longer affect the estimated effect α^k.

4.4 An Empirical Illustration: Spillovers and Contraceptive Reforms

I consider two empirical examples to motivate and demonstrate spillover-robust DD estimation. I focus on two localised contraceptive reforms in different countries. The first is the legalisation of abortion in Mexico city in April of 2007, and the second the expansion of morning after pill availability in certain municipalities of Chile in 2008. Both reforms were sharp, resulting in a large jump in reported rates of contraceptive access, and arrived to only certain areas of the country. In both cases, the geographic location of the reform was defined by the nature of local municipal-level policies, resulting in seperate policies in different municipalities in the country.[9]

[9] I refer to geographic units in each case as municipalities. In Mexico these are referred to as *municipios*, or in the case of Mexico City *delegaciones* and are the level below the State (there are 2,473 in total). In Chile these are known as *comunas*, (of which there are 346) and are also the level below the state.

Contraceptive reform provides a useful test of a spillover-robust DD methodology. Firstly, the arrival is plausibly exogenous at the level of the treated woman.[10] Secondly, the incentives to access contraceptives, especially post-coital treatments such as the morning after pill and abortion is high. Even if a woman is geographically excluded from a treatment municipality, given that the economic and psychic costs of an undesired birth are very high, considerable incentives will exist to travel from a non-treatment area to a treatment area in order to access fertility control policies. Thirdly, contraceptive information may also be important in determining contraceptive behaviour, and this information may travel through (local) friendship networks.

Some further details regarding each reform are provided in the sections below. In each case we estimate traditional difference-in-differences parameters under the assumption that spillovers do not exist (and hence the SUTVA holds), and then augment these estimates with the estimator discussed in the previous sections.

4.4.1 Abortion Reform in Mexico

On April 24, 2007 Mexico City passed a law which which legalised abortion under all circumstances in the first 12 weeks of pregnancy (see for example Fraser (2014) for a discussion). This was a radical change from previous laws which outlawed abortion in all but the extreme circumstances of rape, to save the mother's life, or in the case of fetal inviability. This law was *only* passed in Mexico City (or *Distrito Federal*), the administrative capital, and a region of Mexico containing approximately 8% of the population.

This reform resulted in free and legal access, with legal abortions being widely used. This service has accounted for slightly than 89,000 abortions between 2007 and 2012 Becker and Díaz-Olavarrieta (2013). Women of all reproductive ages have accessed abortion (ages 11-50), with slightly more than 20% of users being teenaged women. In this paper I focus on the effect of legal abortion usage on teenagers, however figures for non-teenagers (showing broadly similar patterns) are provided in appendices 4.D and 4.E. A more comprehensive discussion of the Mexico abortion reform is provided in appendix 4.B.

The effect of this reform on the number of teenage births is examined.[11] In order to do

[10] Both reforms in question were due to legislative changes which were eventually upheld by the supreme court of the country. Additional details regarding the Chile reform are described in appendix 3.A and additional details on the Mexican reform are provided in appendix 4.B.

[11] Numbers of births are used rather than rates given the difficulty in obtaining precise measures of the number of women of each age living in each municipality in each year. Although censal counts of women *are* available at the municipality level, this is only for census years (2000, 2005 and 2010).

so, data from various sources is collected. Microdata on all registered births is collated from yearly vital statistics registers provided by the Mexican National Institute of Statistics and Geography for the years 2001–2010. This is crossed with a range of municipality×year varying measures including spending on medical staff by municipalities, educational investments and stocks, municipal involvement in the *Seguro Popular* program,[12] and access to other types of contraceptives. This results in data on 22.20 million births, of which 1.47 million are to teenage mothers. Basic descriptive statistics of births in Mexico DF, births in municipalities close (<30 km) from Mexico DF, births in other states, along with municipal controls are provided in table 4.1.

Table 4.1: Descriptive Statistics (Mexico)

	Observations	Mean	Std. Dev.	Min.	Max.
Treatment	24550	0.00	0.04	0	1
Close to Treatment	24550	0.00	0.05	0	1
Number of Births (Mexico DF)	160	11744.75	8835.83	1550	34729
Number of Births (Close to DF)	250	12419.40	9254.72	1550	39745
Number of Births (Other Areas)	24300	785.99	3153.99	0	86659
Year (2001-2010)	24550	2005.50	2.87	2001	2010
Number of Medical Staff	24550	57.97	250.81	0	6212
Number of Classrooms	24550	303.51	1000.80	0	19280
Number of Libraries	24550	4.27	16.95	0	708
Municipal Income	24550	75.51	254.56	0	6615
Municipal Spending	24550	82.71	271.05	0	6615
Regional Unemployment Rate	24550	2.93	1.46	0	9

NOTES: Observations are for 2,455 municipalities in 10 years. Number of births refers to total counts for all women aged 15-49 in each municipality within the given area. Municipal income and municipal spending refer to tax receipts and outlays, and are expressed in millions of pesos.

Table 4.2 provides estimates of the effect of the reform on the total number of births by teenagers in treatment municipalities. In column 1 we estimate a specification similar to (4.10): the traditional DD estimate which does not account for spillovers.[13] This is then extended in columns 2 to 5 to account for spillovers (if necessary). These additional columns show the iterative estimation process entailed in the spillover-robust DD process, as described in section 4.3 and equation (4.7). In the case that spillovers exist, and that these spillovers work in the same direction as the treatment effect itself, we should expect that we can reject the null that $\beta < 0$ for coefficient estimates on 'Close' controls. If however, $\hat{\beta}$ is not significantly different to zero, this suggests that areas 'close to treatment' are not different from areas far away from treatment, and that augmenting the specification to account for local spillovers is unnecessary.

[12] *Seguro Popular* is one of the largest publicly-funded health insurance programs in the world, offering coverage to all individuals not covered by private (employer financed) health insurance (Bosch and Campos-Vazquez, 2014). This covers (among many other things) basic antenatal care and contraceptive access.

[13] Rather than estimating (4.10) precisely as written, we estimate a more flexible specification including time varying controls, full time and municipality fixed effects, and municipal trends. The intution however is unchanged.

Table 4.2: Treatment Effects and Spillovers: Mexico (15-19 year olds)

	N Birth (1)	N Birth (2)	N Birth (3)	N Birth (4)	N Birth (5)
Treatment	-125.3***	-126.0***	-127.0***	-127.2***	-127.2***
	(45.33)	(45.36)	(45.33)	(45.32)	(45.32)
Close 1		-119.9**	-120.7**	-120.9**	-120.9**
		(52.69)	(52.87)	(52.88)	(52.88)
Close 2			-40.51**	-40.70**	-40.70**
			(19.92)	(19.92)	(19.92)
Close 3				-9.295	-9.296
				(15.62)	(15.62)
Close 4					-0.0524
					(13.95)
Mean	1,632	1,632	1,632	1,632	1,632
Regions×Time	24,550	24,550	24,550	24,550	24,550

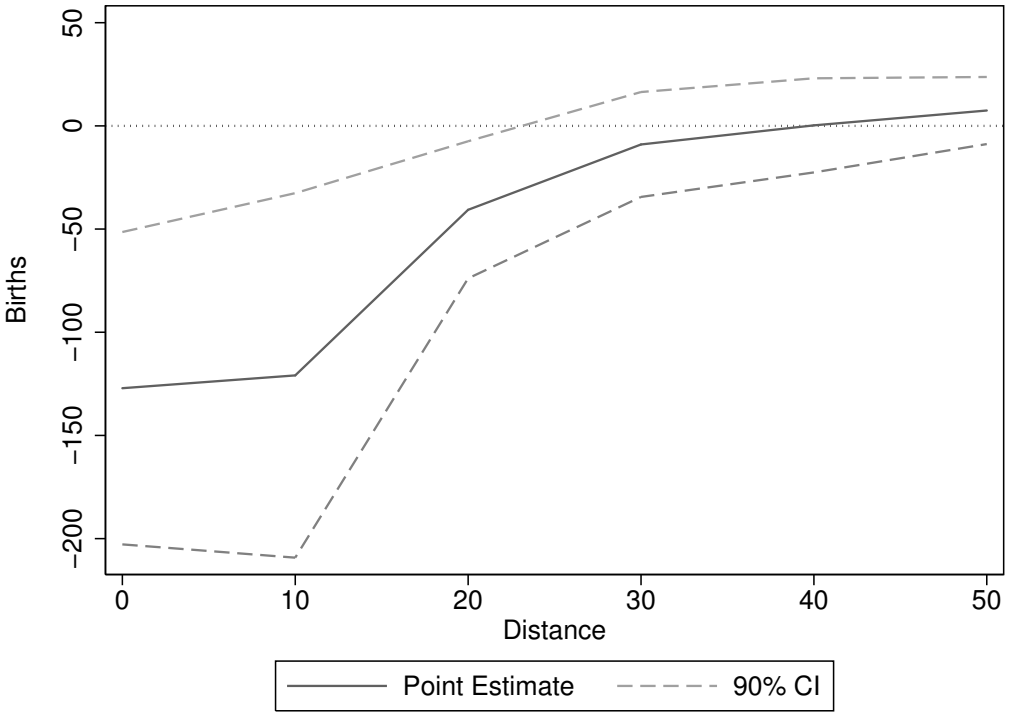
NOTES: Each column represents a separate difference-in-differences regression including full time and municipal fixed effects and linear trends by municipality. Standard errors are clustered at the level of the geographic region of treatment (municipality). Close variables are included in bins of 10km, so Close 1 refers to distances of [0,10)km, Close 2 refers to [10,20)km, and so forth. The dependent variable is a count of all births in the municipality, and is estimated by OLS. Further details regarding controls can be found in section 4.4.1.

Estimates from table 4.2 suggest that, firstly, the effect of the abortion policy is significant in magnitude. It reduces births among teenagers by 125 births per municipality per year. When comparing this to the average level of 1632.1 in treatment municipalities, this is a sizeable (and statistically significant) effect. When augmenting to control for local spillovers in columns 2 to 5, it appears that municipalities 'close to' treatment also are affected by the reform. For those municipalities within 10km of treatment municipalities (but not themselves treated), the effect is a highly statistically significant reduction of approximately 120 births. Column 3 extends to include a range of close controls. Here it becomes apparent that statistically significant effects remain at least up to areas between 10 and 20km from the nearest treatment, and negative point estimates only disappear when travelling greater than 30km away from treatment.

In examining estimates of living in treatment and close to treatment areas, it is worth noting that although spillover estimates are significantly different to zero over ranges of approximately 30 km, the correction for spillovers does not result in statistically significant changes in estimates of α: the effect of living in Mexico DF, after the introduction of the reform (though coefficients move as theoretically hypothesised). Considering the relative small number of municipalities

which are 'close' to Mexico DF (only 0.5% of total municipalities, containing 1.5% of total births in Mexico lie within a 30km radius of Mexico DF), this is not remarkably surprising, as the attenuation bias caused by these municipalitis is small. This is a point I return to discuss more extensively when examining the case of Chile, where treatment is geographically disperse, and a much larger proportion of the population lies 'close' to treatment areas.

Figure 4.1: Treatment and close to treatment effects: 15-19 year olds Mexico



NOTES TO FIGURE: Each point represents a treatment effect for the group living $d \in [0, 50]$ km from the nearest treatment municipality. As such, the point at 0 includes all municipalities to directly receive treatment (Mexico DF). Standard errors are clustered at the level of the municipality. Dotted lines display the 90% confidence interval for all estimates.

Figure 4.3 presents a graphical representation of estimates of a vector of β coefficients from equation (4.7).[14] While the largest effect of treatment is felt in the treatment municipality itself, effects clearly remain even outside of treatment municipalities, suggesting that the spillover robust specification is necessary to estimate causal effects $\hat{\alpha}$ and $\hat{\beta}$.

[14]While here we focus on teenaged girls, appendix 4.D presents similar graphical results for other age groups.

4.4.2 Emergency Contraceptive Reform in Chile

After considerable juridical challenges against the legality of emergency (post-coital) contraception in the country, a Chilean constitutional tribunal in 2008 issued a summary expressly allowing the morning after pill[15] to be prescribed to women. However, this finding was limited to municipal health centres, which are administered by mayors and local governing councils. This resulted in a period of approximately 4 years where the morning after pill was available to women *only* if the mayor of her municipality deemed it appropriate. The reform eventually resulted in morning after pill availability in approximately 150 of the 346 municipalities of the country. Further figures and details of the reform are discussed in chapter 3. A description of the constitutional details of the reform are provided in appendix 3.A, and summary statistics for areas with and without the emergency contraceptive pill are provided in table 3.1.

As for the case of the Mexico abortion reform, a 'traditional' DD specification is estimated, and compared with a spillover-robust DD estimator as proposed in section 4.3. A generalised version of (4.10) is estimated (where full year and municipal fixed effects are added, and municipal linear trends and time-varying controls are included), and compared to an identical version of the equation robust to spillovers between treatment and close-to-treatment areas (4.7). If the traditional DD approach adequately captures the treatment—or in other words, if the SUTVA holds globally—then we should see two things. Firstly, our estimate of α from (4.10) should not be significantly different to that from (4.7). Secondly, the coefficient on each $R^k(i,t)$ should not be significantly different to zero. Formally, if we cannot reject the null that $\beta = 0$, this is evidence against the need for spillover robust DD in this case.

Table 4.3 presents estimates from the Chile reform. In this case the variable $Y(i,t)$ represents the probability of giving birth at time t, a binary outcome taking either 0 or 1 for each individual aged 15-19 years. Column 1 presents an estimate where treatment is defined as having the morning after pill available in the municipality where a woman lives one year prior to the realised birth outcome (birth versus no birth). The lag of one year accounts for the mechanical delay in realisations of $Y(i,t)$ due to child gestation. This alone suggests important effects of the reform: having the reform available in the municipality of residence of the woman is associated with a 4.5% reduction[16] in births the following year. However, in columns (2) to (5), we see that naïve estimates which fail to account for (local) spillovers *understate* the true importance of the

[15]The morning after pill is a hormonal treatment composed of progestin and estrogen which acts to prevent ovulation after sexual intercourse in which alternative forms of contraceptives were not used, or believed to have failed.

[16]Each binary model is estimated by logistic regression and odds ratios are reported. Hence, the percentage reduction in the outcome of interest for a coefficient of -0.046 is calculated as $1 - \exp(-0.046) = 0.045$, or 4.5%. In the remainder of this section, coefficients will always be converted to percentage reductions of the outcome variable when discussed.

Table 4.3: Treatment Effects and Spillovers: Chile (15-19 year olds)

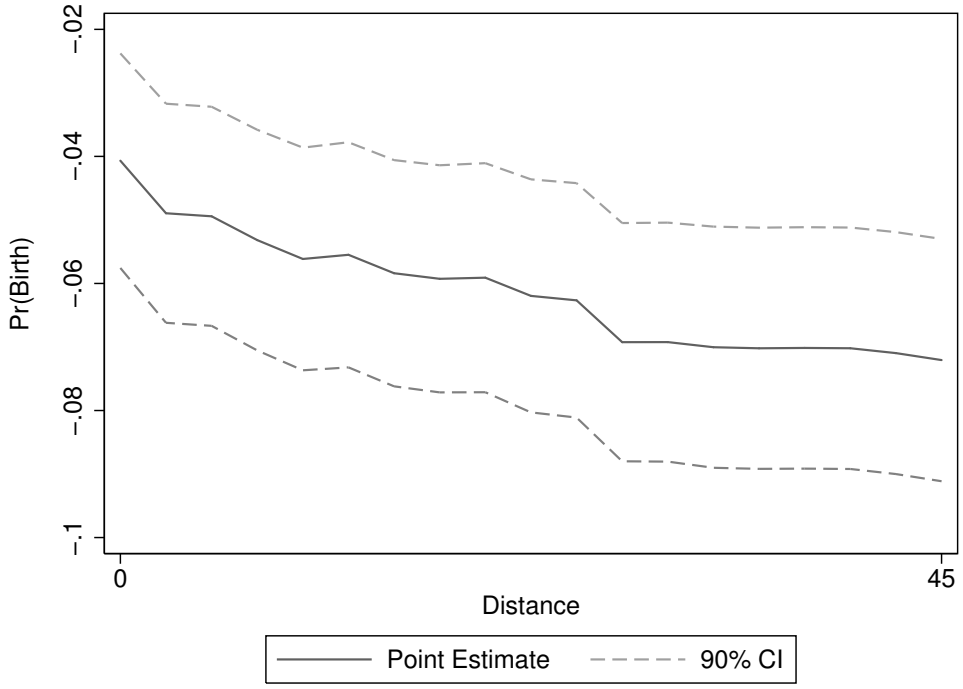
	Pr(Birth) (1)	Pr(Birth) (2)	Pr(Birth) (3)	Pr(Birth) (4)	Pr(Birth) (5)
Treatment	−0.046***	−0.058***	−0.066***	−0.073***	−0.074***
	(0.011)	(0.013)	(0.014)	(0.014)	(0.015)
Close 1		−0.049***	−0.056***	−0.062***	−0.062***
		(0.015)	(0.014)	(0.014)	(0.014)
Close 2			−0.040*	−0.047*	−0.048**
			(0.023)	(0.024)	(0.024)
Close 3				−0.038*	−0.038*
				(0.023)	(0.023)
Close 4					−0.014
					(0.023)
Mean	0.052	0.052	0.052	0.052	0.052
Regions× Time	1,929	1,929	1,929	1,929	1,929

NOTES: Each column represents a separate difference-in-differences regression including full time and municipal fixed effects and linear trends by municipality. Standard errors are clustered at the level of the geographic region of treatment (municipality). Close variables are included in bins of 10km, so Close 1 refers to distances of [0,10)km, Close 2 refers to [10,20)km, and so forth. Models are estimated using a binary (logit) model for birth versus no birth. Coefficients are expressed as log odds.

reform. Column 2 suggests that for teenages living very close to the reform area, the reform appears to be nearly as important (a 4.8% versus a 5.6% reduction in pregnancy rates), even though their municipality is not directly treated. Columns 3-5 progressively include additional 'close' binary variables, up to a distance of 40km. These tests suggest that the effect of the reform is able to travel around 30km, after which point marginal areas are not significantly affected by the reform, and the estimate of the treatment effect in other areas is not affected by additional distance controls. The spillover distance of this reform is reasonably similar to the effects of the Mexico abortion reform discussed in the previous section. Similar tests are run with women aged over 20, as well as using alternative distance measures (distance by road and travel time by road) in appendices 4.C and 4.E.

These results clearly suggest that we *can* reject the null that $\beta = 0$, as a number of 'close' coefficients are significant, in some cases even up to $p = 0.01$. However, tests directly on α do not allow for us to reject that values estimated for various models are significantly different. Examining estimates $\hat{\alpha}$ more carefully suggests that as we move further away from the reform, the effect size monotonically decreases (figure 4.2). This is precisely in-line with what we would expect if SUTVA were violated locally, and the cost (both psychic and economic) of travelling to treatment municipalities increased with distance. This figure suggests that traditional DD

Figure 4.2: Treatment Effects: 15-19 year olds Chile



NOTES TO FIGURE: Each point represents the estimated treatment effect on the treated ($\hat{\alpha}$), conditioning on close controls for $d \in [0, 45]$ km from the nearest treatment municipality. As such, the point at 0 includes all municipalities with the exception of treatment municipalities in the control group. The point at 2.5 controls for spillovers up to 2.5 km (removing these areas from the control group), and so forth at other distances. Standard errors are clustered at the level of the municipality. Dotted lines display the 90% confidence interval for all estimates.

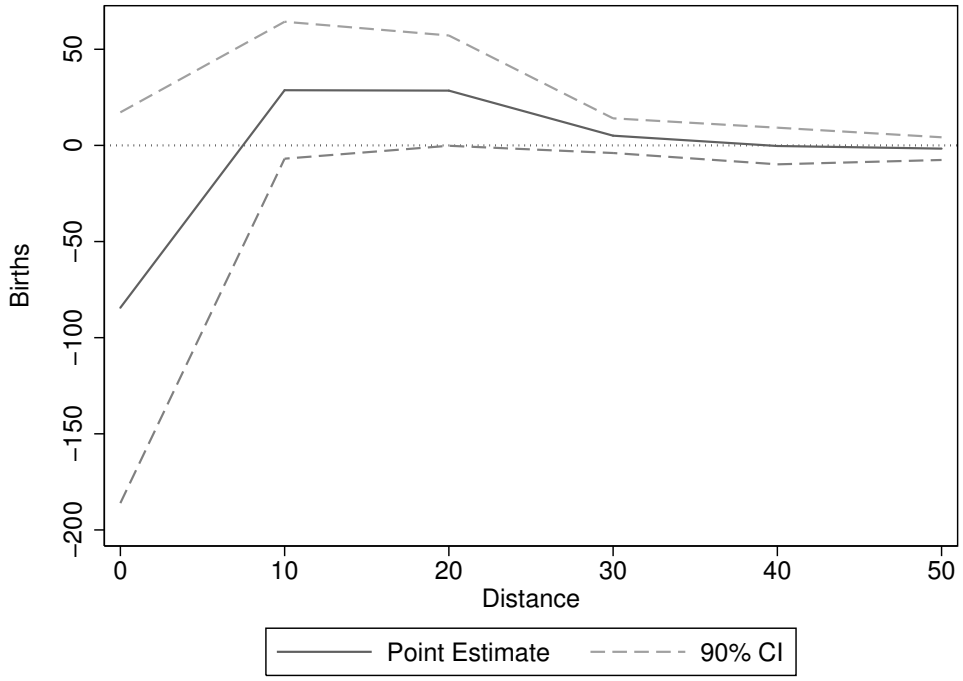
estimates are attenuated when the presence of spillovers are not accounted for, and that the bias in estimates of α are corrected once controlling adequately for spillover distance. The reported estimates of $\hat{\alpha}^k$ in figure 4.2 demonstrate the result derived in section 4.3.1 that if spillovers occur, if they are monotonic in distance, and if they are of the same direction of the treatment itself, then: $|E[\hat{\alpha}^0]| < |E[\hat{\alpha}^1]| < \cdots < |E[\hat{\alpha}^{d/h}]| = |\alpha|$, where d is the maxiumum distance at which spillovers occur, and h is the bandwidth measure, which in the above is 2.5 km.

4.4.3 Running Additional Placebo Tests

Typically, DD estimates are presented along with placebo tests which define 'false' lagged reforms. In other words, by examining outcomes entirely *before* the policy of interest has been implemented, null results are presented as evidence in favour of an appropriately specified functional form of the DD set-up.

In the case of the spillover-robust DD estimate, there are now (at least) two relevant placebos which should be tested. Firstly, the reform must not have any effect on outcomes *before* treatment in treatment municipalities. This is precisely the same as the 'traditional' placebo test described above. Secondly however, the reform should have no effect on predetermined outcomes in municipalites *close* to treatment municipalities. Below we present an example of such placebo tests from the Mexico City abortion reform. Now, as well as having a treatment estimate not significantly different from 0 (ie confidence intervals at $distance = 0$), the same result should hold for close municipalities ($distance > 0$).

Figure 4.3: Treatment and Close, Placebo Tests: 15-19 year olds Mexico



NOTES TO FIGURE: Each point represents a placebo treatment effect for the group living $d \in [0, 50]$ km from the nearest treatment municipality three years *prior* to the reform. All births were realised entirely before the reform began. Standard errors are clustered at the level of the municipality. Dotted lines display the 90% confidence interval for all estimates.

A more demanding series of placebo tests involves the estimation of a full event study based on the DD specification. In this case, instead of estimating a single treatment effect for all periods following the arrival of the natural experiment in question, a binary variable for living in a treatment area is interacted with a series of lags and leads around the date of the reform. This allows for a direct test of the timing of effect. In a Granger (1969) causality framework, any difference between treatment and control states should only emerge following the introduction of the reform: not prior to the date of the reform.

In traditional DD this leads to estimations of event study where coefficients and standard errors are plotted which compare treated to control areas. Insignificant differences prior to the reform and significant differences posterior to the reform are evidence in favour of the parallel trend assumption, and that the reform causes the effect, rather than the other way around[17]. In the case of spillover robust DD estimates, there are now two logical tests to employ. These (seperately) test both parallel trend assumptions (assumption 1 and 2). Both treated *and* close to treated areas can be compared with control areas in an event study framework.

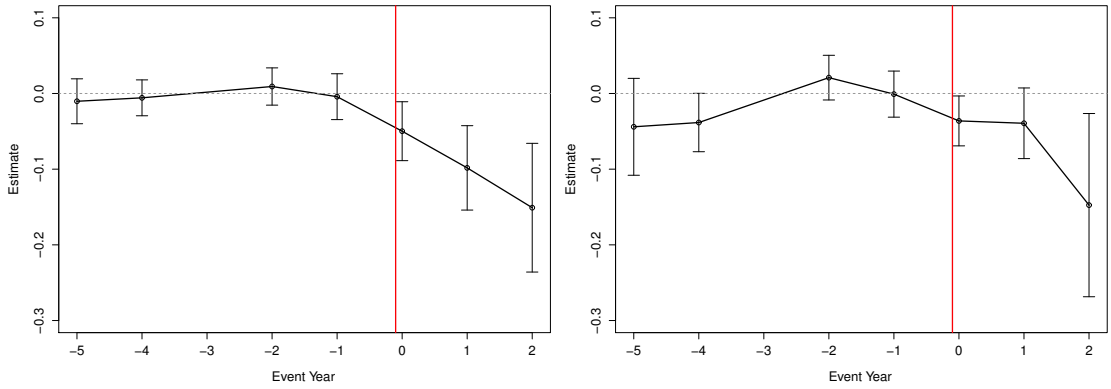
Figures 4.4a and 4.4b present these event studies for the case of Chile. The omitted base year is 3 birth cohorts prior to the reform (a group which gave birth entirely before the year in which the emergency contraceptive pill arrived to Chile). Comparing treatment to control municipalities (4.4a), the parallel trend assumption appears to be valid, with all estimates being close to zero and tightly estimated. Only in years following the reform does the effect diverge from zero, with coefficients at least providing some evidence that the effect of the reform has grown as knowledge of the morning after pill has become more widespread. For close-to-treatment areas the event study is slightly noisier, however once again the divergence between these areas and control areas occurs only *after* the vertical line signalling the first cohort affected by the reform. In this case there is more variation in the magnitude of pre-reform coefficients, though at a 95% confidence level, equality with zero cannot be rejected in any case (evidence broadly in favour of assumption 2).

4.5 Conclusion

Echoing Bertrand et al. (2004), "Differences-in-Differences (DD) estimation has become an increasingly popular way to estimate causal relationships". It is important to consider the assumptions underlying these estimators. In this paper we examine how DD estimates perform when the stable unit treatment value assumption does not hold locally. Such a situation may be common in estimates of the causal effect of policy where compliance is imperfect. If policies entail a benefit to recipients, and if recipients living 'close to' treatment areas who are themselves untreated can somehow cross regional boundaries to receive treatment, we may be concerned that, locally at least, SUTVA is violated.

[17]This is also a test for the presence of phenomena similar to Ashenfelter's dip (Ashenfelter, 1978; Heckman and Smith, 1999), where a reform or program may be the result of a poor outcome prior to the program. Ashenfelter's dip refers to the fact that earnings are often seen to fall prior to entry into labour market training programs, though similar phenomena may ocurr where public policy responds to particularly concerning social indicators such as high rates of teenage pregnancy.

Figure 4.4: Chile Event Study for ATT and ATC



(a) Full Event Study: Treatment

(b) Full Event Study: Close to Treatment

NOTES TO FIGURE: In each figures the horizontal dotted line represents an effect size of 0. The vertical solid line represents the first birth cohort affected by the reform. Each point on the plot represents the effect of living in a treatment (panel A) or close to treatment (panel B) municipality x years before or after the reform took effect. Error bars represent 95% confidence intervals of these estimates.

In this paper I derive a set of conditions by which DD estimates can produce unbiased estimates even in the absence of the SUTVA holding between all units. It is shown that under a weaker set of conditions, both the average effect on the treated and the average effect on the 'close to treated' can be estimated in a DD-type framework. It is suggested that in the absence of this correction for local violations of SUTVA that (if spillovers actually *do* occur) the true effect of the policy is likely to be attenuated.

Using two empirical examples from recent contraceptive policy expansions, it is shown that this is—at least in these cases—an important consideration for the estimation of treatment effects, and effects on nearby neighbourhoods. For both Chile and Mexico, it is shown that pregnancy rates in neighbourhoods located close to areas where contraceptive reforms took place had subsequent reductions in rates of teenage pregnancy. What's more, in Chile (but not in Mexico), the correction for spillovers results in a significant reduction in estimated treatment effects on the treated, correcting an attenuation bias when control units are partially treated. This is a useful reflection on this methodology: where treatment is geographically disperse, and hence many people live close to treatment areas (as in Chile), correcting for failures of the SUTVA is likely to be particularly important. In cases where treatment is only available in a reduced geographic area (such as Mexico), the degree of importance of spillovers are likely to be considerably less when considering estimates of average effects on treated areas.

These tests are easy to run, and indeed a software package that automates this methodology is released with this paper. Given the nature of the assumptions underlying identification in many DD models in the literature, tests of this nature should be included in a basic suite of falsification tests. While the examples in this paper are illustrated using geographic spillovers, spillover-robust DD estimation is certainly not limited to only geographic cases. How (and whether) treatment travels between units should be of fundamental concern to many applications in the economic literature.

4.A Proofs

Proof of Proposition 1. $Y(i,t)$ is generated according to (4.1), and from (4.7), a regression of $Y(i,t)$ on $D(i,t)$ and $C(i,t)$ can be estimated. It is assumed that we have at a representative sample of size N consisting of $\{Y(i,t), D(i,t), X(i,t) : i = 1, \ldots, N, t = 0, 1\}$. By assumption 4A, the assignment rule δ forms $C(i,t)$ allowing for the estimation of (4.7). By definition, α in this regression is equal to:

$$\begin{aligned}\alpha = &\{E[Y(i,1)|D(i,1) = 1, R(i,1) = 0] - E[Y(i,1)|D(i,1) = 0, R(i,1) = 0]\} \\ &- \{E[Y(i,0)|D(i,1) = 1, R(i,1) = 0] - E[Y(i,0)|D(i,1) = 0, R(i,1) = 0]\},\end{aligned}$$

and from assumption 3, each of the expectation terms exists, as there are both fully treated and completely untreated units. Using the potential outcomes framework, we are free to re-write the above expression as:

$$\begin{aligned}\alpha = &\{E[Y^1(i,1)|D(i,1) = 1, R(i,1) = 0] - E[Y^0(i,1)|D(i,1) = 0, R(i,1) = 0]\} \\ &- \{E[Y^0(i,0)|D(i,1) = 1, R(i,1) = 0] - E[Y^0(i,0)|D(i,1) = 0, R(i,1) = 0]\},\end{aligned}$$

given that only in the case where $t = 1$ and $D(i,1) = 1$ we observe the potential outcome where the individual receives treatment: $Y^1(i)$. Using the linearity of the expectations operator, this can finally be re-written as:

$$\alpha = E[Y^1(i,1) - Y^0(i,0)|D(i,1) = 1, R(i,1) = 0] - E[Y^0(i,1) - Y^0(i,0)|D(i,1) = 0, R(i,1) = 0].$$

Now, from assumption 1, we can appeal to parallel trends, and replace the second expectation term in the above expression with $E[Y^0(i,1) - Y^0(i,0)|D(i,1) = 1, R(i,1) = 0]$:

$$\alpha = E[Y^1(i,1) - Y^0(i,0)|D(i,1) = 1, R(i,1) = 0] - E[Y^0(i,1) - Y^0(i,0)|D(i,1) = 1, R(i,1) = 0].$$

Expanding the expectations operator and cancelling out the second term in each of the above items gives:

$$\alpha = E[Y^1(i,1)|D(i,1) = 1, R(i,1) = 0] - E[Y^0(i,1)|D(i,1) = 1, R(i,1) = 0].$$

which finally, once again by the linearity of expectations, can be combined to give $\alpha = E[Y^1(i,1) - Y^0(i,1)|D(i,1) = 1, R(i,1) = 0]$, which can be rewritten as $\alpha = E[Y^1(i,1) - Y^0(i,1)|D(i,1) = 1]$ given that $D(i,1) = 1 \implies R(i,1) = 0$. Combining (4.8) and $\alpha = E[Y^1(i,1) - Y^0(i,1)|D(i,1) = 1]$ we thus have that $\alpha = ATT$ as required.

Turning to the ATC, the same set of steps can be followed for β on the coefficient $R(i,t)$, however now instead of assumption 1 we must rely on parallel-trend assumption 2. This leads to $\beta = E[Y^1(i,1) - Y^0(i,1)|R(i,1) \neq 1]$, and from (4.9) and the previous expression it holds that that $\beta = ATC$. ■

Proof of Proposition 2. With the representative sample $\{Y(i,t), D(i,t), \mathbf{X}(i,t) : i = 1, \ldots, N, t = 0, 1\}$, assumption 4B implies that $\mathbf{X}(i,t)$ can be $C(i,t)$ using assignment rule δ. The remainder of the proof follows the same steps as the proof for proposition 1. ■

4.B Additional Details: 2007 Mexico Abortion Reform

On April 26, 2007 the legislative assembly of the Federal District of Mexico City (Mexico DF), voted to legalise abortion (termed legal interruption of pregnancy) whenever requested by the woman up to 12 weeks of gestation, reforming article 144 of the penal code of Mexico DF. This immediately permitted women from DF to request (free) legal interruption of pregnancy in public health clinics, with a large influx of requests (Contreras et al., 2011). On August 29, 2008 this decision was ratified by the Supreme Court of Mexico.

As well as decriminalising abortion, the law dictated that Mexico DF Department of Health facilities offer free abortion to residents of DF, and on a variable pay scale for women from other areas of the country (Becker and Díaz-Olavarrieta, 2013). Prior to the April 2006 findings, abortion was illegal in Mexico DF (and all of Mexico) in all but a very limited set of circumsances (depending on the state, these circumstances include none, some, or all of rape, fetal inviability or grave danger to the health of the mother). Along with free pregnancy terminations at Ministry of Health clinics, following the reform private health centres were also allowed to provide abortions.

Abortion services were widely accessed following the reform. Between April of 2007 and the end of 2011, 80,000 abortions were performed. These were accessed by women over the entire age range of the fertility distribution, reasonably closely mirroring mother's age at birth in birth data (figure 4.5), though with a slightly higher rate for younger women. Prior to April 2007 very few legal abortions were performed (n line with the restrictions listed above). Between 2001 and 2007 only 62 legal abortions were performed, though clandestine abortion was very common (Becker and Díaz-Olavarrieta, 2013). Further details regarding the reform, demand and subsequent state decisions can be found in (Becker and Díaz-Olavarrieta, 2013), and references therein.

4.C Measuring Distance to Treatment Clusters

Principal measures of distance from treatment is calculated by taking a Euclidean distance from the centroid of non-treatment clusters, to the centroid of the nearest cluster which did receive treatment. However, alternative measures may more accurately capture the true distance of an individual to treatment. As a robustness check, two alternative measures of distance to treatment are calculated and used.

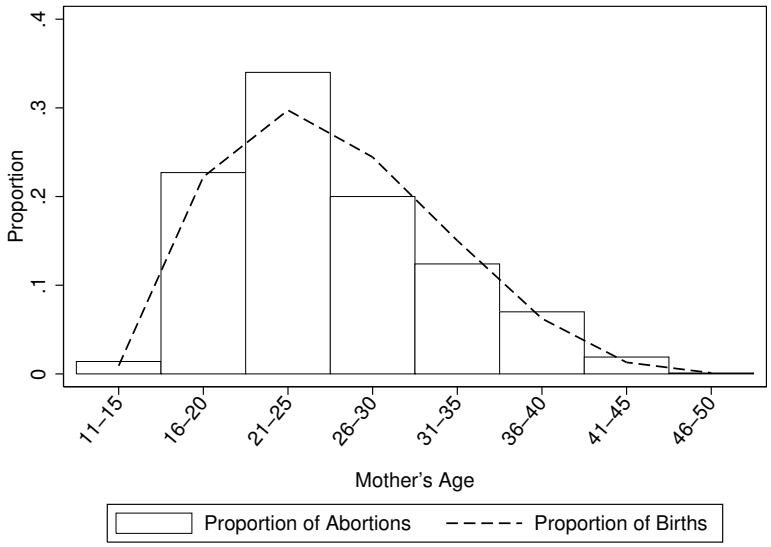
Firstly, I collated the shortest distance over roads from non-treatment to treatment areas. This was calculated using repeated calls to the Google Distance Matrix API[18], which finds the shortest path over roads. In the case of Chile, this requires calculating the distances between all 346 municipalities ($346^2/2 = 59,858$ distance pairs), while in the case of México this requires calculating only the distance from each municipality outside of Mexico DF to each municipality inside Mexico DF ($2457 \times 16 = 39,312$). Secondly, rather than distance in kilometres, as in Euclidean or road distance, a measure of travel *time* was calculated. As a proxy for total travel time, travel time by car was caclulated between areas. This was similarly generated using calls to Google Maps, resulting in one value for each municipality pair. In each case "distance to treatment" is then the minimum value to the nearest treatment area, which varies by municipality and year.

These alternative measures of distance do not majorly affect the quantitative implication of findings in either Chile or Mexico. Appendix figure 4.6 is the analogue of figure 4.2, using travel time rather than Euclidean distance between municipalities as a measure of spillover distance. Results from both figures suggest a treatment effect of approximately -0.075 once accounting for spillovers of 30 minutes travel time or 30km of distance respectively. Regression results for all age groups and all measures did not result in significantly different estimates of the effect of treatment in any case.

[18]Full details can be found at: https://developers.google.com/maps/documentation/distancematrix/#api_key. I have made the computational routine used available on the web at: https://github.com/damiancclarke/spillovers/blob/master/source/distCalc/queryDist.py.

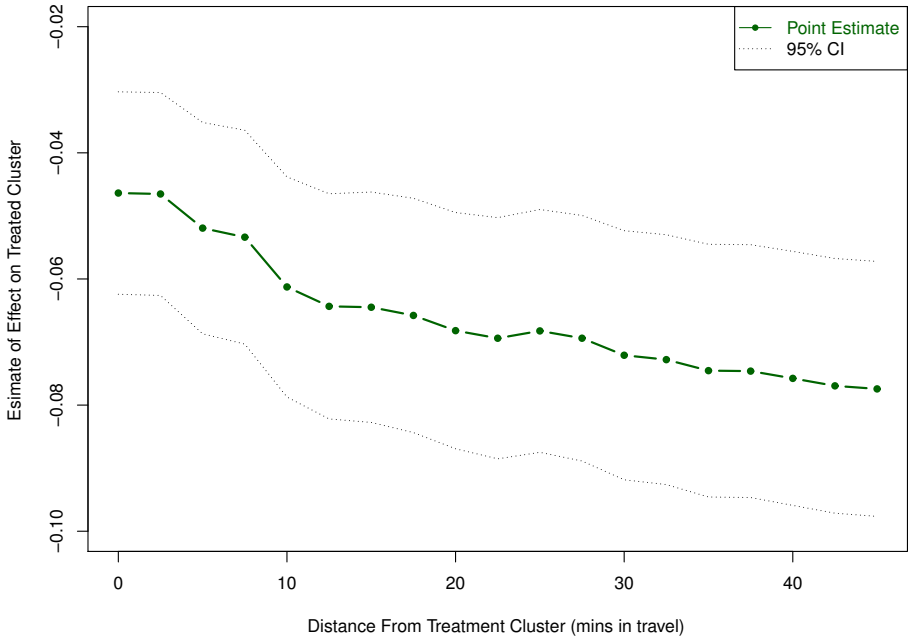
4.D Appendix Figures

Figure 4.5: Birth and Abortion Descriptives: Mexico



NOTES TO FIGURE: Total births are plotted between 2001 and 2010. Abortions are plotted from the date of reform (April 26, 2007) until 2011. The total quantity of births is 22.20 million (all of Mexico), and total abortions are 69,861 (Mexico DF only). Births are calculated from administrative data (INEGI) and abortions from administrative data (Secretary of Health, Mexico DF).

Figure 4.6: Estimate of Average Treatment Effect when Controlling for Travel Time



4.E Appendix Tables

Table 4.4: Treatment Effects and Spillovers: Mexico (20-34 year olds)

	N Birth (1)	N Birth (2)	N Birth (3)	N Birth (4)	N Birth (5)
Treatment	-1,157**	-1,167**	-1,176**	-1,178**	-1,179**
	(468.4)	(469.2)	(469.2)	(469.3)	(469.3)
Close 1		-1,623**	-1,631**	-1,633**	-1,634**
		(655.4)	(656.6)	(656.7)	(656.8)
Close 2			-460.7**	-462.2**	-463.2**
			(192.0)	(192.0)	(192.0)
Close 3				-87.40	-88.38
				(129.3)	(129.3)
Close 4					-94.58
					(156.1)
Mean	10,394	10,394	10,394	10,394	10,394
Regions×Time	24,550	24,550	24,550	24,550	24,550

NOTES: Refer to table 4.2.

Table 4.5: Treatment Effects and Spillovers: Mexico (35-49 year olds)

	N Birth (1)	N Birth (2)	N Birth (3)	N Birth (4)
Treatment	-207.8**	-208.9**	-209.7***	-209.8***
	(80.96)	(81.05)	(81.08)	(81.09)
Close 1		-175.2**	-176.0**	-176.1**
		(71.70)	(71.79)	(71.80)
Close 2			-37.39**	-37.49**
			(17.25)	(17.25)
Close 3				-5.058
Mean	1,415	1,415	1,415	1,415
Regions×Time	24,550	24,550	24,550	24,550

NOTES: Refer to table 4.2.

Table 4.6: Treatment Effects and Spillovers: Chile (20-34 year olds)

	Pr(Birth) (1)	Pr(Birth) (2)	Pr(Birth) (3)	Pr(Birth) (4)
Treatment	-0.031^{***}	-0.040^{***}	-0.042^{***}	-0.043^{***}
	(0.007)	(0.008)	(0.009)	(0.010)
Close 1		-0.034^{***}	-0.035^{***}	-0.036^{***}
		(0.012)	(0.012)	(0.012)
Close 2			-0.007	-0.008
			(0.017)	(0.017)
Close 3				-0.006
				(0.012)
Mean	0.085	0.085	0.085	0.085
Regions× Time	1,929	1,929	1,929	1,929

NOTES: Refer to notes in table 4.3.

Table 4.7: Treatment Effects and Spillovers: Chile (35-49 year olds)

	Pr(Birth) (1)	Pr(Birth) (2)
Treatment	-0.010	-0.011
	(0.010)	(0.011)
Close 1		-0.007
		(0.013)
Mean	0.021	0.021
Regions× Time	1,929	1,929

NOTES: Refer to notes in table 4.3.

Bibliography

A. Abadie. Semiparametric Difference-in-Differences Estimators. *Review of Economic Studies*, 72(1):1–19, 2005.

J. M. Agüero and M. S. Marks. Motherhood and Female Labor Force Participation: Evidence from Infertility Shocks. *The American Economic Review*, 98(2):500–504, 2008.

J. M. Agüero and M. S. Marks. Motherhood and Female Labor Supply in the Developing World: Evidence from Infertility Shocks. *Journal of Human Resources*, 46(4):800–826, 2011.

A. Aizer and F. Cunha. The Production of Human Capital: Endowments, Investments and Fertility. NBER Working Papers 18429, National Bureau of Economic Research, Inc, Sept. 2012.

H. Alderman, M. Lokshin, and S. Radyakin. Tall claims: Mortality selection and the height of children. Policy Research Working Paper 5846, The World Bank, Oct 2011.

D. Almond and B. Mazumder. Fetal Origins and Parental Responses. *Annual Review of Economics*, 5(1):37–56, 05 2013.

D. Almond, K. Y. Chay, and D. S. Lee. The Costs of Low Birth Weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, August 2005.

D. Almond, B. Mazumder, and R. van Ewijk. Fasting During Pregnancy and Children's Academic Performance. NBER Working Papers 17713, National Bureau of Economic Research, Inc, Dec. 2011.

J. G. Altonji, T. E. Elder, and C. R. Taber. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, 113(1):151–184, February 2005.

E. O. Ananat and D. M. Hungerman. The power of the pill for the next generation: Oral contraception's effects on fertility, abortion, and maternal and child characteristics. *The Review of Economics and Statistics*, 94(1):37–51, 2012.

E. O. Ananat, J. Gruber, and P. Levine. Abortion legalization and life-cycle fertility. *Journal of Human Resources*, 42(2), 2007.

E. O. Ananat, J. Gruber, P. B. Levine, and D. Staiger. Abortion and selection. *The Review of Economics and Statistics*, 91(1):124–136, February 2009.

M. Angelucci and G. D. Giorgi. Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption? *American Economic Review*, 99(1):486–508, March 2009.

M. Angelucci and V. D. Maro. Program Evaluation and Spillover Effects. SPD Working Papers 1003, Inter-American Development Bank, Office of Strategic Planning and Development Effectiveness (SPD), May 2010.

J. Angrist, V. Lavy, and A. Schlosser. Multiple experiments for the causal link between the quantity and quality of children. *Journal of Labor Economics*, 28(4):pp. 773–824, 2010.

J. D. Angrist and W. N. Evans. Schooling and labor market consequences of the 1970 state abortion reforms. NBER Working Papers 5406, National Bureau of Economic Research, Inc, Jan. 1996.

J. D. Angrist and W. N. Evans. Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, 88(3):450–77, June 1998.

L. M. Argys and S. L. Averett. The Effect of Family Size on Education: New Evidence from China's One Child Policy. IZA Discussion Papers 9196, Institute for the Study of Labor (IZA), Aug. 2015.

O. Ashenfelter. Estimating the Effect of Training Programs on Earnings. *The Review of Economics and Statistics*, 60(1):47–57, Feb 1978.

O. Ashenfelter and D. Card. Using the Longitudinal Structure of Earnings to Estimate the Effects of Training Programs. *Review of Economics and Statistics*, 67(4):648–660, November 1985.

M. J. Bailey. More Power to the Pill: The Impact of Contraceptive Freedom on Women's Life Cycle Labor Supply. *The Quarterly Journal of Economics*, 121(1):289–320, 2006.

M. J. Bailey. Erratum and Addendum. Posted at *The Quarterly Journal of Economics*, September 2009.

M. J. Bailey. "Momma's Got the Pill": How Anthony Comstock and Griswold v. Connecticut Shaped US Childbearing. *American Economic Review*, 100(1):98–129, March 2010.

M. J. Bailey. Reexamining the Impact of Family Planning Programs on US Fertility: Evidence from the War on Poverty and the Early Years of Title X. *American Economic Journal: Applied Economics*, 4(2):62–97, April 2012.

M. J. Bailey. Fifty Years of Family Planning: New Evidence on the Long-Run Effects of Increasing Access to Contraception. *Brookings Papers on Economic Activity*, 46(1 (Spring):341–409, 2013.

M. J. Bailey, B. Hershbein, and A. R. Miller. The opt-in revolution? contraception and the gender gap in wages. *American Economic Journal: Applied Economics*, 4(3):225–54, July 2012.

M. J. Bailey, M. Guldi, and B. J. Hershbein. Further evidence on the internal validity of the early legal access research design. *Journal of Policy Analysis and Management*, 32(4):899–904, 2013.

S. Baird, A. Bohren, C. McIntosh, and B. Ozler. Designing experiments to measure spillover effects. Policy Research Working Paper Series 6824, The World Bank, Mar. 2014.

D. J. P. Barker. Fetal origins of coronary heart disease. *BMJ*, 311(6998):171–74, 1995.

D. Becker and C. Díaz-Olavarrieta. Decriminalization of Abortion in Mexico City: The Effects on Women's Reproductive Rights. *American Journal of Public Health*, 103(4):590–593, 2013.

G. S. Becker. An Economic Analysis of Fertility. In *Demographic and Economic Change in Developed Countries*, NBER Chapters, pages 209–240. National Bureau of Economic Research, Inc, June 1960.

G. S. Becker and H. G. Lewis. On the interaction between the quantity and quality of children. *Journal of Political Economy*, 81(2):S279–88, Part II, 1973.

G. S. Becker and N. Tomes. Child endowments and the quantity and quality of children. *Journal of Political Economy*, 84(4):S143–62, August 1976.

G. S. Becker and N. Tomes. Human capital and the rise and fall of families. *Journal of Labor Economics*, 4(3):S1–S39, July 1986.

S. O. Becker, F. Cinnirella, and L. Woessmann. The Trade-off Between Fertility and Education: Evidence From Before The Demographic Transition. *Journal of Economic Growth*, 15(3): 177–204, 2010.

J. Behrman, P. Taubman, and R. Pollak. Parental Preferences and Provision for Progeny. *Journal of Political Economy*, 90(1):52–73, 1982.

J. R. Behrman and M. R. Rosenzweig. Returns to birthweight. *The Review of Economics and Statistics*, 86(2):586–601, May 2004.

Y. Ben-Porath. Fertility Response to Child Mortality: Micro Data from Israel. *Journal of Political Economy*, 84(4):S163–78, August 1976.

Y. Ben-Porath and F. Welch. Chance, child traits, and choice of family size. Report R-1117-NIH/RF, RAND Corporation, December 1972.

M. Bertrand, E. Duflo, and S. Mullainathan. How Much Should We Trust Differences-in-Differences Estimates? *Quarterly Journal of Economics*, 119(1):249–275, February 2004.

S. Bhalotra and D. Clarke. The Twin Instrument and the Estimation of the Quality-Quantity Trade-off. Mimeo, University of Essex and University of Oxford, 2015.

S. Bhalotra and S. Rawlings. Gradients of the Intergenerational Transmission of Health in Developing Countries. *The Review of Economics and Statistics*, 95(2):660–672, May 2013.

S. Bhalotra and A. Venkataramani. Shadows of the Captain of the Men of Death: Early Life Health Interventions, Human Capital Investments, and Institutions. Mimeo, University of Essex, 2014.

S. Black, P. J. Devereux, and K. Salvanes. Does grief transfer across generations? In-utero deaths and child outcomes. NBER Working Papers 19979, National Bureau of Economic Research, Inc, Mar. 2014.

S. E. Black, P. J. Devereux, and K. G. Salvanes. 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, 2005.

S. E. Black, P. J. Devereux, and K. G. Salvanes. Staying in the Classroom and out of the Maternity Ward? The Effect of Compulsory Schooling Laws on Teenage Births. *The Economic Journal*, 118(530):1025–1054, 2008.

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

H. Bleakley and F. Lange. Chronic Disease Burden and the Interaction of Education, Fertility, and Growth. *The Review of Economics and Statistics*, 91(1):52–65, February 2009.

M. Bosch and R. Campos-Vazquez. The Trade-Offs of Welfare Policies in Labor Markets with Informal Jobs: The Case of the "Seguro Popular" Program in Mexico. *American Economic Journal: Economic Policy*, 6(4):71–99, 2014.

M. Bougma, T. LeGrand, and J.-F. Kobiané. Fertility Decline and Child Schooling in Urban Settings of Burkina Faso. *Demography*, 52(1):281–313, 2015.

N. Braakmann and J. Wildman. Fertility treatments and the use of twin births as an instrument for fertility. MPRA Paper 54106, University Library of Munich, Germany, Mar. 2014.

S. G. Bronars and J. Grogger. The economic consequences of unwed motherhood: Using twin births as a natural experiment. *The American Economic Review*, 84(5):1141–1156, 1994.

K. F. Butcher and A. Case. The effect of sibling sex composition on women's education and earnings. *The Quarterly Journal of Economics*, 109(3):531–63, August 1994.

J. Cáceres-Delpiano. The impacts of family size on investment in child quality. *Journal of Human Resources*, 41(4):738–754, 2006.

J. Cáceres-Delpiano. Keeping the best for last. Impact of fertility on mother's employment. Evidence from developing countries. Economics Working Papers we086832, Universidad Carlos III, Departamento de Economía, Nov. 2008.

D. Card and A. B. Krueger. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review*, 84(4):772–93, September 1994.

L. Casas Becerra. La saga de la anticoncepción de emergencia en Chile: avances y desafíos. Serie documentos electrónicos no. 2, FLACSO-Chile Programa Género y Equidad, November 2008.

K. K. Charles and S. Melvin, Jr. Abortion Legalization and Adolescent Substance Use. *Journal of Law and Economics*, 49(2):481–505, 2006.

A. Chevalier and T. K. Viitanen. The Long-Run Labour Market Consequences of Teenage Motherhood in Britain. *Journal of Population Economics*, 16(2):323–343, 2003.

P.-A. Chiappori and S. Oreffice. Birth control and female empowerment: An equilibrium analysis. *Journal of Political Economy*, 116(1):113–40, 2008.

F. Christensen. The pill and partnerships: the impact of the birth control pill on cohabitation. *Journal of Population Economics*, 25(1):29–52, January 2012.

H. Chun and J. Oh. An instrumental variable estimate of the effect of fertility on the labour force participation of married women. *Applied Economics Letters*, 9(10):631–634, 2002.

CIA. *The World Fact Book*. Central Intelligence Agency, Washington D.C., 2014.

D. Conley and R. Glauber. Parental Educational Investment and Children's Academic Risk: Estimates of the Impact of Sibship Size and Birth Order from Exogenous Variation in Fertility. *The Journal of Human Resources*, 41(4):pp. 722–737, 2006.

T. G. Conley. GMM estimation with cross sectional dependence. *Journal of Econometrics*, 92 (1):1–45, September 1999.

T. G. Conley, C. B. Hansen, and P. E. Rossi. Plausibly Exogenous. *The Review of Economics and Statistics*, 94(1):260–272, February 2012.

X. Contreras, M. van Dijk, T. Sanchez, and P. Smith. Experiences and Opinions of Health-Care Professionals Regarding Legal Abortion in Mexico City: A Qualitative Study. *Studies in Family Planning*, 42(3):183–190, Sep 2011.

P. J. Cook, A. M. Parnell, M. J. Moore, and D. Pagnini. The effects of short-term variation in abortion funding on pregnancy outcomes. *Journal of Health Economics*, 18(2):241 – 257, 1999.

D. R. Cox. *Planning of Experiments*. John Wiley & Sons Inc., New York, New York, 1958.

J. P. Cristia. The Effect of a First Child on Female Labor Supply: Evidence from Women Seeking Fertility Services. *The Journal of Human Resources*, 43(3):487–510, 2008.

J. Currie, L. Nixon, and N. Cole. Restrictions on Medicaid Funding of abortion: effects on birthweight and pregnancy resolutions. *The Journal of Human Resources*, 31(1):159–187, Winter 1996.

G. B. Dahl and E. Moretti. The Demand for Sons. *The Review of Economic Studies*, 75(4): 1085–1120, 2008.

C.-J. Dalgaard and H. Strulik. Physiology and Development: Why the West is Taller than the Rest. *The Economic Journal*, pages n/a–n/a, 2015. doi: 10.1111/ecoj.12275.

H.-A. Dang and H. Rogers. The decision to invest in child quality over quantity: household size and household investment in education in Vietnam. Policy Research Working Paper Series 6487, The World Bank, June 2013.

J. E. Darroch. Trends in Contraceptive Use. *Contraception*, 87(3):259–263, March 2013.

M. Dayioğlu, M. G. Kirdar, and A. Tansel. 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, 06 2009.

D. N. De Tray. Child quality and the demand for children. *Journal of Political Economy*, 81(2): S70–95, March 1973.

R. Deheija, C. Pop-Eleches, and C. Samii. From Local to Global: External Validity in a Fertility Natural Experiment. Mimeo, Available for download at http://users.nber.org/∼rdehejia/papers/dehejia_pop-eleches_samii.pdf, 2015.

DEIS Ministerio de Educación, Gobierno de Chile. Sistema nacional de servicios de salud. http://intradeis.minsal.cl/Mapas/, December 2013. Base de Establecimientos.

L. M. Demers. The morning after pill. *The New England Journal of medicine*, 284(18):1034–6, 1971.

S. Desai. When Are Children from Large Families Disadvantaged? Evidence from Cross-National Analyses. *Population Studies*, 49(2):pp. 195–210, 1995.

A. DiCenso, G. Guyatt, A. Willan, and L. Griffith. Interventions to reduce unintended pregnancies among adolescents: systematic review of randomised controlled trials. *British Medical Journal*, 324(7351):1426–34, 2002.

C. Dides, M. C. Benavente, and J. Manuel Moran. Entrega de la píldora anticonceptiva de emergencia en el sistema de salud municipal, chile: Estado de situación. Documento de trabajo, FLACSO-Chile. Programa Género y Equidad., March 2009.

C. Dides C., C. Benavente R., and I. Sáez A. Entrega de la píldora anticonceptiva de emergencia en el sistema de salud municipal de chile. Estado de situación, FLACSO-Chile / CLAE, Consorcio Latinoamericano de Concepción de Emergencia., 2010.

C. Dides C., C. Benavente R., I. Sáez A., L. Nicholls S., and T. Correa D. Provisión de anticoncepción de emergencia en el sistema de salud municipal de chile: Estado de situación. Serie de documents, no. 2, Universidad Central de Chile, 2011.

C. Dides Castillo. *Voces en emergencia: el discurso conservador y la píldora del día despúes.* FLACSO, Chile, 2006.

E. Duflo. Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *American Economic Review*, 91(4):795–813, September 2001.

C. P. Durrance. The effects of increased access to emergency contraception on sexually transmitted disease and abortion rates. *Economic Inquiry*, 51(3):1282–1695, 2013.

C. Ellerson, K. Blanchard, A. Webb, A. Bigrigg, and S. Haskell. Emergency contraception. *The Lancet*, 352(9138):1477, 1998.

S. Enke. The economic aspects of slowing population growth. *The Economic Journal*, 76(301): 44–56, 1966.

S. Enke. Economic Consequences of Rapid Population Growth. *Economic Journal*, 81(324): 800–11, 1971.

E. Fitzsimons and B. Malde. Empirically probing the quantity-quality model. IFS Working Papers W10/20, Institute for Fiscal Studies, Sep 2010.

E. Fitzsimons and B. Malde. Empirically Probing the Quantity-Quality Model. *Journal of Population Economics*, 27(1):33–68, Jan 2014.

B. M. Fleisher and J. Rhodes, George F. Fertility, Women's Wage Rates, and Labor Supply. *The American Economic Review*, 69(1):14–24, 1979.

J. Fletcher. The effects of teenage childbearing on the short- and long-term health behaviors of mothers. *Journal of Population Economics*, 25(1):201–218, October 2012.

J. M. Fletcher and B. L. Wolfe. Education and labor market consequences of teenage childbearing: Evidence using the timing of pregnancy outcomes and community fixed effects. *Journal of Human Resources*, 44(2), 2009.

B. Fraser. Tide begins to turn on abortion access in South America. *The Lancet*, 383(9935): 2113–2114, June 2014.

A. T. Geronimus and S. Korenman. The Socioeconomic Consequences of Teen Childbearing Reconsidered. *The Quarterly Journal of Economics*, 107(4):1187–1214, 1992.

C. Goldin. The quiet revolution that transformed women's employment, education, and family. Working paper no, 11953, NBER, 2006.

C. Goldin and L. F. Katz. The power of the pill: Oral contraceptives and women's career and marriage decisions. *Journal of Political Economy*, 110(4):730–770, August 2002a.

C. Goldin and L. F. Katz. More Power to the Pill: The Impact of Contraceptive Freedom on Women's Life Cycle Labor Supply. *Journal of Political Economy*, 110(4):730–770, 2002b.

C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica*, 37(3):424–38, July 1969.

R. Gronau. The Effect of Children on the Housewife's Value of Time. *Journal of Political Economy*, 81(2):S168–S199, 1973.

T. Gross, J. Lafortune, and C. Low. What happens the morning after? the costs and benefits of expanding access to emergency contraception. Documentos de Trabajo 425, Instituto de Economia. Pontificia Universidad Católica de Chile., 2013.

T. Gross, J. Lafortune, and C. Low. What Happens the Morning After? The Costs and Benefits of Expanding Access to Emergency Contraception. *Journal of Policy Analysis and Management*, 33(1):70–93, 2014.

J. Gruber, P. Levine, and D. Staiger. Abortion Legalization and Child Living Circumstances: Who is the "Marginal Child"? *The Quarterly Journal of Economics*, 114(1):263–291, 1999.

M. Guldi. Fertility Effects of Abortion and Birth Control Pill Access for Minors. *Demography*, 45(4):17–827, 2008.

T. Haavelmo. The Statistical Implications of a System of Simultaneous Equations. *Econometrica*, 11(1):1–12, January 1943.

T. Haavelmo. The Probability Approach in Econometrics. *Econometrica*, 12(Supplement):iii–vi+1–115, January 1944.

J. G. Hall. Twinning. *The Lancet*, 362(6736):735–43, August 2003.

E. A. Hanushek. The trade-off between child quantity and quality. *Journal of Political Economy*, 100(1):84–117, February 1992.

J. Heckman and R. Pinto. Causal analysis after haavelmo. *Econometric Theory*, 31:115–151, 2 2015.

J. Heckman and J. A. Smith. The Pre-Programme Earnings Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies. *The Economic Journal*, 109(2):313–348, July 1999.

J. Heckman, H. Ichimura, J. Smith, and P. Todd. Characterizing Selection Bias Using Experimental Data. *Econometrica*, 66(5):1017–1098, September 1998a.

J. J. Heckman and E. Vytlacil. Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, 73(3):669–738, 05 2005.

J. J. Heckman, L. Lochner, and C. Taber. General-Equilibrium Treatment Effects: A Study of Tuition Policy. *American Economic Review*, 88(2):381–86, May 1998b.

H. Holmlund. Estimating Long-Term Consequences of Teenage Childbearing: An Examination of the Siblings Approach. *The Journal of Human Resources*, 40(3):716–743, 2005.

V. J. Hotz, C. H. Mullin, and S. G. Sanders. Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analysing the Effects of Teenage Childbearing. *Review of Economic Studies*, 64(4):575–603, October 1997.

V. J. Hotz, S. W. McElroy, and S. G. Sanders. Teenage Childbearing and Its Life Cycle Consequences: Exploiting a Natural Experiment. *The Journal of Human Resources*, 40(3):683–715, 2005.

G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

J. P. Jacobsen, J. W. Pearce III, and J. L. Rosenbloom. The effects of childbearing on married women's labor supply and earnings: Using twin births as a natural experiment. *Journal of Human Resources*, 34(3):449–474, 1999.

S. Jayachandran, A. Lleras-Muney, and K. V. Smith. Modern Medicine and the 20th-Century Decline in Mortality: Evidence on the Impact of Sulfa Drugs. *American Economic Journal: Applied Economics*, 2(2):118–46, 2010.

T. Joyce. If only policy analysis were so easy. *Journal of Policy Analysis and Management*, 32 (4):897–898, 2013. ISSN 1520-6688.

T. Kane and D. Staiger. Teen Motherhood and Abortion Access. *Quarterly Journal of Economics*, 111(2):467–506, 1996a.

T. J. Kane and D. Staiger. Teen motherhood and abortion access. *The Quarterly Journal of Economics*, 111(2):467–506, May 1996b.

M. S. Kearney and P. B. Levine. Subsidized contraception, fertility, and sexual behavior. *The Review of Economics and Statistics*, 91(1):137–151, February 2009.

M. S. Kearney and P. B. Levine. Why is the Teen Birth Rate in the United States So High and Why Does It Matter? *The Journal of Economic Perspectives*, 26(2):141–166, 2012.

J. Kim and A. Aassve. Fertility and its Consequence on Family Labour Supply. IZA Discussion Papers 2162, Institute for the Study of Labor (IZA), June 2006.

S. Kumar and A. Kugler. Testing the Children Quantity-Quality Trade-Off in India. MPRA Paper 42487, University Library of Munich, Germany, Sept. 2011.

D. S. Lee. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102, 07 2009.

J. Lee. Sibling size and investment in children's education: an Asian instrument. *Journal of Population Economics*, 21(4):855–875, October 2008.

D. I. Levine and G. Painter. The Schooling Costs of Teenage Out-of-Wedlock Childbearing: Analysis with a within-School Propensity-Score-Matching Estimator. *The Review of Economics and Statistics*, 85(4):884–900, 2003.

P. Levine, A. Trainor, and D. Zimmerman. The effect of Medicaid abortion funding restrictions on abortions, pregnancies and births. *Journal of Health Economics*, 15(5):555–578, 1996a.

P. Levine, D. Staiger, T. Kane, and D. Zimmerman. *Roe v Wade* and American Fertility. *American Journal of Public Health*, 89(2):199–203, 1999.

P. B. Levine, A. B. Trainor, and D. J. Zimmerman. The effect of medicaid abortion funding restrictions on abortions, pregnancies and births. *Journal of Health Economics*, 15(5):555–578, October 1996b.

H. Li, J. Zhang, and Y. Zhu. The Quantity-Quality Trade-off of Children in a Developing Country: Identification using Chinese Twins. *Demography*, 45:223–243, 2008.

Z. Li, J. Gindler, and H. Wang. Folic acid supplements during early pregnancy and likelihood of multiple births: a population-based cohort study. *The Lancet*, 361:380–84, 2003.

C. F. Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, February 2013.

V. Maralani. The Changing Relationship between Family Size and Educational Attainment over the Course of Socioeconomic Development: Evidence from Indonesia. *Demography*, 45(3): 693–717, 2008.

B. Mazumder and Z. Seeskin. Skipping breakfast and the sex ratio at birth. Mimeo, Northwestern, 2014.

C. McIntosh. Estimating Treatment Effects from Spatial Policy Experiments: An Application to Ugandan Microfinance. *The Review of Economics and Statistics*, 90(1):15–28, February 2008.

E. Miguel and M. Kremer. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica*, 72(1):159–217, 01 2004.

A. Miller. The effects of motherhood timing on career path. *Journal of Population Economics*, 24(3):1071–1100, July 2011.

A. R. Miller. Motherhood delay and the human capital of the next generation. *American Economic Review*, 99(2):154–58, May 2009.

G. Miller. Contraception as Development? New Evidence from Family Planning in Colombia. *The Economic Journal*, 120(545):709–736, 2010.

D. L. Millimet and L. Wang. Is the Quantity-Quality Trade-Off a Trade-Off for All, None, or Some? *Economic Development and Cultural Change*, 60(1):pp. 155–195, 2011.

R. Moffitt. Remarks on the analysis of causal relationships in population research. *Demography*, 42(1):pp. 91–108, 2005.

A. Nevo and A. M. Rosen. Identification with Imperfect Instruments. *The Review of Economics and Statistics*, 94(3):659–671, August 2012.

G. Noé, H. B. Croxatto, A. M. Salvatierra, V. Reyes, C. Villarroel, C. M. noz, G. Morales, and A. Retamales. Contraceptive efficacy of emergency contraception with levonorgestrel given before or after ovulation. *Contraception*, 84(5):486–492, November 2011.

N. Novikova, E. Weisberg, F. Z. Stanczyk, H. B. Croxatto, and I. S.Fraser. Effectiveness of levonorgestrel emergency contraception given before or after ovulation—a pilot study. *Contraception*, 75(2):112–118, October 2007.

M. J. O'Dowd and E. E. Philipp. *The History of Obstetrics and Gynaecology*. Parthenon, New York, 1994.

V. Ponczek and A. P. Souza. New evidence of the causal effect of family size on child quality in a developing country. *Journal of Human Resources*, 47(1):64–106, 2012.

C. Pop-Eleches. The Supply of Birth Control Methods, Education and Fertility: Evidence from Romania. Manuscript, Columbi Univ., Dept. Econ., 2005.

C. Pop-Eleches. The Impact of an Abortion Ban on Socioeconomic Outcomes of Children: Evidence from Romania. *Journal of Political Economy*, 114(4):744–773, 2006.

N. Qian. Quantity-quality and the one child policy: The only-child disadvantage in school enrollment in rural China. NBER Working Papers 14973, National Bureau of Economic Research, Inc, May 2009.

C. Quintana-Domeque and P. Ródenas-Serrano. Terrorism and Human Capital at Birth: Bomb Casualties and Birth Outcomes in Spain. IZA Discussion Papers 8671, Institute for the Study of Labor (IZA), Nov. 2014.

D. C. Ribar. Teenage Fertility and High School Completion. *The Review of Economics and Statistics*, 76(3):413–424, 1994.

D. C. Ribar. The Socioeconomic Consequences of Young Women's Childbearing: Reconciling Disparate Evidence. *Journal of Population Economics*, 12(4):547–565, 1999.

M. R. Rosenzweig and P. T. Schultz. Fertility and investments in human capital: Estimates of the consequence of imperfect fertility control in Malaysia. *Journal of Econometrics*, 36(1-2): 163–184, 1987.

M. R. Rosenzweig and T. P. Schultz. The Demand for and Supply of Births: Fertility and its Life Cycle Consequences. *The American Economic Review*, 75(5):992–1015, December 1985.

M. R. Rosenzweig and K. I. Wolpin. Testing the quantity-quality fertility model: The use of twins as a natural experiment. *Econometrica*, 48(1):227–40, January 1980a.

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

M. R. Rosenzweig and K. I. Wolpin. Evaluating the Effects of Optimally Distributed Public Programs: Child Health and Family Planning Interventions. *The American Economic Review*, 76(3):470–482, 1986.

M. R. Rosenzweig and K. I. Wolpin. Sisters, Siblings, and Mothers: The Effect of Teen-Age Childbearing on Birth Outcomes in a Dynamic Family Context. *Econometrica*, 63(2):303–326, 1995.

M. R. Rosenzweig and J. Zhang. Do population control policies induce more human capital investment? twins, birth weight and China's one-child policy. *Review of Economic Studies*, 76 (3):1149–1174, 07 2009.

D. B. Rubin. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

D. B. Rubin. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1):34–58, January 1978.

D. B. Rubin. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, 75(371):591–593, September 1980.

C. Sanhueza. Tamaño de familia y orden de nacimiento en Chile: Usando gemelos como experimento natural. *Revista de Analisis Económico – Economic Analysis Review*, 24(2):133–152, Dec 2009.

P. T. Schultz. Population policies, fertility, womens human capital, and child quality. In P. Schultz and J. Strauss, editors, *Handbook of Development Economics*, volume 4, pages 3249–3303. Amsterdam: Elsevier Science B.V., 2008.

B. L. Shepard and L. Casas Becerra. Abortion policies and practices in Chile: Ambiguities and dilemmas. *Reproductive Health Matters*, 15(30):202–210, 2007.

Task Force on Postovulatory Methods of Fertility Regulation. Randomised controlled trial of levonorgestrel versus the yuzpe regimen of combined oral contraceptives for emergency contraception. *The Lancet*, 352:428–33, 1998.

The World Bank. Data bank. http://data.worldbank.org, 2015. Consulted 31 May, 2015.

UK Family Planning Association. Contraception: past, present and future. Fact Sheet, April 1996. Accessed 10/01/2014. Available online at http://www.fpa.org.uk/about/info/contraceptionpastpresentandfuture.htm.

United Nations. World population policies 2009. Technical report, Department of Economic and Social Affairs: Population Division, 2010.

H. von Hertzen, G. Piaggio, A. Peregoudov, J. Ding, J. Chen, S. Song, G. Bártfai, E. Ng, K. Gemzell-Danielsson, A. Oyunbileg, S. Wu, W. Cheng, F. Lüdicke, A. Pretnar-Darovec, R. Kirkman, S. Mittal, A. Khomassuridze, D. Apter, and W. R. G. on Post-ovulatory Methods of Fertility Regulation. Low dose mifepristone and two regimens of levonorgestrel for emergency contraception: a who multicentre randomised trial. *The Lancet*, 360(9348):1803–1810, December 2002.

S. Whitaker. The impact of legalized abortion on high school graduation through selection and composition. *Economics of Education Review*, 30(2):228–246, April 2011.

R. J. Willis. A New Approach to the Economic Theory of Fertility Behavior. *Journal of Political Economy*, 81(2):S14–S64, 1973.

T. Zajonc. Boundary Regression Discontinuity Design. Dissertation, Harvard University, May 2012.

A. Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348–368, June 1962.