CSAE Working Paper WPS/2016-38

The Twin Instrument

Sonia Bhalotra* Damian Clarke†

December, 2016

Abstract

Twin births are often construed as a natural experiment in the social and natural sciences on the premise that the occurrence of twins is quasi-random. We present new population-level evidence that challenges this premise. Using individual data for more than 18 million births (more than 500,000 of which are twins) in 72 countries, we demonstrate that indicators of the mother's health and health-related behaviours and exposures are systematically positively associated with the probability of a twin birth. The estimated associations are sizeable, evident in richer and poorer countries, and evident even in a sample of women who do not use IVF. The positive selection of women into twinning implies that estimates of impacts of fertility on parental investments and on women's labour supply that use twin births to instrument fertility will tend to be downward biased. This is pertinent given the emerging consensus that these relationships are weak. Using two large samples, one for developing countries and one for the United States, and focusing upon twin-instrumented estimates of the quantity–quality trade-off, we demonstrate the nature of the bias and estimate bounds on the true parameter.

JEL codes: J12,J13,C13,D13,I12.

Keywords: Twins; fertility; maternal health; miscarriage; quantity–quality trade-off; parental investment; bounds

*Department of Economics and ISER, The University of Essex. Contact: srbhal@essex.ac.uk

†Deparment of Economics, Universidad de Santiago de Chile. Contact: damian.clarke@usach.cl

Introduction

Twins have intrigued humankind for more than a century (Thorndike, 1905). In behavioural genetics, demography and psychology, monozygotic twins are studied to assess the importance of nurture relative to nature (Thorndike, 1905; Boomsma et al., 2002; Polderman et al., 2015; Phillips, 1993; Bouchard and Propping, 1993; McClearn et al., 1997; Nisén et al., 2013). In the social sciences, twin births are also used to denote an unexpected increase in family size which assists causal identification of the impact of fertility on investments in children and on women's labour supply (Rosenzweig and Wolpin, 2000, 1980a; Bronars and Grogger, 1994; Black et al., 2005). A premise of studies that use twin differences or the twin instrument is that twin births are quasi-random. We present new population-level evidence that challenges this premise. Using 18,652,028 births in 72 countries, of which 539,544 (2.89%) are twins, we show that the likelihood of a twin birth varies systematically with maternal condition. We document that this association is meaningfully large, and widespread.

The association of twin births and maternal condition is evident in richer and poorer countries, and we show that it holds for all available markers of maternal condition including health stocks and health conditions prior to pregnancy (height, body mass index, diabetes, hypertension, asthma, kidney disease), health-related behaviours in pregnancy (healthy diet, smoking, alcohol, drug consumption), exposure to stress in pregnancy, and measures of the availability of medical professionals and prenatal care. The underlying hypothesis is that twins are more demanding of maternal resources than singletons and so conditions that challenge maternal health, be they long-standing under-nutrition (marked by height) or pregnancy behaviours (like smoking), are more likely to result in miscarriage of twins. We substantiate this mechanism using United States Vital Statistics data (containing 14 to 16 million births).

Previous research has documented that twins have different endowments from singleton children, for example, twins are more likely to have low birth weight and congenital anomalies (Hall, 2003; Rosenzweig and Zhang, 2009; Almond et al., 2005). We focus not on differences between twins and singletons but rather on differences between mothers of twins and singletons, which indicate whether occurrence of twin births is quasi-random. It is known that twin births are not strictly random, occurring more frequently among older mothers, at higher parity and in certain races and ethnicities (Hall, 2003; Bulmer, 1970), but as these variables are typically observable, they can be adjusted for. Similarly, it is well-documented that women using artificial reproductive technologies (ART) are much more likely to give birth to twins (Vitthala et al., 2009) and as ART-use is recorded in many birth registries, it can also be controlled for and a conditional randomness assumption upheld. The reason that our finding is potentially a major challenge

is that maternal condition is multi-dimensional and almost impossible to fully measure and adjust for. For instance, foetal health is potentially a function of whether pregnant women skip breakfast (Mazumder and Seeskin, 2014), whether they suffer bereavement in pregnancy (Persson and Rossin-Slater, Forthcoming; Black et al., 2016), their exposure to air pollution (Chay and Greenstone, 2003), and a host of other such variables.

After comprehensively documenting in Part 1 of the paper that the occurrence of twin births is correlated with maternal health, in Part 2 we trace the implications of this for research that has exploited the assumed randomness of twin births. We argue that studies using twins to isolate exogenous variation in fertility will tend to under-estimate the impact of fertility on parental investments in children, and on women's labour supply. This is pertinent given the ambiguity of the available evidence. Recent studies using the twin instrument challenge a long-standing theoretical prior of Becker (1960); Becker and Lewis (1973); Becker and Tomes (1976) in rejecting the presence of a quantity–quality trade-off in developed countries (Black et al., 2005; Angrist et al., 2010), but our estimates suggest that this rejection could in principle arise from ignoring the positive selection of women into twin birth. Similarly, research using the twin instrument tends to find that additional children have small if any influence on women's labour force participation (see (Lundborg et al., 2014)). But, again, these estimates are likely to be downward biased as these studies do not account for the positive selection of twin mothers on health-related indicators. The results of studies in Economics, Psychology, Education and Biology that instead exploit the genetic similarity of twins will not be biased but will tend to have more restricted external validity than previously assumed.

To illustrate the problem, we focus upon the quantity–quality (QQ) trade-off. We provide estimates for the United States (using about 225,000 births, drawn from the United States National Health Interview Surveys (NHIS) for 2004-2014) and for a pooled sample of developing countries (containing more than 1 million births in 68 countries over 20 years, available from the Demographic and Health Surveys). These data are chosen because they contain information on child outcomes and maternal health. Consistently using these two samples allows us to assess the generality of our findings, and it allows that the QQ trade-off, as well as the violation of the exclusion restriction that concerns us, are different in richer vs poorer countries.

First, we show that negative selection of women into high fertility[1] and positive selection of women into twin birth imply that the OLS and IV estimates provide bounds on the true parameter. We show that these bounds can be tightened by controlling for available indicators of maternal health. In particular, we display routine twin-IV estimates of the QQ trade-off and

[1] We test and find evidence for this using partial controls for socio-economic status and health.

our results replicate the common finding that there is no discernible trade-off. We then adjust for available maternal health related characteristics. We find that, even though we have only a small subset of the range of relevant maternal health indicators in our data sets, conditioning upon them leads to emergence of a QQ trade-off. This demonstrates the force of our contention.

Since the adjustment is partial, in principle the exclusion restriction continues to be violated. Given that the first stage (twins predicting fertility) is powerful, we next estimate bounds on the IV estimates on the premise that twin births are plausibly if not strictly exogenous (Conley et al., 2012). With a view to improving the precision and relevance of these bounds, we estimate rather than assume a measure of the extent of the violation of the exclusion restriction. The data requirements for this are non-trivial—we need data on two generations, with an exogenous shock to maternal health in the first generation, and measures of child quality in the second generation. For this, we exploit natural experiments in the United States and Nigeria.

We now present some of the effect sizes and place them in perspective with reference to research on other determinants of the outcomes. The twin-IV estimator produces coefficients that are small and not significantly different from zero. Conditioning upon available characteristics produces statistically significant coefficients consistent with a trade-off, for instance, in samples with at least three births, this is 0.05 s.d. for years of education in developing countries, and 0.06 s.d. for an index of child health in the United States, and in the sample with at least two births it is 0.10 s.d. for grade retention in the US. The Conley bounds, in general, confirm the presence of a trade-off. The lower bound is 0.05 to 0.06 s.d. for education in developing countries, 0.13 to 0.24 s.d. for education in the USA and 0.02 to 0.09 of a s.d. for child health in the USA.[2] Observe that the trade-off is no smaller in the USA than in developing countries. This is important given that the recent studies arguing there is no trade-off are set in richer countries, and a natural reconciliation of these results with earlier studies proposed is that the trade-off may exist but only in poorer countries where a larger share of families is credit constrained.

Consider the estimated 0.05 s.d. increase in education in developing countries stemming from one less birth in the family. In that sample this corresponds to 0.17 years of completed education This is similar to the gains in completed education flowing from high-profile experiments providing information on returns to education (Jensen, 2010), or a school-based de-worming programme (Baird et al., 2016).[3] So the adjusted estimates are of a size that it is not prudent to dismiss. Moreover, our estimates indicate the change in investment (education) for one additional birth but, as fertility rates remain high in many developing countries, the total effect can be large.

[2] All of these figures are negative; we do not display the negative sign as this is implied by the language of a trade-off.

[3] In section 3.1.3 we provide comparisons with other studies too.

To take stock, this paper makes two main contributions. First, it establishes that mothers of twins are selectively healthy. In Part 1, we link this in with a growing literature in economics. Second, it shows that the widely used twin-instrument for fertility is invalid, but suggests how with the use of available measures of maternal health and the estimation of bounds, inference can be made in the IV setting. The second contribution puts back on the stage the issue of a potential human capital cost to fertility. Governments actively devise policies to influence fertility, for instance, countries like China have penalized fertility, while many countries including Italy and Canada have incentivized it, often with non-linear rules.[4] Advocates of policies encouraging smaller families argue that large families invest less in the quality of each child, limiting human capital accumulation and living standards (Galor and Weil, 2000; Moav, 2005).

Any human capital costs of fertility are naturally of greater concern when fertility is high and when a large share of it is unwanted. In 2015 the average number of births per woman in low income countries was 5 and, comparing actual with stated desired fertility, we estimate the share of unwanted births is as high as 60 per cent in some countries, with a mean of 27 per cent. Unwanted fertility is not unique to poorer countries. For instance, despite access to contraceptive methods, 21 percent of all pregnancies in 2011 in the United States ended in elective abortion (Guttmacher Institute, 2016). Moreover, there is a strong trend in IVF use, and up to 40% of IVF successes result in multiple births to women who wanted one child (Kulkarni et al., 2013), creating a growing set of unwanted children. Irrespective of whether additional births are wanted or unwanted, re-establishing the evidence that they lead to diminished investments in human capital in children is important, possibly more so recently in view of growing evidence of the long run dynamic benefits of childhood investments (Campbell et al., 2014; Heckman et al., 2013).

While schooling rates have been increasing globally, of 163 developing countries, only 47 have achieved universal primary education (World Bank, 2009) and in two-thirds of sub-Saharan African countries, more than 30 per cent of students who start primary school are expected to drop out. Credit constraints have been identified as a factor (UNESCO, 2011), and these tend to tighten as the number of dependent children increases. In high income countries, although all children stay in school until the legislated minimum school leaving age, grade retention is a significant marker of educational progress. Systematic reviews examining research over almost a century conclude that grade retention is one of the strongest predictors of high school dropout,

[4] As discussed in Mogstad and Wiswall (2016), families with children receive special treatment under the tax and transfer provisions in 28 of the 30 Organization for Economic Development and Cooperation countries (OECD (2002)). Many of these policies are designed such that they reduce the cost of having a single child more than the cost of having two or more children, in effect promoting smaller families. For example, welfare benefits or tax credits are, in many cases, reduced or even cut off after reaching a certain number of children.

and is associated with lower earnings in adulthood (Jimerson, 1999, 2001; Jimerson et al., 2002). It is estimated that over 2.4 million (5-10%) students are retained every year in the US. Rising through the twenty-five years up to 2003, this was estimated to cost over 13 billion dollars per year just to pay for the extra year of schooling (i.e. ignoring its long run costs) (Anderson et al., 2002). Retention rates are higher among boys, ethnic minorities and children of less educated parents (Warren et al., 2014). Passage in 2002 of the No Child Left Behind Act in the US, with its emphasis on mastery of minimum grade-level competencies as a condition for promotion, has renewed discussion of grade retention in public policy making. The third measure of child quality we use is a subjective measure of child health. Beyond its intrinsic value, the long term health and socio-economic payoff to improved child health is estimated to be large Almond and Currie (2011). Although our estimates pertain to implications of fertility for parental investments, as discussed, we expect that the true career costs of children among women are larger than estimated in twin-IV studies, consistent with, for instance, Adda et al. (2016).

This paper is laid out as follows. In Part 1, we present evidence from several countries and time periods, using multiple indicators of maternal health that twin births are not random, we show that selective miscarriage is a mechanism by which this association emerges, and we discuss other potential mechanisms. In Part 2, we develop the implications of this result for studies in which twin births are used to instrument fertility, focusing upon estimates of responses of parental investments (proxied by indicators of child quality) to an additional birth. We demonstrate the substantive significance of the twin-IV bias. Using natural experiments, we estimate the degree of failure of the exclusion restriction and, using these estimates, construct bounds on the true parameter.

Part I – Twin Births and Maternal Health

In this section, we first detail the evidence that twin births are positively correlated with indicators of maternal health, health-related behaviours, health facilities, and stress in pregnancy. We also present a positive association of twin births with the mother's education (even in a sample of non-ART users), which we argue is consistent with education facilitating access to and uptake of new health-related information (Kenkel, 1991; Lleras-Muney and Cutler, 2010). We demonstrate the generality of this result using administrative data and large representative survey data from several countries across several years, and a variety of indicators of health. In poor countries, many women are severely under-nourished, antenatal services are weak and, as a result, foetal health is worse, as indicated for instance by birth weight or neonatal mortality; see Bhalotra

and Rawlings (2013). It is therefore immediately plausible that many women may not have the fitness to carry twin conceptions to birth. In richer countries, while chronic under-nutrition is less prevalent, many women engage in risky behaviours and/or experience medical problems in pregnancy (see Appendix Table A1). Using the variables for which we have comparable data in richer and poorer countries, we show that the associations of interest are evident at a range of levels of country income (GDP). While other critiques have been directed at twin studies (Rosenzweig and Zhang, 2009), the systematic tendency for maternal health and pregnancy behaviours to influence the distribution of twins across families appears not to have been previously documented, and certainly not using population-level data.[5]

Second, we discuss the mechanisms by which this association may arise. We estimate the relevance of selective maternal survival (which is only a potential issue in the DHS data) and show that this cannot explain away our results. Instead we provide evidence that twin conceptions are disproportionately more likely to miscarry among women with adverse health indicators. We are agnostic on the question of whether twin conceptions are random. Our findings add a novel twist to a recent literature documenting that a mother's health and her environmental exposure to nutritional or other stresses during pregnancy influence birth outcomes, with many studies documenting lower birth weight (Currie and Moretti, 2007; Bernstein et al., 2005; Quintana-Domeque and Ródenas-Serrano, 2014). If birth weight is the intensive margin, we may think of miscarriage as the extensive margin response, or the limiting case of low birth weight. Similarly, other studies have demonstrated that weaker maternal condition is associated with a lower probability of male birth (Trivers and Willard, 1973; Almond and Edlund, 2007). Our results arise from a similar process and, intersecting our hypothesis with Trivers-Willard, we show that twin births are more likely to be female.

1 Methodology

To test the "as good as random" assumption regarding twin births, we estimate conditional regressions of the form:

$$twin_{bjt} = \gamma_0 + \gamma_1 Health_{bjt} + \mu_b + \lambda_t + \varepsilon_{bjt}. \quad (1)$$

Here, *twin* is an indicator of whether a birth of order b born to woman j at age t is a twin. We control for fixed effects for mother's age and parity, given that these are known to influence the

[5] Rosenzweig and Zhang (2009) argue that the twin-IV estimator fails to account for parents' reinforcing investments in better endowed (singleton) children, and that it ignores scale-economies. Our critique is independent of these considerations.

probability of twin births (for example Hall (2003), Rosenzweig and Wolpin (1980a)). Where births are observed over multiple years, races or geographic areas, we include the relevant fixed effects. If, as is commonly assumed, twin birth is an event which is conditionally as good as random, the coefficients on maternal health variables $Health_{jt}$ should not be statistically distinguishable from zero. Our test is equivalent to a test of balance of characteristics of 'treated' (with twins) and 'control' (without twins) mothers. In the principal specification, equation 1 is estimated separately for each health indicator. However we also display estimates of a version in which all available measures of health are entered together. Standard errors are always clustered at the level of the mother.

For ease of exposition, we maintain subscript t for the woman's age but the health indicators are measured either during or before pregnancy, to avoid the potential concern of reverse causality, i.e. that twin births cause greater depletion of the mother's health than singleton births. Notice, though, that if we were to have to work with data in which an indicator of mother's health was only available after birth (as is the case for BMI in the DHS data), then this would make it *harder* for us to establish that mothers of twins are selectively healthy. In this sense, accounting for feedback from twin birth to mother's health would only strengthen the evidence we present. A particularly nice case that we discuss below uses an exogenous measure of environmental stress in pregnancy.

We add controls for education (and wealth, which is available for the developing country and the Chilean data) to allow for the fact that education may encourage and wealth may facilitate health-seeking behaviours, and also to make sure that $Health$ is not simply proxying for socio-economic status. For the two indicators for which we have comparable data across countries, we shall further test whether the association of twinning with the indicator is a function of country income. For the United States, where the data permit this distinction, we re-estimate the equation excluding all women who report having used assisted reproductive technologies. For the developing country sample, on the premise that ART was not available prior to 1990, we split the birth data into pre- and post-1990 samples. This way we can be sure that our results are independent of ART use.

We perform an alternative test that exploits within-mother variation. This essentially involves testing whether women who produce twins had healthier births before the twin birth, as this would be a measure of their pre-determined health. For each $n = \{2, 3, 4\}$ we estimate:

$$InfantDeath_{ij,b<n,t} = \alpha_0 + \alpha_1 Twin_{j,b=n,t} + \mu_b + \lambda_t + \nu_{ijbt}, \tag{2}$$

where we restrict the sample to children i of mother j who are born at least one year before

birth order $b < n$. Thus, for $n = 2$, the independent variable $Twin$ takes the value of one if the mother gives birth to twins on her second birth, and zero if she gives birth to a singleton on her second birth. We then regress a binary variable for infant mortality of children born at birth order 1 on the mother's twin status at birth order 2. We generalize this to higher birth orders. Infant mortality is widely used as a marker of health and it has the advantage that it is largely predetermined (except for children born within a year of their older sibling). If twinning is as good as random, we should observe that no pre-determined variables predict twinning, and so we cannot reject the hypothesis $H_0 : \alpha_1 = 0$, while if 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, and hence $\alpha_1 < 0$. Once again, standard errors are clustered at the level of the mother, and age and birth order fixed effects are included.

2 Data and Descriptive Statistics

We sought data that satisfied the following requirements. First, that the samples are representative and, given the relative rarity of twins, large. Second, that the data include birth records and that they distinguish singleton from multiple births. Third, that they contain indicators of the mother's health and/or health-related behaviours. Datasets fulfilling these criteria include administrative birth data from the United States, Spain and Sweden, and household survey data from Chile, the United Kingdom, and 68 developing countries (the Demographic Health Surveys, or DHS) for births that occur during 1972-2012. Together these data contain fifteen indicators of health. Administrative data for births in other countries including, for instance, India, Mexico and Argentina do not contain measures of maternal health. Geographic coverage of the data we use is mapped in online appendix figure A1.

We consistently restrict the sample to women aged 18-49 years old, and exclude the small proportion of births in which more than two children are born at once (triplets and higher order multiple births). Given not only a positive association of ART with the likelihood of twin births (Vitthala et al., 2009), but also that ART users are typically more educated and wealthy (Lundborg et al., 2014), it is important to demonstrate that our hypothesis holds independently of ART use. Since the US Vital Statistics data indicate ART use for every birth in post-2009 data where the updated birth certificate is used, we present estimates using the universe of mothers giving birth between 2009-2013, who reported *not* using ART. This involves removing approximately 1.6% of births from the sample.

A description of each dataset and its coverage is provided in online data appendix C and

summary statistics are in appendix table A1. Estimates of equation 2 require the complete fertility history including the survival status of all children preceding each twin or singleton birth, which only the DHS provide, for 68 developing countries. In the left-hand panel of Appendix Table A2 we present summary statistics by birth type from the DHS data (we discuss data in the right-hand panel in part II of this paper).

Not all birth registers that include maternal health indicators also include information on foetal deaths, but the US Vital Statistics data do. We pooled all births and foetal deaths recorded in US administrative data between 1998 and 2002. We stopped in this year because, from 2003, a considerable re-definition of birth certificate data meant that foetal death and birth data did not share similar controls and the coverage varied by state. Prior to 2002 however we are able to observe for all states whether a mother smokes or drinks during pregnancy, whether she suffered from anemia prior to pregnancy, and her educational level.

Twin births make up 2.84% of all births in the United States, 2.55% of all births in Sweden and Chile, 2.37% in county Avon of the UK and 2.10% of births in the DHS sample (see table A1). Twins are not as rare as we may think: about 1 in 80 live births and hence 1 in 40 babies is a twin. The proportion of twin births tends to be higher in richer countries, consistent with our contention that twin births occur more often to women with the resources to carry twin births to term. Similarly Appendix table A2 presents a simple comparison of means of characteristics of mothers and children in families with and without at least one twin birth. The comparison of mother characteristics for both the USA and developing countries shows, consistent with our hypothesis, that healthy mothers (as proxied by height, BMI and probability of being underweight) are more likely to give birth to twins. We test formally for the equality of means of a wider set of maternal health indicators in appendix table A3, though these tests are unconditional, and do not account for twin mothers tending to be older and for twins being more likely at higher birth parities. We present conditional tests in the next section.[6]

[6]Plots of the frequency of twin births over time in the US and in the developing country sample, along with aggregate health in the population as proxied by female life expectancy are in appendix figure A2. However, given that these plots cover both the pre- and post-IVF period, any correlation between trends in twinning and aggregate health may be incidental.

3 Results: Twinning

3.1 Twin Births and Maternal Condition

Various indicators of maternal health: conditions, behaviours, facilities In table 1 we present estimates of equation 1 for four richer countries and for a sample that pools data for 68 developing countries. We use multiple indicators for maternal health before pregnancy, health-related behaviours and exposure to stress during pregnancy, and availability of prenatal and medical facilities in the local area. The indicators available vary across the five samples that we analyse but we find broadly consistent results across indicators and across samples. All independent variables are standardised as Z-scores so that the estimates can be cast as the effects of increasing by 1 standard deviation (sd) the independent variable of interest. Unstandardised results are presented in appendix table A5.

We observe that being underweight, having any of a range of morbidities prior to conception, and adopting risky behaviours in pregnancy, each significantly reduces the probability of a twin birth. A healthy diet in pregnancy, height (an indicator of the stock of health (Silventoinen, 2003; Bhalotra and Rawlings, 2013)) and greater access to prenatal and medical care raise the likelihood of giving birth to twins. Maternal education is also a significant predictor of twins, consistent with education promoting health (Kenkel, 1991; Lleras-Muney and Cutler, 2010). Statistical significance of the health indicators in Table 1 is robust to running regressions which condition on all available indicators of the mother's health, including education (table A4). The effects are sizeable, with a 1 sd improvement in the indicator tending to increase the likelihood of twinning by 6-12% in most cases, although there is variation, with smaller effects from fresh fruit consumption and larger effects from height. The rest of this section will elaborate these findings.

Estimates for the USA As discussed, our estimates for the US are restricted to women who do not use any assisted reproductive technologies (ART). We estimate that a 1 sd increase in rates of smoking in each trimester is associated with a 0.20-0.25 pp reduction in the risk of a twin birth (p <0.001 in each case). These results all hold even when correcting test statistics for large sample sizes and the increasing likelihood of rejecting a null, as outlined in Leamer (1978); Deaton (1997). Smoking in the third trimester imposes the largest reduction, consistent with evidence that adverse effects of smoking on birth weight are largest in the third trimester (Bernstein et al. (2005); also see supplementary table A6). These effects of smoking are about 10% of the mean rate of twinning.

Diabetes and hypertension prior to pregnancy have similar standardized effects, reducing the likelihood of twin birth by between 0.2 and 0.3 pp.[7] Height and education have larger standardized effects, of 0.63 and 0.81 pp respectively, and being underweight is associated with a 0.15 pp decrease in the probability of twins. Estimates for the 1.6 percent of women using ART are in table A7 and are, with the exception of being underweight, larger and statistically significant for every indicator, underlining the additional sensitivity of birth outcomes in this group.

Estimates for Sweden, Avon, Chile and Low Income Countries Analysis of birth registers from Sweden for the years 1993-2012 indicates similar standardised effect sizes for smoking, diabetes, height and being underweight to those for US women. Although the effect of hypertension is smaller, at close to 0.10 pp., the similarity of the results for Sweden and the USA is striking. The Swedish data additionally record asthma prior to conception, which we estimate reduces the risk of twin births by 0.015 pp. Survey data from Avon county UK 1991-1992 and Chile 2006-2009 again exhibit patterns similar to those identified for Sweden and the USA for anthropometric indicators of health, risky behaviours and pre-pregnancy illnesses. All results are in table 1 so here we only discuss estimates for indicators specific to these additional datasets. The UK data contain unique information on eating healthily during pregnancy and our estimates indicate that the standardised effect of this is a 0.54 pp increase in the likelihood of having twins. The Chilean data record drug use during pregnancy and we find that frequent drug use reduces the probability of twins by 0.16 pp, an estimate similar to that for frequent alcohol consumption in this sample.

In the sample that pools data for 68 developing countries in 1961-2012, we observe some measures of health stocks including height, weight and body mass index, and we observe the local availability of prenatal care and access to medical professionals. These variables are all measured as the leave-one-out rate of healthcare access in the mother's cluster of residence since we are interested in availability rather than use to avoid the concern that mothers conceiving twins may be more likely to actively seek birth attendance. These measures are useful because health service coverage in low-income countries is far from universal. We observe that taller and heavier women are more likely to twin (mirroring the findings from other contexts), and estimate that a 1 sd increase in availability of doctors or nurses is associated with a 0.092 pp and 0.065 pp increase in the likelihood of twins respectively.

Quasi-experimental variation in maternal stress: Spain Using the methodology and data described in Quintana-Domeque and Ródenas-Serrano (2014), we estimated a similar regression

[7] We do not include measures of morbidities during gestation such as pre-eclampsia as these may be endogenous, dependent upon the higher biological demands of a twin conception.

using ETA bombing in Spain as a plausibly exogenous shock to maternal stress during pregnancy (a proxy for $Health$). We find that an additional bomb casualty in the province of residence of a pregnant woman decreases the likelihood that she will have a twin birth by 0.01% and 0.012% (p <0.01 and 0.05 respectively); see table 2. This effect is larger and only statistically significant during the second and third trimesters, similar to the effects of smoking by trimester documented in table 1.[8]

Survival of pre-twin births as a marker of mother's health As discussed earlier, an alternative test of the quasi-randomness of twin births consists of examining an indicator of the health stocks of mothers, proxied by the early life survival of their children born *before* twin births. This is a natural measure of maternal health, capturing a woman's ability to produce surviving children, which is exactly what we hypothesize is challenged by twins. Estimates of equation 2 are presented in table 3. We see that mothers who went on to have second-born twins had much lower rates of infant mortality for their *first* births than women who had second-born singletons, and that this generalizes to higher parities.[9] Additionally, we observe that introducing controls for mother's health (height and weight) and socioeconomic status (completed education) reduces the size of the estimated infant mortality differential between twin and non-twin mothers, suggesting that these controls proxy for the positive selection of women into twinning. In the first row these controls explain approximately 15% of the selection into twinning, however even conditional on education and anthropometrics, twin mothers have 1.7 fewer infant deaths than non-twin mothers per 100 live births.

3.2 Cross-Country Comparisons and the Role of Income

Cross-country comparison of the association of twinning with mother's height Figure 1 shows that in 68 of 70 countries for which data on women's heights can be matched to birth records, twin mothers are on average significantly taller than non-twin mothers.[10] Height is the indicator of health most widely measured in birth and demographic data and several studies show that it responds to infection and nutritional scarcity in the growing years, for instance individuals exposed to famine and war have been shown to have lower stature in adulthood, other things

[8] Quintana-Domeque and Ródenas-Serrano (2014) find that the same exposure reduces average birth weight by approximately 0.3 grams (trimester 1), and increases the likelihood of low birth weight by 0.14%.

[9] Infant mortality is defined as the death of a child before their first birthday. We remove from the sample any children who were still alive but 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.

[10] Each estimate reflects the mean difference between twin and non-twin mothers, conditioning on age and parity fixed effects. As the comparison is within country, it nets out country differences including differences in the genetic pool (Deaton, 2007).

equal (Silventoinen, 2003; Bozzoli et al., 2009; Wang et al., 2010; Akresh et al., 2012). Moreover, previous research has shown widespread associations of short stature among mothers with the risk of low birth weight and infant mortality among their children (Bhalotra and Rawlings, 2013).

Twinning and maternal health: is there an income gradient? Since many women in poorer countries are under-nourished, it seems plausible that their resources are particularly challenged in carrying twins to term. As a result, we may expect that income growth and poverty reduction attenuate the association of mother's health and twin births. On the other hand, risky behaviours in pregnancy may be increasing in income, so the gradient will depend upon the health indicator that is analysed. To assess this, we need a comparable index of mother's health for countries that span a range of income levels. As height is widely available, we plot the point estimates from figure 1 against GDP per capita in figure 2. The estimates lie above the zero line, indicating that the relationship persists in high income countries.

Mother's education In table 1, it is noteworthy that in all samples where maternal education is recorded it is significantly positively related to the likelihood of twinning, even in the non-ART sample in the US. In the online appendix we present similar plots to the height figures 1 and 2 displaying education differences between twin and non-twin mothers in all countries in the sample (figures A3 and A4). We observe that in the overwhelming majority of countries, twin mothers have greater education than non-twin mothers. When we plot these estimates against GDP per capita using all countries in the sample, we find little evidence to suggest that twin selection decreases as countries become more developed. If anything, there is a weakly positive correlation between country income and the education differential between twin and non-twin mothers. We interpret this as capturing the finding cited earlier that the effects of education on health care access and uptake are most substantial in environments in which health-care technologies are changing rapidly.

4 Mechanisms

In this section we consider three alternative hypotheses for the process determining the results contained in the preceding section. We shall refer to these as the conception, gestation and maternal survival mechanisms. First, healthier mothers may be more likely to *conceive* twins on account of an underlying genetic or biological process, such as that mediated by the follicle stimulating hormone Hall (2003). Second, conditional upon conceiving twins, healthier mothers may be more likely to take both fetuses to term. Third, 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.

Any of these processes is sufficient to violate the "as good as random" assumption insofar as they imply that observing twins will depend upon possibly unmeasured maternal behaviours and characteristics. Nonetheless, we may be interested in determining which of these is the relevant channel. For example, if twins are less likely *only* due to selective maternal death, then as mothers become more likely to survive childbirth (ie as maternal mortality declines), threats to validity become considerably less relevant. We cannot directly test the conception hypothesis since the relevant data are unavailable. However we shall present tests for the second and third hypothesis and argue that our conclusions in Part 2 of the paper are robust to the first mechanism being at play.

4.1 Selective foetal death

The gestation hypothesis is that carrying twins to term is more demanding than carrying singletons to term, and so stressors of maternal health will lead to selective miscarriage of twins. It has been documented that the biological demands of twin pregnancies are higher than the demands of non-twin pregnancies (Shinagawa et al., 2005; Kahn et al., 2003) and also that, in general, healthier mothers are less likely to miscarry (García-Enguídanosa et al., 2002). What we contribute here is to test the natural intersection of these hypotheses, and estimate the extent to which miscarriage is most frequent among less-healthy women carrying twins. The estimated equation is:

$$\begin{aligned} FoetalDeath_{ijt} &= \gamma_0 + \gamma_1 Twin_{ijt} + \gamma_2 Health_{jt} + \gamma_3 Twin \times Health_{ijt} + \\ &\quad \lambda_t + \phi_a + \mu_b + u_{ijt}. \end{aligned} \tag{3}$$

$FoetalDeath_{ijt}$ is a binary variable (multiplied by 1,000) indicating whether a birth was taken to term (coded as 0) or resulted in a miscarriage (coded as 1). As above, i indicates a conception leading to birth or foetal death, j a mother, and t is year, Health is an indicator of the mother's health, Twin is an indicator for whether the conception is a twin or a singleton and, as before, fixed effects for year (λ_t), mother's age (ϕ_a) and birth order (μ_b) are included. This is then regressed on the twin status of the pregnancy (1 if twins, 0 if singleton), a variable recording an indicator of maternal health, and an interaction between twin conception and maternal health. The coefficient of interest γ_3 is the differential effect of the variable $Health_{jt}$ on twin conceptions. The available measures of health are behaviours observed entirely before birth (smoking or drinking during

pregnancy) or health conditions observed entirely before conception (anemia). If $\gamma_3 = 0$, then the twin fetus is as likely to miscarry as the singleton fetus when exposed to health (dis)amenity $Health_{jt}$.

The results are in table 4. The evidence confirms previous research showing that the spontaneous abortion rate among twins (at 1 in 8 conceptions), is about three times that among singletons (Boklage, 1990). We can reject the null hypothesis that $H_0 : \gamma_3 = 0$. This adds new evidence of steeper gradients in indicators of mother's health for twins than for singletons. For example, a 1 standard deviation increase in rates of smoking during pregnancy whilst carrying a singleton elevates the risk of miscarriage by 1.39 foetal deaths per 1,000 live births. The corresponding risk elevation among mothers pregnant with twins is an increase of 2.55 foetal deaths (1.394+1.154), which is almost twice the risk. Alcohol consumption in pregnancy is similarly almost twice as risky for women carrying twins, and the risks associated with anemia are about three times as high. We also show that a college education modifies the difference in miscarriage probabilities more than three times as much when the mother is carrying twins than when she is carrying a singleton. Often one of two twins miscarries. In such cases, if the survivor is recorded as a singleton birth then we will tend to under-estimate the importance of maternal condition. In other words, our contention holds *a fortiori*.

Overall, these results establish a plausible mechanism for the associations that we document in table 1. Here we have modelled miscarriage conditional upon the conception being twin or singleton. If in fact maternal health raises the chances of a twin conception, then this will reinforce our contention. If, instead, maternal health is for some undocumented reason negatively associated with twin conception, then our findings hold despite this and they under-estimate the importance of, for instance, behaviours or stress exposures during pregnancy on the chances of producing live twin births.[11]

Trivers and Willard (1973) made an argument similar to ours but pertaining to the distribution of sons across women (Trivers and Willard, 1973; Almond and Edlund, 2007). They observed

[11] In biomedical research it is primarily monozygotic (MZ) twins that are thought to be strictly randomly allocated across families. It is recognized that there may be genetic predispositions toward dizygotic twins, and that the risk of giving birth to dizygotic twins (DZ) is elevated among women with high levels of the follicle stimulating hormone (FSH), which is often more prevalent among taller and heavier women (Li et al. 2003, Hall 2003). Since dizygotic twins constitute about two-thirds of all twins, this could in principle contribute to explaining the associations we document with height and BMI but there is no previous evidence that twinning is *also* associated with behavioural and stress markers (and these associations are unlikely to derive from FSH levels). Moreover, the biomedical literature has not documented these associations in any population level data, let alone across countries, time and indicators. Our results also constitute a challenge to the assumption that monozygotic twinning is random (conditional on age and parity). Although we are unable to distinguish MZ from DZ twins in our data, our point is that even if twin conception is random, it must be the case that, on average, women in "good condition" are more able to carry twins to term.

that since the male foetus is more vulnerable to adverse health conditions, sons are more likely to be born of healthy mothers. As for twins, so for sons, selective miscarriage is the suggested mechanism. Intersecting our hypothesis with theirs, we investigated whether male twins were more likely to miscarry than female twins other things equal. We used the large data sets in table 1 (US, Sweden and the developing country data). We find that twin births are more likely to be female ($p<0.001$). This affords a further test of our hypothesis and a validation of the Trivers-Willard hypothesis.[12]

4.2 Selective maternal survival

A potential concern is that if the less-healthy women among those who delivered twins died in childbirth, the data may not contain those women, in which case we would have a biased representation in which twin births in the data are selectively associated with more healthy women. In fact this concern does not apply to the administrative US and Sweden data where *all* births are recorded and where we see clear associations of twinning and maternal health so it cannot be the only explanation of our findings. Similarly, in the UK and Chile data sets, the survey design ensures that representative coverage is not affected by maternal death.[13] However this issue does arise in use of the DHS data which ask mothers to report fertility histories so if the mother is not alive, the children are not in the sample even if the children are alive. Also, concerns about selection on account of maternal mortality are most pertinent in the DHS developing country sample given that maternal mortality is considerably higher in poorer countries, the lifetime risk of a maternal death being 1 in 41 in low income countries as compared with 1 in 3300 in high income countries.

To assess the magnitude of this bias in the DHS estimates, we follow Alderman et al. (2011) and simulate estimates under the assumption that less-healthy women who died in childbirth were all carrying twins, and more healthy women who died in childbirth were not carrying twins. This is of course an extreme assumption that puts our results to the harshest test. We simulate the presence of the women who died and test whether this (over-)correction for maternal survival selection causes the association of twin births and maternal health to disappear.

[12] We found an older biological literature which recognizes that males are under-represented among twins, and even more under-represented among triplets (James, 1975; Bulmer, 1970), but this literature does not explicitly link in with Trivers-Willard.

[13] In ALSPAC data from the UK, women were prospectively enrolled when pregnant entirely before exposure to considerable maternal mortality risk, and children were subsequently followed over their lives. In the ELPI data from Chile, a representative sample was chosen *after* birth, however the sampling unit was at the level of the child, rather than the mother, so children would be represented even in the case that their mother was no longer living.

A data challenge here is that we do not observe the health of women who died in childbirth, indeed, the original problem is that we do not observe these women at all. However the DHS data record, for all surviving women, not only their height, BMI and pregnancy outcomes but also, for each of their sisters, whether or not their sisters died and whether the cause was related to childbirth. We thus have the maternal mortality status of all sisters of every female respondent. Most DHS countries are in Africa and, given high fertility, there are often many sisters and as respondents are 15-49 at the date of survey, a fair fraction of sisters will have experienced child birth and been exposed to the high risks of maternal mortality that characterise Africa. We proceed by assuming that the respondent's health (indicated by her height and BMI) is a reasonable proxy for the health of her sisters who she reports died. This assumption is validated in Figure 3 which shows that maternal mortality is much higher among sisters of women with lower stature or BMI, conditional upon country and year fixed effects, a quadratic in mother's age and age at first birth. In particular, sisters of women shorter than the mean height of 155.5cm are considerably more likely to have suffered maternal death, and this is particularly so for women shorter than 145cm.

To test the potential importance of selective maternal survival in explaining twin selection, we inflate the sample by the number of women who, according to our sister method calculations, 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 the estimates of equation (1) under the extreme assumption described above—that all less-healthy women who died were pregnant with twins, while all healthy women who died were not. We create a range of different binary distinctions of 'healthy vs less-healthy', using the available individual data on height and BMI. These results are presented in table 5. The first column repeats the baseline estimates, with no sample bias adjustment, but using the smaller sample of DHS countries for which the sister method to estimate maternal mortality is feasible. In this sample, a 1 point increase in BMI is associated with a 0.046% increase in the probability of twinning. The remaining columns attempt to adjust for sample selection as described. As expected, the adjustment reduces the importance of positive maternal health in predicting twinning, with the coefficient on BMI falling from 0.0460 to 0.0437 in column 2. In subsequent columns we conduct a similar exercise, but use successively less conservative assumptions to define "less-healthy" women. Even in the final column, where the entire bottom half of the anthropometric distribution is assumed less-healthy, the coefficient on both height and BMI remains positive and significant in the simulated sample.[14] Overall, these results establish

[14]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 A10, 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 5, providing further evidence that selective

that maternal mortality selection does not drive the DHS results. As discussed, this sort of selection is ruled out by construction in the four data sets we have for richer countries.

Part II – The Quantity–Quality Trade-off

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 quantity is increasing in child quality and *vice versa*. By comparing families that unexpectedly produced an additional child with families that produced only singleton births, it is argued that the quantity–quality trade-off can be identified.

However, consistent estimation of the trade-off rests on the untestable premise 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 or stress exposures during pregnancy or on her endowments prior to pregnancy. As comprehensively documented in Part 1 of this paper, this is at odds with the evidence.

In this section of the paper, we first present a brief review of the literature on the QQ trade-off, underlining that the recent consensus seems to be that a trade-off does not exist, at least for the commonly used measures of child quality. We then present the essentials of the 2SLS approach in which twins are used to instrument fertility so that we can formalize the bias that we assert in the Introduction. We then demonstrate the nature of the bias and the extent to which it is rectified using available measures of maternal health from two data sets, one for the United States (the NHIS) and one that pools developing countries (the DHS). We then explain how we estimate bounds. Importantly, we estimate the extent of the failure of the exclusion restriction using two natural experiments, one set in the United States, and one in Nigeria. The estimated bounds are, in general, negative, consistent with a QQ trade-off.

maternal survival is not enough to explain the correlation between maternal health and twinning.

The Quantity–Quality Trade-off

A long-standing theoretical result in the literature on human capital formation and, in particular, family responses to child-bearing is the existence of a quality-quantity trade-off (Becker, 1960; Becker and Lewis, 1973; Willis, 1973; De Tray, 1973; Becker and Tomes, 1976). This is consistent with casual evidence. An empirical regularity that has been noted in cross-sectional and time series data is that children from large families have weaker educational outcomes (Hanushek, 1992; Blake, 1989; Galor, 2012). Using the data we shall analyse further in this section from the USA and developing countries, we replicate this pattern of growth in human capital accumulation being concurrent with fertility decline in online appendix figures A5 and A6.

However, the empirical literature seeking to identify the quantity–quality trade-off within households is ambiguous. Early work including Hanushek (1992) and Rosenzweig and Wolpin (1980a) documented significant negative effects of additional births within a family on average child educational outcomes. More recent work contains mixed results. Using IV or difference-in-differences approaches, recent results include a significantly positive relationship (Qian, 2009), a significant negative relationship (Grawe, 2008; Ponczek and Souza, 2012a; Lee, 2008; Becker et al., 2010; Bougma et al., 2015) and no significant relationship (Black et al., 2005; Angrist et al., 2010; Fitzsimons and Malde, 2010; Åslund and Grönqvist, 2010).[15] More recently, it has been argued that where the usual twin-IV approach identifies no significant relationship, allowing for non-linear and non-monotonic effects of family fertility on children's education leads to emergence of a negative relationship (Brinch et al., Forthcoming; Mogstad and Wiswall, 2016).

The twin instrument has been widely used in other contexts too. It has been used to estimate the effects of childbearing on women's labour force participation (Rosenzweig and Wolpin, 1980b; Jacobsen et al., 1999; Angrist and Evans, 1998), and the consequences of out of wedlock births on marriage market outcomes, poverty and welfare receipt (Bronars and Grogger, 1994). Within-twin comparisons have also been widely used in the economics, medical, biology and psychology literature. In this paper we focus upon the use of twin births as an instrument for total fertility, although we observed in Part I of the paper that our point about twins being born to positively selected mothers raises a challenge to the external validity of studies relying upon twin differences.

Early studies in the IV literature recognize that twins are only as good as random conditional on maternal age and parity (Rosenzweig and Wolpin, 1980a). Some studies restrict the sample to births that occur before the introduction of fertility treatments (Cáceres-Delpiano, 2006; Angrist

[15] A fuller description of these empirical results, their magnitude and their context is provided in Clarke (2016).

et al., 2010), or they drop families undergoing fertility treatment (Braakmann and Wildman, 2014). The more recent wave of twin studies refurbish the controls to include the mother's race and educational attainment. In some cases the validity of the conditional randomness assumption is directly probed by regressing the occurrence of twin birth on observable family outcomes, or testing for the equality of means of characteristics such as mother's education between twin and non-twin families (see Black et al. (2005), Li et al. (2008), Rosenzweig and Zhang (2009)).[16] However, as is well known and acknowledged in each case, any such tests are at best partial evidence in support of instrumental validity. Importantly, no previous study has attempted to control for maternal health conditions or behaviours.

As discussed in Part 1, the twin instrument has previously been criticised for other reasons. For instance, a recent critique has focused upon parental behaviours responding to twins rather than, as here, on the likelihood that parental behaviours affect the likelihood of twinning. In particular, Rosenzweig and Zhang (2009) highlight that twins have lower birth endowments (also see Almond et al. (2005)). This is also evident in our data (see appendix figures A7 and A8). They argue that if parents reinforce endowments then this behaviour may obscure an underlying QQ trade-off (this is examined in Angrist et al. (2010) and Fitzsimons and Malde (2014)). We remain agnostic on this. Our critique is in principle orthogonal to this critique, providing a different reason that an underlying QQ trade-off may be obscured. Since our critique pertains to omitted variables indicating maternal (and hence foetal) health, we test our critique by working with this particular omitted variables bias as discussed next.

1 Methodology

1.1 The Quantity–Quality Trade-off: OLS Estimates

Empirical analyses of the quantity–quality trade-off attempt to produce consistent estimates of α_1 in the following population equation:

$$quality_{ij} = \alpha_0 + \alpha_1 quantity_j + \boldsymbol{X}\boldsymbol{\alpha_x} + \varepsilon_{ij}. \tag{4}$$

Here, *quality* is a measure of human capital attainment of, or investment in, child i in family j, and *quantity* is fertility or the number of siblings of child i. A significant QQ trade-off implies that $\alpha_1 < 0$. Relevant family and child level controls are included, denoted $\boldsymbol{X}$. As has been

[16] Appendix table A12 provides an inventory of some highly cited studies of the QQ trade-off in which twins are used to instrument fertility, including information on the controls included in each case.

extensively discussed in a previous literature, estimation of α_1 using OLS will result in biased coefficients given that child quality and quantity are jointly determined (Becker and Lewis, 1973; Becker and Tomes, 1976), and unobservable parental behaviours and attributes influence both fertility decisions, and investments in children's education (Qian, 2009).

The direction of the OLS bias is determined by the sign on the conditional correlation between $quantity_j$ and the unobserved error term: $E[quantity_j \cdot \varepsilon_{ij}|\boldsymbol{X}]$.[17] If mothers with weaker preferences for child quality have more children, OLS estimates will overstate the true QQ trade-off, and the converse will hold for positive selection into fertility. For the case of negative selection, which is what most previous studies suggest:

$$E[quantity_j \cdot \varepsilon_{ij}|\boldsymbol{X}] < 0 \quad \Leftrightarrow \quad E[\hat{\alpha}_1] < \alpha_1.$$

If all variables which predict selection into fertility could be observed and controlled for, OLS estimates would converge to the true value of α. Define $\varepsilon = \boldsymbol{H} + \boldsymbol{S} + \varepsilon^*$, where we partition the stochastic error term into a vector of observable measures of mother's health capital ($\boldsymbol{H}$), socioeconomic variables ($\boldsymbol{S}$), and all other unobserved components (ε^*). Assuming no covariance between the three components of the error term,[18] the step-by-step removal of selection variables will result in the estimated coefficient becoming continually closer to the true parameter. For negative selection on fertility this implies:

$$E[\hat{\alpha}_1] < E[\hat{\alpha}_1^H] < E[\hat{\alpha}_1^{S+H}] < \alpha_1.$$

In the above, $\hat{\alpha}_1$ refers to the estimated coefficient on *quantity* in (4) when running a naive regression without controls. The coefficients $\hat{\alpha}_1^H$ and $\hat{\alpha}_1^{S+H}$ refer to (respectively) coefficients when the model is augmented to control for observable health capital $\boldsymbol{H}$, and then also observable socioeconomic status $\boldsymbol{S}$.

Hypothesis 1: If there is negative selection into fertility then OLS will over-estimate the mag-

[17] In the simplest case of a single included (endogenous) variable with both $quality_{ij}$ and $quantity_j$ reformulated as deviations from their means, the omitted bias formula gives: $E[\hat{\alpha}_1] = \alpha_1 + \sum_{i=1}^{N} quantity_i \cdot \varepsilon_{ij} / \sum_{i=1}^{N} quantity_j \cdot quantity_j$ and given that the denominator of the second term is strictly positive, the bias $E[\hat{\alpha}_1] - \alpha_1$ is directly proportional and of the same sign as $\sum_{i=1}^{N} quantity_i \cdot \varepsilon_{ij}$.

[18] This assumption can be loosened with little implication for the analysis which follows. If we do not assume additive separability, covariance terms between each error component must be included when considering movements in the estimate $\hat{\alpha}_1$. 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 other unobserved variables which positively affect child quality are also likely to be positive, the omission of the covariance terms does not effect the inequalities discussed below. This is something which we test empirically later in this paper. Recent work by Gelbach (2016) examines more formally how estimated covariates depend on the step-wise inclusion of additional variables in a regression.

nitude of the true trade-off. Addition of variables predicting selection as controls will lead to the OLS estimate approaching the population value from below.[19]

1.2 The Twin Instrument for Fertility

Following the seminal work of (Rosenzweig and Wolpin, 1980a), fertility has been instrumented with the incidence of twin births on the premise that they constitute an exogenous shock to family size.[20] The 2SLS specification can be written as:

$$quantity_j = \pi_0 + \pi_1 twin_j + \boldsymbol{X}\boldsymbol{\pi}_{\boldsymbol{x}} + \nu_{ij}, \tag{5a}$$

$$quality_{ij} = \beta_0 + \beta_1 \widehat{quantity}_j + \boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{x}} + \eta_{ij}. \tag{5b}$$

where $twin_j$ is an indicator for whether the nth birth in family j is a twin birth. As described further in section 2.2, a series of samples are constructed, referred to as the $n+$ groups, and consisting of children born before birth n in families with at least n births. The idea is that children born prior to birth n (the subjects) are randomly assigned either one (control group) or two (treatment group) siblings at the nth birth, and this allows us to estimate causal impacts of the additional birth on investments in or outcomes of these children. The twins themselves are excluded from the estimation sample.[21] If twins are a valid instrument, the parameter β_1 is consistent and hence equal to the parameter α_1 from the population equation 4.

In Part I of this paper we provide compelling evidence that challenges the validity of the twin instrument and that implies that a range of omitted variables for maternal health may contaminate η_{ij}. If mothers who invest more in their pregnancies (for instance by averting smoking)

[19] A natural corollary of this is that if fertility were positively selected (Fort et al., 2016; Myrskylä et al., 2009), then OLS would under-estimate the magnitude of the true trade-off and adding variables predicting fertility selection as controls would lead to the estimate approaching its true value from above. However, unobservable heterogeneity will remain and the size of the residual bias cannot be determined without imposing some additional assumption about the degree of selection on unobservables. Later we will see that our estimates indicate negative selection into fertility.

[20] Other approaches include instrumenting with the gender mix of the first two children (Angrist and Evans, 1998; Conley and Glauber, 2006), or exploiting policy experiments such as the introduction and subsequent relaxation of the one child policy in China (Qian, 2009; Argys and Averett, 2015), the eradication of hookworm (Bleakley and Lange, 2009) and the introduction of antibiotics in the US (Bhalotra et al., 2016).

[21] This takes care of the concern that since twins tend to be born with weaker endowments (e.g. birth weight), they will tend to have systematically different quality outcomes. 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 gestation period than singletons. In our data samples similar endowment differences are observed. For example, appendix figure A7 documents the much larger average reported birth size of twins versus singletons in DHS data. Birth weight figures show similar patterns from United States administrative data (appendix figure A8).

also invest more in their children after birth, then the twin-IV estimates will be inconsistent. We demonstrated positive selection of mothers of twins in Part I using measures of the mother's health and education. This implies:

$$\text{plim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} twin_j \cdot \eta_j > 0 \Leftrightarrow \text{plim}_{N\to\infty} \hat{\beta}_1 > \beta_1.$$

As in the preceding OLS discussion, we can partition $\eta_{ij} = \boldsymbol{H} + \boldsymbol{S} + \eta^*_{ij}$, and additional controls can be added in a step-wise manner to the twin-IV regressions. Assuming again, for simplicity, no correlation between the error terms, removing sources of bias from η_{ij} in sequence will result in the following series of estimated coefficients:

$$\text{plim}\,\hat{\beta}_1 > \text{plim}\,\hat{\beta}_1^S > \text{plim}\,\hat{\beta}_1^{S+H} > \beta_1.$$

The estimated parameter will approach the population parameter from above. Since, as discussed in Part 1, all determinants of twin birth are virtually impossible to account for, twin-IV will under-estimate the true magnitude of the QQ trade-off, leading to hypothesis 2.

Hypothesis 2: If women who produce live twin births are positively selected, the twin-IV estimator will under-estimate the true QQ trade-off, although addition of predictors of twins as controls will lead to the estimate approaching the true value from above.

1.3 Bounds Derived by Bias Adjustment

If women with high fertility are negatively selected and women with twin births are positively selected then the true parameter will lie between the OLS and twin-IV estimates, and bias adjustment will tighten these bounds. In particular:

$$E[\hat{\alpha}_1] < E[\hat{\alpha}_1^H] < E[\hat{\alpha}_1^{S+H}] < \alpha_1 = \beta_1 < \text{plim}\,\hat{\beta}_1^{S+H} < \text{plim}\,\hat{\beta}_1^H < \text{plim}\,\hat{\beta}_1.$$

Hypothesis 3: If there is negative selection of women into (higher) fertility and positive selection of women into twin births, then the OLS and twin-IV estimators will provide bounds on the true parameter and we can tighten these bounds by adjusting for the influence of variables which predict selection.[22]

[22] If instead women exhibiting high fertility are positively selected then both sets of estimates will

An alternative and frequently employed method of bias adjustment in OLS estimates with selection is provided by Altonji et al. (2005) with extensions in Oster (2013). In this case, rather than simply inferring the sign of bias from coefficient movements, the likelihood that $\alpha_1 = 0$ is inferred from comparing parameter estimates with and without controlling for observables. Following Altonji et al. (2005), if selection on observables is informative for selection on unobservables, then observing only a small difference in estimated coefficients between the baseline estimate ($\hat{\alpha}_1$) and the estimate obtained after conditioning upon all relevant observables ($\hat{\alpha}_1^{S+H}$) provides evidence that any unobservable predictors will have small impacts on the estimated coefficient. Using this insight, we can determine how much selection on unobservables is required compared to selection on observables to explain away the QQ trade-off. More recently, Oster (2013) suggests that the explanatory power of unobservables and observables should be considered together and suggests that as well as making an assumption regarding the relative importance of unobservables and observables in the process of interest, a maximum R-squared, R_{max} should be defined, which is the R-squared of a (theoretical) model where the variable of interest is regressed on all observable and unobservable variables.

Following Altonji et al. (2005) and Oster (2013) we are thus able to use our OLS estimates $\hat{\alpha}_1$ and $\hat{\alpha}_1^{S+H}$ to ask how important unobservables must be with respect to observables for the true QQ trade-off to be zero. This leads to a single estimate of the ratio of unobservables to observables rather than an estimate of the parameter estimate itself. However, in inverting the logic, we can also construct bounds on the estimate of α_1. As Oster (2013) documents, when defining the relative importance of observables and observables as well as the R_{max} outlined above, we arrive at a set estimator for α_1. We follow Oster and Alonji et al. in defining equal selection, and follow Oster in defining a maximum R-squared of 1.3 times the R-squared from a model estimated only with the full set of observable controls.[23] Under these assumptions, we

approach the population parameter from above, and so no lower bound estimate will exist:

$$\begin{aligned} \operatorname{plim} \hat{\beta}_1 > \operatorname{plim} \hat{\beta}_1^S > \operatorname{plim} \hat{\beta}_1^{S+H} > \beta_1 \\ E[\hat{\alpha}_1] > E[\hat{\alpha}_1^S] > E[\hat{\alpha}_1^{S+H}] > \alpha_1. \end{aligned}$$

[23] This criterion is defined based on analysis of randomised studies in which unbiased treatment effects are known by design. As stated in Oster (2013), pp 5–6:

> There is considerable variation across papers in the robustness of stability claims, but this does not suggest an appropriate general value for the bound on R_{max}. There are many possible ways to calculate such a bound. In Section 5, I suggest one: that randomized results might provide a bounding value. I use a sample of randomized papers, also from top journals, to derive cutoff values of Π which would allow at least 90% of randomized results to survive: this value is $\Pi = 1.3$. To the extent that this is an attractive methodology for generating bounds on R_{max} it suggests that researchers might calculate a bias-adjusted treatment effect bound using a value of $R_{max} = 1.3\tilde{R}$. In the sample of non-randomized results considered, about 45% would survive this standard.

can derive bounds on the OLS estimates.

1.4 Estimating IV Bounds under Plausible Exogeneity

In this section we describe an alternative approach to inference for IV models developed by Conley et al. (2012) for cases when the instrument is plausible but fails the exclusion restriction.[24] They provide an operational definition of plausibly (or approximately) exogenous instruments, defining a parameter γ that reflects how close the exclusion restriction is to being satisfied in the following model (adapted to the QQ model for this paper):

$$quality_{ij} = \delta_0 + \delta_1 quantity_j + \gamma twin_j + \boldsymbol{X}\boldsymbol{\delta_x} + \vartheta_{ij}. \tag{6}$$

Since the parameters δ_1 and γ are not jointly identified, prior information or assumptions about γ are used to obtain estimates of the parameter of interest, δ_1. The IV exclusion restriction is equivalent to imposing ex-ante that γ is precisely equal to zero. As Conley et al. (2012) lay out, rather than assuming strict exogeneity, one can define plausible exogeneity as a situation in which γ is likely near, but not precisely equal to zero. By estimating or imposing some (weaker) restriction on γ, this buys the identifying information to bound the parameter of interest, even when the IV exclusion restriction does not hold exactly. Conley et al.'s methods are ideally suited to the empirical application of this paper. They show that their bounds are most informative when the instruments are strong, and the twin instrument is strong (evidence below).

In Part I of this paper, we provide evidence that leads us to suspect that γ will not equal zero. 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.[25]

Conley et al. (2012) show that bounds for the IV parameter β_1 from equation 5b can be generated under a series of assumptions regarding γ. These include a simple assumption regarding the support of γ (their "Union of Confidence Intervals" approach), or a fully specified prior for the distribution of γ (their "Local to Zero approach"). In the latter case, a correctly specified prior often leads to tighter bounds. This suggests the following:

[24] Nevo and Rosen (2012) propose a bounds estimate for IV of a broadly 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 $\rho_{z_j u}$, which is analogous to the sign of γ in Conley et al.'s framework which we apply here.

[25] If one or other of these conditional correlations is equal to zero, IV estimates will not be inconsistent. We have shown that twin mothers are healthier but for this to challenge the exclusion restriction, we would also have to have that maternal health has a direct impact on child quality. Below we discuss how this can be estimated.

Hypothesis 4: If we have an unbiased estimate of γ, a measure of the extent of the invalidity of the exclusion restriction, we can estimate bounds on the QQ trade-off.

From equation 6, γ represents the conditional effect of being born of a twin mother on child quality:

$$\gamma = \left.\frac{\partial quality_{ij}}{\partial twin_j}\right|_X$$

In practice, bounds identification based on γ only pushes the identification problem back by one step, as consistent bounds rely on having an unbiased estimate of γ, which is not trivial.

So as to obtain a consistent estimate of γ, albeit from different samples, we exploit quasi-experimental changes in maternal health ($health_j$) and use these to obtain consistent estimates of the impact of maternal health on (a) child quality and (b) the probability of a twin birth. We then 'scale' the first by the second. First, we estimate

$$\left.\frac{\partial quality_{ij}}{\partial health_j}\right|_X = \phi_q.$$

Under the assumption that the change in health is quasi-experimental, this is a causal estimate of a 1 unit change in $health_j$ on child quality. Since γ is the effect of maternal health scaled by the difference in health between twin and non-twin mothers we also estimate :

$$\left.\frac{\partial health_j}{\partial twin_j}\right|_X = \phi_t.$$

With these two parameters in hand, we obtain a causal estimate of γ as:

$$\gamma = \left.\frac{\partial quality_{ij}}{\partial twin_j}\right|_X = \left.\frac{\partial quality_{ij}}{\partial health_j}\right|_X \times \left.\frac{\partial health_j}{\partial twin_j}\right|_X = \phi_q \times \phi_t. \tag{7}$$

As it involves the estimated quantities $\widehat{\phi}_q$ and $\widehat{\phi}_q$, γ will be subject to sampling uncertainty: $\widehat{\gamma} = \widehat{\phi}_q \times \widehat{\phi}_t$. Thus, the estimate $\widehat{\gamma}$ will have a distribution. If we can estimates both $\widehat{\gamma}$ and its distribution, this gives us the consistent prior for the full distribution of γ required in Conley et al.'s LTZ approach. We estimate the distribution using resampling (bootstrap) methods, using which we can compare the analytical distribution with a series of known distributions[26], or indeed use the analytical distribution of $\widehat{\gamma}$ directly in the bounds estimate of β_1.[27] We provide a full

[26] If, for example, we determine that γ is normally distributed, estimation then proceeds by imposing the prior distribution for γ as:

$$\gamma \sim \mathcal{N}(\widehat{\mu}_\gamma, \widehat{\sigma}^2_\gamma). \tag{8}$$

[27] Conley et al. (2012) discuss a simulation-based algorithm (p. 265) for estimation, which can be used given any prior, including non-normal priors, for the distribution of γ. In practice, our preferred

description of this resampling process in online appendix D.

Implementing this approach imposes fairly strong data requirements. We require data that capture differential exposure of women to a quasi-experimental change in their pre-pregnancy health, together with measures of the quality of their children. In addition, we need information on the prevalence of twin births in this sample of women. In the follow subsections, we describe two studies, one set in the United States, and the other in Nigeria, which offer a large and representative sample of women with birth data and intergenerational linkage (i.e. the capacity to identify their births, and some outcomes for their births), and in which we observe the incidence of a quasi-experimental shock to maternal health. In the United States the shock is the introduction of antibiotics in 1937 and in Nigeria it is the Biafra war that raged through 1967-1970. We show how we exploit these cases to estimate γ and its distribution.

1.4.1 Estimating γ: A case from the United States

The first antibiotics, sulfonamide drugs, were introduced across the United States in 1937, following clinical trials in London and New York and there was nothing else on the stage until penicillin was introduced during the Second War. There was immediate and widespread uptake and the drugs were hailed as a "miracle" (Lesch, 2006). Their arrival was associated with a sharp drop in a range of infectious diseases that were treatable by these drugs (Jayachandran et al., 2010). In particular, pneumonia, the leading cause of death among children after congenital causes, fell sharply and this decline was largest among infants (Bhalotra and Venkataramani, 2014). Although there are no direct measures of the adult health of individuals exposed to the antibiotics at birth, it is plausible that infant health improvements persist and generate improvements in adult health; some evidence of this is in (Almond et al., 2011; Butikofer and Salvanes, 2015; Hjort et al., 2016; Bhalotra et al., 2015). What is pertinent for our purposes is whether any improvements in the adult health of women are such as to influence the quality of their children. We therefore estimate this reduced form using the identification strategy of Bhalotra and Venkataramani (2014) but with outcomes of the children of exposed women rather than the outcomes of the women themselves as dependent variables. Identification exploits the timing of this shock to health at birth together with the fact that the largest drops in pneumonia occurred in states with the highest initial burdens of disease. This assumes that states with high vs low burdens of penumonia did not have different trends in the outcomes before the introduction of antibiotics. To demonstrate that this is the case we estimated an event study (see appendix

estimates are based on the entire empirical distribution.

figure A9).[28]

Let m signify the mother, and $m+1$ signify her children. Using the United States micro-census files, we estimate:

$$quality_{stc}^{m+1} = \alpha + \phi_1^q(Post_t \times basePneumonia_s^m) + \theta_{rs} + \eta_{rt} + \varphi \mathbf{X}_{st}^m + \lambda_{rc} + (\theta_s \times t) + \varepsilon_{stc} \quad (9)$$

where ϕ_1^q is an estimate of the change in child quality associated with the mother's exposure to antibiotics in her infancy. The pre-intervention mean pneumonia mortality rate at the state level, s, is denoted $basePneumonia_s^m$ and interacted with $(Post_t)$, which indicates birth cohorts 1937 and after. We control for race-specific fixed effects for census year t, mother's birth cohort c, and mother's birth state s as well as state-specific linear time trends. The coefficient of interest is of similar size and significance conditional upon the state and time varying controls (health and education infrastructure, state income) and upon a vector of rates of mortality from control diseases (diseases not treatable with sulfanomides) interacted with the indicator post (these results are available upon request).

The second step is to estimate the association of the health shock experienced by women at their birth with the probability that they have a twin birth. This is an experimental analogue of the associations we present in Part 1 of this paper. We take the conditional average rate of baseline pneumonia in the state of residence for all women who give birth to a twin, and the similar conditional rate for non-twin mothers, using the same controls as in equation 9. In other words, we calculate

$$\phi_1^t = \overline{bP_s}_{twin_j=1}|_X - \overline{bP_s}_{twin_j=0}|_X = \left.\frac{\partial bP_s}{\partial twin_j}\right|_X.$$

In view of our findings in Part 1 our expectation is that women with lower exposure to pneumonia at birth will be more likely to have twins, and hence $\phi_1^t < 0$.

As discussed, with these two quantities in hand, we can estimate γ by taking their product:

$$\phi_1^q \times \phi_1^t = \frac{\partial quality_{ij}}{\partial bP_s} \times \left.\frac{\partial bP_s}{\partial twin_j}\right|_X = \left.\frac{\partial quality_{ij}}{\partial twin_j}\right|_X = \gamma_{US}. \quad (10)$$

We can plug this into our estimates of the bounds on β_1 using following Conley et al. (2012), as described earlier.

[28] Bhalotra and Venkataramani (2014) demonstrate parallel trends for first generation outcomes; we demonstrate this for second generation outcomes.

1.4.2 Estimating γ in Nigeria

Since we shall proceed to analyse alternative estimators of the QQ trade-off in developing countries and not only in the US, we obtained an estimate of γ from Nigeria. Here, we exploit the exposure of individuals through their growing years to the Nigerian civil war. This was the first modern war in sub-Saharan Africa after independence and one of the bloodiest. It raged in Biafra, the secessionist region in the South-East of Nigeria from 6 July 1967 to 15 January 1970, killing between 1 to 3 million people and causing widespread malnutrition and devastation. The war created a virtual famine in the Southeast, where it was fought, and the effects of under-nutrition were potentially reinforced by trauma and the increased incidence of infections. Akresh et al. (2012, 2016) investigate long run effects of war exposure, exploiting the differential exposure of the Christian Igbo community resident in Biafra relative to other ethnic groups (in other states), interacted with the timing of the war. They show that women exposed to the war were shorter as adults, and more likely to be over-weight. As height and obesity are measures of health, they thus establish that the war was a shock to maternal health. We use their identification strategy to estimate impacts on children's education of the mother being exposed to the war in utero, using a continuous measure for the number of months exposed.

The estimated equation is :

$$quality_{ites}^{m+1} = \alpha + \phi_2^q war_{te}^m + \alpha_t + \theta_e + \lambda_s + \mu_e t + u_{ites} \tag{11}$$

for woman i of ethnicity e born in year t and state s. The indicator of *quality* is a z-score (standardized by age and gender) for the years of education of children in generation $m+1$ and $\widehat{\phi_2^q}$ is the reduced form effect on this of the maternal health shock created by the war. Analogous to the US case, we thus estimate $\phi_2^t = \overline{war}_{twin=1} - \overline{war}_{twin=0} = \left.\frac{\partial war}{\partial twin}\right|_X$, so that we can estimate γ, the twin-mediated effect of maternal health on child-quality as:

$$\phi_2^q \times \phi_2^t = \frac{\partial quality_{ij}}{\partial war_s} \times \left.\frac{\partial war_s}{\partial twin_j}\right|_X = \left.\frac{\partial quality_{ij}}{\partial twin_j}\right|_X = \gamma_{Nigeria}. \tag{12}$$

2 Data and Estimation Samples

2.1 Data

We shall consistently estimate OLS and twin-IV estimates employing microdata from the US and from a sample of 68 developing countries. In order to estimate the (health and SES augmented) specification 4, we require information on sibling-linked births, measures of child quality and characteristics of the mother that include indicators of her health in addition to the more commonly available age, race and education. The data we use are chosen to satisfy these requirements. These are the United States National Health Interview Surveys (NHIS), which have been fielded in an identical way from 2004-2014, and the Demographic and Health Surveys (DHS) for 68 countries, which have been applied over 20 years using a broadly similar design. Further details on the data are provided in online appendix C. In appendix tables A8 and A9 we show that the twin non-randomness argument described in part I also holds in these datasets, even when focusing on births occurring in a pre-IVF time period.

In both data sets, children are included in the sample if aged between 6 and 18 years when surveyed. While ideally we would observe completed education, to our knowledge no large datasets are available measuring child's completed education, mother's total fertility, *and* a wide range of maternal health measures taken before the birth of the child. We would have liked to use the data used in recent prominent studies of the QQ trade-off (Black et al., 2005; Angrist et al., 2010), but the Israeli data do not contain indicators of maternal condition or maternal behaviours, and the Norwegian data are not publicly accessible.

A measure of child 'quality' available in both data sets is educational attainment. Since the children are 6-18 and in the process of acquiring education, we use an age-standardized z-score. In the DHS, the reference group consists of children in the same country and birth cohort, while in the NHIS, it consists of children with the same month and year of birth. Thus coefficients are expressed in standard deviations. While in the developing country setting relative school progress is an appropriate measure of child human capital given high rates of dropout and/or over-age school entry, this is not the case in the USA. In these data, grade-retention is a relevant measure of educational progress. It is estimated that between 2 and 6% of children are held back at least one grade in primary school (Warren et al., 2014). The NHIS also provides a subjectively assessed binary indicator of child health (excellent or not), which we model as an additional indicator of child quality.[29]

[29] While we would also like to analyze a health measure in the developing country sample, anthropometrics are only available for births that occur within five (or fewer) years of the survey, and infant

Appendix table A2 provides summary statistics for the DHS and NHIS data. Fertility and maternal characteristics are described at the level of the mother, while child education, and health outcomes are described at the level of the child. Twin births make up 1.98% of all births in the DHS sample, and 2.57% in the NHIS sample (a similar share to that in the US birth certificate data described in figure A2a). As expected, twin families are larger than non-twin families. Figure 4 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 in both the DHS sample (figure 4a) and the US sample (figure 4b). This establishes the relevance (power) of the twin instrument for fertility.

2.2 Estimation Samples

Studies that instrument fertility with the occurrence of a twin birth leverage the unexpected additional child to study impacts on outcomes of siblings born before the additional child. Define families with at least two birth events as 2+ families. In this group, we shall compare families in which twins occur at the second birth event (treated group) with families in which a singleton occurs at second order (control group). The subjects, for whom we measure indicators of child quality (proxies for parental investment) are the first-born children. Following Black et al. (2005), we similarly construct a 3+ sample which consists of families with at least three birth events and then we compare outcomes for the first two births across families that have a twin birth at order three (treated) and families that have a single birth at order three (control). Many existing studies, such as Angrist et al. (2010), focus upon the 2+ and 3+ samples. Given higher fertility rates in the developing country sample that we analyse, we also include 4+ families in which twins occur at fourth order and outcomes are studied for the first three births.

Restricting the sample to families with at least n births in this way primarily ensures that we avoid selection on preferences over family size. It also addresses the potential problem that, since the likelihood of a twin birth is increasing in birth order (see figures A10 and A11)[30], increasing family size raises the chances of having a twin birth.

mortality is unsuitable as the twin-IV estimator involves analysing child quality for children born *prior to* twins who will have already been fully exposed to infant mortality risk by the time the twins were born.

[30]Such a relationship is an empirical regularity in all data examined. Rosenzweig and Wolpin (1980a) report rates which increase by parity in USA (see also appendix figure A10). In DHS data a similar pattern is observed, as presented in appendix figure A11.

3 Results: Estimates of the QQ trade-off

We initially present the routine OLS and twin-IV estimates since, under the assumptions about selection into fertility discussed in section 1.1, these provide bounds on the true parameter. In each case, we show how these estimates are modified upon addition of available controls for the mother's health. So as to ascertain that the indicators of health are not simply proxying for socio-economic status, we then introduce controls for mother's education.[31] Our expectation is that the introduction of controls will tighten the bounds, diminishing the size of the trade-off estimated by OLS and increasing the size of the trade-off estimated by IV.

The former would confirm the hypothesis of negative selection into fertility and the latter would confirm positive selection into twin birth, affording a direct test of our hypothesis that the twin-IV estimator is biased downward by virtue of twins being born to healthier mothers. Having presented the standard results, illustrated the direction of partial bias-adjustment, and established OLS-IV bounds, we move on to estimating bounds on the twin-IV estimates under the assumption that the twin instrument is plausibly exogenous if not strictly exogenous. In estimating the Conley et al. (2012) bounds we make a substantive contribution by estimating the degree of violation of the twin exclusion restriction by exploiting natural experiments in two quite different settings, in both of which there are exogenous changes to maternal health that we leverage to estimate both how this modifies the probability of a twin birth and how it influences measures of child quality.

3.1 The QQ Trade-off

3.1.1 OLS Estimates

Samples and controls OLS results for the developing country and the US samples are presented in Table 6. To avoid confounding of the estimates by fertility preferences, results are presented for three $n+$ samples in which twinning occurs at birth order n, $\forall\ n \in \{2, 3, 4\}$ and the outcomes are studied for all pre-twin births i.e. the first-born to the $n-1$ born child. We consistently control for fixed effects for age of the child, age of mother at survey date, age of mother at birth and race. In the developing country sample we also condition on country and survey year fixed effects, and we show results with birth order controls. The available controls for mother's health

[31] As discussed in Part 1, to the extent that educated women exhibit healthier behaviours (Currie and Moretti, 2003; Lleras-Muney and Lichtenberg, 2005), education may influence twin births via its impact on health-related behaviours that we do not have the data to capture directly.

are height, BMI and cluster-level health service availability in the developing country sample, and BMI and a self-reported assessment of own health on a Likert scale in the US sample. In both samples, the control for socioeconomic status is years of education of the mother (see Appendix Table A2 for summary statistics of these variables) and in the developing country sample we also control for the wealth quintile of the mother.

Estimates with and without partial adjustment: Developing countries Consider the estimates for developing countries in Panel A of table 6. The baseline specifications show that an additional child in the family is associated with a reduction of 12.3 to 15.2% of a standard deviation in a child's educational progress relative to their birth cohort. The introduction of observable controls: first for mother's health and then for mother's health and education progressively reduces the estimated trade-off to nearly half of the initial value (between 6.6 and 8.5% of a standard deviation). This confirms negative fertility selection. This suggests an upper bound on the QQ trade-off of between 6.6 and 8.5% of a standard deviation. The estimated education-fertility trade-off is decreasing in the birth order at which twins (the additional child) occur, i.e. it is largest in the 2+ sample and smallest in the 4+ sample. In contrast to the case in Black et al. (2005), the controls for birth order do not eliminate the trade-off (see appendix table A13).

Bounds on the OLS estimates: Developing countries Although the introduction of relevant observables diminishes the trade-off, consistent with a diminishing of the OLS-bias, the adjusted OLS estimates remain potentially biased on account of multitude unobservables. So as to gain some sense of the size of the residual bias, we follow Altonji et al. (2005) and estimate how important additional unobservable controls must be compared to observable controls for the true effect α_1 to be zero. We also present the Oster statistic (see table 6) and, for both, we rely upon the third column of each group that includes the richest set of observables. The Altonji et al. statistic suggests that unobservable characteristics of the mother would need to be about 1.2 times as important as observables for the OLS estimate of the QQ trade-off to be entirely driven by selection into fertility. This assumes that selection on unobservables scales with selection on observables. We can relax this assumption if we are willing to impose a maximum R-squared on the theoretical reduced form regression (4) when including all relevant observable and unobservable elements. As discussed in section 1.3 we follow Oster (2013)'s suggestion of inflating the R-squared from the specification with full observable controls by 30% to arrive to our maximum theoretical R-squared. The Oster bounds place the estimate for the QQ trade-off between 2% and 8% of a standard deviation in educational attainment.

Estimates and bounds: the USA These estimates are in Panel B of table 6 presents the OLS estimates with and without controls for maternal characteristics, along with the Altonji

and Oster statistics for the US sample. The baseline estimates imply that an additional child within the family is associated with a 2.3 to 4.4% standard deviation reduction in education progress. Again, in line with negative selection into fertility, controlling for maternal health and socioeconomic status tends to halve the size of this trade-off, to between 1.0 and 2.4% of a s.d. (in the 4+ sample, which is a relatively small sample, the coefficient of 1.0% of a s.d. is not statistically significant). A similar pattern of results emerges using child health. The unadjusted OLS estimates imply that an additional sibling reduces the health of older siblings by between 1.1 and 2.8%. As for education, there is a clear diminishing of the estimated trade-off upon adjustment for maternal characteristics. It falls by half or a third, to between 0.3 and 1.7%. The trade-off is similar for the 2+ and 3+ samples and smaller and insignificant in the 4+ sample. However, for health, this "gradient" is reversed and the largest child health–fertility trade-off is in the 4+ sample and the smallest in the 2+ sample. Note that results conditional upon birth order are in appendix table A14 and as for the DHS sample, the trade-off is robust to these controls. The Altonji et al. ratio ranges from a large value of 2, providing quite strong evidence of a significant trade-off, to values as low as 0.5, suggesting that if unobservables are only half as important as observables, the true effect would be 0. In each of the specifications for health, Oster's bounds imply the presence of a QQ trade-off, albeit small, but in two of three specifications for educational attainment, the bounds include zero, implying that for these samples we cannot sign the relationship.

3.1.2 IV Estimates with the Twin Instrument

IV estimates using the twin instrument are in tables 7 (DHS) and 8 (US), the first-stage estimates are in panel A and the second stage estimates in panel B.

IV Estimates: Developing Countries. The first stage estimates demonstrate the well-known power of the twin instrument. It consistently passes weak instrument tests (the Kleibergen-Paap rk statistic and its p-value is presented in panel A). The point estimates indicate that the incidence of twins raises total fertility by about 0.7 to 0.8 births. That this estimate is always less than one is in line with other estimates in the twin literature and is evidence of partial reduction of future fertility following twin births (compensating behaviour). Consistent with this, the first stage coefficient is increasing in parity. In panel B, the first column ("Base") for each parity group presents estimates of $\hat{\beta}_1$ from equation 5a using the current state of the art twin-IV 2SLS estimator. In each of the three samples, in line with the findings of recent studies (Angrist et al. (2010); Black et al. (2005); Cáceres-Delpiano (2006); Fitzsimons and Malde (2014); Åslund and Grönqvist (2010), we find no significant QQ trade-off. This is not simply because IV estimates are

less precise than OLS estimates (as emphasized in Angrist et al. (2010)), rather, the coefficients are much smaller.

Consistent with our hypothesis and the evidence we present in Part 1 that twin mothers are positively selected on health (and education), we see that introducing controls for maternal selectors of twinning, a QQ trade-off emerges in the 3+ and 4+ samples, even though the available controls are almost certainly a partial representation of the range of relevant facets of maternal health stocks, health-related behaviours and environmental influences on foetal health. The bias adjustment is meaningful and statistically significant. In the 3+ sample, the commonly estimated specification produces a point estimate of 2.8% which is not statistically significant, and partial bias adjustment raises this to 4% (conditional on maternal health indicators) or 4.6% (if mother's education is also included). In the 4+ sample, the corresponding figures are 2.7% and 3.7%.

As discussed in section 3.1.1, the controls for health are the mother's height, her BMI and indicators of the availability of prenatal care and of qualified nurses and doctors in her local area. In the developing country data, BMI is recorded at the survey date rather than pre-pregnancy, and the survey date marks different durations since the last pregnancy for different women. For this reason, BMI as measured is a noisy proxy for the underlying longer-term BMI of the woman. As long as any noise is similar for treated women (women who have, at some point, had twins) and women who have only ever had singletons (control women), this does not directly bear upon our hypothesis. However, one may be concerned that carrying twins stresses the mother's nutritional status, resulting in lower BMI after pregnancy. Since we demonstrated in Part 1 that women with lower BMI are less likely to have twins (and, in the 2SLS regressions in table 7, controlling for BMI raises the trade-off), it follows that any selective depletion of twin mothers captured in BMI at the survey date will lead us to under-estimate the QQ trade-off. This makes our estimates conservative.

IV Estimates: United States The first stage estimates for the USA sample (table 8) are very similar to those for the developing country sample, with a twin birth at parity 2, 3 or 4 leading to an additional 0.7 to 0.8 total births. The second stage estimates also follow a similar pattern insofar as the baseline specification indicates no significant relationship between twin-mediated increases in fertility and either the indicator of school progression, or the indicator of child health. However, upon the introduction of controls for maternal health and education, the coefficient describing the QQ trade-off tends to increase. In the case of education, it grows more negative in each sample and is statistically significant in the 2+ sample, with a point estimate of 10.2%. When child quality is indicated by health, the point estimate in the 2+ sample remains insignificant but in the 3+ and 4+ samples it grows more negative and in the 3+ sample it is statistically significant at 5.9% of a s.d. Notice that the USA samples range between about 21,000

and 61,000 individuals while the developing country data samples range between about 260,000 and 400,000, so we have more limited statistical power with the US data. It is well recognised that IV estimates are much less precise than their OLS counterparts, and in the case of the twin instrument, only a small proportion (∼2%) of the population of births are twins. Nevertheless, partial bias adjustment shows, in US data, a statistically significant QQ trade-off for education in the 2+ sample (comprising about 50% of the total sample) and for health in the 3+ sample (comprising about a third of the total sample).

3.1.3 Heterogeneity: Unadjusted and Adjusted IV Estimates

Theoretical statements of the QQ model tend to assume, for simplicity, that all children in a family have the same endowments and receive the same parental investment. More recent work (for example the theoretical work of Aizer and Cunha (2012) and empirical papers by Rosenzweig and Zhang (2009); Brinch et al. (Forthcoming); Mogstad and Wiswall (2016)) relax this assumption. Among other things, this allows for reinforcing or compensating behaviours in parental investment choices (Almond and Mazumder, 2013). This implies allowing the coefficient β_1 to vary across 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.

We find heterogeneity in the trade-off by gender. In developing countries, the trade-off is larger and only statistically significant for girl children (see appendix tables A15 and A16). This is consistent with a fairly widespread preference for sons (Filmer et al., 2009) and with lower returns to primary education for girls (although higher returns to secondary education; see Patrinos (2008)) in these countries. There is, similarly, evidence that the trade-off is stronger among females in Brazil (Ponczek and Souza, 2012b). However, in the US sample (tables A17 and A18) we find the opposite pattern. This may not indicate a preference for girls but rather, be understood in terms of boys being more vulnerable to early life health risks (Gluckman and Hanson, 2005; Low, 2000) and more prone to grade repetition Warren et al. (2014) than girls, in which case equal inputs may produce unequal outcomes across the genders.[32]

Using the multi-country developing country sample, we investigated heterogeneity by country income level, following the World Bank classification of countries (at the date of survey) into low income and middle income countries (see Appendix tables A19 and A20). The estimated trade-off is considerably larger in middle-income countries, at 7% of a standard deviation, only

[32]Comparability across our samples is limited by the fact that, in the developing country sample, health is not an outcome. The stylized fact is that, on average, boys complete more years of schooling in developing countries while, in the US, boys perform less well in terms of grade-retention.

slightly smaller than the corresponding OLS estimate. In fact, even after (partial) adjustment for selection on observables, there is no significant trade-off in the low-income sample, and the point estimates are in the region of 2-3% of a standard deviation.[33]

Our results imply that an additional birth in a family is associated with 0.17 fewer years of completed education (developing countries) or 0.22 fewer grades progressed (USA). In a widely cited study, Jensen (2010) shows that providing students with information on the returns to secondary school in their area led, on average, to their completing 0.20-0.35 more years of school over the next four years. In a similarly high-profile experiment, Baird et al. (2016) find the deworming in school led to an increase of 0.26 years of schooling and Bhalotra and Venkataramani (2013) find that a 1 s.d. decrease in under-5 diarrheal mortality (11 deaths per 1000 live births) is associated with girls growing up to achieve an additional 0.38 years of schooling, while both studies find no increase in school years for boys. Almond (2006) finds that foetal exposure to influenza in 1918 was associated with 0.126 years (1.5 months) less schooling at the cohort-level and Bhalotra and Venkataramani (2014) show that exposure to antibiotic-led reductions in pneumonia in infancy resulted in individuals completing 0.7 additional years of education in adulthood relative to unexposed cohorts. The PROGRESA cash transfer in Mexico is estimated to have generated a 0.66 increase in years of schooling (Schultz, 2004).

3.2 Bounding the QQ Trade-off

3.2.1 Generalised Bounds

As discussed, the adjusted twin-IV results will not provide consistent estimates of β_1 via 2SLS as there are almost certainly omitted indicators of maternal health. Rather than discard the twin-IV estimates altogether, we harness their power in predicting fertility using Conley et al.

[33] In this paper we focus nearly exclusively on the internal validity of twins estimates (IV consistency). In recent work, Dehejia et al. (2015) examine the external validity of the Angrist and Evans (1998) approach of using the sex composition of the first two births as an instrument for continued fertility and women's labor force participation. They show that controlling for micro and macro level covariates reduces extrapolation prediction error, but that macro covariates dominate. Our findings in this section parallel theirs insofar as they show that there is heterogeneity in impacts of fertility across contexts. While any conclusive analysis of the difference we find between middle and low income countries is beyond the scope of this paper, we propose these hypothesis for further consideration in future work. First, if returns to human capital are low, then the payoff to reducing the quantity of children in order to afford higher investments in their quality may be low. Second, the low income countries in the sample are mostly in Africa, where fostering and child labour are prevalent. If children contribute to relaxing budget constraints, the premise of the QQ trade-off needs to be re-drawn. If children are fostered, the biological mother will determine the quantity (number) of children but if some children are sent away to be fostered by other households, investments in both the movers and stayers may be different. Polygamy may similarly complicate matters to the extent that it creates multiple decision-makers.

(2012) bounds to assess the empirical significance of the omitted variables.

As outlined in section 1.4, this involves the definition of a prior belief over the sign and magnitude that the coefficient on twin birth (γ) would take in equation 6. To begin, we examine bounds on the estimate of β_1 under a range of values for γ. These range from 0 (in which case the instrument is valid and having a twin vs a singleton mother is not correlated with child quality) to 0.1, or 10% of a standard deviation, in which case the validity assumption on the instrument is violated, and having a twin mother has a positive effect on child quality. These results are displayed in figure 5 (for developing countries) and figure 6 (for the USA) for the 3+ samples; results for the 2+ and 4+ samples are in online appendix figures A12 and A13. The bounds at each point of the figures correspond to the assumption that $\gamma \sim U(0, \delta)$ with δ displayed on the x-axis. Thus, when $\delta = 0$, γ is exactly 0, and the bounds collapse to the 95% confidence interval for the traditional IV estimate. However, as δ increases, the exclusion restriction on the IV moves increasingly away from zero. We observe, firstly, a widening of the estimated bounds as the size of the exogeneity error increases[34], and secondly that the upper bound becomes increasingly negative, moving in the direction of finding a QQ trade-off.[35] In both figures the vertical red line displays our preferred estimate for γ, the estimation of which we discuss further below. For developing countries and for the US (when the outcome is a measure of child health, but not for education, where the estimates are considerably less precise) we observe baseline IV results with bounds that are not informative of the sign of the trade-off when the exclusion restriction is assumed to hold exactly. However, as γ grows, the bounds do quickly become informative, suggesting that with a γ as low as 0.002 in the US or 0.008 in developing countries, a significant QQ trade-off emerges. While using an interval of values for γ has the advantage of being unrestrictive (0.1 is a very large value), the bounds are quite wide. As discussed in section 1.4, we next attempt to estimate γ with a view to improving the precision (and empirical relevance) of the IV-bounds.

3.2.2 Estimating the Violation of the Exclusion Restriction

We estimate γ and its distribution using the two maternal health shocks described in section 1.4.1 (for the US) and section 1.4.2 (for the developing country sample). Estimates for each of the components of γ are presented in table 9.

[34] As Conley et al. (2012) discuss, the degree of failure of the exclusion restriction is analogous to sampling uncertainty related to the IV parameter β_1. As the exclusion restriction is increasingly relaxed, the "exogeneity error" related to the instrument inflates the traditional variance-covariance matrix.

[35] This is in line with the twin-IV estimates becoming more negative upon including controls that mitigate the omitted variable bias which leads to violation of the exclusion restriction.

The United States. In panel A, we use quasi-experimental variation in the exposure of women to antibiotics in their birth year in early twentieth century America to estimate impacts of mother's health on children's education, cast as a Z-score, with the standardization using the birth cohort distribution. Following equation 9 (and Bhalotra and Venkataramani (2014)), we estimate that the reduced form effect of the mother's exposure is an increase in the child's completed education of 4.97% of a standard deviation, or approximately 0.15 years of education.[36] This estimate is the quantity ϕ_1^q in equations 9 and 10. In the second column, we show estimates that imply that, conditional upon health and fertility controls, mothers who produce twin births are, on average, in states with 12.5% *lower* rates of pneumonia. This augments the evidence presented in Part 1 of this paper, adding a further case of twin births being a function of health conditions. Following equation 10, in column 3 we interact $\widehat{\phi}^q$ and $\widehat{\phi}^t$ to form a consistent estimate for γ of 0.0062 (or 0.62% of a standard deviation). Bootstrapping this distribution results in an estimated variance of 0.0027. The empirical distribution estimated from 100 bootstrap replications is displayed in figure 7a, overlaid with an analytical normal distribution with the same mean and variance. When comparing our estimate of γ to IV estimates discussed in section 3.1, we see that the direct effect of having a (healthier) twin mother on child quality is considerably smaller than the (point) estimates of the effect of fertility on child outcomes. While it is reassuring that the violation of the exclusion restriction is estimated as small, in that it implies that the instrument is "close to" being exogenous (in Conley et al. (2012)'s terminology), the evidence we provide shows that it is nevertheless sufficient to generate substantively different conclusions regarding the QQ trade-off.

Nigeria. We repeat the procedure for estimating the violation of the exclusion restriction using quasi-experimental variation in the mother's foetal exposure to the Biafra war that was fought in Nigeria in 1967-1970. Results are in panel B of table 9. The first column presents an estimate of ϕ_2^q from equation 11. Children of mothers exposed to the war in utero have 1.54% of a standard deviation less education, equivalent to 0.052 years (compared to children of mothers unexposed to the war in utero).[37] As discussed earlier, the second-generation impact must be scaled by the difference in rates of exposure of women who (at some point) give birth to twins and those that only ever have singleton births. The second column shows that, on average, twin mothers come from states and cohorts that are 26.7% less likely to have suffered war. Together these estimates imply a small positive estimate of γ of 0.004 (or 0.4% of a standard deviation in education outcomes), not dissimilar to the value estimated using a shock in early twentieth century America. The bootstrapped distribution of γ based on 100 replications is displayed in

[36] The results from Bhalotra and Venkataramani (2014) suggest that exposure to sulfa drugs increased schooling of the first generation (the mothers) by 0.7 years. Our estimates suggest that the trickle down to the next generation was smaller (by more than a factor of four), but still significant.

[37] This is not directly relevant here but, again, notice that the second-generation effect is smaller than the impact on the first generation, which is 0.6 years of education (Akresh et al., 2016).

figure 7b (bootstrap variance 0.0022).

3.2.3 Bounds on the QQ Trade-off When γ is Known

With causally estimated values for γ in hand for both the developing country and US context, we are able to pin down the bounds described in figures 5-6. See table 10. In the left-hand columns we present the UCI approach in which we assume that $\gamma \in [0, 2\hat{\gamma}]$. This assumption is chosen such that the true $\hat{\gamma}$ in each case will lie precisely in the middle of the confidence interval, following Conley et al. (2012)'s empirical example. For the LTZ approach, we use the additional information available from the estimates we presented in the previous sub-section, and assume that γ is distributed precisely according to the estimated empirical distribution.[38]

In all cases, our preferred bounds estimates are those in the right-hand columns of table 10, as these are more efficient based on the estimated bootstrap distribution. For the developing country sample, estimates of the QQ trade-off in determining educational attainment, in the 3+ and 4+ samples, are bounded between slightly less than zero and 6% of a standard deviation and the mid-point of these bounds falls at 2.6% and 3.7% of a standard deviation respectively. An additional sibling thus *does* appear to depress a child's educational attainment, and this is of the order of magnitude of 3-5% of a standard deviation

In panel B, we provide bounds on the IV-estimates for the US sample. While the mid-point of the bounds is virtually always negative (health in the 2+ group is the only exception), the bounds are most informative for the 2+ (education) and 3+ (health) samples. The bounds are much tighter for child health measures, suggesting that an additional birth reduces the likelihood that a previous sibling is reported as being in excellent health by between 0.2 to 7%.[39]

Using Appendix Table A2, we can convert these standardised estimates into years of education. The effect on education of first and second-borns from having a fertility shock at the third birth, or on first to third-borns from a fertility shock at the fourth birth is estimated to be approximately 4% of a standard deviation in the developing country sample. Using the standard deviation in the sample of 3.8 years, this implies an average effect of around 0.15 years of education per additional sibling at the age of 13 years (the average age in the sample). In the case of the US estimates, for the same 2+ and 3+ groups the average estimated effect of 8% of a

[38] Results are robust to assuming the less-correct but more computationally simple: $\gamma \sim \mathcal{N}(\mu_{\hat{\gamma}}, \sigma_{\hat{\gamma}})$ and are available upon request.

[39] The US sample is much smaller than the developing country sample, so the estimates are less precise, making it more difficult to construct tight or informative bounds.

standard deviation equates to a marginal effect of 0.22 years of education by the age of 11 years. On average the likelihood of being reported as being in excellent health falls by 1.6% following an additional birth among the same group. As discussed in the Introduction, these are quite large effects relative to the educational gains documented as arising from a range of different policy interventions.

Conclusion and Discussion

Twin births are not random. Based on a considerable body of evidence compiled from vital statistics and survey data from low- and high-income countries in different time periods, we demonstrate that mothers with greater health stocks, mothers who engage in positive health-related behaviours, and mothers living in less stressful environments or in regions with better prenatal and public health services are all significantly more likely to have twins. We show that mothers who have twin births are healthier *prior* to the occurrence of the twin birth. We argue that the mechanism is selective foetal death and we substantiate this, showing that maternal health/healthy behaviours act to raise the likelihood of taking twins to term, conditional upon conception of twins.

As discussed earlier, twin birth is a marker of foetal health and our findings, which are unusually rich in the number of indicators and countries for which they obtain, may be seen as highlighting the relevance of maternal health for foetal health (proxied by foetal survival). Recent research demonstrating long run socio-economic returns to investing in foetal and infant health, improving the pre-school environment and raising parenting quality has stimulated policy interventions across the world that are motivated to enhance the potential for nurture to lift up the trajectories of children, especially when born into disadvantaged circumstances (Heckman et al., 2010; Almond and Currie, 2011; Carneiro et al., 2015). Our results point to the significance of, for instance, nutrition, stress and prenatal care for mothers in achieving these goals.

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 capital). While OLS estimates are biased on account of negative selection of women into fertility, twin-IV estimates are biased in the opposite direction by positive selection into twin birth. This is important because existing evidence from the QQ literature is mixed. In particular, recent prominent studies find that the trade-off is frequently not statistically different from zero. We show that even partially correcting for twin endogeneity is sufficient to push estimates of the trade-off up by about 3%-5% 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 be as high as 8% of a standard deviation, though it is typically centered around 3-5% of a standard deviation.

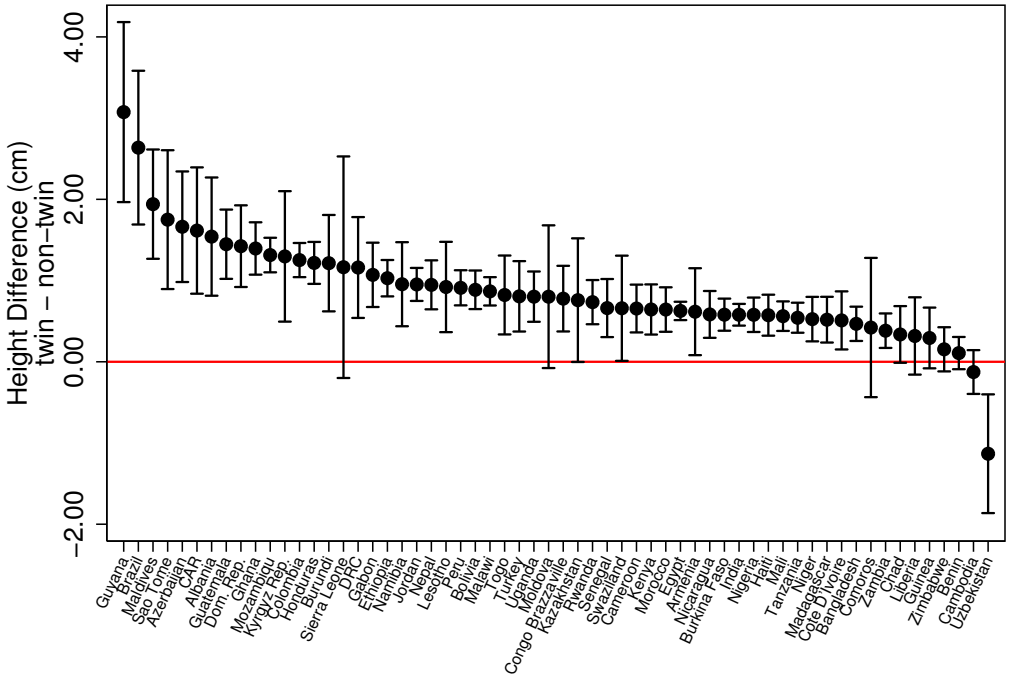
We conclude that additional unexpected births do have quantitatively important effects on their siblings' educational outcomes. A 5% of a standard deviation increase is equivalent to an additional 0.17 years in the classroom. As detailed in the Introduction, the implications of these findings are far-reaching, not only in terms of vindication of Beckerian theory but because they guide fertility control policies.[40] We constructed 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 A6a and A6b; also see *eg* Hanushek (1992)). It is of considerable relevance to researchers and to policy makers to determine whether these trends contain a causal component.

Also, by the same arguments as traced in this paper, the impact of fertility on women's labour supply is probably larger than the estimates in twin-IV studies suggest. This is topical when educational attainments of women in rich and poorer countries alike are over-taking those of men and transforming the work-family balance, with consequences for women's autonomy, marital stability and child outcomes (Rendall, 2010; Newman and Olivetti, 2016; Lundberg et al., 2016).

[40] 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).

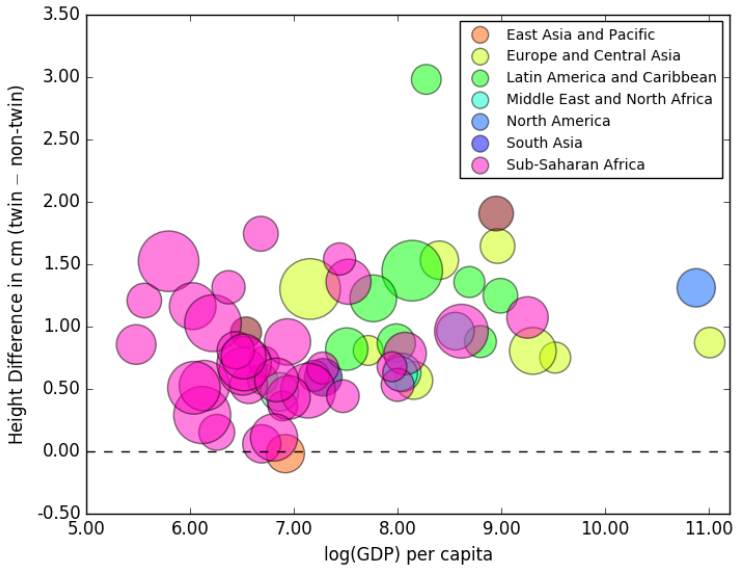
Figures

Figure 1: Height Differential By Twin and non-Twin Mothers by Country



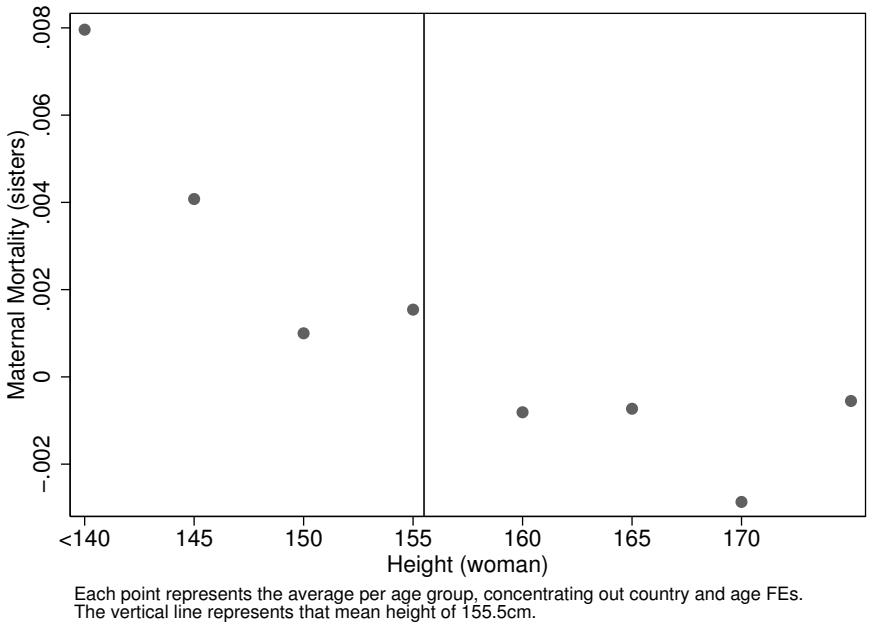
Note to figure 1: Point estimates of the average difference in height between mothers of twin and singleton births are presented along with the 95% confidence intervals for each country for which the required microdata are available. Sources of data are described in section 2. When based on survey data, each point is weighted to be nationally representative, and if based on vital statistics data, the universe of births is included. The difference-in-mean estimates are conditioned upon total fertility, mother's age and child year of birth.

Figure 2: Height Differential By Twin and non-Twin Mothers by Country and GDP



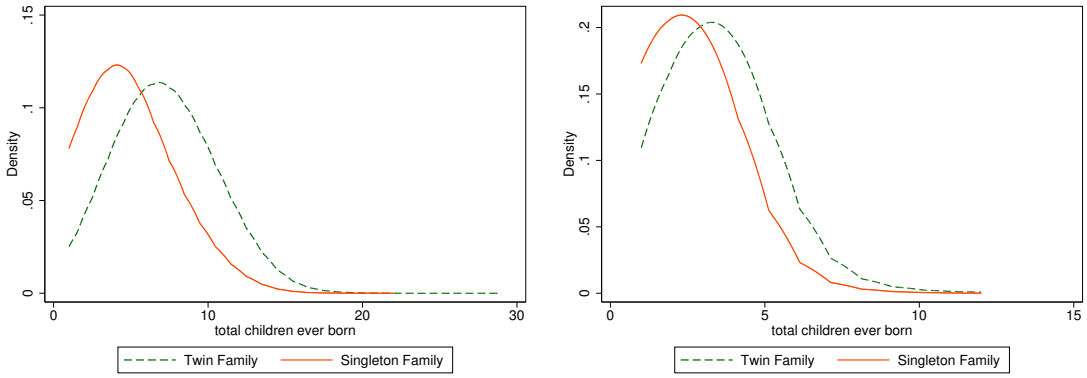
Note to figure 2: The correlation of the average height differential between twin and singleton mothers in a country with the country's log GDP percapita is plotted. Estimates for the height differential are calculated using the same controls and methodology as in figure 1. Each circle represents a country and the size of the circle indicates the proportion of births in the country that are twins. Circles above the horizontal dotted line imply that mothers of twins are taller on average. The global correlation between the height difference and GDP conditional on continent fixed effects is 0.259 (t-statistic 1.95).

Figure 3: Height and Selective Survival



Note to figure 3: Data consists of all women in the DHS from countries where the maternal mortality module was applied in surveys. Heights are based on measures for all mothers at the time of the survey, and rates of maternal mortality are calculated based on the survival status of each sister of surveyed women.

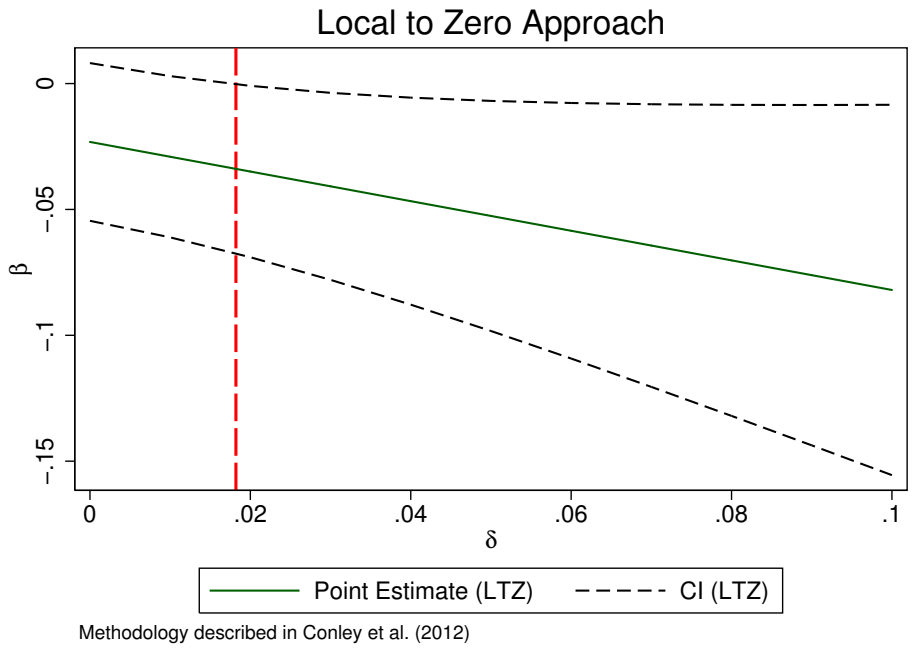
Figure 4: Twins shift the fertility distribution outward



(a) Developing Countries (b) United States

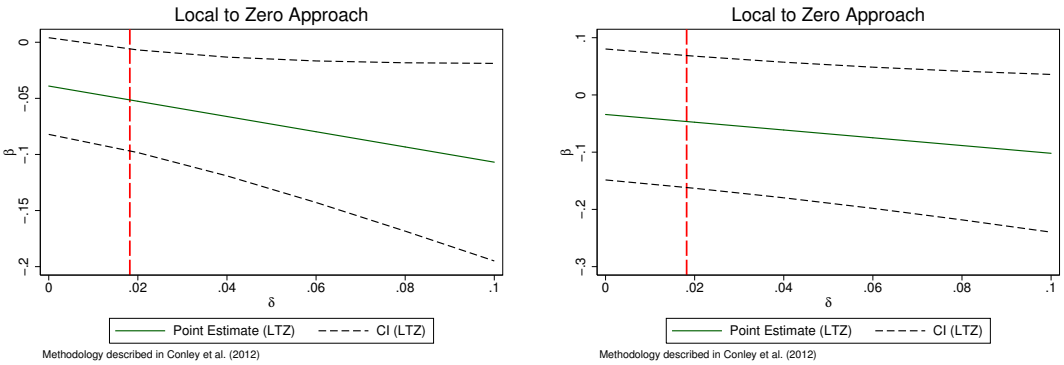
Note to figure 4: Densities of family size come from the full estimation samples from DHS and NHIS data. Kernel densities are plotted (bandwidth equals two in all cases), and present the frequency of the total number of children per family by family type.

Figure 5: Plausibly Exogenous Bounds: School Z-Score (Developing Countries 3+)



Note to figure 5: Confidence intervals and point estimates are calculated according to Conley et al. (2012) using DHS data and specifications described in section 3.2. 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 structural equation, 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 section 1.4 and table 10.

Figure 6: Plausibly Exogenous Bounds: (USA 3+)

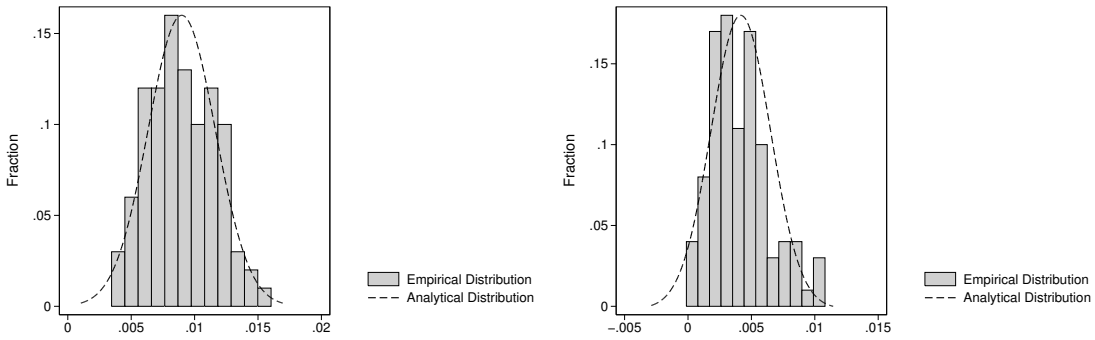


(a) Excellent Health

(b) Education Z-Score

Notes to figure 6: See notes to figure 5. An identical approach is employed, however now using USA (NHIS) data.

Figure 7: Bootstrap Estimates of $\hat{\gamma}$



(a) United States (Sulfonamide)

(b) Nigeria (Biafra Civil War)

NOTES TO FIGURE 7: The empirical distribution is generated by performing J=100 bootstrap replications to estimate ϕ^t and ϕ^q for each of Nigeria and USA (see complete discussion in section 1.4). The overlaid analytical distribution in each figure is a normal distribution $\sim \mathcal{N}(\mu_{\hat{\gamma}}, \sigma_{\hat{\gamma}})$. The estimates for ϕ^t and ϕ^q and γ are displayed in Table 9.

Tables

Table 1: Effects of maternal health on twin births

Health Behaviours / Access			Health Stocks and Conditions		
Variable	Estimate	[95% CI]	Variable	Estimate	[95% CI]
Panel A: United States [N =13,962,330, % Twin = 2.84]					
Smoked Before Pregnancy	-0.105***	[-0.132,-0.078]	Height	0.626***	[0.599,0.653]
Smoked Trimester 1	-0.205***	[-0.232,-0.178]	Underweight	-0.151***	[-0.176,-0.126]
Smoked Trimester 2	-0.247***	[-0.272,-0.222]	Obese	0.035**	[0.008,0.062]
Smoked Trimester 3	-0.250***	[-0.275,-0.225]	Diabetes	-0.288***	[-0.317,-0.259]
Education	0.805***	[0.772,0.838]	Hypertension	-0.217***	[-0.250,-0.184]
Panel B: Sweden [N =1,240,621, % Twin = 2.55]					
Smoked (12 weeks)	-0.266***	[-0.301,-0.231]	Height	0.617***	[0.592,0.642]
Smoked (30-32 weeks)	-0.285***	[-0.312,-0.258]	Underweight	-0.140***	[-0.173,-0.107]
			Obese	-0.113***	[-0.137,-0.089]
			Asthma	-0.015*	[-0.033,0.003]
			Diabetes	-0.253***	[-0.278,-0.228]
			Kidney Disease	-0.079***	[-0.101,-0.057]
			Hypertension	-0.099***	[-0.121,-0.077]
Panel C: United Kingdom (Avon) [N =10,463, % Twin = 2.37]					
Healthy Foods	0.538***	[0.256,0.820]	Height	0.399***	[0.115,0.683]
Fresh Fruit	0.019	[-0.281,0.319]	Underweight	-0.161	[-0.439,0.117]
Alcohol (Infrequently)	-0.099	[-0.373,0.175]	Obese	-0.046	[-0.322,0.230]
Alcohol (Frequently)	-0.358**	[-0.630,-0.086]	Diabetes	-0.056	[-0.328,0.216]
Passive Smoke	0.047	[-0.243,0.337]	Hypertension	-0.480***	[-0.752,-0.208]
Smoked during Pregnancy	-0.162	[-0.448,0.124]			
Education	0.416*	[-0.002,0.834]			
Panel D: Chile [N =26,527, % Twin = 2.55]					
Smoked during Pregnancy	-0.327***	[-0.572,-0.082]	Underweight	-0.183*	[-0.399,0.033]
Drugs (Infrequently)	0.002	[-0.253,0.257]	Obese	-0.258***	[-0.446,-0.070]
Drugs (Frequently)	-0.161***	[-0.196,-0.126]			
Alcohol (Infrequently)	-0.072	[-0.362,0.218]			
Alcohol (Frequently)	-0.172***	[-0.213,-0.131]			
Education	0.529***	[0.168,0.858]			
Panel E: Developing Countries (DHS) [N =2,052,338, % Twin = 2.10]					
Doctor Availability	0.092***	[0.059,0.125]	Height	0.282***	[0.251,0.313]
Nurse Availability	0.065***	[0.034,0.096]	Underweight	-0.092***	[-0.117,-0.067]
Prenatal Care Availability	0.109***	[0.082,0.136]	Obese	0.062***	[0.031,0.093]
Education	0.153***	[0.122,0.184]			

Each coefficient represents a separate regression of child's birth type (twin or singleton) on the mother's health behaviours and conditions. In each sample, all mothers aged 18-49 are included. Twins (depdendent variable) is mutliplied by 100 and the independent variables are standardised as Z-scores so coefficients are interpreted as the percentage point change in twin births associated with a 1 standard deviation increase in the variable of interest. All models include fixed effects for age and birth order, and where possible, for wealth (panels A and D) and for gestation of the birth in weeks (panels A and B). Standard errors are clustered by mother, and asterisks indicate statistical significance: *p<0.1 **p<0.05 ***p<0.01. Conditional results and unstandardised results are included as online appendix tables A4 and A5.

Table 2: Twinning and Stress *in Utero*

Dependent Variable: Twins×100	(1)	(2)	(3)
ETA Bomb casualities 1st trimester of pregnancy	0.002	-0.002	-0.002
	(0.006)	(0.006)	(0.004)
ETA Bomb casualities 2nd trimester of pregnancy	-0.010***	-0.010***	-0.010***
	(0.004)	(0.004)	(0.004)
ETA Bomb casualities 3rd trimester of pregnancy	-0.012*	-0.013*	-0.013**
	(0.007)	(0.008)	(0.006)
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 the Quintana-Domeque and Ródenas-Serrano (2014) sample 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 3: Test of hypothesis that women who bear twins have better prior health

Dependent Variable: Infant Mortality×100	(1) Base	(2) +H	(3) +S&H	Mean
Treated (2+)	-2.071***	-1.773***	-1.770***	9.758
	(0.223)	(0.235)	(0.236)	
Treated (3+)	-4.544***	-4.391***	-4.382***	10.157
	(0.211)	(0.225)	(0.226)	
Treated (4+)	-4.292***	-4.030***	-4.054***	10.827
	(0.186)	(0.192)	(0.192)	

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+, and 4+ 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 (section 2.2). 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, +H and +S&H controls are described in table 6. *p<0.1; **p<0.05; ***p<0.01

Table 4: Fetal Deaths, Twinning, and Health Behaviours

Dependent Variable: Fetal Death× 1,000	(1) Smokes	(2) Drinks	(3) No College	(4) Anemic	(5) N Cigs	(6) N Drinks	(7) Years Educ
Twin	9.907***	10.368***	8.991***	11.337***	9.939***	10.354***	19.630***
	[0.123]	[0.119]	[0.145]	[0.117]	[0.121]	[0.119]	[0.552]
Health (Dis)amenity	1.394***	4.924***	1.683***	0.608***	0.108***	0.602***	-0.242***
	[0.066]	[0.260]	[0.038]	[0.131]	[0.005]	[0.038]	[0.007]
Twin × Health	1.154***	3.559**	3.573***	1.303**	0.061*	0.756***	-0.674***
	[0.416]	[1.754]	[0.218]	[0.641]	[0.032]	[0.206]	[0.040]
Constant	5.195***	5.476***	4.268***	5.949***	5.214***	5.482***	8.277***
	[0.022]	[0.021]	[0.028]	[0.020]	[0.021]	[0.021]	[0.088]
Observations	13,660,400	13,809,830	15,909,836	16,158,564	13,679,142	13,828,573	15,909,836

Each column represents a regression of whether a birth ends in a fetal death (multiplied by 1,000) on twins, a health behaviour or health stock, and the interaction between twins and the health variable. The health variable in each column is indicated in the column title. Regressions including controls for mother's age, child birth year and total fertility fixed effects are presented in appendix table A11. Heteroscedasticity robust standard errors are displayed in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 5: Can Selective Maternal Survival Explain Twinning Rates?

Dependent Variable: Twins×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 6. 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. Heteroscedasticity robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 6: OLS Estimates: Developing Country and US

Dependent Variable:	2+			3+			4+		
Child Quality	Base	+H	+S&H	Base	+H	+S&H	Base	+H	+S&H
Panel A: Developing Country Results									
Dependent Variable = School Z-Score									
Fertility	-0.152***	-0.130***	-0.085***	-0.142***	-0.120***	-0.077***	-0.123***	-0.103***	-0.066***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	259966	259966	259966	395721	395721	395721	409615	409615	409615
R-squared	0.11	0.13	0.19	0.09	0.12	0.19	0.08	0.11	0.19
Altonji et al. Ratio		1.269			1.185			1.158	
Oster Bounds		[-0.085,-0.036]			[-0.077,-0.026]			[-0.066,-0.020]	
Panel B: US Results									
Dependent Variable = School Z-Score									
Fertility	-0.044***	-0.032***	-0.025***	-0.040***	-0.032***	-0.024***	-0.023*	-0.017	-0.010
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.013)	(0.013)	(0.013)
Observations	61267	61267	61267	47308	47308	47308	21352	21352	21352
Altonji et al. Ratio		1.316			2.019			0.764	
Oster Bounds		[-0.025,-0.005]			[-0.024,0.000]			[-0.010,0.019]	
Dependent Variable = Excellent Health									
Fertility	-0.011***	-0.005**	-0.003	-0.016***	-0.009***	-0.007***	-0.028***	-0.018***	-0.017***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)
Observations	70277	70277	70277	53393	53393	53393	24358	24358	24358
Altonji et al. Ratio		0.447			0.798			1.569	
Oster Bounds		[-0.003,-0.000]			[-0.007,-0.004]			[-0.017,-0.014]	

OLS regressions described in equation 4 are presented using developing country (DHS) and US (NHIS) data. The 2+, 3+ and 4+ samples are defined in the estimation sample section of the paper (section 2.2). Base controls consist of fixed effects for child's age and year of birth, child gender, mother's age at birth, and a cubic for mother's age at time of survey. For the USA sample, mother's race fixed effects are included. For DHS data, country fixed effects are also included. Additional socioeconomic controls consist of mother's education and (for DHS data) wealth quintile fixed effects, and health controls include a continuous measure of mother's BMI, and for DHS, mother's height and coverage of prenatal care at the level of the survey cluster. For USA data, we include controls for mother's self assessed health on a Likert scale. Refer to section 3.1.1 for discussion of the Altonji et al. (2005) ratio and Oster (2013) bounds. Standard errors are clustered by mother.*p<0.1; **p<0.05; ***p<0.01

Table 7: Developing Country IV Estimates

	2+			3+			4+		
	Base	+H	+S&H	Base	+H	+S&H	Base	+H	+S&H
Panel A: First Stage									
Dependent Variable = Fertility									
Twins	0.789***	0.830***	0.841***	0.799***	0.824***	0.834***	0.842***	0.859***	0.865***
	(0.030)	(0.029)	(0.028)	(0.026)	(0.026)	(0.025)	(0.027)	(0.026)	(0.026)
Observations	259966	259966	259966	395721	395721	395721	409615	409615	409615
Kleibergen-Paap rk statistic	686.11	825.95	873.68	914.63	1018.39	1071.47	1004.99	1084.13	1071.55
p-value of rk statistic	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Panel B: IV Results									
Dependent Variable = School Z-Score									
Fertility	-0.005	-0.017	-0.014	-0.028	-0.040*	-0.046**	-0.027	-0.037*	-0.037**
	(0.028)	(0.027)	(0.026)	(0.022)	(0.021)	(0.020)	(0.023)	(0.021)	(0.019)
Observations	259966	259966	259966	395721	395721	395721	409615	409615	409615
R-Squared	0.04	0.08	0.15	0.04	0.08	0.16	0.03	0.07	0.15

Panels A and B present coefficients and standard errors for the first and second stages in equations 5a and 5b. The 2+ subsample refers to all first born children in families with at least two births. 3+ refers to first- and second-borns in families with at least three births, and 4+ refers to first- to third-borns in families with at least four births. Panel A presents the first-stage coefficients of twinning on fertility for each group. Base controls consist of child age, mother's age, and mother's age at birth fixed effects plus country and year-of-birth FEs. Additional socioeconomic controls consist of mother's education and wealth quintile fixed effects, and health controls include a continuous measure of mother's height and BMI and coverage of prenatal care at the level of the survey cluster. 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. In panel B 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+ groups respectively). Standard errors are clustered by mother.*p<0.1; **p<0.05; ***p<0.01

Table 8: US IV Estimates

	2+			3+			4+		
	Base	+H	+S&H	Base	+H	+S&H	Base	+H	+S&H
Panel A: First Stage									
Dependent Variable = Fertility (School Z-Score Second Stage)									
Twins	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)
Dependent Variable = Fertility (Excellent Health Second Stage)									
Twins	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)
Kleibergen-Paap rk statistic	632.46	708.35	735.59	234.43	244.11	245.40	97.04	98.25	104.98
p-value of rk statistic	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Panel B: IV Results									
Dependent Variable = School Z-Score									
Fertility	-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)
Dependent Variable = Excellent Health									
Fertility	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)
Observations (Education)	61267	61267	61267	47308	47308	47308	21352	21352	21352
Observations (Health)	70277	70277	70277	53393	53393	53393	24358	24358	24358

NOTES: Regressions in each panel and the definition of the 2+, 3+ and 4+ groups are identical to table 7 and are described in notes to table 7. This table presents the same regressions however now 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 for schooling variables, and for children aged between 1-18 years for health variables. The Kleibergen-Paap rk statistic for the first stage regressions is displayed for the regression using the education sample only. Qualitatively similar results are observed for the health sample. Descriptive statistics for each variable can be found in table A2. Standard errors are clustered by mother. *p<0.1; **p<0.05; ***p<0.01

Table 9: Estimates of γ Using Maternal Health Shocks

	$\frac{\partial Educ}{\partial Health}$	$\frac{\partial Health}{\partial Twin}$	$\gamma = \frac{\partial Educ}{\partial Twin}$	γ (bootstrap)
Panel A: United States				
Estimate	0.0497***	0.125***	0.0062	0.0062
	(0.0181)	(0.0181)		(0.0027)
Observations	943,038	943,038		
R-squared	0.011	0.069		
Panel B: Nigeria				
Estimate	-0.0154**	-0.267**	0.0040	0.0040
	(0.00637)	(0.00637)		(0.0022)
Observations	26,205	26,205		
R-squared	0.022	0.991		

NOTES: Regression results for panel A use the 5% sample of 1980 US census data and follow the specifications in Bhalotra and Venkataramani (2014). Regression results from panel B are based on all Nigerian DHS data in which children can be linked to their mothers. Specifications and samples are identical to those described in Akresh et al. (2012). The estimate of γ is formed by taking the product of the column 1 and column 2 estimates. A full description of this process, along with the non-pivotal bootstrap process to estimate the standard error of γ is provided in section 1.4, and online appendix D.

Table 10: 'Plausibly Exogenous' Bounds

	UCI: $\gamma \in [0, 2\hat{\gamma}]$		LTZ: Empirical Distribution γ	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
Panel A: DHS				
Two Plus	-0.0651	0.0109	-0.0504	0.0165
Three Plus	-0.0637	-0.0049	-0.0506	-0.0005
Four Plus	-0.0735	-0.0210	-0.0597	-0.0143
Panel B: USA (Education)				
Two Plus	-0.2194	-0.0028	-0.1656	-0.0035
Three Plus	-0.1256	0.0835	-0.1304	0.0622
Four Plus	-0.4359	0.1054	-0.2403	0.0196
Panel B: USA (Health)				
Two Plus	-0.0244	0.0618	-0.0235	0.0374
Three Plus	-0.1160	-0.0185	-0.0753	-0.0028
Four Plus	-0.1069	0.0237	-0.0930	0.0039

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 described in Conley et al. (2012) and section 1.4 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 γ follows the empirical distribution estimated in each case. The consistent estimation of $\hat{\gamma}$ and its entire distribution is discussed in section 1.4, and estimates for γ are provided in table 9.

References

J. Adda, C. Dustmann, and K. Stevens. The Career Costs of Children. CESifo Working Paper Series 6158, CESifo Group Munich, 2016.

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.

R. Akresh, S. Bhalotra, M. Leone, and U. Osili. War and Stature: Growing Up During the Nigerian Civil War. *American Economic Review (Papers & Proceedings)*, 102(3):273–77, 2012.

R. Akresh, S. Bhalotra, M. Leone, and U. Osili. First and Second Generation Impacts of the Nigeria-Biafra War. Mimeo, 2016.

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. Is the 1918 Influenza Pandemic Over? Long-Term Effects of *In Utero* Influenza Exposure in the Post-1940 U.S. Population. *Journal of Political Economy*, 114(4):672–712, 2006.

D. Almond and J. Currie. Killing Me Softly: The Fetal Origins Hypothesis. *Journal of Economic Perspectives*, 25(3):153–172, 2011.

D. Almond and L. Edlund. Trivers–Willard at birth and one year: evidence from US natality data 1983–2001. *Proceedings of the Royal Society of London B: Biological Sciences*, 274(1624): 2491–2496, 2007.

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, H. W. Hoynes, and D. W. Schanzenbach. Inside the War on Poverty: The Impact of Food Stamps on Birth Outcomes. *The Review of Economics and Statistics*, 93(2):387–403, May 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.

G. E. Anderson, A. D. Whipple, and S. R. Jimerson. Grade Retention: Achievement and Mental Health Outcomes. Document available at http://www.wrightslaw.com/info/fape.grade.retention.nasp.pdf, National Association of School Psychologists, Oct. 2002.

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. 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. Åslund and H. Grönqvist. Family size and child outcomes: Is there really no trade-off? *Labour Economics*, 17(1):130–39, 2010.

S. Baird, J. H. Hicks, M. Kremer, and E. Miguel. Worms at Work: Long run Impacts of a Child Health Investment. *The Quarterly Journal of Economics*, 131(4):1637–1680, 2016.

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.

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.

I. M. Bernstein, J. A. Mongeon, G. J. Badger, L. Solomon, S. H. Heil, and S. T. Higgins. Maternal smoking and its association with birth weight. *Obstetrics and Gynecology*, 106(5): 986–991, 2005.

S. Bhalotra and S. Rawlings. Gradients of Intergenerational Transmission of Health in Developing Countries. *The Review of Economics and Statistics*, 95(2):660–672, 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. Bhalotra, M. Karlsson, and T. Nilsson. Infant Health and Longevity: Evidence from a Historical Trial in Sweden. Discussion Paper 8969, IZA, April 2015.

S. Bhalotra, S. Telaligac, and A. Venkataramani. Fertility, Mortality Risk and Returns to Human Capital: Quasi-Experimental Evidence from 20th Century America. Mimeo, University of Essex, 2016.

S. R. Bhalotra and A. Venkataramani. Cognitive Development and Infectious Disease: Gender Differences in Investments and Outcomes. IZA Discussion Papers 7833, Institute for the Study of Labor (IZA), Dec. 2013.

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. Does Grief Transfer across Generations? Bereavements during Pregnancy and Child Outcomes. *American Economic Journal: Applied Economics*, 8(1):193–223, January 2016.

J. Blake. *Family Size and Achievement*. University of California Press, Berkeley, 1989.

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.

C. E. Boklage. Survival probability of human conceptions from fertilization to term. *International Journal of Fertility*, 35(2):79–94, 1990.

D. Boomsma, A. Busjahn, and L. Peltonen. Classical Twin Studies and Beyond. *Nature Reviews Genetics*, 3(11):872–882, November 2002.

T. J. Bouchard and P. Propping. *Twins as a Tool of Behavioural Genetics*. John Wiley & Sons, Chichester, United Kingdom, 1993.

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

C. Bozzoli, A. Deaton, and C. Quintana-Domeque. Adult height and childhood disease. *Demography*, 46(4):647–669, November 2009.

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.

C. Brinch, M. Mogstad, and M. Wiswall. Beyond LATE with a Discrete Instrument. *Journal of Political Economy*, xx(x):xxx–xxx, Forthcoming.

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.

M. G. Bulmer. *The Biology of Twinning in Man*. Oxford Clarendon Press, Oxford, UK, 1970.

A. Butikofer and K. G. Salvanes. Disease Control and Inequality Reduction: Evidence from a Tuberculosis Testing and Vaccination Campaign. Discussion Paper 28/2015, NHH Dept. of Economics, November 2015.

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

F. Campbell, G. Conti, J. J. Heckman, S. H. Moon, R. Pinto, E. Pungello, and Y. Pan. Early Childhood Investments Substantially Boost Adult Health. *Science*, 343(6178):1478–1485, 2014.

P. Carneiro, K. Løken, and K. G. Salvanes. A flying start? maternity leave benefits and long-run outcomes of children. *Journal of Political Economy*, 123(2):365–412, 2015.

K. Chay and M. Greenstone. The Impact of Air Pollution on Infant Mortality: Evidence from Geographic Variation in Pollution Shocks Induced by a Recession. *The Quarterly Journal of Economics*, 118(3):1121–1167, 2003.

D. Clarke. Fertility and Causality. CSAE Working Paper Series 2016-32, Centre for the Study of African Economies, University of Oxford, Nov. 2016.

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, C. B. Hansen, and P. E. Rossi. Plausibly Exogenous. *The Review of Economics and Statistics*, 94(1):260–272, February 2012.

J. Currie and E. Moretti. Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings. *The Quarterly Journal of Economics*, 118(4):1495–1532, 2003.

J. Currie and E. Moretti. Biology as Destiny? Short- and Long-Run Determinants of Intergenerational Transmission of Birth Weight. *Journal of Labor Economics*, 25(2):231–264, 2007.

D. N. De Tray. Child quality and the demand for children. *Journal of Political Economy*, 81(2): S70–95, March 1973.

A. Deaton. *The Analysis of Household Surveys – A Microeconometric Approach to Development Policy*. The Johns Hopkins University Press, 1997.

A. Deaton. Height, health, and development. *Proceedings of the National Academy of Sciences*, 104(33):13232–13237, August 2007.

R. Dehejia, C. Pop-Eleches, and C. Samii. From Local to Global: External Validity in a Fertility Natural Experiment. Working Paper 21459, National Bureau of Economic Research, August 2015.

D. Filmer, J. Friedman, and N. Schady. Development, Modernization, and Childbearing: The Role of Family Sex Composition. *World Bank Economic Review*, 23(3):371–398, 2009.

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.

M. Fort, N. Schneeweis, and R. Winter-Ebmer. Is Education Always Reducing Fertility? Evidence from Compulsory Schooling Reforms. *The Economic Journal*, pages n/a–n/a, 2016. ISSN 1468-0297. doi: 10.1111/ecoj.12394.

O. Galor. The demographic transition: causes and consequences. *Cliometrica, Journal of Historical Economics and Econometric History*, 6(1):1–28, January 2012.

O. Galor and D. N. Weil. Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond. *The American Economic Review*, 90(4):806–828, 2000.

A. García-Enguídanosa, M. Calleb, J. Valeroc, S. Lunaa, and V. Domínguez-Roja. Risk factors in miscarriage: a review. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 102(2):111–119, May 2002.

J. B. Gelbach. When do Covariates Matter? And Which Ones, And How Much? *Journal of Labor Economics*, 34(2):509–543, 2016.

P. Gluckman and M. Hanson. *The Fetal Matrix: Evolution, Development and Disease*. Cambridge University Press, Cambridge, United Kingdom, 2005.

N. D. Grawe. The quality–quantity trade-off in fertility across parent earnings levels: a test for credit market failure. *Review of Economics of the Household*, 6(1):29–45, 2008.

Guttmacher Institute. Induced Abortion in the United States. Fact sheet, Guttmacher Institute, Sept. 2016. URL https://www.guttmacher.org/sites/default/files/factsheet/fb_induced_abortion_3.pdf.

J. G. Hall. Twinning. *The Lancet*, 362(9385):735–743, 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, R. Pinto, and P. Savelyev. Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review*, 103(6):2052–86, October 2013.

J. J. Heckman, S. H. Moon, R. Pinto, P. A. Savelyev, and A. Yavitz. The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, 94(1):114–128, 2010.

J. Hjort, M. Sølvsten, and M. Wüst. Universal Investment in Infants and Long-run Health. Mimeo, Technical Report, 2016.

J. P. Jacobsen, J. W. P. 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.

W. H. James. Sex ratio in twin births. *Annals of Human Biology*, 2(4):365–378, 1975.

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.

R. Jensen. The (Perceived) Returns to Education and the Demand for Schooling. *The Quarterly Journal of Economics*, 125(2):515–548, 2010.

S. R. Jimerson. On the Failure of Failure: Examining the Association Between Early Grade Retention and Education and Employment Outcomes During Late Adolescence. *Journal of School Psychology*, 37(3):243 – 272, 1999.

S. R. Jimerson. Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review*, 30(3):313 – 330, 2001.

S. R. Jimerson, P. Ferguson, A. D. Whipple, G. E. Anderson, and M. J. Dalton. Exploring the Association Between Grade Retention and Dropout: A Longitudinal Study Examining Socio-Emotional, Behavioral, and Achievement Characteristics of Retained Students. *The California School Psychologist*, 7(1):51–62, 2002.

B. Kahn, L. H. Lumey, P. A. Zybert, J. M. Lorenz, J. Cleary-Goldman, M. E. D'Alton, and J. N. Robinson. Prospective risk of fetal death in singleton, twin, and triplet gestations: Implications for practice. *Obstetrics & Gynecology*, 102(4):685–92, 2003.

D. S. Kenkel. Health Behavior, Health Knowledge, and Schooling. *Journal of Political Economy*, 99(2):287–305, 1991.

A. D. Kulkarni, D. J. Jamieson, H. W. J. Jones, D. M. Kissin, M. F. Gallo, M. Macaluso, and E. Y. Adashi. Fertility Treatments and Multiple Births in the United States. *New England Journal of Medicine*, 369(23):2218–2225, 2013.

E. E. Leamer. *Specification Searches – Ad Hoc Inference with Nonexperimental Data*. John Wiley & Sons, Inc., 1978.

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 childrens education: an Asian instrument. *Journal of Population Economics*, 21(4):855–875, October 2008.

J. E. Lesch. *The First Miracle Drugs: How the Sulfa Drugs Transformed Medicine*. Oxford University Press, Oxford, 2006.

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.

A. Lleras-Muney and D. Cutler. Understanding Differences in Health behaviors by Education. *Journal of Health Economics*, 29(1):1–28, 2010.

A. Lleras-Muney and F. Lichtenberg. The Effect Of Education On Medical Technology Adoption: Are The More Educated More Likely To Use New Drugs? *Annales d'Economie et Statistique*, 79/80, 2005.

B. S. Low. *Why Sex Matters: A Darwinian Look at Human Behavior*. Princeton University Press, Princeton, NJ, 2000.

S. Lundberg, R. A. Pollak, and J. Stearns. Family Inequality: Diverging Patterns in Marriage, Cohabitation, and Childbearing. *Journal of Economic Perspectives*, 30(2):79–102, 2016.

P. Lundborg, E. Plug, and A. Wurtz Rasmussen. Fertility Effects on Labor Supply: IV Evidence from IVF Treatments. Discussion Paper 8609, IZA, 2014.

B. Mazumder and Z. Seeskin. Skipping breakfast and the sex ratio at birth. Mimeo, 2014.

G. E. McClearn, B. Johansson, S. Berg, N. L. Pederson, F. Ahern, S. A. Petrill, and R. Plomin. Substantial genetic influence on cognitive abilities in twins 80 or more years old. *Science*, 276 (5318):1560–1563, June 1997.

O. Moav. Cheap Children and the Persistence of Poverty. *The Economic Journal*, 115(500): 88–110, 2005.

M. Mogstad and M. Wiswall. Testing the Quantity-Quality Model of Fertility: Linearity, Marginal Effects, and Total Effects. *Quantitative Economics*, 7(1):157–192, 2016.

M. Myrskylä, H.-P. Kohler, and F. C. Billari. Advances in development reverse fertility declines. *Nature*, 460:741–743, 2009.

A. Nevo and A. M. Rosen. Identification with Imperfect Instruments. *The Review of Economics and Statistics*, 94(3):659–671, August 2012.

A. F. Newman and C. Olivetti. Career Women and the Durability of Marriage. mimeo, Boston University, 2016.

J. Nisén, P. Martikainen, J. Kaprio, and K. Silventoinen. Educational Differences in Completed Fertility: A Behavioral Genetic Study of Finnish Male and Female Twins. *Demography*, 50(4): 1399–1420, August 2013.

E. Oster. Unobservable Selection and Coefficient Stability: Theory and Validation. Working Paper 19054, National Bureau of Economic Research, May 2013.

H. Patrinos. Returns to Education: The Gender Perspective. In M. Tembon and L. Fort, editors, *Girls' Education in the 21st Century: Gender Equality, Empowerment, and Economic Growth*, Directions in Development: Human Development, chapter 4, pages 209–240. World Bank, 2008.

P. Persson and M. Rossin-Slater. Family Ruptures, Stress, and the Mental Health of the Next Generation. *American Economic Review*, xx(x):xxx–xxx, Forthcoming.

D. I. W. Phillips. Twin studies in medical research: Can they tell us whether diseases are genetically determined? *The Lancet*, 341(8851):1008–1009, April 1993.

T. J. C. Polderman, B. Benyamin, C. A. de Leeux, P. F. Sullivan, A. van Bochoven, P. M. Visscher, and D. Posthuma. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7):702–709, May 2015.

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, 2012a.

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, 2012b.

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.

M. Rendall. Brain versus brawn: the realization of women's comparative advantage. IEW – Working Papers 491, Institute for Empirical Research in Economics - University of Zurich, Sept. 2010.

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):pp. 328–348, 1980b.

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

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.

T. P. Schultz. School subsidies for the poor: evaluating the Mexican Progresa poverty program. *Journal of Development Economics*, 74(1):199–250, 2004.

S. Shinagawa, S. Suzuki, H. Chihara, Y. Otsubo, T. Takeshita, and T. Araki. Maternal basal metabolic rate in twin pregnancy. *Gynecologic and Obstetric Investigation*, 60(3):145–48, 2005.

K. Silventoinen. Determinants of variation in adult body height. *Journal of Biosocial Science*, 35(2):263–285, April 2003.

E. L. Thorndike. Measurement of Twins. *The Journal of Philosophy, Psychology and Scientific Methods*, 2(2):547–553, Sep 1905.

R. L. Trivers and D. E. Willard. Natural selection of parental ability to vary the sex ratio of offspring. *Science*, 179(4068):90–92, 1973.

United Nations. World population policies 2009. Technical report, Department of Economic and Social Affairs: Population Division, 2010.

S. Vitthala, T. A. Gelbaya, D. R. Brison, C. T. Fitzgerald, and L. G. Nardo. The risk of monozygotic twins after assisted reproductive technology: a systematic review and meta-analysis. *Human Reproduction Update*, 15(1):45–55, Jan-Feb 2009.

Y. Wang, X. Wang, Y. Kong, J. H. Zhang, and Q. Zeng. The Great Chinese Famine leads to shorter and overweight females in Chongqing Chinese population after 50 years. *Obesity*, 18 (3):588–592, Sep 2010.

J. R. Warren, E. Hoffman, and M. Andrew. Patterns and Trends in Grade Retention Rates in the United States, 1995–2010. *Educational Researcher*, 43(9):433–443, 2014.

R. J. Willis. A New Approach to the Economic Theory of Fertility Behavior. *Journal of Political Economy*, 81(2):S14–S64, 1973.

Acknowledgements We are grateful to Paul Devereux, James Fenske, Judith Hall, Martin Karlsson, Cheti Nicoletti, Carol Propper, Margaret Stevens, Atheen Venkataramani, Marcos Vera-Hernandez, Frank Windmeijer, Emilia Del Bono, Climent Quintana-Domeque, Pedro Ródenas, Libertad González, Hanna Mühlrad, Anna Aevarsdottir, Martin Foureaux Koppensteiner, Ryan Palmer and Pietro Biroli along with various seminar audiences and discussants for helpful comments and/or sharing data. Any remaining errors are our own.

ONLINE APPENDIX

For the paper:

THE TWIN INSTRUMENT
Sonia Bhalotra and Damian Clarke

Contents

A Appendix Figures **A2**

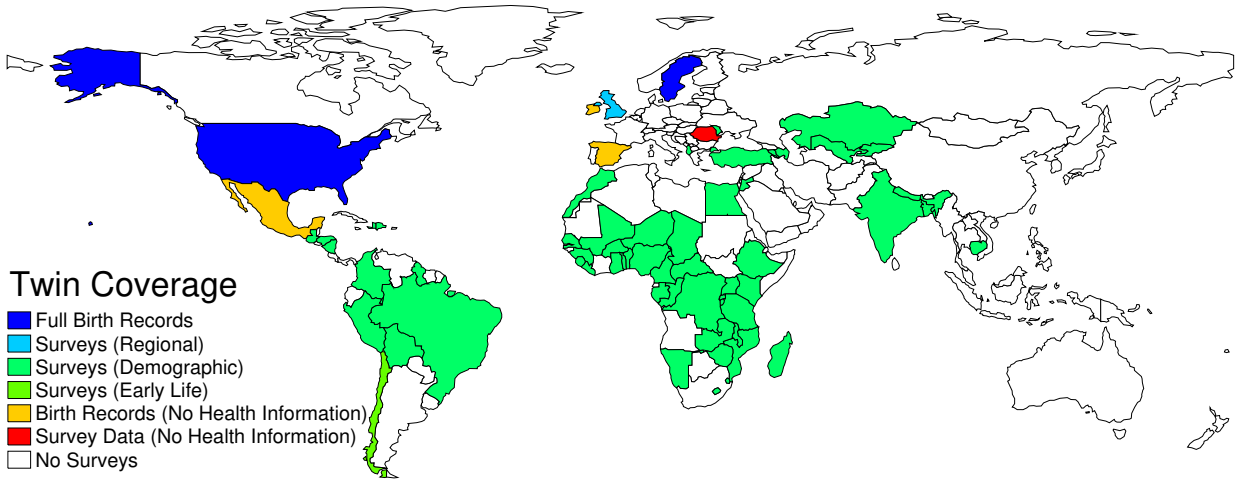
B Appendix Tables **A9**

C Data Appendix **A32**
C.1 Regressions of Twinning on Maternal Health A32
C.2 The DHS A32
C.3 The NHIS A33

D Resampling and Simulation Based Estimation of γ **A34**
D.1 Bootstrap Confidence Intervals A34
D.2 Simulation-Based Estimation for non-Normal Distributions A34

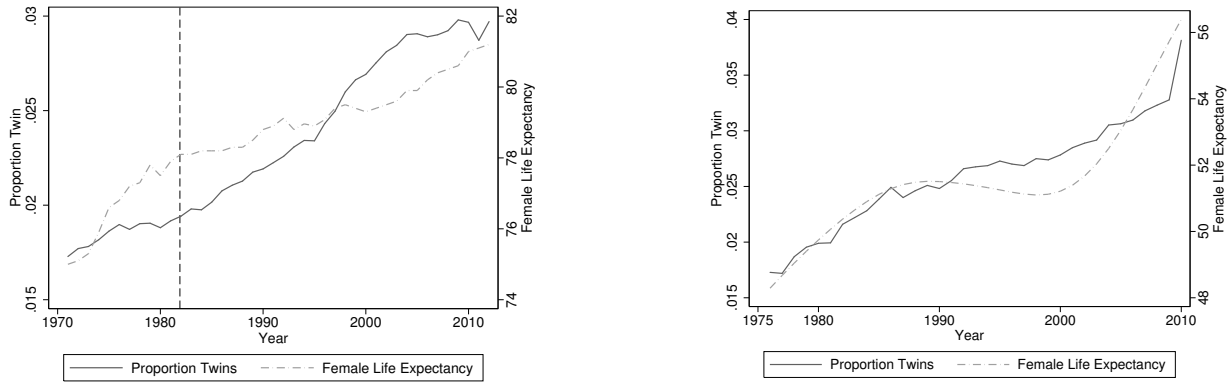
A Appendix Figures

Figure A1: Coverage of data containing indicators of twin births and maternal health by country and data type



Different colours represent different types of data (surveys, national vital statistics, or no data collected). Each data type is described in the figure legend.

Figure A2: Proportion of Twins of All Births

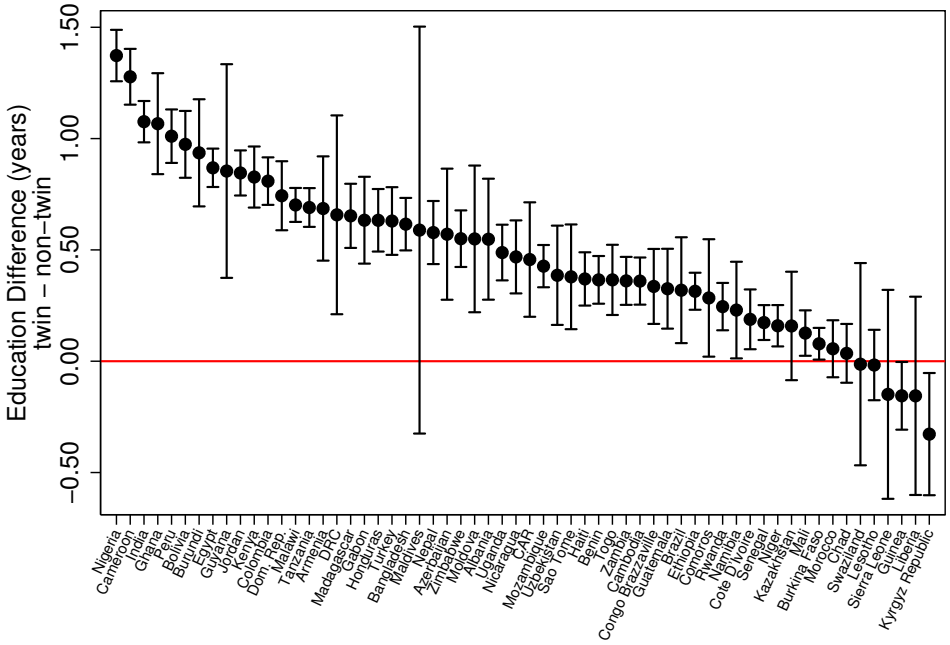


(a) USA

(b) Sub-Saharan Africa

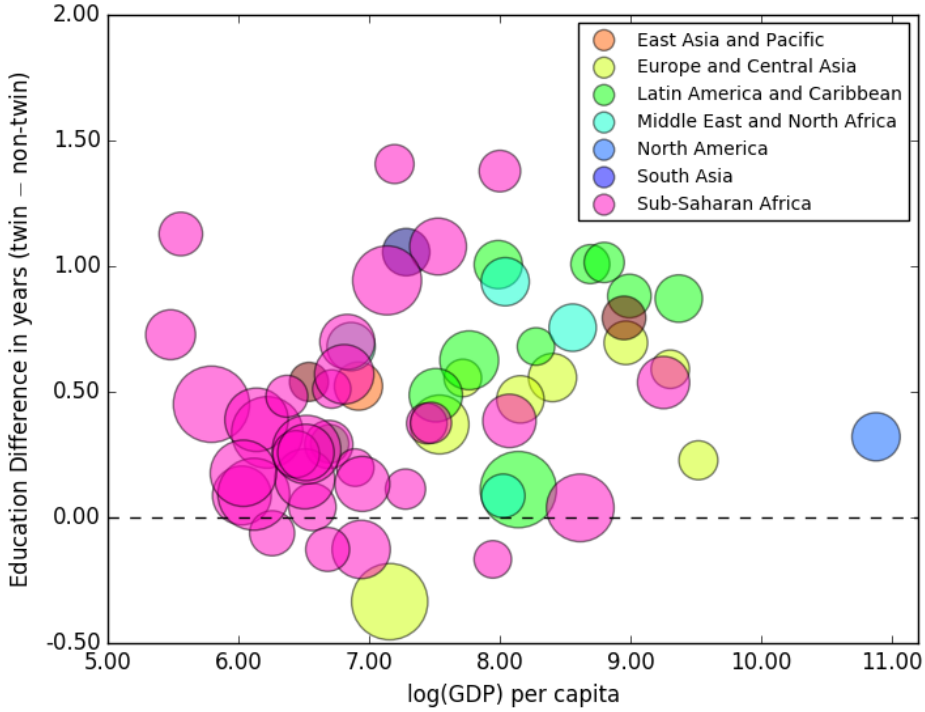
Note to figure A2: Female life expectancy data comes from the World Bank Data Bank. The proportion of twins in panel (a) is calculated from NVSS Birth Certificate data. The proportion of twins in panel (b) is calculated from DHS data using all countries in sub-Saharan Africa (and smoothed using a 2,1,2 moving average). The vertical dotted line in panel (a) represents the first reported use of IVF in the USA. In each case mothers up to the age of 35 are included, to avoid concerns regarding high IVF use at older ages. Results are quantitatively similar if using all mothers. Panel (a) is based on 132,783,003 live births, and panel (b) is based on 1,208,373 live births.

Figure A3: Completed Education Differential By Twin and non-Twin Mothers by Country



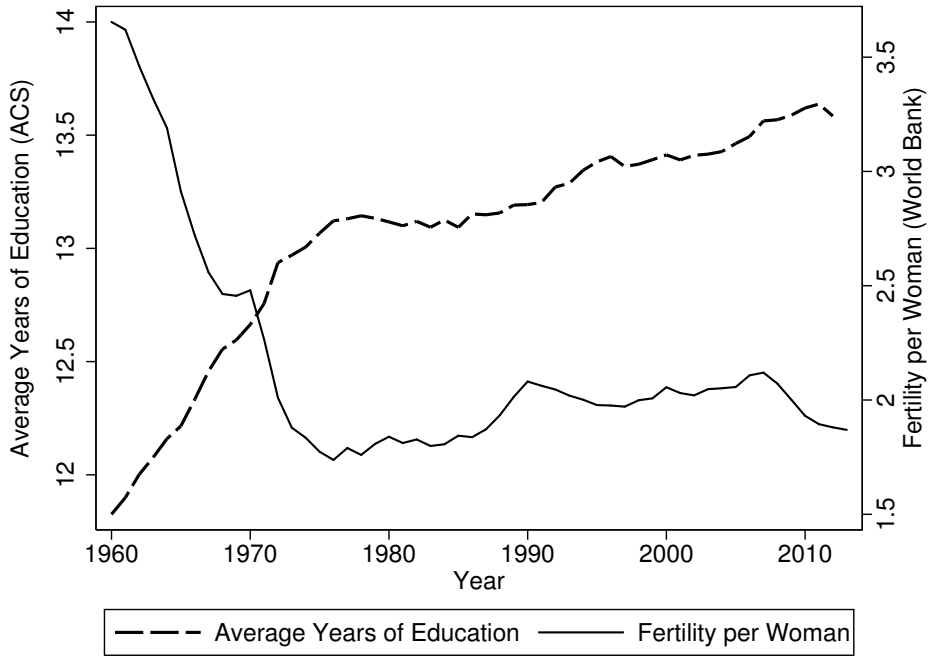
Note to figure A3: Refer to figure 1.

Figure A4: Completed Education Differential By Twin and non-Twin Mothers by Country and GDP



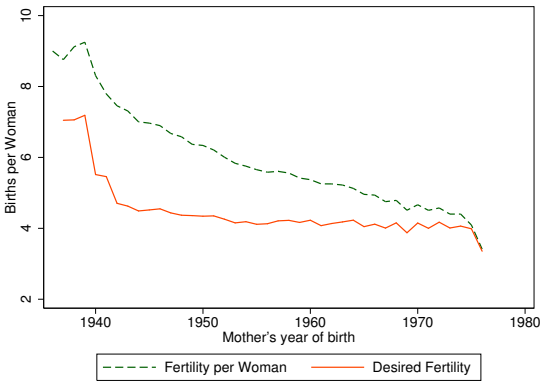
Note to figure A4: Refer to figure 2

Figure A5: Education and Fertility Trends (USA)

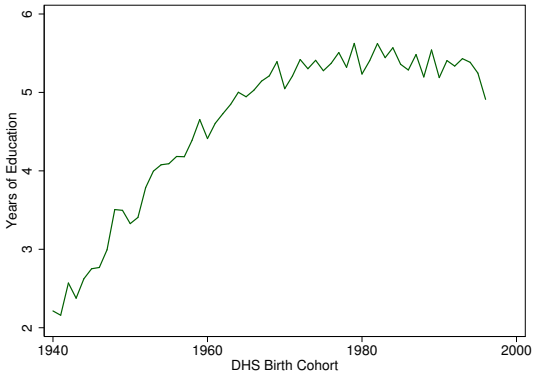


Note to figure A5: Trends in fertility and education are compiled from the World Bank databank and the American Community Surveys (ACS), respectively. Trends in fertility are directly reported by the World Bank as completed fertility per woman were she exposed to prevailing rates in a given year for her whole fertile life. Education is calculated using all women aged over 25 years in the ongoing ACS (2001-2013) collected by the United States Census Bureau. The figure presents average completed education for all women aged 25 in the year in question.

Figure A6: Education and Fertility (Developing Countries)



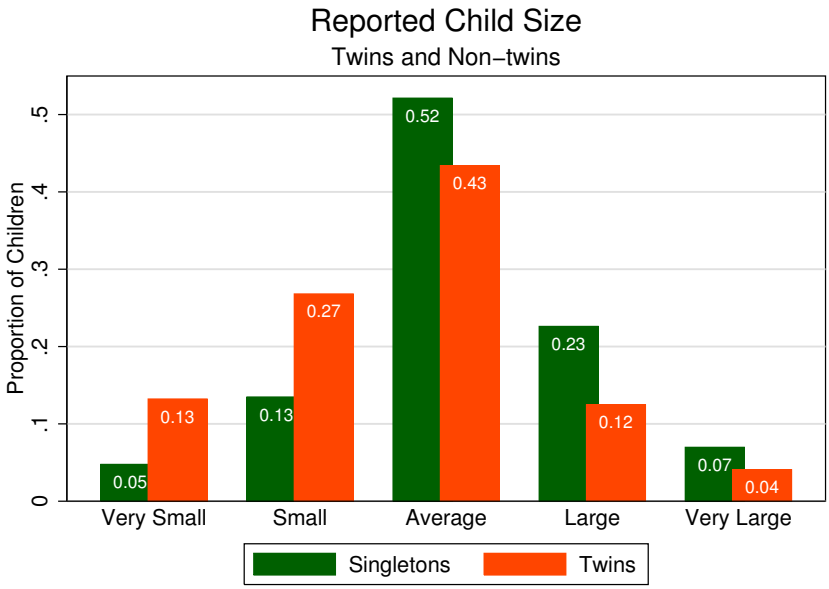
(a) Trends in Fertility



(b) Trend in Education

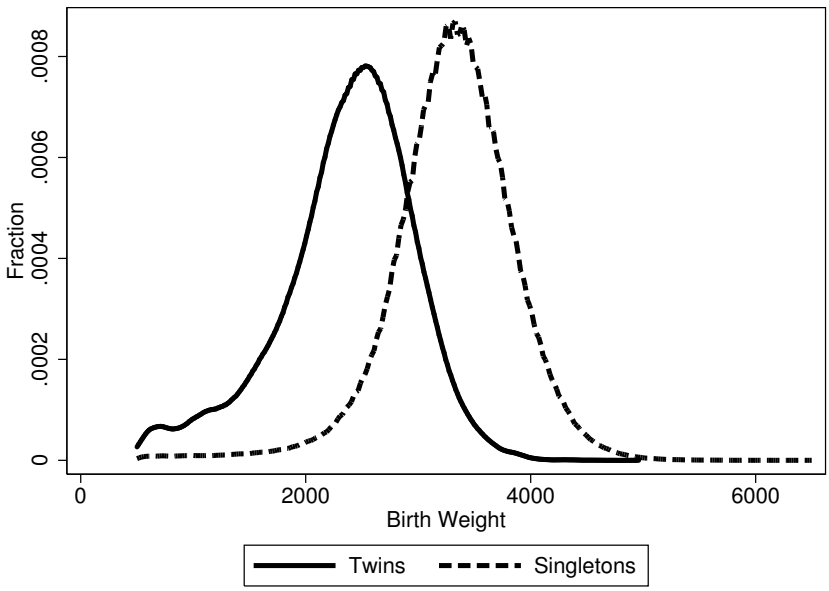
Note to figure A6: Cohorts are made up of all individuals from the DHS who are aged over 35 years (for fertility), and over 15 years (for education). In each case the sample is restricted to those who have approximately completed fertility and education respectively. Full summary statistics for these variables are provided in table A2, and a full list of country and survey years are available in table A21.

Figure A7: Birth Size of Twins versus Singeltons (Developing Countries)



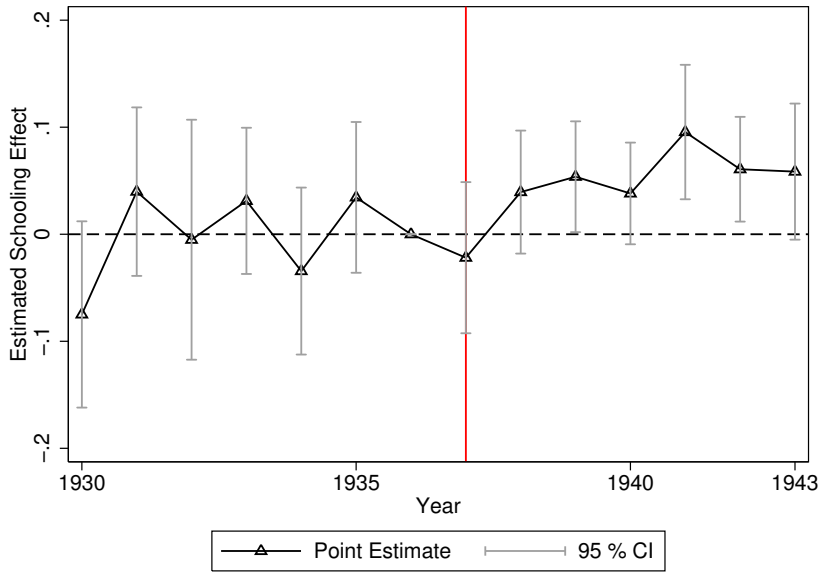
Note to figure A7 Estimation sample consists of all surveyed births from DHS countries occurring within 5 years prior to the date of the survey. For each of these births, all mothers retrospectively report the (subjective) size of the baby at the time of birth.

Figure A8: Birth Weight of Twins versus Singeltons (USA)



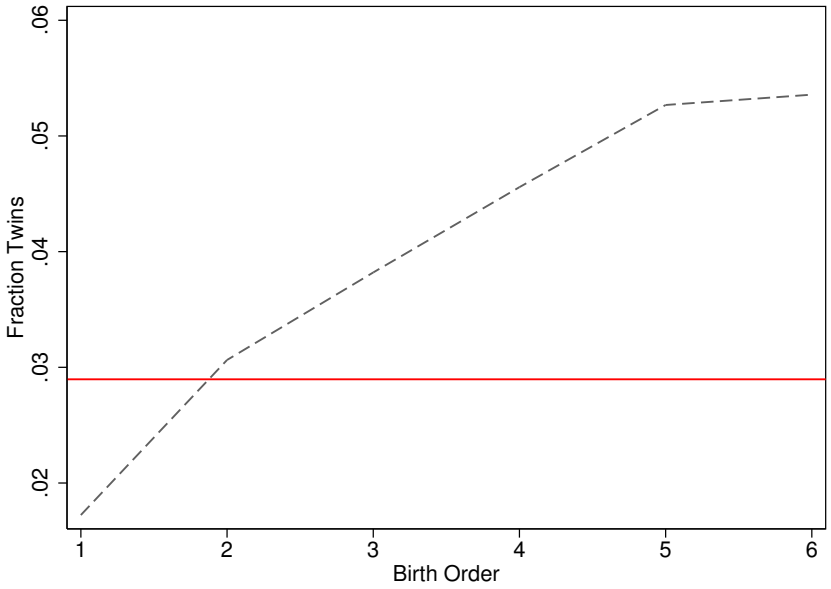
Note to figure A8 Estimation sample consists of all non-ART births from NVSS data between 2009 and 2013. Birthweights below 500 grams and above 6,500 grams are trimmed from the sample.

Figure A9: Test of Parallel Trends of Second Generation Sulfa Effects for γ



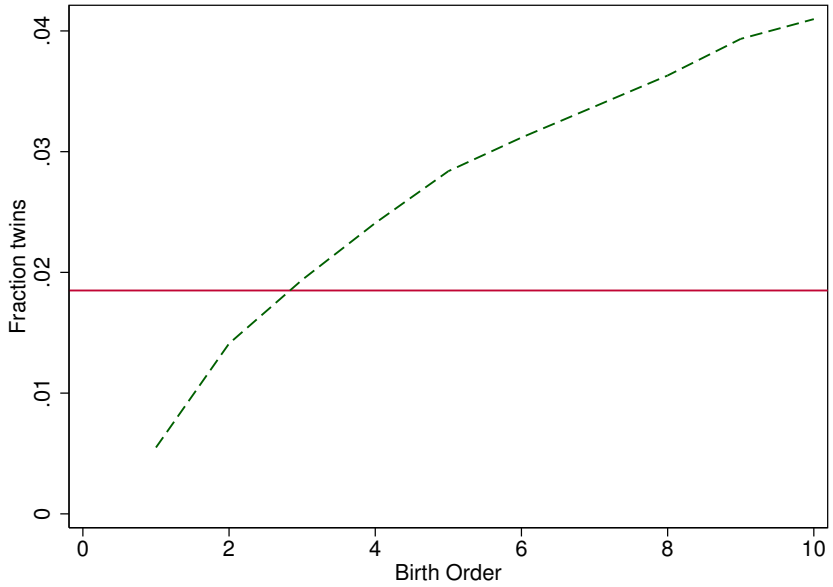
Note to figure A9: Graph replicates specification (9) from the paper, however now interacting *basePneumonia* with each mother's birth year, rather than a single *Post* dummy starting from 1937. Each coefficient and confidence interval displays the differential effect of a child's mother being born in a high- or low-pneumonia state by birth year surrounding the sulfa reform. The year preceding the arrival of sulfa reform is omitted (1936) and post sulfa estimates and confidence intervals represent the differential impact of sulfa drugs on second generation (educational) outcomes of children of affected women. Standard errors are clustered by state.

Figure A10: Proportion of Twins by Birth Order (United States)



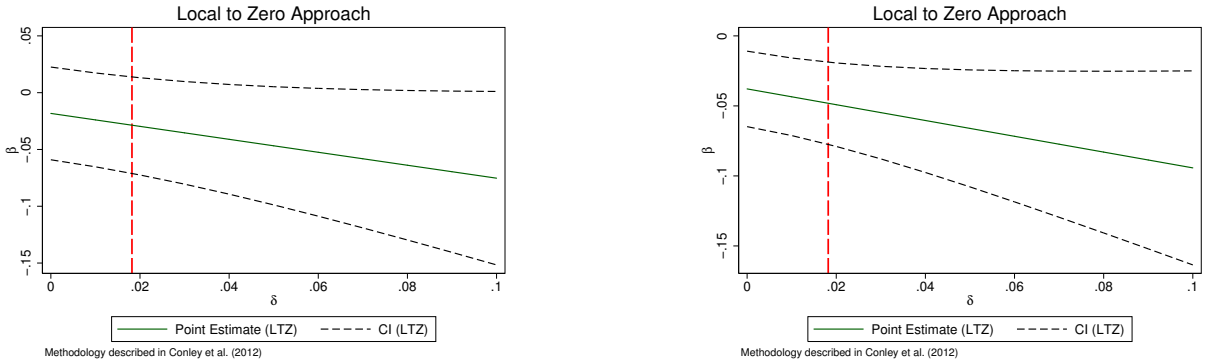
Note to figure A10 The fraction of twin births are calculated from the full sample of non-ART users in NVSS data from 2009-2013. The solid line represents the average fraction of twins in the full sample (2.89%), while the dotted line presents twin frequency by birth order. The dotted line joins points at each birth order. Birth orders greater than 6 are removed from the sample given that these account for less than 0.5% of all recorded births.

Figure A11: Proportion of Twins by Birth Order (Developing Countries)



Note to figure A11 The fraction of twin births are calculated from the full sample of DHS data. The solid line represents the average fraction of twins in the full sample (1.85%), while the dotted line presents twin frequency by birth order. The dotted line joins points at each birth order $\in \{1, \ldots, 10\}$. The fraction of singleton births is $1-$frac(twin).

Figure A12: Plausibly Exogenous Bounds: School Z-Score (Developing Countries 2+ and 4+)

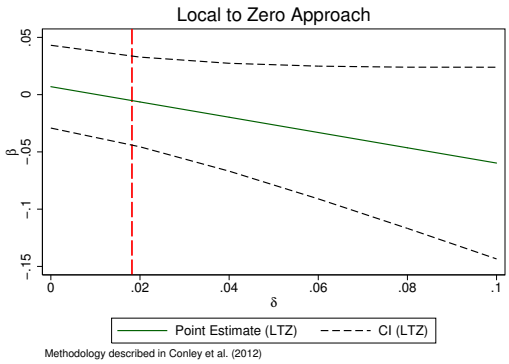


(a) Two Plus

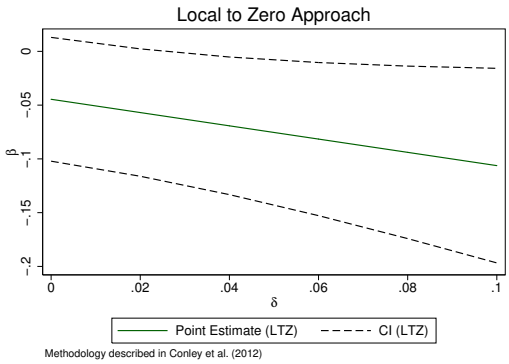
(b) Four Plus

Note to figure A12: Refer to notes to figure 5 of the main text.

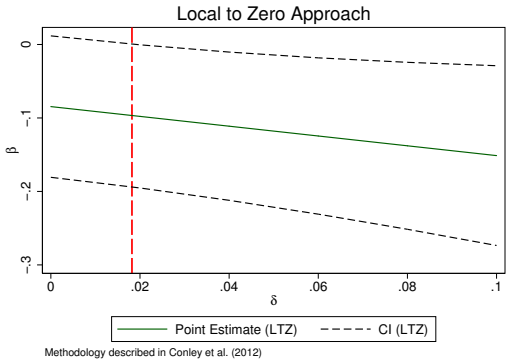
Figure A13: Plausibly Exogenous Bounds: (USA 2+ and 4+)



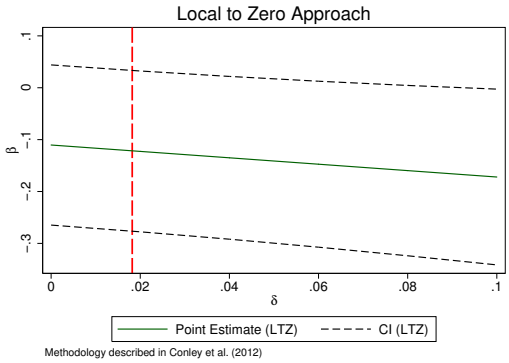
(a) Excellent Health (2+)



(b) Excellent Health (4+)



(c) Education Z-Score (2+)



(d) Education Z-Score (4+)

Note to figure A13: Refer to notes to figure 5 of the main text.

B Appendix Tables

Table A1: Summary Statistics (Twin Regressions)

	N	Mean	Std.Dev.	Min	Max
Panel A: United States					
Mother's height (cm)	1,363,558	163.00	7.26	91.44	198.12
Mother's education (years)	1,363,558	4.19	1.79	1.00	9.00
Mother Smoked Before Pregnancy	1,363,558	0.12	0.32	0.00	1.00
Mother Smoked in 1st Trimester	1,363,558	0.09	0.28	0.00	1.00
Mother Smoked in 2nd Trimester	1,363,558	0.08	0.26	0.00	1.00
Mother Smoked in 3rd Trimester	1,363,558	0.07	0.26	0.00	1.00
Mother had pre-pregnancy diabetes	1,363,558	0.01	0.09	0.00	1.00
Mother had pre-pregnancy hypertension	1,363,558	0.01	0.12	0.00	1.00
Mother is underweight (pre-pregnancy)	1,363,558	0.06	0.23	0.00	1.00
Mother is obese (pre-pregnancy)	1,363,558	0.20	0.40	0.00	1.00
Percent Twin Births	1,363,558	2.84	16.62	0.00	100.00
Mother's Age in years	1,363,558	28.09	5.78	18.00	49.00
Panel B: Sweden					
Pre-pregnancy asthma	1,240,621	0.07	0.25	0	1
Pre-pregnancy diabetes	1,240,621	0.01	0.07	0	1
Pre-pregnancy kidney disease	1,240,621	0.01	0.07	0	1
Pre-pregnancy hypertension	1,240,621	0.01	0.06	0	1
Smoked at 12 weeks gestation	1,240,621	0.09	0.29	0	1
Smoked at 30-32 weeks gestation	1,240,621	0.07	0.27	0	1
Height	1,240,621	166.38	6.35	100	200
Underweight (BMI $<$ 18.5) Prior to Pregnancy	1,240,621	0.02	0.15	0	1
Obese (BMI $\geq$ 30) Prior to Pregnancy	1,240,621	0.11	0.31	0	1
Percent Twin Births	1,240,621	2.55	15.77	0	100
Mother's Age in Years	1,240,621	29.96	5.11	18	49
Panel C: United Kingdom (Avon)					
Underweight (BMI $<$ 18.5) Prior to Pregnancy	10,463	0.04	0.21	0.00	1.00
Obese (BMI $\geq$ 30) Prior to Pregnancy	10,463	0.05	0.22	0.00	1.00
Mother's height (cm)	10,463	164.10	6.68	129.54	200.66
Pre-pregnancy diabetes	10,463	0.00	0.06	0.00	1.00
Pre-pregnancy hypertension	10,463	0.04	0.20	0.00	1.00
Pre-pregnancy infections (total)	10,463	3.26	1.24	0.00	7.00
Frequent Healthy Food in Pregnancy	10,463	0.21	0.40	0.00	1.00
Frequent Fresh Fruit in Pregnancy	10,463	0.30	0.46	0.00	1.00
Infrequent Alcohol Consumption in Pregnancy	10,463	0.19	0.39	0.00	1.00
Frequent Alcohol Consumption in Pregnancy	10,463	0.06	0.24	0.00	1.00
Exposed to Passive Smoke in Pregnancy	10,463	0.43	0.50	0.00	1.00
Smoked during Pregnancy	10,463	0.17	0.38	0.00	1.00
Mother's education (years)	10,463	12.29	1.83	10.00	16.00
Percent Twin Births	10,463	2.37	15.21	0.00	100.00
Mother's Age in years	10,463	28.67	4.61	18.00	45.00
Panel D: Chile					
Mother Smoked During Pregnancy	14,050	0.10	0.30	0.00	1.00

Drugs During Pregnancy (Sporadically)	14,050	0.01	0.07	0.00	1.00
Drugs During Pregnancy (Regularly)	14,050	0.00	0.05	0.00	1.00
Alcohol During Pregnancy (Sporadically)	14,050	0.07	0.25	0.00	1.00
Alcohol During Pregnancy (Regularly)	14,050	0.00	0.06	0.00	1.00
Mother Obese Prior to Pregnancy	14,050	0.02	0.14	0.00	1.00
Mother Low Weight Prior to Pregnancy	14,050	0.07	0.25	0.00	1.00
Mother's Education in Years	14,050	10.83	3.59	0.00	16.00
Percent Twin Births	14,050	2.55	15.76	0.00	100.00
Mother's Age in Years	14,050	27.89	6.61	18.00	49.00
Panel E: Developing Countries					
Mother's Height (cm)	2,052,338	155.82	7.11	73.40	230.50
Mother is underweight	2,052,338	0.12	0.32	0.00	1.00
Mother is obese	2,052,338	0.11	0.32	0.00	1.00
Mother's Education	2,052,338	4.22	4.56	0.00	27.00
Prenatal Care in Area (% Doctor)	2,052,338	0.35	0.29	0.00	1.00
Prenatal Care Area (% Nurse)	2,052,338	0.41	0.26	0.00	1.00
Prenatal Care in Area (% Any)	2,052,338	0.79	0.20	0.00	1.00
Percent Twin Births	2,052,338	2.10	14.33	0.00	100.00
Mother's age in years	2,052,338	34.15	7.54	18.00	49.00

Each panel presents descriptive statistics of data from each sample analysed in Figure 1 of the paper. Panel A comes from the United States Vital Statistics System for all non-ART users from 2009-2013, Panel B consists of all births from the Swedish Medical Birth Register from 1993-2012, and Panel C comes Avon Longitudinal Study of Parents and Children. Further details on data are available in Part 1 section 2 and appendix C. All variables are either binary measures, or with units indicated in the variable name.

Table A2: Summary Statistics

	Developing Countries			United States		
	Single	Twins	All	Single	Twins	All
Mother's Characteristics						
Fertility	3.592	6.489	3.711	1.925	3.107	1.955
	(2.351)	(2.724)	(2.436)	(1.001)	(1.176)	(1.022)
Age	31.18	35.49	31.36	36.05	36.88	36.07
	(8.095)	(7.385)	(8.113)	(8.396)	(7.997)	(8.387)
Education	4.823	3.582	4.772	12.54	12.70	12.54
	(4.721)	(4.330)	(4.712)	(2.326)	(2.232)	(2.323)
Height	155.6	157.4	155.7	-	-	-
	(7.075)	(7.050)	(7.083)	-	-	-
BMI	23.31	23.69	23.32	27.45	28.01	27.47
	(4.819)	(5.004)	(4.827)	(6.628)	(7.247)	(6.645)
Pr(BMI)<18.5	0.124	0.100	0.123	0.0206	0.0168	0.0205
	(0.330)	(0.300)	(0.329)	(0.142)	(0.128)	(0.142)
Excellent Health	-	-	-	0.320	0.325	0.320
	-	-	-	(0.466)	(0.468)	(0.466)
Children's Outcomes						
Age	13.24	13.36	13.24	11.19	10.77	11.18
	(4.912)	(4.873)	(4.909)	(3.892)	(3.902)	(3.893)
Education (Years)	4.495	4.006	4.465	5.135	4.633	5.123
	(3.811)	(3.680)	(3.804)	(3.835)	(3.736)	(3.833)
Education (Z-Score)	0.00652	-0.0967	0.000	0.00236	-0.0980	0.0000
	(1.000)	(0.992)	(1.000)	(1.001)	(0.949)	(1.000)
Infant Mortality	0.0587	0.137	0.0592	-	-	-
	(0.235)	(0.137)	(0.236)	-	-	-
Excellent Health	-	-	-	0.526	0.536	0.526
	-	-	-	(0.499)	(0.499)	(0.499)
Fraction Twin			0.0198			0.0257
			(0.139)			(0.0158)
Birth Order Twin			4.419			2.196
			(2.448)			(1.064)
Observations	2,167,094	43,771	2,210,865	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 born to the 897,130 mothers responding to any publicly available Demographic and Health Survey or the 88,178 mothers responding to the National Health Interview Survey from 2004 to 2014. 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 birth and country cohort for DHS, and birth quarter cohort for NHIS (mean 0, std deviation 1). Infant mortality refers to the proportion of children who die before 1 year of age. Maternal height is reported in centimetres, and BMI is weight in kilograms over height in metres squared. For a full list of DHS country and years of survey, see appendix table A21.

Table A3: Test of Balance of Observables: Twin versus Non-Twin

	Singleton Family	Twin Family	Diff. (Diff. SE)	t-stat
Panel A: Developing Countries				
Mother Education's	4.823	3.582	-1.241	-49.64
			(0.025)	
Father's Education	6.099	4.985	-1.114	-39.78
			(0.028)	
Mother's BMI	23.305	23.685	0.380	14.62
			(0.026)	
Mother is underweight	0.124	0.100	-0.024	-12.00
			(0.002)	
Mother's Height (cm)	155.611	157.410	1.799	47.34
			(0.038)	
Panel B: USA				
Mother's Education	12.532	12.742	0.210	4.67
			(0.045)	
Father's Education	12.821	12.968	0.146	3.95
			(0.037)	
Mother's BMI	27.652	28.123	0.472	2.51
			(0.188)	
Mother is underweight	0.020	0.016	-0.004	-1.00
			(0.004)	
Mother in Excellent Health	0.323	0.338	0.014	1.56
			(0.009)	

NOTES: Panel A is estimated from DHS data, and panel B uses NHIS data. The first two columns display means, while the third column displays the difference and its standard error, estimated using a two-tailed t-test. Reported t-statistics of the difference are based on an unconditional t-test. Education is measured in years and underweight refers to a BMI<18.5. Full descriptive statistics and definitions are provided in table A2.

Table A4: Effects of maternal health on twin births (conditional results)

Health Behaviours / Access			Health Conditions		
Variable	Estimate	[95% CI]	Variable	Estimate	[95% CI]
Panel A: United States					
Smoked Before Pregnancy	0.225***	[0.174,0.276]	Height	0.560***	[0.533,0.587]
Smoked Trimester 1	-0.070*	[-0.144,0.004]	Underweight	-0.180***	[-0.205,-0.155]
Smoked Trimester 2	-0.176***	[-0.276,-0.076]	Obese	0.102***	[0.073,0.131]
Smoked Trimester 3	-0.171***	[-0.259,-0.083]	Diabetes	-0.263***	[-0.292,-0.234]
Education	0.678***	[0.645,0.711]	Hypertension	-0.205***	[-0.238,-0.172]
Panel B: Sweden					
Smoked (12 weeks)	0.049*	[-0.008,0.106]	Height	0.612***	[0.587,0.637]
Smoked (30-32 weeks)	-0.307***	[-0.352,-0.262]	Underweight	-0.148***	[-0.181,-0.115]
			Obese	-0.082***	[-0.106,-0.058]
			Asthma	-0.005	[-0.023,0.013]
			Diabetes	-0.243***	[-0.268,-0.218]
			Kidney Disease	-0.066***	[-0.088,-0.044]
			Hypertension	-0.082***	[-0.104,-0.060]
Panel C: United Kingdom (Avon)					
Healthy Foods	0.537***	[0.253,0.821]	Height	0.408***	[0.122,0.694]
Fresh Fruit	-0.116	[-0.422,0.190]	Underweight	-0.191	[-0.469,0.087]
Alcohol (Infrequently)	0.069	[-0.251,0.389]	Obese	-0.008	[-0.286,0.270]
Alcohol (Frequently)	-0.398**	[-0.716,-0.080]	Diabetes	-0.065	[-0.337,0.207]
Passive Smoke	0.193	[-0.121,0.507]	Hypertension	-0.479***	[-0.751,-0.207]
Smoked during Pregnancy	-0.165	[-0.475,0.145]			
Education	0.384*	[-0.039,0.807]			
Panel D: Chile					
Smoked during Pregnancy	-0.256**	[-0.501,-0.011]	Underweight	-0.172	[-0.388,0.044]
Drugs (Infrequently)	0.018	[-0.243,0.279]	Obese	-0.256***	[-0.444,-0.068]
Drugs (Frequently)	-0.096***	[-0.141,-0.051]			
Alcohol (Infrequently)	-0.037	[-0.331,0.257]			
Alcohol (Frequently)	-0.115***	[-0.160,-0.070]			
Education	0.486***	[0.155,0.817]			
Panel E: Developing Countries					
Doctor Availability	0.036	[-0.009,0.081]	Height	0.269***	[0.238,0.300]
Nurse Availability	0.045**	[0.006,0.084]	Underweight	-0.085***	[-0.110,-0.060]
Prenatal Care Availability	0.042**	[0.003,0.081]	Obese	0.046***	[0.015,0.077]
Education	0.083***	[0.050,0.116]			

Regressions replicate table 1, however all variables are included in one regression. Refer to additional notes to table 1.

Table A5: Effects of maternal health on twin births (unstandardised variables)

Health Behaviours / Access			Health Conditions		
Variable	Estimate	[95% CI]	Variable	Estimate	[95% CI]
Panel A: United States					
Smoked Before Pregnancy	-0.325***	[-0.409,-0.241]	Height	0.086***	[0.082,0.090]
Smoked Trimester 1	-0.724***	[-0.818,-0.630]	Underweight	-0.644***	[-0.750,-0.538]
Smoked Trimester 2	-0.936***	[-1.036,-0.836]	Obese	0.088**	[0.017,0.159]
Smoked Trimester 3	-0.969***	[-1.069,-0.869]	Diabetes	-3.365***	[-3.718,-3.012]
Education	0.447***	[0.429,0.465]	Hypertension	-1.857***	[-2.141,-1.573]
Panel B: Sweden					
Smoked (12 weeks)	-0.704***	[-0.798,-0.610]	Height	0.099***	[0.095,0.103]
Smoked (30-32 weeks)	-1.030***	[-1.132,-0.928]	Underweight	-0.716***	[-0.887,-0.545]
			Obese	-0.411***	[-0.497,-0.325]
			Asthma	-0.085	[-0.189,0.019]
			Diabetes	-3.737***	[-4.113,-3.361]
			Kidney Disease	-1.359***	[-1.731,-0.987]
			Hypertension	-1.872***	[-2.286,-1.458]
Panel C: United Kingdom (Avon)					
Healthy Foods	1.333***	[0.635,2.031]	Height	0.059***	[0.016,0.102]
Fresh Fruit	0.039	[-0.594,0.672]	Underweight	-0.794	[-2.164,0.576]
Alcohol (Infrequently)	-0.259	[-0.976,0.458]	Obese	-0.218	[-1.512,1.076]
Alcohol (Frequently)	-1.567***	[-2.759,-0.375]	Diabetes	-0.951	[-5.573,3.671]
Passive Smoke	0.096	[-0.494,0.686]	Hypertension	-2.536***	[-3.975,-1.097]
Smoked during Pregnancy	-0.433	[-1.199,0.333]			
Education	0.225*	[0.000,0.450]			
Panel D: Chile					
Smoked during Pregnancy	-1.084***	[-1.894,-0.274]	Underweight	-0.753*	[-1.641,0.135]
Drugs (Infrequently)	0.021	[-3.439,3.481]	Obese	-1.734***	[-3.000,-0.468]
Drugs (Frequently)	-3.053***	[-3.717,-2.389]			
Alcohol (Infrequently)	-0.274	[-1.383,0.835]			
Alcohol (Frequently)	-2.783***	[-3.436,-2.130]			
Education	0.140***	[0.052,0.228]			
Panel E: Developing Countries					
Doctor Availability	0.321***	[0.203,0.439]	Height	0.040***	[0.036,0.044]
Nurse Availability	0.253***	[0.130,0.376]	Underweight	-0.288***	[-0.366,-0.210]
Prenatal Care Availability	0.549***	[0.410,0.688]	Obese	0.196***	[0.098,0.294]
Education	0.033***	[0.025,0.041]			

Regressions replicate table 1, however all variables are unstandardised. Refer to additional notes to table 1.

Table A6: Smoking and birthweight

Dependent Variable: Birthweight	All Births	Non-Twin Births	Twin Births
Smokes 3 Months Prior to Pregnancy	-98.36***	-100.8***	-57.41***
	(1.172)	(1.189)	(5.591)
Smokes Trimester 1	-140.3***	-144.1***	-92.67***
	(1.316)	(1.333)	(6.440)
Smokes Trimester 2	-163.0***	-167.8***	-106.9***
	(1.390)	(1.407)	(6.940)
Smokes Trimester 3	-168.3***	-173.2***	-109.8***
	(1.417)	(1.434)	(7.137)
Average Birthweight	3,283.5	3,311.5	2,369.7
Observations	1,411,556	1,370,368	40,151

Each cell represents a multivariate OLS regression of smoking behaviour on birthweight using the sample of USA birth data used in table 1. All specifications follow those reported in table 1. Smoking in each period is a binary measure, and birthweight is measured in grams.

Table A7: Effects of maternal health on twin births (ART only)

Health Behaviours / Access			Health Conditions		
Variable	Estimate	[95% CI]	Variable	Estimate	[95% CI]
United States: ART Only [N =217,703, % Twin = 35.39]					
Smoked Before Pregnancy	-1.113***	[-1.266,-0.960]	Height	1.880***	[1.719,2.041]
Smoked Trimester 1	-1.175***	[-1.324,-1.026]	Underweight	0.105	[-0.050,0.260]
Smoked Trimester 2	-1.123***	[-1.274,-0.972]	Obese	-2.204***	[-2.365,-2.043]
Smoked Trimester 3	-1.042***	[-1.193,-0.891]	Diabetes	-1.318***	[-1.489,-1.147]
Education	2.426***	[2.254,2.598]	Hypertension	-1.592***	[-1.766,-1.418]

Results are reported following the specifications in table 1, for USA only (where ART usage is observed for all births). The sample period and specification is identical to those in table 1, however now only Artificial Reproductive Technology users are included in the regression.

Table A8: Probability of Giving Birth to Twins

Twin*100	(1) All	(2) Income	(3) Income	(4) Time	(5) Time	(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 A9: 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 A10: 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 A11: Fetal Deaths, Twinning, and Health Behaviours (Conditional)

	(1) Smokes	(2) Drinks	(3) No College	(4) Anemic	(5) N Cigs	(6) N Drinks	(7) Years Educ
Twin	9.648***	10.080***	8.621***	10.972***	9.679***	10.063***	19.150***
	[0.199]	[0.194]	[0.222]	[0.191]	[0.197]	[0.194]	[0.946]
Health (Dis)amenity	0.867***	4.219***	1.229***	0.337**	0.071***	0.525***	-0.122***
	[0.072]	[0.346]	[0.044]	[0.133]	[0.006]	[0.082]	[0.008]
Twin × Health	0.856	3.322	3.548***	1.307	0.040	0.755	-0.667***
	[0.721]	[3.526]	[0.369]	[1.093]	[0.058]	[0.890]	[0.067]
Constant	4.717***	4.909***	4.152***	5.276***	4.727***	4.926***	6.311***
	[0.051]	[0.051]	[0.048]	[0.049]	[0.051]	[0.051]	[0.117]
Observations	13,660,400	13,809,830	15,909,836	16,158,564	13,679,142	13,828,573	15,909,836

Each column represents a regression of whether a birth ends in a fetal death (multiplied by 1,000) on twins, a health behaviour or health stock, and the interaction between twins and the health variable. The health variable in each column is indicated in the column title. Each regression also controls for mother's age fixed effects, total number of mother's birth, and the year of birth. Coefficients from the regression are reported, and heteroscedasticity robust standard errors are displayed in parentheses. ***p-value<0.01, **p-value<0.05, *p-value<0.01.

Table A12: 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	Estimates 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 A13: OLS Estimates with and without Birth Order Controls (Pooled DHS Data)

	No Birth Order FEs			Birth Order FEs			
	(1) Base	(2) +S	(3) +S+H	(4) No Fertility	(5) Base	(6) +S	(7) +S+H
Total number of children in the family	-0.117***	-0.101***	-0.067***		-0.128***	-0.108***	-0.072***
	[0.001]	[0.001]	[0.001]		[0.001]	[0.001]	[0.001]
Birth Order 2				-0.175***	-0.057***	-0.062***	-0.040***
				[0.004]	[0.004]	[0.003]	[0.003]
Birth Order 3				-0.352***	-0.099***	-0.109***	-0.071***
				[0.005]	[0.005]	[0.005]	[0.005]
Birth Order 4				-0.493***	-0.099***	-0.117***	-0.075***
				[0.006]	[0.006]	[0.006]	[0.006]
Birth Order 5				-0.596***	-0.061***	-0.088***	-0.057***
				[0.007]	[0.007]	[0.007]	[0.007]
Birth Order 6				-0.687***	-0.017*	-0.056***	-0.043***
				[0.008]	[0.009]	[0.009]	[0.008]
Birth Order 7				-0.749***	0.051***	-0.005	-0.014
				[0.009]	[0.010]	[0.010]	[0.010]
Birth Order 8				-0.784***	0.140***	0.068***	0.033***
				[0.010]	[0.012]	[0.012]	[0.011]
Birth Order 9				-0.838***	0.207***	0.114***	0.055***
				[0.012]	[0.014]	[0.014]	[0.013]
Birth Order $\geq$10				-0.855***	0.396***	0.270***	0.164***
				[0.014]	[0.016]	[0.016]	[0.015]
Observations	1128703	1128703	1128703	1128703	1128703	1128703	1128703

Table A14: OLS Estimates with and without Birth Order Controls (USA)

	No Birth Order FEs			Birth Order FEs			
	(1) Base	(2) +S	(3) +S+H	(4) No Fertility	(5) Base	(6) +S	(7) +S+H
Total Number of Children in the family	-0.030***	-0.028***	-0.024***		-0.038***	-0.025***	-0.022***
	(0.004)	(0.004)	(0.004)		(0.004)	(0.004)	(0.004)
Birth Order 2				0.004	0.023***	-0.032***	-0.032***
				(0.006)	(0.006)	(0.008)	(0.008)
Birth Order 3				-0.013	0.040***	-0.057***	-0.057***
				(0.010)	(0.010)	(0.015)	(0.015)
Birth Order 4				-0.008	0.083***	-0.047*	-0.043*
				(0.019)	(0.020)	(0.025)	(0.025)
Birth Order 5				0.041	0.177***	0.018	0.024
				(0.041)	(0.041)	(0.045)	(0.045)
Birth Order 6				-0.036	0.147*	-0.048	-0.035
				(0.080)	(0.079)	(0.081)	(0.081)
Birth Order 7				-0.063	0.162	-0.052	-0.036
				(0.156)	(0.157)	(0.156)	(0.158)
Birth Order 8				0.203	0.461	0.207	0.231
				(0.501)	(0.500)	(0.498)	(0.498)
Birth Order 9				-0.225**	0.065	-0.221*	-0.177
				(0.107)	(0.116)	(0.124)	(0.135)
Birth Order $\geq$10				-0.155***	0.193***	-0.147***	-0.111
				(0.048)	(0.044)	(0.054)	(0.071)
Observations	163931	163931	163931	163931	163931	163931	163931

Table A15: Developing Country Estimates: OLS, Bounds, and IV (Girls Only)

	2+			3+			4+		
	Base	+H	+S&H	Base	+H	+S&H	Base	+H	+S&H
Panel A: OLS Results									
Fertility	-0.168***	-0.143***	-0.095***	-0.159***	-0.134***	-0.088***	-0.141***	-0.117***	-0.077***
	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	127290	127290	127290	196489	196489	196489	204420	204420	204420
R-squared	0.13	0.16	0.22	0.11	0.14	0.22	0.10	0.14	0.22
Altonji et al. Ratio		5.72	1.301		5.36	1.239		4.875	1.203
Oster Ratio		1.003	0.559		1.319	0.650		1.397	0.679
Panel B: IV Results									
Fertility	-0.023	-0.040	-0.032	-0.046	-0.056*	-0.058**	-0.031	-0.048*	-0.053**
	(0.038)	(0.037)	(0.036)	(0.030)	(0.029)	(0.026)	(0.029)	(0.028)	(0.024)
Observations	127290	127290	127290	196489	196489	196489	204420	204420	204420
R-Squared	0.05	0.10	0.18	0.05	0.09	0.18	0.03	0.09	0.18
Panel C: First Stage									
Twins	0.813***	0.860***	0.874***	0.792***	0.826***	0.833***	0.797***	0.822***	0.831***
	(0.043)	(0.041)	(0.041)	(0.034)	(0.033)	(0.033)	(0.032)	(0.032)	(0.032)
Observations	127290	127290	127290	196489	196489	196489	204420	204420	204420
Kleibergen-Paap rk statistic	358.80	436.48	460.51	550.68	622.29	650.60	608.78	675.38	695.44
p-value of rk statistic	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Refer to notes to table 7 in the paper.*p<0.1; **p<0.05; ***p<0.01

Table A16: Developing Country Estimates: OLS, Bounds, and IV (Boys Only)

	2+			3+			4+		
	Base	+H	+S&H	Base	+H	+S&H	Base	+H	+S&H
Panel A: OLS Results									
Fertility	-0.136***	-0.116***	-0.074***	-0.126***	-0.105***	-0.066***	-0.107***	-0.089***	-0.055***
	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	132676	132676	132676	199232	199232	199232	205195	205195	205195
R-squared	0.09	0.11	0.17	0.08	0.10	0.17	0.07	0.10	0.17
Altonji et al. Ratio		5.8	1.194		5.0	1.1		4.944	1.058
Oster Ratio		0.790	0.503		0.944	0.551		1.107	0.567
Panel B: IV Results									
Fertility	0.016	0.010	0.006	-0.009	-0.024	-0.032	-0.022	-0.026	-0.021
	(0.040)	(0.039)	(0.037)	(0.029)	(0.028)	(0.026)	(0.028)	(0.026)	(0.024)
Observations	132676	132676	132676	199232	199232	199232	205195	205195	205195
R-Squared	0.02	0.06	0.13	0.03	0.06	0.14	0.02	0.06	0.13
Panel C: First Stage									
Twins	0.761***	0.796***	0.802***	0.805***	0.823***	0.835***	0.886***	0.896***	0.898***
	(0.042)	(0.041)	(0.040)	(0.034)	(0.033)	(0.032)	(0.033)	(0.033)	(0.033)
Observations	132676	132676	132676	199232	199232	199232	205195	205195	205195
Kleibergen-Paap rk statistic	321.23	383.23	404.98	576.75	637.37	677.83	706.41	744.47	724.44
p-value of rk statistic	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Refer to notes to table 7 in the paper.*p<0.1; **p<0.05; ***p<0.01

Table A17: US Estimates: OLS and IV (Female Only)

	2+			3+			4+		
	Base	+H	+S&H	Base	+H	+S&H	Base	+H	+S&H
Panel A: OLS Results									
School Z-Score	-0.040***	-0.030***	-0.023***	-0.043***	-0.036***	-0.029***	-0.037***	-0.031**	-0.024*
	(0.007)	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.013)	(0.013)	(0.013)
Altonji et al. Ratio		3.040	1.353		5.231	2.116		5.082	1.832
Oster Ratio		1.322	0.853		1.756	1.101		1.518	0.828
Excellent Health	-0.015***	-0.009***	-0.007**	-0.018***	-0.009***	-0.007**	-0.033***	-0.024***	-0.022***
	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.006)	(0.004)	(0.004)
Altonji et al. Ratio		1.475	0.864		1.033	0.631		5.379	1.937
Oster Ratio		3.269	1.804		2.600	1.543		8.581	5.484
Panel B: IV Results									
School Z-Score	0.071	0.057	0.050	0.098	0.081	0.080	-0.149	-0.158	-0.162
	(0.075)	(0.070)	(0.069)	(0.107)	(0.104)	(0.104)	(0.137)	(0.136)	(0.135)
Excellent Health	0.011	0.022	0.019	0.038	-0.008	-0.008	-0.016	-0.034	-0.040
	(0.039)	(0.032)	(0.032)	(0.054)	(0.045)	(0.045)	(0.084)	(0.067)	(0.066)
Panel C: First Stage									
School Z-Score	0.624***	0.654***	0.662***	0.668***	0.676***	0.679***	0.766***	0.770***	0.777***
	(0.037)	(0.037)	(0.036)	(0.045)	(0.044)	(0.044)	(0.090)	(0.090)	(0.087)
Excellent Health	0.680***	0.713***	0.721***	0.692***	0.699***	0.703***	0.761***	0.770***	0.778***
	(0.037)	(0.036)	(0.036)	(0.042)	(0.041)	(0.041)	(0.085)	(0.085)	(0.082)
Kleibergen-Paap rk statistic	343.92	385.88	396.49	269.67	288.08	287.74	80.77	82.55	89.86
p-value of rk statistic	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Observations (Education)	29900	29900	29900	23318	23318	23318	10720	10720	10720
Observations (Health)	34380	34380	34380	26368	26368	26368	12209	12209	12209

NOTES: Refer to table 8. *p<0.1; **p<0.05; ***p<0.01

Table A18: US Estimates: OLS and IV (Male Only)

	2+			3+			4+		
	Base	+H	+S&H	Base	+H	+S&H	Base	+H	+S&H
Panel A: OLS Results									
School Z-Score	-0.048***	-0.035***	-0.028***	-0.039***	-0.029**	-0.022*	-0.007	-0.001	0.005
	(0.008)	(0.009)	(0.009)	(0.011)	(0.011)	(0.011)	(0.021)	(0.022)	(0.022)
Altonji et al. Ratio		2.728	1.417		2.843	1.306		0.186	-0.067
Oster Ratio		1.128	0.843	0.554	0.470		0.011	-0.092	
Excellent Health	-0.007**	-0.001	0.001	-0.015***	-0.008**	-0.007*	-0.024***	-0.014***	-0.013***
	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.006)	(0.005)	(0.005)
Altonji et al. Ratio		0.183	0.447		1.127	0.888		1.396	1.180
Oster Ratio		0.186	-0.065		3.359	1.930		3.071	2.477
Panel B: IV Results									
School Z-Score	-0.227**	-0.233**	-0.233**	-0.090	-0.095	-0.096	-0.137	-0.137	-0.145
	(0.105)	(0.094)	(0.095)	(0.079)	(0.080)	(0.080)	(0.228)	(0.237)	(0.231)
Excellent Health	0.033	0.036	0.035	-0.087*	-0.088**	-0.087**	0.067	-0.015	-0.019
	(0.037)	(0.028)	(0.028)	(0.046)	(0.038)	(0.038)	(0.063)	(0.062)	(0.061)
Panel C: First Stage									
School Z-Score	0.678***	0.740***	0.739***	0.794***	0.796***	0.798***	0.822***	0.821***	0.842***
	(0.036)	(0.037)	(0.036)	(0.072)	(0.071)	(0.071)	(0.097)	(0.096)	(0.096)
Excellent Health	0.702***	0.764***	0.763***	0.806***	0.807***	0.808***	0.824***	0.825***	0.845***
	(0.032)	(0.033)	(0.033)	(0.067)	(0.066)	(0.066)	(0.092)	(0.091)	(0.091)
Kleibergen-Paap rk statistic	349.79	400.61	415.24	121.88	125.85	126.29	72.46	73.26	76.84
p-value of rk statistic	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Observations (Education)	31367	31367	31367	23990	23990	23990	10632	10632	10632
Observations (Health)	35897	35897	35897	27025	27025	27025	12149	12149	12149

NOTES: Refer to table 8. *p<0.1; **p<0.05; ***p<0.01

Table A19: Developing Country Estimates: OLS, Bounds, and IV (Middle Income Countries)

	2+			3+			4+		
	Base	+H	+S&H	Base	+H	+S&H	Base	+H	+S&H
Panel A: OLS Results									
Fertility	-0.145***	-0.129***	-0.085***	-0.152***	-0.133***	-0.091***	-0.145***	-0.124***	-0.085***
	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	102175	102175	102175	148374	148374	148374	144950	144950	144950
R-squared	0.11	0.12	0.17	0.10	0.11	0.17	0.10	0.12	0.18
Altonji et al. Ratio		8.063	1.417		7.0	1.492		5.905	1.417
Oster Ratio		0.747	0.518		1.127	0.654		1.429	0.728
Panel B: IV Results									
Fertility	-0.068	-0.077	-0.076	-0.064*	-0.075**	-0.064*	-0.043	-0.056	-0.060*
	(0.051)	(0.050)	(0.048)	(0.037)	(0.036)	(0.035)	(0.040)	(0.038)	(0.034)
Observations	102175	102175	102175	148374	148374	148374	144950	144950	144950
R-Squared	0.06	0.08	0.13	0.05	0.07	0.14	0.04	0.07	0.14
Panel C: First Stage									
Twins	0.734***	0.780***	0.794***	0.758***	0.811***	0.825***	0.784***	0.829***	0.844***
	(0.049)	(0.046)	(0.045)	(0.045)	(0.044)	(0.043)	(0.046)	(0.044)	(0.043)
Observations	102175	102175	102175	148374	148374	148374	144950	144950	144950
Kleibergen-Paap rk statistic	221.25	292.66	317.88	285.22	345.36	374.11	291.16	353.89	384.99
p-value of rk statistic	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Refer to notes to table 7 in the paper.*p<0.1; **p<0.05; ***p<0.01

Table A20: Developing Country Estimates: OLS, Bounds, and IV (Low Income Countries)

	2+			3+			4+		
	Base	+H	+S&H	Base	+H	+S&H	Base	+H	+S&H
Panel A: OLS Results									
Fertility	-0.157***	-0.128***	-0.083***	-0.138***	-0.112***	-0.070***	-0.114***	-0.094***	-0.058***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	157791	157791	157791	247347	247347	247347	264665	264665	264665
R-squared	0.12	0.15	0.22	0.09	0.13	0.21	0.08	0.12	0.20
Altonji et al. Ratio		4.414	1.122		4.308	1.029		4.7	1.036
Oster Ratio		0.938	0.523		1.194	0.570		1.417	0.605
Panel B: IV Results									
Fertility	0.027	0.014	0.023	-0.009	-0.024	-0.037	-0.018	-0.027	-0.025
	(0.032)	(0.031)	(0.030)	(0.028)	(0.026)	(0.024)	(0.027)	(0.024)	(0.022)
Observations	157791	157791	157791	247347	247347	247347	264665	264665	264665
R-Squared	0.02	0.08	0.17	0.03	0.08	0.18	0.02	0.08	0.17
Panel C: First Stage									
Twins	0.836***	0.862***	0.871***	0.816***	0.828***	0.835***	0.870***	0.874***	0.873***
	(0.038)	(0.037)	(0.036)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.033)
Observations	157791	157791	157791	247347	247347	247347	264665	264665	264665
Kleibergen-Paap rk statistic	493.50	548.01	574.50	640.83	679.76	694.71	731.27	753.29	699.75
p-value of rk statistic	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Refer to notes to table 7 in the paper.*p<0.1; **p<0.05; ***p<0.01

Table A21: Full Survey Countries and Years (DHS)

		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: 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.

C Data Appendix

Additional data details are provided in two parts: the first (subsection C.1) describes the various databases recording twin births and maternal health using in Part 1 of the paper. The second, (subsections C.2 and C.3) describes the datasets which are used to produce main IV and OLS results of the QQ trade-off. These are based on DHS and NHIS data described briefly in section 2 and further below. 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.

C.1 Regressions of Twinning on Maternal Health

In regressions in Part 1 of the paper that examine the characterstics of mothers and the relationship these characteristics and twin births and miscarriage, we consult a number of other datasets as described in the data section of the paper. These are the following:

- United States National Vital Statistics Birth Data
- United States National Vital Statistics Foetal Death Data
- Spanish Vital Statistics (INE)
- The Swedish Medical Birth Register
- Longitudinal Early Life Survey, Chile (ELPI)
- The Avon Longitudinal Study of Parents and Children (ALSPAC)

In the case of the first 4 datasets (administrative records of births and/or fetal deaths), we use all recorded instances for mothers aged 18-49, focusing on twins as our outcome variable of interest. The only exception is the United States Vital Statistics data, in which case we observe Artificial Reproductive Technology (ART) use, and remove the 1.6% of ART users from the estimation sample. 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 table 1. Finally, for the ALSPAC survey follows mothers and their children who were born in the early 1990's in the county of Avon, UK. We use all mothers from the principal wave of enrolled children.

C.2 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 the online appendix (table A21). 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.[1]

C.3 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 of Columbia each year.[2] 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.

[1] 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.

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

D Resampling and Simulation Based Estimation of γ

D.1 Bootstrap Confidence Intervals

The methodology to estimate γ in equations (10) and (12) is described in section 1.4 of the paper. In the case of Conley et al.'s UCI approach, this estimate is then sufficient to produce bounds on β_1, assuming that: $\gamma \in [0, 2\hat{\gamma}]$. 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. In the case of the more precise LTZ approach (our preferred method) 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. 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 known 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 to 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 $\ln\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$.

D.2 Simulation-Based Estimation for non-Normal Distributions

Estimates of the full distribution of γ are presented in Figures 7a and 7b. These are the estimated $\hat{\gamma}_j$ from $j \in \{1, \ldots, 100\}$ bootstrap replications for γ in Nigeria and the United States. In all cases, when the underlying empirical distribution is tested for equality against the overlaid analytical distribution (uniform, normal, log normal, χ^2), the normal distribution provides the best fit of the analytical with the empirical distribution.[3]

However, the underlying distribution appears to not be perfectly normal, and it appears doubtful that this would be the case asymptotically. Fortunately, Conley et al. (2012) describe a simulation-based estimation method to calculate γ in the case of a non-normal distribution for γ. We have followed this methodology using the empirical distribution calculated bootstrapping for γ. This code has been publicly released as `plausexog` for Stata (Clarke, 2014). The simulation-based estimation procedure is described fully in Conley et al. (2012) p. 265 as a five step algorithm. The procedure consists of taking repeated draws from the variance-covariance matrix estimated using IV with the plausibly exogenous instrument, and in each case adding to it a draw from the distribution of γ, scaled by a quantity which depends on the strength of the instrument. Conley et al. refer to the underlying distibution of γ as F, and the scale parameter as A, where $A = (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z)$. These repeated draws then lead to a large number of estimates for β, the parameter of interest, and a 95% confidence interval is taken by forming $[\hat{\beta} - c_{1-\alpha/2}, \hat{\beta} + c_{\alpha/2}]$, where c are percentiles of the distribution of simulated estimates.

Thus, as well as estimating the LTZ case where we assume that γ is distributed $\sim \mathcal{N}(\mu_{\hat{\gamma}}, \sigma^2_{\hat{\gamma}})$, we can estimate a version fully utilizing the bootstrapped distribution of $\hat{\gamma}$ described in the previous sub-section. In this case, we use as F, the distribution of γ, the empirically estimated distribution of γ. The simulation based algorithm then consists of taking $b \in 1, \ldots, B$ draws from the empirically estimated F, as well as B draws from the

[3] In the US, We cannot reject that γ is normal with a p-value of 0.782. In this case, although we can't reject that γ is log normal, the p-value is much lower, at 0.203. Values for Nigeria suggest a quantitatively similar result.

variance-covariance matrix, and defining the 95% confidence interval based on the 2.5 and 97.5% quintiles of the resulting simulated values for β.

References

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.

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.

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

D. Clarke. PLAUSEXOG: Stata module to implement Conley et al's plausibly exogenous bounds. Statistical Software Components, Boston College Department of Economics, May 2014.

T. G. Conley, C. B. Hansen, and P. E. Rossi. Plausibly Exogenous. *The Review of Economics and Statistics*, 94(1):260–272, February 2012.

E. Fitzsimons and B. Malde. Empirically probing the quantity-quality model. *Journal of Population Economics*, 27(1):33–68, Jan 2014.

D. S. Lee. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102, 07 2009.

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.

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.

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.