I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11878

The Twin Instrument: Fertility and Human Capital Investment

Sonia Bhalotra
Damian Clarke

OCTOBER 2018

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11878

The Twin Instrument: Fertility and Human Capital Investment

Sonia Bhalotra
University of Essex and IZA

Damian Clarke
Universidad de Santiago de Chile and IZA

OCTOBER 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Twin Instrument: Fertility and Human Capital Investment[*]

Twin births are often used to instrument fertility to address (negative) selection of women into fertility. However recent work shows positive selection of women into twin birth. Thus, while OLS estimates will tend to be downward biased, twin-IV estimates will tend to be upward biased. This is pertinent given the emerging consensus that fertility has limited impacts on women's labour supply, or on investments in children. Using data for developing countries and the United States, we demonstrate the nature and size of the bias in the twin-IV estimator of the quantity-quality trade-off and estimate bounds on the true parameter.

JEL Classification: J12, J13, C13, D13, I12

Keywords: twins, fertility, maternal health, quantity-quality trade-off, parental investment, bounds, IV

Corresponding author:
Sonia Bhalotra
ISER & Department of Economics
University of Essex
Wivenhoe Park
Colchster CO4 3SQ
United Kingdom
E-mail: srbhal@essex.ac.uk

[*] We are grateful to Paul Devereux, James Fenske, Judith Hall, Christian Hansen, Martin Karlsson, Toru Kitagawa, Magne Mogstad, Rohini Pande, Adam Rosen, Paul Schulz, Margaret Stevens, Atheen Venkataramani and Marcos Vera-Hernandez, along with various seminar audiences and discussants for helpful comments and/or sharing data. Any remaining errors are our own. An earlier version of this paper appeared as "The Twin Instrument", IZA DP 10405, December 2016. A revised version of Part 1 of this working paper is forthcoming at The Review of Economics and Statistics. This paper contains a revised version of Part 2 of the working paper.

1 Introduction

Following Becker (1960), fertility has been modeled jointly with investments in children and with women's labour force participation. In line with the average tendency for negative selection into fertility, linear least squares estimates of associations of fertility with children's human capital, and with women's employment tend to be upward biased. Since the pioneering work of Rosenzweig and Wolpin (1980a,b), a considerable literature has attempted to address selection by using twins to instrument fertility (see Appendix Table A1). The premise is that twin births are quasi-random, so that the event of a twin birth constitutes a "natural" natural experiment (Rosenzweig and Wolpin, 2000).

In a recent paper (Bhalotra and Clarke, forthcoming), we presented new population-level evidence that challenges this premise. Using individual data for 17 million births in 72 countries, we demonstrated that indicators of the mother's health, her health-related behaviours, and the prenatal health environment are systematically positively associated with the probability of a twin birth. The estimated associations are large, evident in richer and poorer countries, evident even among women who do not use IVF, and hold for sixteen different measures of health. We provided evidence that selective miscarriage is the likely mechanism. The upshot of our findings is that women who have twin births are positively selected on unobservables related to health. If, as is plausible (and we will demonstrate), those unobservables are correlated with child human capital or with women's labour force participation, then twin-instrumented estimates of the relationship between fertility and child outcomes, or women's labour supply will tend to be upward biased, moving towards a null-estimate.

This is pertinent as it could resolve the ambiguity of the available evidence on these relationships. Recent studies using the twin instrument reject the presence of a quantity–quality (QQ) fertility trade-off (Black et al., 2005; Angrist et al., 2010), challenging a long-standing theoretical prior of Becker (1960); Becker and Lewis (1973); Becker and Tomes (1976). Similarly, research using the twin instrument finds that additional children have relatively little influence on women's labour market participation, at least after the first few years (Rosenzweig and Wolpin, 1980a; Bronars and Grogger, 1994; Jacobsen et al., 1999; Vere, 2011). In principle, addressing the omission of maternal health related variables could adjust for the downward bias in these studies, and provide a true estimate of the trade-offs. In practice, maternal health is multi-dimensional and almost impossible to fully measure and adjust for. To take a few examples, foetal health is potentially a function of whether pregnant women skip breakfast (Mazumder and Seeskin, 2015), whether they suffer bereavement in pregnancy (Black et al., 2016), and fetal exposure to air pollution (Chay and Greenstone, 2003).

In this paper we investigate how inference in a literature concerned with causal effects of fertility on human capital can proceed with partial adjustment and bounding. We first illustrate the hypothesized direction of the bias of the twin-IV estimator, by introducing available controls for maternal health in the estimation. Since this adjustment is necessarily partial, we proceed to estimate bounds on the IV estimates. Given that the first stage (twins predicting fertility) is powerful, we follow Conley et al. (2012) in estimating bounds on the premise that twin births are plausibly if not strictly exogenous. In a sensitivity check, we also estimate bounds under the different assumptions of Nevo and Rosen (2012), again using twin births as an "imperfect instrumental variable".

We provide estimates for the US using about 225,000 births, drawn from the US 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, or DHS. 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 relationship of interest, as well as the violation of the exclusion restriction that concerns us, are different in richer vs. poorer countries.

We start by briefly demonstrating, on the particular data samples used in this analysis, our earlier result that the probability of twin birth is significantly positively associated with indicators of maternal health. We then set the stage by showing the routine OLS and twin-IV estimates on our data samples. The OLS estimates suggest a fertility-human capital trade-off and, following (Altonji et al. (2005)) to gauge the importance of unobservables, we conclude that accounting for unobservables is unlikely to dissolve the trade-off. The twin-IV estimates replicate, in our samples, the finding in recent studies that there is no discernible trade-off. However, adjusting for available maternal health related characteristics, even though these are only a small subset of the range of relevant indicators, leads to emergence of a QQ trade-off. This finding generalizes to recent non-linear models of the QQ trade-off (Brinch et al., 2017; Mogstad and Wiswall, 2016), holding even when the impact of fertility is allowed to vary by parity. For instance, in samples with at least three births, an additional child is associated with lower human capital outcomes for the first two births: this is estimated as 0.05 s.d. for years of education in developing countries, and 0.06 s.d. for an index of child health in the US, and in the sample with at least two births it is 0.10 s.d. for grade progression in the US (or 0.22 fewer grades progressed).

The bounds also confirm the presence of a trade-off at certain parities for education or health outcomes.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 -0.10 of a s.d. for child health in the USA.[1] 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. This said, the US sample is considerably smaller than the developing country sample and bounds are correspondingly wider. As a result, bounds are uniformly more informative in the developing country sample.

[1] Each is a range because the coefficient varies with whether twins occur at the second, third or fourth birth order. We place these effect sizes in perspective in section 5.2.4.

The results indicate that marginal increases in fertility often lead to diminished investments in the human capital of children, and the trade-off is not negligibly small. This is important, especially in view of growing evidence of the long run dynamic benefits of childhood investments (Heckman et al., 2013). These estimates put 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.[2] Moreover, advocates of policies encouraging smaller families rest their case on larger families investing less in the quality of each child, limiting human capital accumulation and living standards (Galor and Weil, 2000; Moav, 2005).

2 The Fertility-Investment Trade-off and the Twin Instrument

A long-standing theoretical result in the literature on human capital formation is the existence of a quantity-quality (QQ) trade-off (Becker, 1960; Becker and Lewis, 1973; Willis, 1973; De Tray, 1973; Becker and Tomes, 1976). The essential idea of these studies is that the shadow price of child quality is increasing in child quantity and vice versa. This provides behavioural micro-foundations consistent with an empirical regularity that has been noted in cross-sectional and time series data, which is that children from large families have weaker educational outcomes (Hanushek, 1992; Blake, 1989; Galor, 2012). We replicate this pattern using our data samples from the USA and developing countries (see Appendix Figures A1 and A2).

However, empirical evidence of a QQ trade-off is ambiguous. Early work including Hanushek (1992) and Rosenzweig and Wolpin (1980a) documented significant negative effects of addi-

[2] 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.

tional births within a family on average child educational outcomes. Using IV or difference-in-differences approaches, recent studies include estimates of a significantly positive relationship (Qian, 2009), a significant negative relationship (Grawe, 2008; Ponczek and Souza, 2012; Lee, 2008; Bougma et al., 2015) and no significant relationship (Black et al., 2005; Angrist et al., 2010; Fitzsimons and Malde, 2010), see the review in Clarke (2018). 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., 2017; Mogstad and Wiswall, 2016). In this paper, we assess twin-IV estimates on two different data samples, examining sensitivity to adjustment for maternal health in linear and non-linear models, and to (small or signed) violations of the exclusion restriction.[3] In this paper we focus nearly exclusively on the internal validity of twins estimates (IV consistency). In recent work, Aaronson et al. (2017); Bisbee et al. (2017) examine the external validity of the twin instrumented or sex-mix instrumented estimates of the impact of fertility on female labour supply.[4]

3 Methodology

3.1 Estimating The Quantity–Quality Trade-off with Twins

Analyses of the QQ trade-off attempt to produce consistent estimates of α_1 in the following population-level equation:

$$quality_{ij} = \alpha_0 + \alpha_1 quantity_j + \boldsymbol{X\alpha_x} + \varepsilon_{ij}. \tag{1}$$

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

[3] The twin instrument has also been used to estimate varying 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).

[4] We note that like Aaronson et al. (2017), our estimates suggest considerable heterogeneity by country income levels. We also observe heterogeneity by child gender.

that $\alpha_1 < 0$. Relevant family and child level controls are included, denoted $\boldsymbol{X}$. As has been 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: $E[quantity_j \cdot \varepsilon_{ij} | \boldsymbol{X}]$. If mothers with weaker preferences for child quality have more children, OLS estimates will overstate the true QQ trade-off.

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. The 2SLS specification can be written as:

$$quantity_j = \pi_0 + \pi_1 twin_j + \boldsymbol{X}\boldsymbol{\pi_x} + \nu_{ij}, \tag{2a}$$

$$quality_{ij} = \beta_0 + \beta_1 \widehat{quantity}_j + \boldsymbol{X}\boldsymbol{\beta_x} + \eta_{ij}. \tag{2b}$$

where $twin_j$ is an indicator for whether the n^{th} birth in family j is a twin birth. As described further in section 4, 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 (subjects) are randomly assigned either one sibling (and make up the control group) or two siblings (and make up the treatment group) at the n^{th} 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.[5] If twins are a valid instrument, the parameter β_1 is consistent and hence in limit equal to the parameter α_1 from the population equation 1.

[5] 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 US, 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. We document similar endowment differences in our data samples (Appendix Figure A3 and A4).

Bhalotra and Clarke (forthcoming) provide evidence that omitted variables for maternal health may contaminate η_{ij}, and in Section 5.1 we document this for the data used in this paper. If mothers who invest more in their pregnancies (for instance by averting smoking before birth) also invest more in their children after birth, then the twin-IV estimates will be inconsistent. There is some evidence for instance in Uggla and Mace (2016) that healthier mothers (indicated by health measures such as used in our earlier work) invest more in children in a range of domains. Positive selection of mothers of twins implies:

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

We can partition the stochastic error term from equation 2b into a vector of observable measures of mother's health capital ($\boldsymbol{H}$), socioeconomic variables ($\boldsymbol{S}$), and all other unobserved components, as $\eta_{ij} = \boldsymbol{H} + \boldsymbol{S} + \eta^*_{ij}$. Assuming a positive (or zero) covariance between the three components of the error term,[6] the step-by-step removal of selection predictors will result in the estimated coefficient becoming continually closer to the true parameter. Thus:

$$\text{plim}\,\hat{\beta}_1 > \text{plim}\,\hat{\beta}_1^S > \text{plim}\,\hat{\beta}_1^{S+H} > \beta_1. \tag{3}$$

The coefficients $\hat{\beta}_1^H$ and $\hat{\beta}_1^{S+H}$ refer to coefficients in a model augmented to control for observable health capital $\boldsymbol{H}$, and then also observable socioeconomic status $\boldsymbol{S}$. Since, as discussed further in section 5.1, all determinants of twin birth are virtually impossible to account for, twin-IV will under-estimate the magnitude of the QQ trade-off, although addition of predictors of twins as controls will lead to the estimate approaching the true value from above.

3.2 Estimating IV Bounds with an Imperfect Instrument

Given that we can never fully control for maternal health even with the full set of *observable* controls, point estimation of the QQ trade-off is not possible. However, under additional

[6] Given that the covariance between elements of $\boldsymbol{S}$ and $\boldsymbol{H}$ is found 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, it is very likely that each covariance term is positive. This is tested later in this paper when examining IV estimates.

assumptions relating to the failure of the IV exclusion restriction, or correlations between the IV, the endogenous variable, and unobservables we *can* bound the QQ trade-off. In order to proceed based on IV even in the presence of twin selection, we follow two procedures to bound the QQ trade-off using the (potentially imperfect) twin instrument.

The first of these is the Nevo and Rosen (2012) "Imperfect IV" procedure. This procedure is ideally suited to the context examined here, as it suggests that if twins are positively selected and if fertility is negatively selected, and if twinning and fertility are positively correlated, then the true parameter will be bounded by the OLS and the IV estimate discussed above.[7] If we are willing to additionally assume that the twin instrument is "less endogenous" than fertility (Nevo and Rosen's assumption 4), we can further tighten the bounds by forming a compound instrument based on the endogenous fertility variable, and the imperfect twin instrument. This instrument, ($V = \sigma_{quantity}Twin_j - \sigma_{Twin}quantity_j$), where σ refers to the standard deviation, can provide tighter bounds on the β_1 parameter where $\widehat{\beta}^{V}_{IV} \leq \beta_1 \leq \widehat{\beta}^{twin}_{IV}$, suggesting end points for a series of IV bounds on the parameter β_1.

The Nevo and Rosen (2012) procedure is straightforward and relies on quite weak assumptions. Namely, to produce bounds in this case, the only additional assumptions we require are that there is negative selection into fertility, a widely accepted stance in the literature (Qian, 2009), and one that is verified in surveys querying fertility preferences, which show that less educated women desire more children (e.g. Bhalotra and Cochrane (2010)); that twins are positively selected, which is shown in Bhalotra and Clarke (forthcoming), that twin births are positively associated with fertility, which we show in the first stage regressions below, and that there is less selection into twin birth than into fertility, which seems reasonable.

The upper bound in the case of Nevo and Rosen is the upper end of the 95% confidence

[7] We can follow the notation of Nevo and Rosen (2012) precisely if we multiply twins by -1, as their assumptions and lemmas are based on identically signed correlations between the endogenous variable and unobservables, and the IV and unobservables. In our case, once twins is multiplied by -1, assumption 3 is met assuming negative fertility selection and positive twin selection: $\rho_{xu}\rho_{zu} \leq 0$, where ρ denotes correlation. In the notation of our paper, x refers to *quantity* in equation 1, z refers to *twin* in equation 2a, u refers to the unobservable stochastic term ε_{ij} in 1. Then, under Nevo and Rosen (2012, Lemma 1), $\sigma_{xz} < 0$, or the negative of twins and fertility will be negatively correlated, and as such $\widehat{\beta}^{twin}_{IV} \leq \beta_1 \leq \widehat{\beta}_{OLS}$.

interval on the original twin IV estimate $\widehat{\beta}_{IV}^{twin}$. From equation 3 we know that positive selection of twins inflates this IV estimate upwards. As such, to offer a more informative identification region at the upper bound, we also implement 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. 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):

$$quality_{ij} = \delta_0 + \delta_1 quantity_j + \gamma twin_j + \boldsymbol{X\delta_x} + \vartheta_{ij}. \tag{4}$$

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. Rather than assuming this holds exactly, one can define plausible exogeneity as a situation in which γ is nearly, but not precisely equal to zero. Estimating or imposing some (weaker) restriction on γ buys the identifying information to bound the parameter of interest, even when the IV exclusion restriction does not hold exactly.[8]

Conley et al.'s methods are ideally suited to the empirical application of this paper because they show that their bounds are most informative when the instruments are strong, and the twin instrument is strong (evidence below). In section 5.1, 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.[9]

[8] Conley et al. (2012) state that "Manski and Pepper (2000) consider treatment effect bounds with instruments that are assumed to monotonically impact conditional expectations, which is roughly analogous to assuming $\gamma \in [0, \infty]$". The procedure we follow here is hence an extension of the Manski and Pepper procedure.

[9] If one or other of these conditional correlations is equal to zero, IV estimates will not be inconsistent. Section 5.1 only shows that twin mothers are healthier than mothers of singletons. To complement this, we also show below a series of positive associations of maternal health and both investments in children and child outcomes. We also discuss how this can be estimated in reduced form from natural experiments in particular settings.

Conley et al. (2012) show that bounds for the IV parameter β_1 from equation 2b 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", or UCI, approach), or a fully specified prior for the distribution of γ (their "Local to Zero", or LTZ, approach). In the latter case, a correctly specified prior often leads to tighter bounds. We follow both strategies, the first is agnostic, placing little structure over the violation of the exclusion restriction by simply allowing a large range for γ, and the second involves estimating γ as a(n auxiliary) model parameter.

In general, the Conley et al. (2012) procedure relies on additional assumptions, as we must form a prior over the magnitude of the failure of the exclusion restriction, while in Nevo and Rosen (2012) we only need to provide the sign.[10] The advantage of the Conley et al. procedure that makes it worthwhile despite its stronger assumptions, is that it potentially returns tighter bounds on both the upper and lower end, while Nevo and Rosen retains the original IV upper bound and only tightens the lower bound using information from the original OLS estimates.

4 Data and Descriptive Statistics

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 1, 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 US NHIS, which have been fielded in an identical way from 2004-2014, and the DHS for 68 countries, which have been applied over 20 years using a broadly similar design.

In both data sets, children are included in the sample if aged between 6 and 18 years

[10] It is worth noting however, that Conley et al.'s procedure allows for cases where the prior over γ is of indeterminate sign, which Nevo and Rosen (2012) does not.

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; Mogstad and Wiswall, 2016), but the Israeli data do not contain indicators of maternal condition or maternal behaviours, and the Norwegian data are not publicly accessible, and additionally contain very few markers of maternal health.

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). Grade retention has also been documented to have substantial subsequent impacts on school drop-out and long term attainment (Manacorda, 2012). 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.[11] Case et al. (2002) have demonstrated that an identical self-reported measure of health predicts mortality and morbidity in the US population. Further details on all variable definitions are provided in Appendix B.

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

[11] 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 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.

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. As expected, twin families are larger than non-twin families. Figure 1 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 distribution for all-singleton families in both the DHS sample (Figure 1a) and the US sample (Figure 1b). This establishes the relevance (power) of the twin instrument for fertility, which is formally assessed below.

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 Appendix Figures A5 and A6), increasing family size raises the chances of having a twin birth. In the DHS sample, 42% of all children are in on of the 2+, 3+ or 4+ samples. In the US sample, this value is

45%. Children will be in none of these samples if they are either high birth order children, or if they are low birth order children who do not have older siblings.

5 Results

5.1 Twin Births and Maternal Condition

In Bhalotra and Clarke (forthcoming) we document that mothers with greater health stocks prior to conception or those who engage in more healthy behaviours or are in a healthier environment during pregnancy are more likely to take twins to term. In other words, twins are born to selectively healthy mothers. In order for this to invalidate twin-IV estimates, two conditions must be satisfied. First, twins must be non-random conditional on observable controls (non-independence) and second, twins must have an impact on the outcome of interest beyond that mediated by fertility (non-excludability). Here we document that this is the case in the two data samples used in this paper, and direct readers to Bhalotra and Clarke (forthcoming) where additional evidence in other contexts is presented.

Using the two data sets analysed in this paper, we regress the probability of a twin birth on indicators of maternal health, holding constant socioeconomic status and demographic characteristics. In the US sample (which is much smaller, limiting statistical power, see Table 1), twinning is positively associated with mother's education and BMI, and negatively associated with the mother's smoking status prior to the birth. The smoking indicator is statistically significant even in the pre-IVF period. In Bhalotra and Clarke (forthcoming) we use the universe of births in the US, between 2010 and 2013, and after removing births assisted by Artificial Reproductive procedures such as IVF, we document negative associations of twinning with diabetes and hypertension before pregnancy, with smoking before and during pregnancy and with being short or underweight before pregnancy.[12]

[12] 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.

In the developing country sample (Table 2), we observe that, conditional on maternal age and country and year of birth fixed effects, twin births are positively associated with the mother's education and health, proxied by her height and body mass index (BMI). This result holds even in a period before IVF became available (column 5), and in both low and middle income countries. We also identify a statistically significant positive impact of public health availability on the likelihood of twinning (column 6).[13]

We also investigated whether the source of twin non-randomness additionally has a direct effect on the outcome of interest. This seems plausible since mothers with better health stocks and mothers engaging in positive behaviours prior to pregnancy are likely to be healthier themselves and have stronger preferences over health and educational investments in children following pregnancy, with direct impacts on child outcomes. Evidence of positive causal effects of maternal health with child health or education is not so easy to find but evidence of associations for health is in Uggla and Mace (2016) and Kahn et al. (2002). We document similar associations using our analysis samples. The US results are in Table 3. We regress available measures of child investment (whether the child has any type of health coverage) and outcomes (whether the child has any health limits, the child's standardised educational achievement, and whether the child is classified by parents as being in excellent health), on the maternal characteristics documented to predict twinning in this sample. In each case, we observe that positive maternal health measures are correlated with a reduced likelihood of having health limitations or not having insurance (columns 1-2), and correlated with positive measures of human capital outcomes (education and self-informed health status; columns 3-4). The developing country results are in Table 4. Maternal height, BMI and education are all positively associated with the likelihood of making more positive antenatal investments in child outcomes (the number of appointments, and the likelihood of giving birth at home rather than in a medical centre). We also see impacts of the same maternal health indicators

[13] We include indicators of prenatal care by doctors or nurses in the mother's DHS cluster, rather than the mother's uptake, as this is potentially endogenous to birth type.

on the child's education.[14]

In summary, there is compelling evidence that mothers of twins are selectively healthy. There is also suggestive evidence that healthier women make greater investments in children and that their children have better human capital outcomes. We will test this more formally when progressively introducing controls in IV models in the following section.

5.2 The QQ Trade-off

We now turn to 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 3.1, 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 also introduce controls for mother's education. 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 IV estimated trade-off. 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.

5.2.1 OLS Estimates

OLS results for both samples are in Table 5. We consistently control for fixed effects for age of the child, age of the mother at birth, and the year of the survey. In the developing country sample we also condition on country fixed effects, and in the US sample on census region and mother's race fixed effects. We additionally show results with birth order controls. The available controls for mother's health are height, BMI and cluster-level health service

[14] The maternal health indicators are also all positively associated with infant survival; the reason this is not displayed is that we do not analyse infant survival as an outcome for the reasons indicated in footnote [11] above.

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 Table A2 for summary statistics of these variables) and in the developing country sample we also control for the wealth quintile of the family.

The introduction of observable controls, first for mother's health and then also for her education progressively reduces the estimated trade-off to nearly half of the initial value in both samples, consistent with negative fertility selection. The adjusted estimates for education in developing countries are between 6.6 and 8.5% of a standard deviation. In the US they are between 1 and 2.5% for education and between 0.3 and 1.7% for health status. The Altonji et al. (2005) statistic for the DHS sample suggests that unobservable characteristics of the mother would need to be about 1 to 1.2 times as important as observables for these estimate of the QQ trade-off to be entirely driven by selection into fertility. The corresponding ratio in the US varies from between 1 to 3. In developing countries, 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 the US, 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. In contrast to the case in Black et al. (2005), the controls for birth order do not eliminate the trade-off (Appendix Tables A3 and A4).

5.2.2 IV Estimates with the Twin Instrument

IV estimates using the twin instrument are in Tables 6 (DHS) and 7 (US), the first-stage estimates are in panel A and the second stage in panel B. In these Tables we present coefficients on the variable of interest (fertility), however provide full output of all coefficients in Appendix Tables A5, A6 and A7.

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 are 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 2a 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 section 5.1 that twin mothers are positively selected on health (and education), we see that upon 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 4.1% (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%.

While one way to compare the base and full control specifications is to test whether each coefficient differs from zero, an alternative test is to compare the estimated coefficients (and standard errors) to each other. We thus also test each coefficient compared to the "Base" coefficient, and present the p-values of this test as “Coefficient Difference” at the foot of panel B. We can often reject equality of the coefficients in the specifications with and without controls for maternal health. Implementing these tests requires that we take account of the correlations between error terms in each model. In order to do this we replicate IV estimates using GMM, which allows us to estimate models simultaneously and hence compare coefficients across models. Additional details related to this test are provided in Appendix C.

IV Estimates: United States The first stage estimates for the US sample (Table 7) are 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 magnitude. 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%. 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. As discussed earlier in this section, it is well recognised that twin-IV estimates are often not precise. So it is quite striking that we find a significant trade-off for education and health. Overall, partial bias adjustment reveals 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).

Recent work suggests that focusing on monozygotic (MZ) rather than dizygotic (DZ) twins may resolve issues related to the heritability of twinning and relationships between twinning and some maternal characteristics (Farbmacher et al., 2016). While we cannot observe whether a twin pair are MZ or DZ in either of our data sources, when we use only same sex twins to construct the twin instrument, as they are considerably more likely to be MZ, we observe a similar pattern, where once again estimates diverge from zero and become significant when controls for maternal health are included. Results for the DHS are in Table A8 and for the NHIS in Table A9.

5.2.3 Non-Linear Models

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 investments. More recent work, for example the theoretical work of Aizer and Cunha (2012), and empirical papers by Rosenzweig and Zhang (2009); Brinch et al. (2017); Mogstad and Wiswall (2016); Bagger et al. (2013) relax this assumption. Among other things, this allows for reinforcing or compensating behaviours in parental investment choices (Almond and Mazumder, 2013). This implies that we should allow the coefficient β_1 to vary across children in the family.

Using DHS data for which we have a sufficiently large sample to split instruments, we re-estimate our regressions following the non-linear marginal fertility models of Brinch et al. (2017); Mogstad and Wiswall (2016). We provide a full discussion of the methodology in Appendix D, and in the analysis below we follow the procedure laid out by Mogstad and Wiswall (2016) precisely. Models of this type loosen the linear marginal effects estimated on fertility, and allow for a one-unit shift in fertility at different birth orders to have potentially varied impacts on existing children.

We report the restricted (linear) and non-restricted (non-linear) IV models in Table 8, and the corresponding first stage results in Appendix D (Appendix Table A14). We report results by the same parity samples as the main IV results presented in Table 6.

In Table 8 we observe, firstly, that as described in Table 6, the linear specifications are universally lower, and often become statistically distinguishable from zero when partially controlling for the selection of twins as compared to the baseline estimate not controlling for twin selection. These results only differ from those reported earlier in that we now restrict the sample to families with 6 children or fewer in line with results reported in Mogstad and Wiswall (2016), which involves a loss of between 5 and 18 percent of the sample depending on the parity sample used. For full descriptives on family size in each parity group refer to Appendix Figure A7. Turning to panel B, we observe a similar non-linear dynamic as that reported in Mogstad and Wiswall (2016). For example, in the two-plus sample, we observe that the twin instrumented estimate of the effect of moving from one to two siblings is large and positive, while the impact of moving from two to three siblings is large and negative. However, most interestingly for the present analysis, the non-linear impacts are nearly universally *larger* in absolute terms when partially controlling for twin selection. As was the case with the linear model, we observe that the marginal fertility effects become nearly everywhere more negative, and in certain cases become statistically different from zero. Thus, our finding that the twin-IV estimator tends to under-estimate the causal effect of fertility on child human capital holds in the linear and non-linear specifications.

5.2.4 IV effect sizes in perspective

Since the QQ trade-off has been called into question, it is important to consider the size of the partially-bias-adjusted estimates and not just their sign and statistical significance. Our results (in the linear model) 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 that de-worming 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).

If we consider grade retention in the US, our estimates suggest that an additional birth results in 0.22 fewer years completed. This is of similar magnitude to estimates of the effect of an additional 1,000 grams of birthweight over the normal birthweight range (a 0.31 increase in years of schooling) in Royer (2009), and estimates of the impact of historical exposure to high rather than low malaria rates (a 0.4 year reduction) in Barreca (2010). Turning to the effects on health, we find that an additional birth (at order 3 or 4) reduces the likelihood that siblings are in excellent health by between 3-6%. Almond and Mazumder (2005) document that in the long-run, the 1918 influenza pandemic increased the likelihood of being in poor or fair health (the inverse of our health measure) by 10%. Overall, the adjusted estimates are of a size that it is not prudent to dismiss. Moreover, our estimates indicate the change in investment (education or health) for one additional birth but, as fertility rates remain high in many developing countries, the total effect can be large.

5.3 Bounding the QQ Trade-off

5.3.1 Generalised Bounds

The adjusted twin-IV results will not provide consistent estimates of β_1 as there are almost certainly omitted indicators of maternal health. Although documenting that observable measures of health (which also impact child quality) are correlated with the instrument does not prove instrumental invalidity, it does suggest that it is highly likely that similar non-observable factors will also be correlated, thus resulting in invalidity. A recent study proposes a formal test of instrument invalidity (Kitagawa, 2015). Using the 2+ sample for the DHS data this test rejects the validity of the twin instrument – see Appendix Figure A8 and Table A10; however this test is sensitive to curse of dimensionality considerations, and so to implement it we had to simplify the specification of controls.[15] We do not report results for the NHIS data because the sample is too small to obtain informative confidence intervals.

Rather than discard the twin-IV estimator altogether, we harness its power in predicting fertility using IV bounds to assess the empirical significance of the omitted variables. As outlined in section 3.2, we begin by estimating Nevo and Rosen (2012) bounds. These are based on the assumptions that twins are positively selected and fertility is negatively selected. Evidence for both of these assumptions is in Tables 5 and 6-7 where it is observed that controlling for education and health results in the OLS coefficients on fertility growing less negative and the IV coefficients on twins growing more negative. It is further assumed that twins is a less endogenous variable than fertility. The bounds are in Table 9 (columns 2-3; IV point estimates are presented for comparison in column 1). These estimates provide a lower bound on the QQ parameter estimated in Tables 6-7 of approximately 5-8% of a standard deviation across the DHS and NHIS samples.[16] As discussed in section 3.2, the upper bound in Nevo and Rosen's bounding procedure is determined by the upper bound of the 95% confidence interval of the original twin IV estimates. As such, estimates which are not significant at 95% confidence levels in Tables 6-7 will once again be non-informative when using the Nevo and Rosen (2012) procedure.

[15] In particular, the inclusion of a large number of fixed effects is prohibitive, and so we replace country and mother year of birth fixed effects with continent and decade of birth fixed effects respectively.

[16] The NHIS data contain only 21,000-61,000 observations (depending on the parity sample), about 10-15% of the DHS sample. As highlighted by Angrist et al. (2010), the twin IV estimator is typically under-powered. When we construct bounds, we further challenge statistical power. So the bounds for the NHIS sample are often imprecise, irrespective of whether we use Conley or Nevo Rosen bounds. As a result, in general here we focus on bounds in the developing country sample, although we present bounds from both settings.

In order to gain additional precision in bounds estimates at the *upper* bound, we also estimate Conley et al. (2012) bounds. As discussed, we need to define a prior belief over the sign and magnitude that the coefficient on twin birth (γ) would take in equation 4. To begin, we assume a range of values for γ from 0 to 0.05, or 5% of a standard deviation, in which case instrument validity is violated, and having a twin mother has a positive effect on child quality conditional on fertility. The results are in Figure 2 for developing countries and Figure 3 for the US for the 3+ samples; results for the 2+ and 4+ samples are in Appendix Figures A9 and A10. We assume $\gamma \sim U(0, \delta)$ 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.

Given that twin IV estimators tend to produce wide confidence intervals (Angrist et al., 2010), Conley et al. (2012) bounds will also tend to be wide. As δ increases, the violation of the exclusion restriction increases. We observe, firstly, a widening of the estimated bounds as the size of the violation increases,[17] and secondly that the upper bound becomes increasingly negative, moving in the direction of finding a QQ trade-off.[18] 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 (0.05 is a very large value for the exclusion restriction), the bounds are quite wide.

With a view to improving the precision and relevance of these bounds, we estimate rather

[17] 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" (in Conley et al.'s terminology) related to the instrument inflates the traditional variance-covariance matrix.

[18] 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.

than assume γ, the measure of the extent of the violation of the exclusion restriction. This is (as usual) the product of two relationships which, here, are the relationship between the probability of a twin birth and maternal health, and the relationship between maternal health and investments in children. 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 US and Nigeria. This is in line with Conley et al. (2012) who illustrate their estimator with examples involving back of the envelope calculations of γ for each case. In Appendix E we detail how we leverage two historical natural experiments involving a shock to the health of women, namely, the Biafra war in Nigeria and the introduction of the first antibiotics to the US, to estimate γ.[19] We also conduct a number of back of the envelope plausibility tests. In general these suggest that γ is around 0.004-0.006, or that having a (positively-selected) twin mother has a direct effect of around 0.4 to 0.6% of a standard deviation in quality outcomes.

As we outline at more length in Appendix E.1 and E.2, the generation of this estimate for γ is based on particular shocks which impact maternal health. We present evidence supporting the assumptions for these estimates in Appendix E.3, and put these in the context of Conley et al.'s methods in Appendix E.4. These reduced form estimates of γ based on exogenous events provide a well founded estimate to use in the Conley et al. (2012) procedure, but one may be concerned about external validity of these estimates, given that they are derived from 1930s America (sulfa drugs) and 1970s Nigeria (Biafra war). We can, however, show that estimates of γ from contemporary DHS data (which are used in the main analysis and hence relevant for estimates of γ) are in fact of the same order of magnitude as our estimates from America and Nigeria. Consider Appendix Table A11, which shows that a one standard deviation change in maternal BMI is associated with a 0.070 s.d. increase in

[19] We are agnostic about intermediate variables ("mediators") and simply show what is key to the violation of the twin instrument, which is that the exogenous shock to maternal health impacts on an indicator of child quality. We then scale this estimate by the difference in health between women who give birth to twins and women who give birth to singletons.

the child's educational Z-score (column 4). We observe that twin mothers in the same data sample have BMI 0.050 s.d. higher than non-twin mothers. Scaling (multiplying) this by the estimated association (0.070×0.050) produces an estimate of gamma (a measure of the violation in the exclusion restriction, or the twin-mediated effect of maternal BMI on child outcomes) of 0.0035 s.d. This is of the same order of magnitude as the value of γ that we estimate from the Biafra (0.0040) and Sulfa (0.0062) case studies. We can calculate a range of such estimates using education and height (as well as BMI), and find values of 0.025 for education (0.215×0.0121) and 0.00196 for height (0.019×0.103). Importantly, all of these values fall within the estimated distributions of γ used to calculate Conley et al. bounds, displayed in Appendix Figure A13.

It is important to note that while degree of the violation of the exclusion restriction is estimated to be relatively small (at 0.4 or 0.6 percent of a standard deviation of the child quality measure, education), γ is obtained after scaling the estimated impact of maternal health shocks/characteristics (that predict twinning) on the final outcomes of interest (child quality indicators). The scaling factor is the difference in the maternal health indicator between mothers of twins and mothers of singletons – this is 0.050 in the BMI example above, it is 0.125 in the sulfa experiment, and 0.267 in the Biafra experiment (all figures from Appendix Table A15). Thus γ is in fact much smaller than the measure of violation that is of interest.

Using these estimates of γ, we are able to pin down the bounds described in Figures 2-3. See Table 9, columns 4-5, where we present the UCI approach in which we assume that $\gamma \in [0, 2\hat{\gamma}]$. This assumption is chosen such that the true $\hat{\gamma}$ described in Appendix E 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 estimates of both γ and its distribution, which allow uncertainty for our estimates of γ and assume that γ is distributed precisely according to the estimated empirical distribution (refer to Appendix E.5).

Our preferred bounds estimates are those in the right-hand columns of Table 9, as these are more efficient, being 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.

For the US sample we see that 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. These indicate that an additional birth reduces the grade progression of an older sibling by 16.6% of a s.d. (upper bound), or 8.3% of a s.d. (mid point), and their likelihood of being reported as being in excellent health by 7% (upper bound) or 3.5% (mid point).

In Figure 4 we plot the estimated coefficients and bounds for the developing country sample altogether so they are readily compared. The corresponding plot of all estimates for the (much smaller) US sample is in Appendix Figure A11). The figures show the OLS and IV estimates, with base controls, health, and health and socioeconomic controls and we show the Nevo–Rosen and Conley et al. bounds for each of the 2+, 3+ and 4+ groups. The informativeness of the bounds is evaluated against the criteria laid out by Hotz et al. (1997): firstly do the bounds enable us to determine if the effect is negative or positive, secondly can we reject the point estimates of linear IV, and thirdly do our bounds allow us to reject the OLS estimate of the causal effect. In general, for the 3+ and 4+ samples in the DHS data, the bounds *are* informative of the (negative) sign of the trade-off, but not for the 2+ sample. In terms of the second and third criteria, we can never exclude the point estimate of the original IV estimate from our bounds, however we often *can* reject the original OLS estimate, which is important given recent evidence that many IV estimates are inaccurate, and frequently include OLS point estimates in their confidence intervals (Young, 2018).

Using summary statistics from Table A2, we can convert standardised estimates from these bounds 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 5% of a standard deviation in the developing country sample.[20] Using the standard deviation in the sample of 3.8 years, this implies an average effect of around 0.19 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 based on the midpoint of bounds estimates is 8% of a standard deviation in grade retention, which 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 4.2% according to the midpoint of bounds following an additional birth among the same group. Overall, these are quite large effects relative to the marginal effects of different policy interventions considered in the literature (see section 5.2.4).

Conclusion and Discussion

This paper demonstrates that twin-IV estimates of the fertility- human capital trade-off tend to be biased downward on account of positive selection of women into twin birth, a problem that has not been previously recognized. 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. Using partial identification to bound the effect of child quantity on child quality suggests that the *true* effect size 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. The estimated 4%-5% of a standard deviation increase is equivalent to an additional 0.15 to 0.19 years in the classroom in the developing country sample, and estimates of approximately 8% of a standard deviation in the US account for 0.22

[20] This estimate is the average midpoint if the bound estimates from the three plus and four plus samples in Table 9 and can be calculated as: $\frac{1}{2} \times [(0.0646 - 0.0067)/2 + 0.0067 + (0.0748 + 0.0235)/2 + 0.0235]$.

more grades progressed on average. 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.

Any human capital costs of fertility are naturally of greater concern not only when fertility is high but also when a large share of it is unwanted. In 2015 the average number of births per woman in low income countries was five 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 US 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. This might exacerbate impacts of additional births on investments in preceding births.

References

D. Aaronson, R. Dehejia, A. Jordan, C. Pop-Eleches, C. Samii, and K. Schulze. The Effect of Fertility on Mothers' Labor Supply over the Last Two Centuries. IZA Discussion Papers 10559, Institute for the Study of Labor (IZA), Feb. 2017.

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.

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 B. Mazumder. The 1918 Influenza Pandemic and Subsequent Health Outcomes: An Analysis of SIPP Data. *American Economic Review*, 95(2):258–262, May 2005.

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.

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.

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.

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

J. Bagger, J. A. Birchenall, H. Mansour, and S. Urza. Education, Birth Order, and Family Size. NBER Working Papers 19111, National Bureau of Economic Research, Inc, June 2013.

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.

A. I. Barreca. The Long-Term Economic Impact of In Utero and Postnatal Exposure to Malaria. *Journal of Human Resources*, 45(4):865–892, 2010.

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. Bhalotra and D. Clarke. Twin Birth and Maternal Condition. *Review of Economics and Statistics*, xx(x):xxx–xxx, forthcoming.

S. Bhalotra and T. Cochrane. Where Have All the Young Girls Gone? Identification of Sex Selection in India. IZA Discussion Papers 5381, Institute for the Study of Labor (IZA), Dec. 2010.

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. 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.

J. Bisbee, R. Dehejia, C. Pop-Eleches, and C. Samii. Local instruments, global extrapolation: External validity of the labor supply–fertility local average treatment effect. *Journal of Labor Economics*, 35(S1):S99–S147, 2017.

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.

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. Brinch, M. Mogstad, and M. Wiswall. Beyond LATE with a Discrete Instrument. *Journal of Political Economy*, 125(4):985–1039, 2017.

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.

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

A. Case, D. Lubotsky, and C. Paxson. Economic Status and Health in Childhood: The Origins of the Gradient. *The American Economic Review*, 92(5):1308–1334, 2002.

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. Children And Their Parents: A Review Of Fertility And Causality. *Journal of Economic Surveys*, 32(2):518–540, April 2018.

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.

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

H. Farbmacher, R. Guber, and J. Vikström. Increasing the credibility of the Twin birth instrument. Working Paper Series 2016:10, IFAU - Institute for Evaluation of Labour Market and Education Policy, June 2016.

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.

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.

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.

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.

V. J. Hotz, C. H. Mullin, and S. G. Sanders. Bounding Causal Effects Using Data From a Contaminated Natural Experiment: Analysis the Effects of Teenage Chilbearing. *Review of Economic Studies*, 64(4):575–603, Oct. 1997.

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.

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

R. S. Kahn, B. Zuckerman, H. Bauchner, C. J. Homer, and P. H. Wise. Women's Health After Pregnancy and Child Outcomes at Age 3 Years: A Prospective Cohort Study. *American Journal of Public Health*, 92(8):1312–1318, 2002.

T. Kitagawa. A test for instrument validity. *Econometrica*, 83(5):2043–2063, 2015.

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.

J. Lee. Sibling size and investment in children's education: an Asian instrument. *Journal of Population Economics*, 21(4):855–875, October 2008.

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.

M. Manacorda. The Cost of Grade Retention. *The Review of Economics and Statistics*, 94 (2):596–606, 2012.

C. F. Manski and J. V. Pepper. Monotone Instrumental Variables: With an Application to the Returns to Schooling. *Econometrica*, 68(4):997–1010, 2000.

B. Mazumder and Z. Seeskin. Breakfast skipping, extreme commutes and the sex composition at birth. *Biodemography and Social Biology*, 61(2):187–208, 2015.

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.

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

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.

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.

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, 2009.

H. Royer. Separated at Girth: US Twin Estimates of the Effects of Birth Weight. *American Economic Journal: Applied Economics*, 1(1):49–85, January 2009.

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

C. Uggla and R. Mace. Parental investment in child health in sub-Saharan Africa: a cross-national study of health-seeking behaviour. *Royal Society Open Science*, 3(2), 2016.

J. Vere. Fertility and parents’ labour supply: new evidence from US census data. *Oxford Economic Papers*, 63(2):211–231, 2011.

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.

A. Young. Consistency without Inference: Instrumental Variables in Practical Application. mimeo, London School of Economics, June 2018.

Tables

Table 1: Probability of Giving Birth to Twins USA (NHIS)

Twin×100	All	Time	
		1982-1990	1991-2013
Mother's Education (Years)	0.060**	0.115	0.056**
	(0.025)	(0.096)	(0.026)
Mother's Height (Inches)	0.012	0.049	0.008
	(0.025)	(0.087)	(0.026)
Mother's BMI	0.010**	0.025	0.009*
	(0.004)	(0.016)	(0.005)
Smoked Prior to Birth	-0.285**	-1.336**	-0.183
	(0.137)	(0.526)	(0.142)
Observations	103,589	6,891	96,698
R-Squared	0.004	0.031	0.004

This table presents regressions of whether each birth is a twin or a singleton on a number of maternal characteristics. All specifications include a full set of mother's age, 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 inches and BMI is weight in kg divided by height in metres squared. Heteroscedasticity-robust standard errors are included in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 2: Probability of Giving Birth to Twins (Developing Countries by Income and Time Period)

Twin×100	All	Income		Time		Prenatal
		Low inc	Middle inc	1990-2013	1972-1989	
Mother's Age	0.540***	0.550***	0.517***	0.601***	0.314***	0.541***
	(0.027)	(0.033)	(0.047)	(0.031)	(0.058)	(0.027)
Mother's Age Squared	-0.007***	-0.007***	-0.007***	-0.008***	-0.003**	-0.007***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Age at First Birth	-0.050***	-0.082***	-0.001	-0.052***	-0.040***	-0.050***
	(0.008)	(0.010)	(0.013)	(0.010)	(0.015)	(0.008)
Mother's Education (years)	0.021***	0.019**	0.013	0.021***	0.019	0.018**
	(0.007)	(0.009)	(0.010)	(0.008)	(0.012)	(0.007)
Mother's Height (cm)	0.059***	0.058***	0.058***	0.063***	0.044***	0.058***
	(0.004)	(0.005)	(0.007)	(0.005)	(0.007)	(0.004)
Mother's BMI	0.047***	0.059***	0.038***	0.045***	0.050***	0.045***
	(0.006)	(0.009)	(0.009)	(0.007)	(0.010)	(0.006)
Prenatal Care (Doctor)						0.333**
						(0.142)
Prenatal Care (Nurse)						0.312**
						(0.142)
Prenatal Care (None)						0.008
						(0.181)
Observations	2,046,907	1,287,585	759,322	1,525,966	520,941	2,043,217
R-Squared	0.006	0.006	0.005	0.006	0.005	0.006

NOTES: This table presents results for the developing country sample splitting by pre- and post-1990. Main specifications for the developing country sample are pooled for all years. 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. Prenatal care variables refer to average levels of coverage in DHS clusters. These prenatal measures are only recorded for births in 5 years preceding each survey wave, and as such, a small number of (small) clusters do not have records available. Standard errors clustered by mothers are presented in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 3: Maternal Health and Child Investments/Outcomes (NHIS)

	No Health Insurance	Health Limits	Education Z-Score	Excellent Health
Mother's Education (Years)	-0.016***	-0.001***	0.019***	0.020***
	(0.001)	(0.000)	(0.002)	(0.001)
Mother's Height (Inches)	-0.001**	-0.002***	0.005***	0.008***
	(0.000)	(0.000)	(0.002)	(0.001)
Mother's BMI	0.000	0.000***	0.000	-0.002***
	(0.000)	(0.000)	(0.000)	(0.000)
Smoked Prior to Birth	0.008***	0.031***	-0.046***	-0.058***
	(0.002)	(0.003)	(0.009)	(0.004)
Observations	103,589	103,589	74,777	103,589
R-Squared	0.047	0.013	0.019	0.033

Regressions are presented of child investments or child outcomes on a number of maternal characteristics. All specifications and variable definitions follow Table 1 and include a full set of mother's age, survey year, region of birth, and mother's race dummies. No Health insurange, health limits and excellent health are binary variables, and models are estimated as linear probability models. Education Z-Score is a standardized score of the child's completed years of education compared with his or her birth year and birth month cohort. Height is measured in inches and BMI is weight in kg divided by height in metres squared. Heteroscedasticity-robust standard errors are included in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 4: Maternal Health and Child Investments/Outcomes (Developing Country Sample)

	Maternal Characteristics			With Cluster-Level Health Measures		
	Home Birth	Antenatal Visits	Education Z-Score	Home Birth	Antenatal Visits	Education Z-Score
Mother's Age	0.026***	0.024***	0.004**	0.025***	0.030***	0.002
	(0.001)	(0.005)	(0.002)	(0.001)	(0.004)	(0.002)
Mother's Age Squared	-0.000***	-0.001***	-0.000***	-0.000***	-0.001***	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age at First Birth	-0.011***	0.070***	0.009***	-0.010***	0.060***	0.008***
	(0.000)	(0.002)	(0.000)	(0.000)	(0.001)	(0.000)
Mother's Education (years)	-0.034***	0.267***	0.075***	-0.029***	0.219***	0.072***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
Mother's Height (cm)	-0.001***	0.020***	0.005***	-0.001***	0.016***	0.005***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
Mother's BMI	-0.013***	0.076***	0.022***	-0.011***	0.060***	0.020***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
Prenatal Care (Doctor)				-0.190***	1.213***	0.063***
				(0.004)	(0.030)	(0.008)
Prenatal Care (Nurse)				-0.039***	0.110***	0.069***
				(0.004)	(0.028)	(0.008)
Prenatal Care (None)				0.336***	-3.673***	-0.421***
				(0.005)	(0.033)	(0.010)
Observations	749,010	615,621	1,128,729	749,006	615,619	1,125,305
R-Squared	0.292	0.334	0.129	0.320	0.385	0.137

Regressions are presented of child investments or child outcomes on a number of maternal characteristics. All specifications and variable definitions follow Table 2 and include a full set of country and year of birth fixed effects. Home birth and antenatal visits are recorded only for children aged 0-4 at the time of the survey, and the standardised education score is recorded only for children aged 6-18 (of school age). Additional notes are available in Table 2. DHS sample weights are used, and standard errors are clustered by mother. *** p<0.01, ** p<0.05, * p<0.1

Table 5: OLS Estimates of the QQ Trade-off: 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***	-0.139***	-0.116***	-0.074***	-0.120***	-0.098***	-0.061***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)
Observations	259,958	259,958	259,958	395,687	395,687	395,687	409,576	409,576	409,576
R-Squared	0.109	0.132	0.193	0.093	0.120	0.190	0.081	0.113	0.188
Altonji et al. Ratio			1.258			1.137			1.035
Panel B: US Results									
Dependent Variable = School Z-Score									
[1em] Fertility	-0.031***	-0.031***	-0.024***	-0.032***	-0.031***	-0.023***	-0.020	-0.018	-0.011
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.013)	(0.013)	(0.013)
Observations	61,267	61,267	61,267	47,308	47,308	47,308	21,352	21,352	21,352
R-Squared	0.027	0.030	0.034	0.027	0.030	0.034	0.041	0.045	0.049
Altonji et al. Ratio			3.202			2.587			1.157
Dependent Variable = Excellent Health									
[1em] Fertility	-0.002	-0.005**	-0.003	-0.010***	-0.008***	-0.007***	-0.024***	-0.018***	-0.016***
	(0.003)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)
Observations	70,277	70,277	70,277	53,393	53,393	53,393	24,358	24,358	24,358
R-Squared	0.033	0.321	0.323	0.041	0.329	0.331	0.054	0.341	0.343
Altonji et al. Ratio			-4.069			1.801			2.126

OLS regressions described in equation 1 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 4). 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. Standard errors are clustered by mother.*p<0.1; **p<0.05; ***p<0.01

Table 6: 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	0.832***	0.842***	0.843***	0.829***	0.838***	0.839***	0.860***	0.866***	0.868***
	(0.029)	(0.029)	(0.028)	(0.025)	(0.025)	(0.025)	(0.026)	(0.026)	(0.026)
Kleibergen-Paap rk statistic	826.32	872.87	928.69	1066.33	1124.76	1158.30	1119.75	1094.96	1141.43
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	-0.004	-0.015	-0.012	-0.028	-0.041*	-0.046**	-0.027	-0.038*	-0.037**
	(0.028)	(0.027)	(0.026)	(0.022)	(0.021)	(0.020)	(0.022)	(0.021)	(0.019)
Observations	259,958	259,958	259,958	395,687	395,687	395,687	409,576	409,576	409,576
Coefficient Difference		0.019	0.354		0.006	0.055		0.108	0.316

Panels A and B present coefficients and standard errors for the first and second stages in equations 2a and 2b. 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 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). The rk test statistic and corresponding p-value reject that the twin instruments are weak in each case. Coefficient Difference in Panel B refers to a test that the coefficient estimate on Fertility in a given model is identical to the estimate on Fertility in the base case. This test takes account of the correlation between errors in the base and augmented regression model (in the spirit of seemingly unrelated regressions), but is estimated by GMM to house the IV models estimated here. Low p-values are evidence against equality of the two estimates. Standard errors are clustered by mother.*p<0.1; **p<0.05; ***p<0.01

Table 7: 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)									
[1em] Twins	0.698***	0.701***	0.704***	0.740***	0.740***	0.743***	0.795***	0.799***	0.813***
	(0.026)	(0.026)	(0.026)	(0.047)	(0.047)	(0.047)	(0.080)	(0.080)	(0.079)
Kleibergen-Paap rk statistic	704.10	710.40	737.62	245.84	249.12	250.47	98.54	99.30	105.52
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	-0.098	-0.099	-0.101*	-0.012	-0.015	-0.017	-0.124	-0.134	-0.142
	(0.061)	(0.061)	(0.060)	(0.067)	(0.067)	(0.067)	(0.152)	(0.152)	(0.149)
Observations	61,267	61,267	61,267	47,308	47,308	47,308	21,352	21,352	21,352
Coefficient Difference		0.818	0.434		0.547	0.394		0.231	0.113
Dependent Variable = Excellent Health									
Fertility	0.009	0.027	0.026	-0.036	-0.058*	-0.057*	0.033	-0.025	-0.031
	(0.025)	(0.021)	(0.021)	(0.039)	(0.032)	(0.032)	(0.060)	(0.053)	(0.052)
Observations	70,277	70,277	70,277	53,393	53,393	53,393	24,358	24,358	24,358
Coefficient Difference		0.164	0.212		0.341	0.366		0.122	0.089

Notes: Regressions in each panel and the definition of the 2+, 3+ and 4+ groups are identical to Table 6 and are described in notes to Table 6. 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 race FEs. Additional socioeconomic controls consist of mother's education fixed effects, and health controls include a continuous measure of mother's BMI, and a Likert scale measure of a mother's self-assessed health. 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 first stage results and tests of instrument strength are displayed for the regression using the education sample only. Qualitatively similar results are observed for the health sample. A description of the Kleibergen-Paap statistic and Coefficient Difference are provided in notes to Table 6. 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 8: Linear and Non-Linear IV Estimates for Marginal Effects with and without Full Twin Controls

	Two-Plus		Three-Plus		Four-Plus	
	Baseline	+S&H	Baseline	+S&H	Baseline	+S&H
Instrument	Non-linear IV	Non-linear IV	Non-linear IV	Non-linear IV	Non-linear IV	Non-linear IV
Panel A: Linear Estimates of Marginal Effects						
Number of Children	0.031	0.011	-0.011	-0.048**	0.012	-0.024
	(0.036)	(0.033)	(0.026)	(0.024)	(0.029)	(0.025)
Panel B: Unrestricted Estimates of Marginal Effects						
Siblings ≥ 2	0.209**	0.0175				
	(0.083)	(0.121)				
Siblings ≥ 3	-0.089	-0.102*	-0.012	-0.063*		
	(0.057)	(0.054)	(0.052)	(0.037)		
Siblings ≥ 4	-0.032	-0.041	-0.036	-0.045	0.014	-0.036
	(0.066)	(0.044)	(0.043)	(0.039)	(0.041)	(0.051)
Siblings ≥ 5	0.055	0.059	0.034	0.036	0.036	0.035
	(0.078)	(0.040)	(0.073)	(0.038)	(0.041)	(0.036)
Observations	245,534	245,534	356,892	356,892	334,924	334,924

Each column and panel presents a separate regression using DHS data. Siblings ≥ 2 refers to the marginal effect of moving from 1 to 2 siblings, Siblings ≥ 3 refers to moving from 2 to 3 siblings, and so forth. Each model includes maternal age, country, survey year and child age fixed effects as well as child's gender. The regressions in columns 2, 4 and 6 are augmented with all socioeconomic and health controls described in Table 5 of the paper. Standard errors are estimated using a block bootstrap sampling each family with replacement, and for each bootstrap replication the both the regression and the constructed instruments are reestimated. First stage regressions are displayed in Table A14.

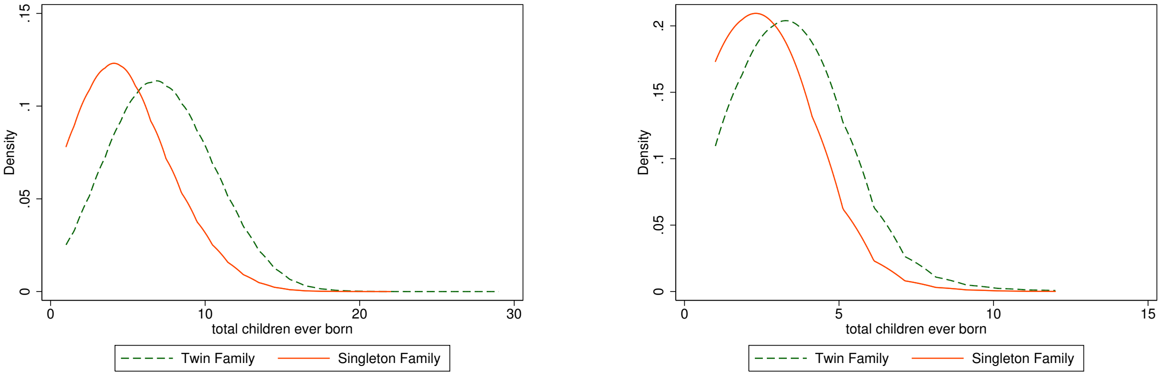
Table 9: Bounds Estimates of the Quantity–Quality Trade-off

	IV	Nevo and Rosen (2012)		Conley et al. (2012)			
		Imperfect IV Bounds		UCI: $\gamma \in [0, 2\hat{\gamma}]$		LTZ: Empirical Distribution γ	
	with Controls	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound
Panel A: DHS							
Two Plus	-0.0124	-0.0842	0.0384	-0.0702	0.0184	-0.0657	0.0132
Three Plus	-0.0456	-0.0759	-0.0068	-0.0690	-0.0012	-0.0646	-0.0067
Four Plus	-0.0371	-0.0632	-0.0001	-0.0800	-0.0197	-0.0748	-0.0235
Panel B: USA (Education)							
Two Plus	-0.1023	-0.0480	0.0164	-0.2195	-0.0026	-0.2101	-0.0113
Three Plus	-0.0164	-0.0448	0.1149	-0.1291	0.0795	-0.1208	0.0709
Four Plus	-0.1488	-0.0709	0.1547	-0.4329	0.1200	-0.4242	0.1132
Panel B: USA (Health)							
Two Plus	0.0267	-0.0843	0.0374	-0.0247	0.0615	-0.0164	0.0534
Three Plus	-0.0539	-0.0764	-0.0072	-0.1107	-0.0137	-0.1027	-0.0219
Four Plus	-0.0298	-0.0638	0.0001	-0.1041	0.0295	-0.0972	0.0217

NOTES: This table presents upper and lower bounds of a 95% confidence interval for the effects of family size on (standardised) children's educational attainment and health (health in USA only). Nevo and Rosen (2012) bounds are presented in columns 2 and 3, and variants of Conley et al. (2012) bounds are presented in columns 4-7. the IV point estimate with full controls is displayed for comparison in column 1. Nevo and Rosen (2012) bounds are based on the assumption that twinning is positively selected and fertility is negatively selected, and twins are "less endogenous" than fertility. Conley et al. (2012) bounds are estimated as described in section 3.2 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 preferred prior for γ ($\hat{\gamma}$) and its distribution is discussed in Appendix E, and estimates for γ are provided in Table A15. Comparisons under a range of priors are presented in Figures 2-3. Each estimate is based on the specifications with full controls from Tables 6 and 7.

Figures

Figure 1: Twins shift the fertility distribution outward

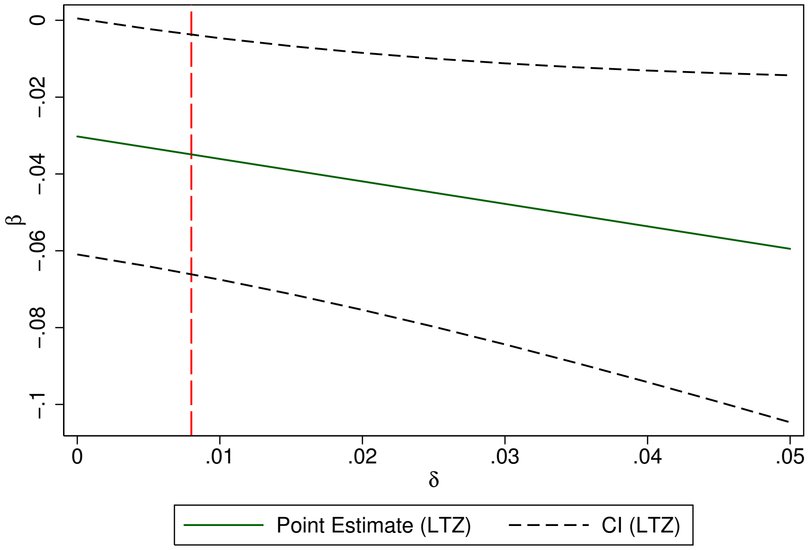


(a) Developing Countries

(b) United States

Note to Figure 1: 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.

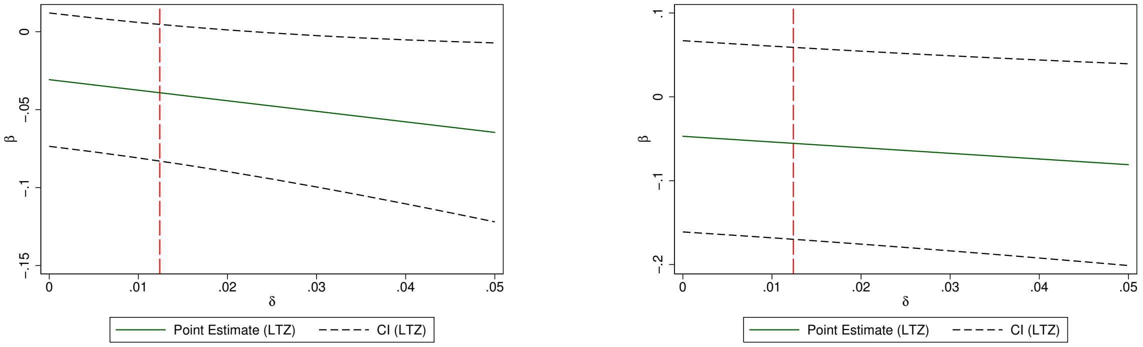
Figure 2: Plausibly Exogenous Bounds: School Z-Score (Developing Countries 3+)



Methodology described in Conley et al. (2012)

Note to Figure 2: Confidence intervals and point estimates are calculated according to Conley et al. (2012) using DHS data and specifications described in section 5.3. 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 the point at which the preferred estimate $\hat{\gamma}$ lies precisely at the centre of the assumed support for γ. Further discussion is provided in section 3.2 and Table 9.

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



Methodology described in Conley et al. (2012)

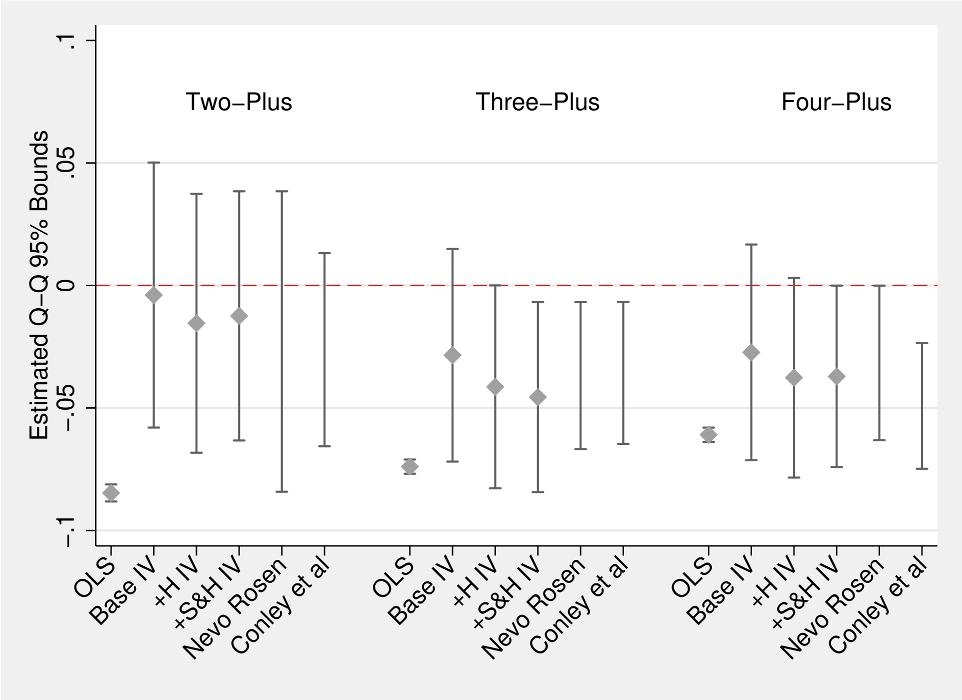
Methodology described in Conley et al. (2012)

(a) Excellent Health

(b) Education Z-Score

Notes to Figure 3: See notes to Figure 2. An identical approach is employed, however now using USA (NHIS) data.

Figure 4: Parameter and Bound Estimates of the Q–Q Trade-off



Note to Figure 4: Each set of estimates refer to the 95% confidence intervals on parameter bounds of the impact of fertility on child education. Two-Plus, Three-Plus and Four-Plus refer to parity specific groups. Base IV refer to the IV estimate most closely following the existing literature, with +H and +S&H presenting IV estimates controlling for maternal health and socioeconomic variables. OLS point estimates are presented along with their 95% confidence intervals, which are quite narrow. OLS estimates include all maternal controls (corresponding to base, and +S&H). Versions without maternal controls are even more negative. The final two sets of bounds in each group are estimated following Nevo and Rosen (2012) and Conley et al. (2012) procedures, and do not have a corresponding point estimate.

ONLINE APPENDIX

For the paper:

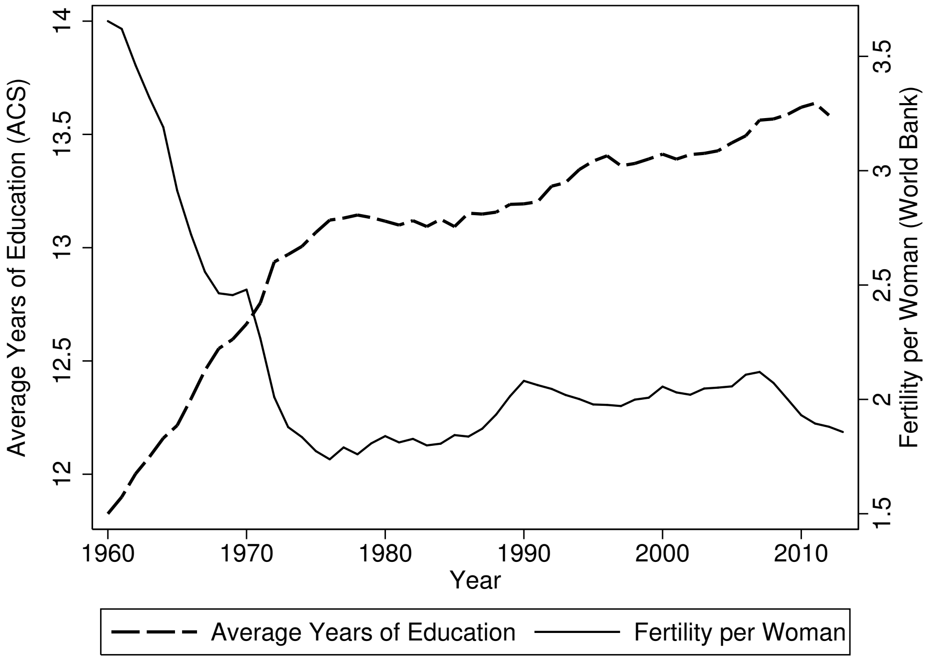
The Twin Instrument: Fertility and Human Capital Investment

Sonia Bhalotra and Damian Clarke

Contents

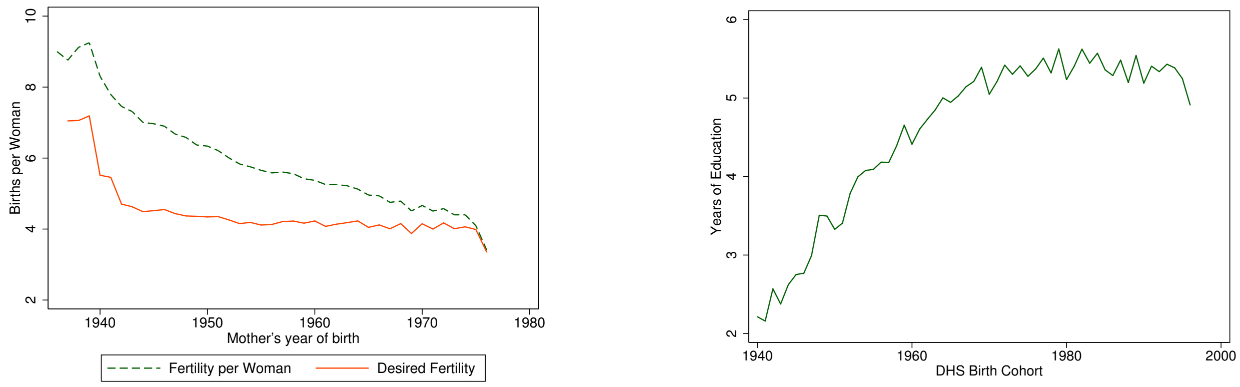
A Appendix Figures and Tables

Figure A1: Education and Fertility Trends (USA)



Note to figure A1: 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 A2: Education and Fertility (Developing Countries)

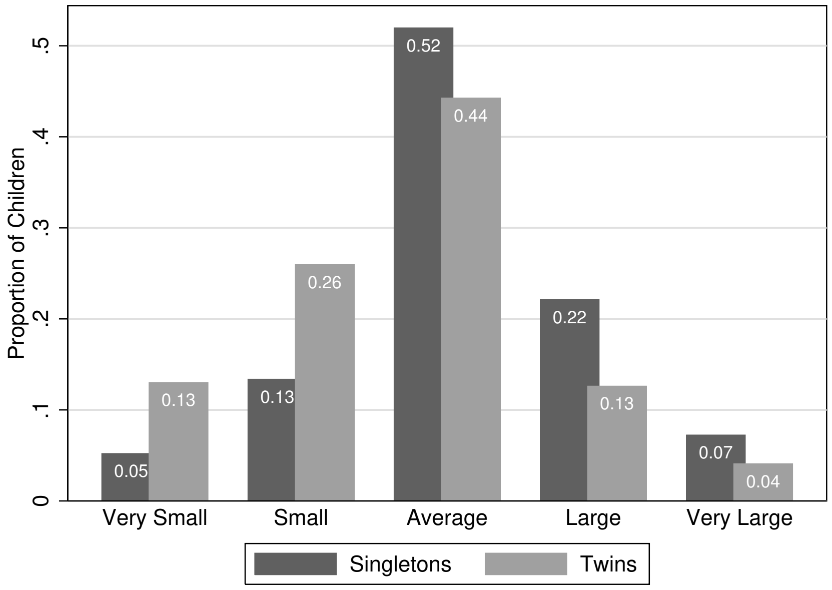


(a) Trends in Fertility

(b) Trend in Education

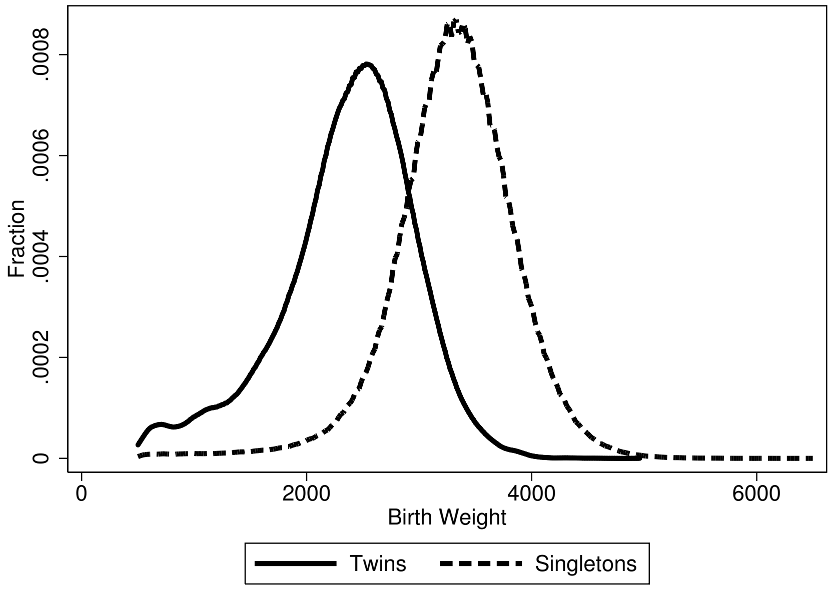
Note to figure A2: Cohorts are made up of all individuals from the DHS who are aged over 35 years (for fertility), and over 15 years (for education). In each case the sample is restricted to those who have approximately completed fertility and education respectively. Full summary statistics for these variables are provided in table A2, and a full list of country and survey years are available in table A12.

Figure A3: Birth Size of Twins versus Singletons (Developing Countries)



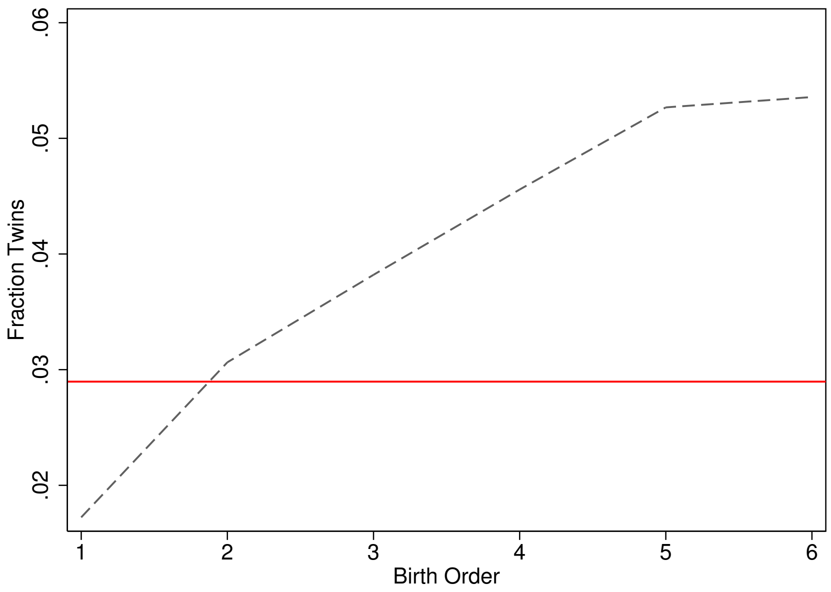
Note to figure A3 Estimation sample consists of all surveyed births from DHS countries occurring within 5 years prior to the date of the survey. For each of these births, all mothers retrospectively report the (subjective) size of the baby at the time of birth.

Figure A4: Birth Weight of Twins versus Singletons (USA)



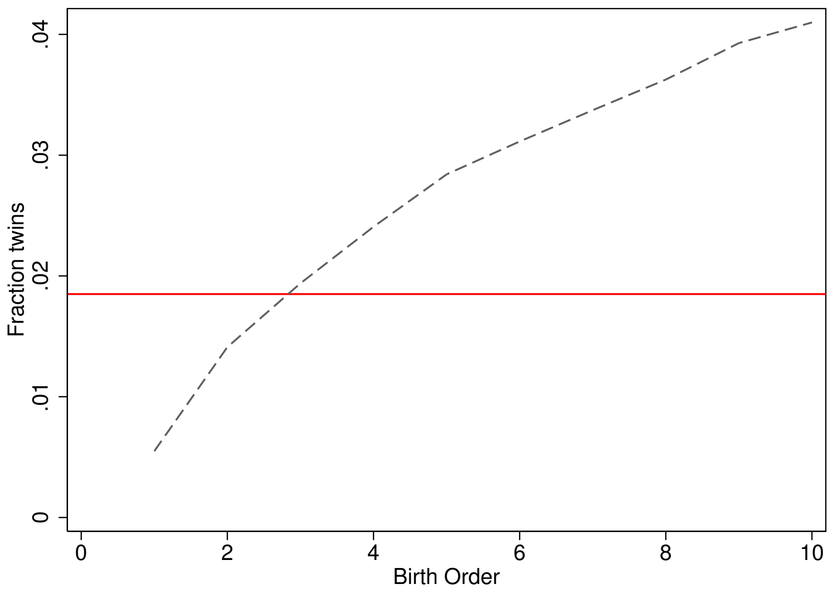
Note to figure A4 Estimation sample consists of all non-ART births from NVSS data between 2009 and 2013. Birthweights below 500 grams and above 6,500 grams are trimmed from the sample.

Figure A5: Proportion of Twins by Birth Order (United States)



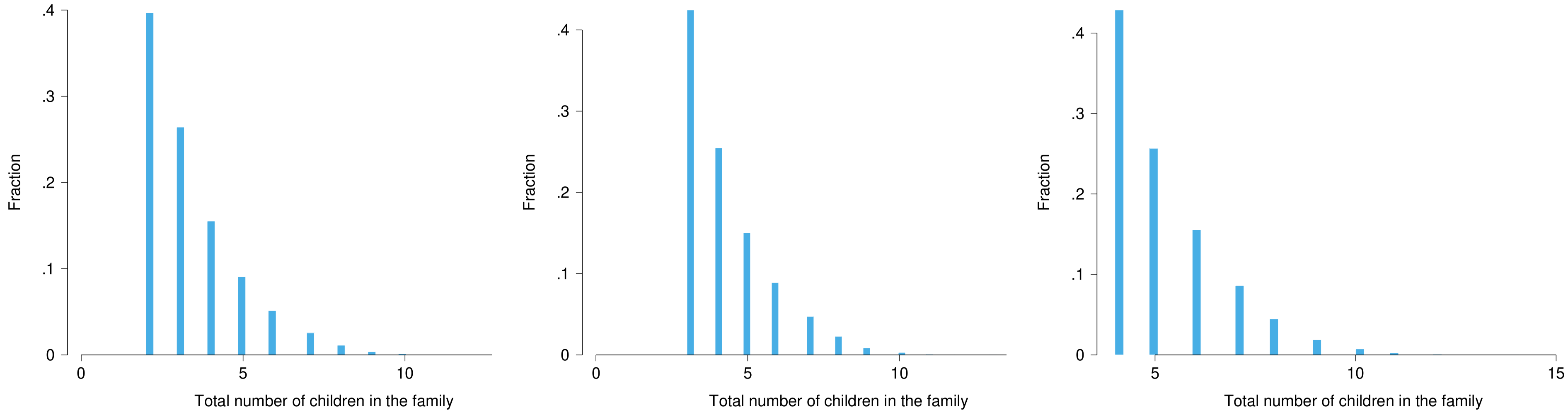
Note to figure A5 The fraction of twin births are calculated from the full sample of non-ART users in NVSS data from 2009-2013. The solid line represents the average fraction of twins in the full sample (2.89%), while the dotted line presents twin frequency by birth order. The dotted line joins points 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 A6: Proportion of Twins by Birth Order (Developing Countries)



Note to figure A6 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 A7: Total Family Size in Analysis Samples



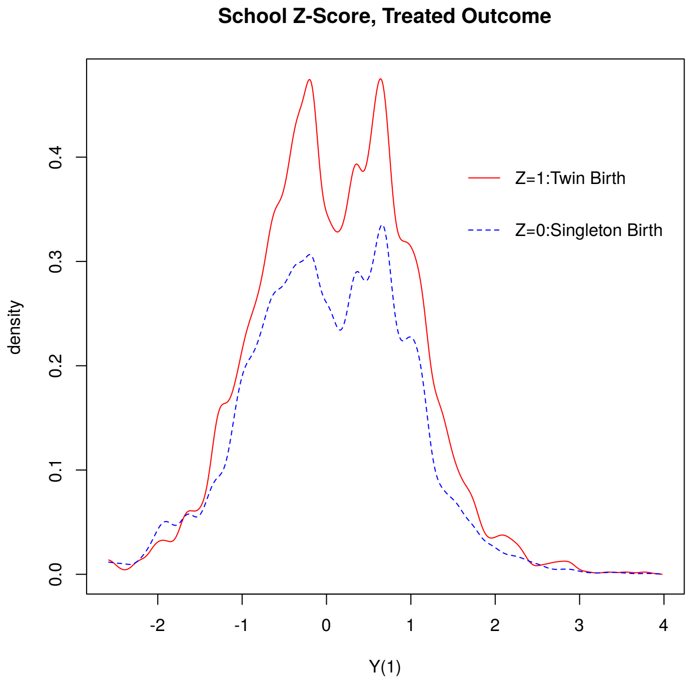
(a) Two-Plus Group

(b) Three-Plus Group

(c) Four-Plus Group

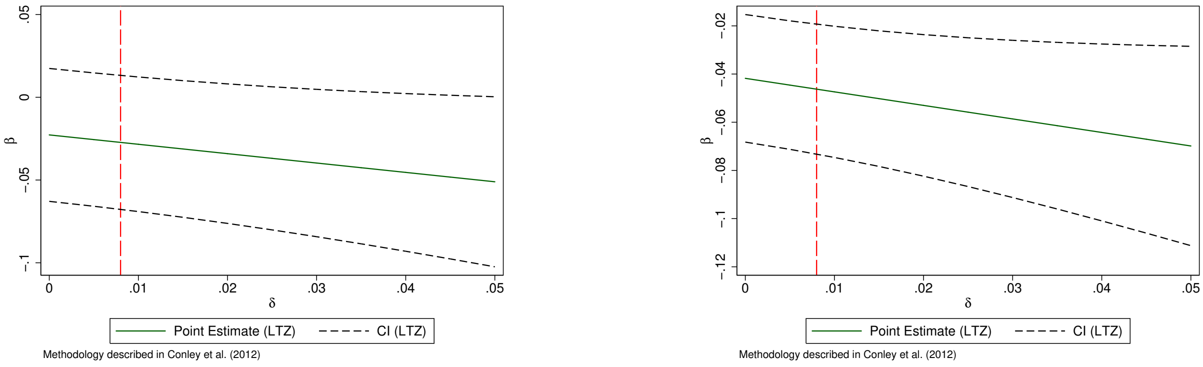
Note to figure A7: Histograms display the total family size of families meeting inclusion criteria for each estimation sample (two-plus, three-plus, and four-plus). By definition, the two-plus sample only includes families with at least two births, the three-plus sample only includes families with at least three births, and the four-plus sample only includes families with at least four births.

Figure A8: Density Test of Instrumental Validity from Kitagawa (2015)



Note to figure A8: Kernel density plots document the sub-densities of the outcome variable of interest in IV regressions (school Z-score) for children preceding twins and for children not preceding twins in the 2+ sample. “Treated” refers to families with at least 3 children, and so both densities document frequencies only for this group. The Kitagawa (2015) test consists of determining whether the two densities intersect, with intersection being evidence of instrumental *in*validity. We follow Kitagawa in using a Gaussian kernel and bandwidth of 0.08. Outliers are suppressed from the graph to ease visualisation of the sub-densities. Results for the full version of the test including controls along with p-values associated with instrumental invalidity are presented in table A10.

Figure A9: Plausibly Exogenous Bounds: School Z-Score (Developing Countries 2+ and 4+)

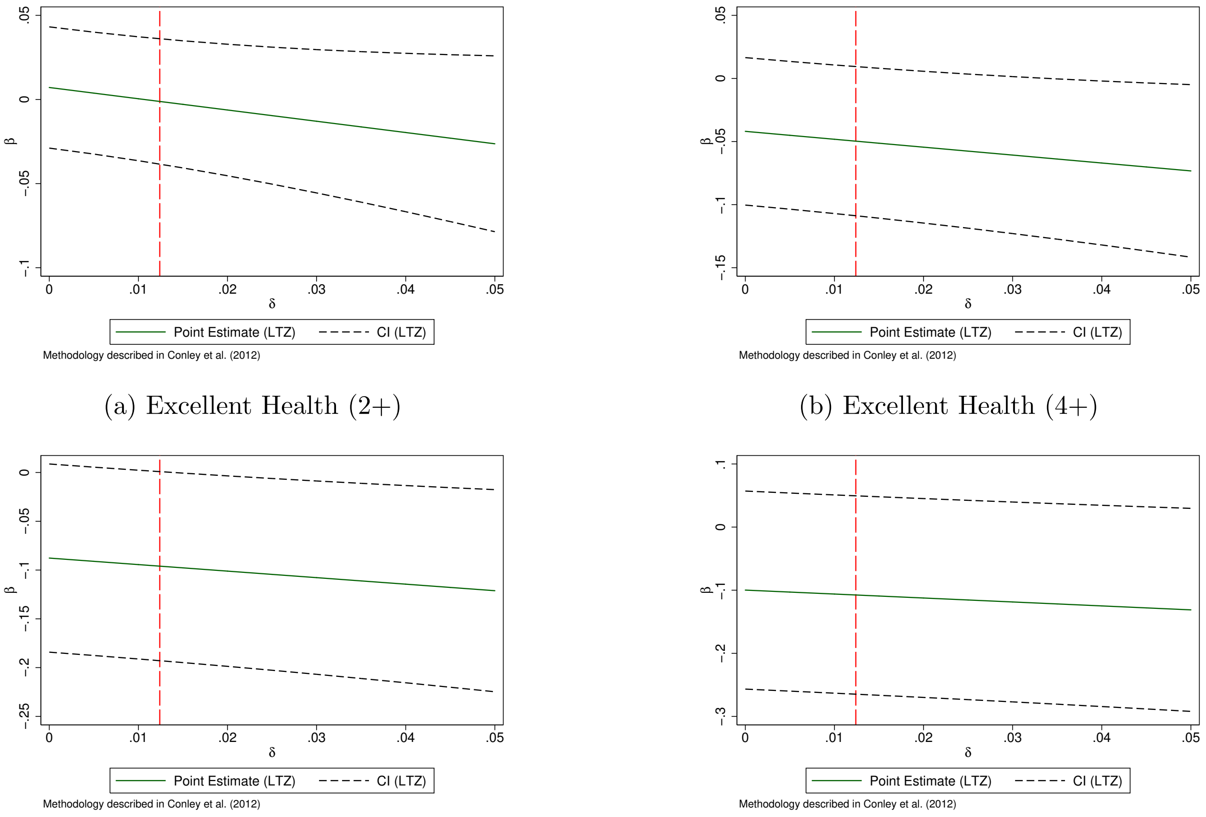


(a) Two Plus

(b) Four Plus

Note to figure A9: Refer to notes to figure 5 of the main text.

Figure A10: Plausibly Exogenous Bounds: (USA 2+ and 4+)



(a) Excellent Health (2+)

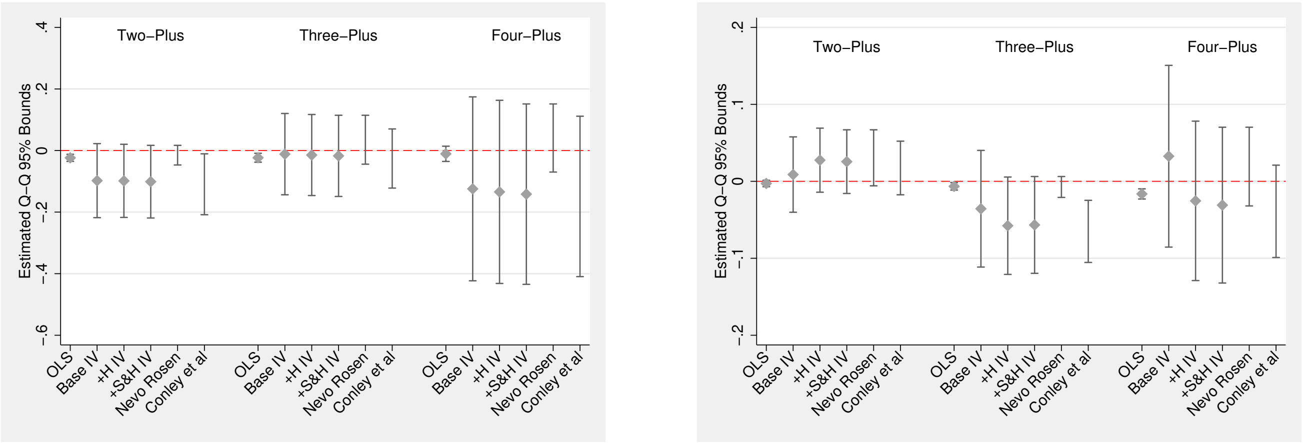
(b) Excellent Health (4+)

(c) Education Z-Score (2+)

(d) Education Z-Score (4+)

Note to figure A10: Refer to notes to figure 5 of the main text.

Figure A11: Parameter and Bound Estimates of the Q–Q Trade-off (USA)



(a) Education Z-Score

(b) Excellent Health

Note to Figure A11: Refer to notes to Figure 4. Identical bounds are presented, but in this case based on NHIS data (with considerably fewer observations).

Table A1: The Quantity–Quality Trade-off and the Twin Instrument: Recent Studies

				Estimates	
Author	Data, Period	Controls Included	Sample	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 (0.003)	-0.038 (0.047)
			Three Plus	-0.076 (0.004)	-0.016 (0.044)
			Four Plus	-0.059 (0.006)	-0.024 (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 (0.000)	0.002 (0.003)
			Three Plus	0.017 (0.001)	0.010 (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 (0.005)	0.174 (0.166)
			Three Plus	-0.143 (0.005)	0.167 (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 (-29.6)†	0.002 (0.18)†
			Three Plus	-0.038 (-21.4)†	-0.024 (-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 (0.001)	-0.019 (0.015)
			Three Plus	-0.020 (0.001)	0.007 (0.025)
			Four Plus	-0.018 (0.002)	-0.032 (0.036)

Author	Data, Period	Controls Included	Sample	Estimates OLS	 IV
(6) Rosenzweig and Zhang (2009)	The Chinese Child Twins Survey (CCTS), 2002-2003. Individuals selected from twins' (aged 7-18) and non-twin households. Outcome (reported here) is years of schooling	Mother's age at time of birth, child gender and age.	Reduced Form Reduced Form + Bwt	-0.307 (1.92)† -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) Two Plus (F) Three Plus (M) Three Plus (F)	-0.233 (0.010) -0.277 (0.015) -0.230 (0.010) -0.283 (0.015)	-0.137 (0.146) -0.372 (0.198) -0.060 (0.164) -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 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	3.094	1.955
	(2.351)	(2.724)	(2.436)	(1.002)	(1.185)	(1.024)
Age	31.18	35.49	36.16	37.24	36.19	
	(8.095)	(7.385)	(8.113)	(8.423)	(8.069)	(8.415)
Education	4.823	3.582	4.772	12.57	12.74	12.58
	(4.721)	(4.330)	(4.712)	(2.310)	(2.220)	(2.308)
Height	155.6	157.4	155.7	-	-	-
	(7.075)	(7.050)	(7.083)	-	-	-
BMI	23.31	23.69	23.32	27.65	28.12	27.66
	(4.819)	(5.004)	(4.827)	(6.715)	(7.326)	(6.732)
Pr(BMI)<18.5	0.124	0.100	0.123	0.0197	0.0159	0.0196
	(0.330)	(0.300)	(0.329)	(0.139)	(0.125)	(0.139)
Excellent Health	-	-	-	0.318	0.324	0.318
	-	-	-	(0.465)	(0.468)	(0.465)
Children's Outcomes						
Age	11.55	11.67	11.56	11.19	10.77	11.18
	(3.287)	(3.278)	(3.286)	(3.891)	(3.901)	(3.891)
Education (Years)	3.584	3.174	3.556	5.151	4.650	5.139
	(3.152)	(3.022)	(3.145)	(3.851)	(3.769)	(3.850)
Education (Z-Score)	0.00423	-0.100	0.000	0.00274	-0.110	0.0000
	(0.982)	(0.981)	(1.000)	(1.001)	(0.950)	(1.000)
Infant Mortality	0.0587	0.137	0.0592	-	-	-
	(0.235)	(0.137)	(0.236)	-	-	-
Excellent Health	-	-	-	0.531	0.541	0.531
	-	-	-	(0.499)	(0.498)	(0.499)
Fraction Twin			0.0203			0.0257
			(0.139)			(0.158)
Birth Order Twin			4.448			2.196
			(2.457)			(1.064)
Observations	2,046,879	41,547	2,005,332	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 874,945 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 A12.

Table A3: 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
Fertility	-0.117*** (0.001)	-0.101*** (0.001)	-0.067*** (0.001)		-0.128*** (0.001)	-0.108*** (0.001)	-0.072*** (0.001)
Birth Order 2				-0.175*** (0.004)	-0.057*** (0.004)	-0.061*** (0.003)	-0.040*** (0.003)
Birth Order 3				-0.352*** (0.005)	-0.099*** (0.005)	-0.109*** (0.005)	-0.071*** (0.005)
Birth Order 4				-0.493*** (0.006)	-0.099*** (0.006)	-0.117*** (0.006)	-0.075*** (0.006)
Birth Order 5				-0.596*** (0.007)	-0.062*** (0.007)	-0.089*** (0.007)	-0.057*** (0.007)
Birth Order 6				-0.688*** (0.008)	-0.018** (0.009)	-0.057*** (0.009)	-0.044*** (0.008)
Birth Order 7				-0.750*** (0.009)	0.051*** (0.010)	-0.005 (0.010)	-0.014 (0.010)
Birth Order 8				-0.786*** (0.010)	0.138*** (0.012)	0.066*** (0.012)	0.031*** (0.011)
Birth Order 9				-0.839*** (0.012)	0.206*** (0.014)	0.113*** (0.014)	0.054*** (0.013)
Birth Order ≥ 10				-0.856*** (0.014)	0.395*** (0.016)	0.268*** (0.016)	0.163*** (0.015)
Observations	1,128,699	1,128,699	1,128,699	1,128,699	1,128,699	1,128,699	1,128,699

Table A4: OLS Estimates with and without Birth Order Controls (USA)

	No Birth Order FEs			Birth Order FEs			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Base	+S	+S+H	No Fertility	Base	+S	+S+H
Fertility	-0.026***	-0.027***	-0.023***		-0.023***	-0.024***	-0.020***
	(0.004)	(0.004)	(0.004)		(0.004)	(0.004)	(0.004)
Birth Order 2				-0.049***	-0.032***	-0.033***	-0.033***
				(0.008)	(0.008)	(0.008)	(0.008)
Birth Order 3				-0.103***	-0.061***	-0.060***	-0.059***
				(0.015)	(0.015)	(0.015)	(0.015)
Birth Order 4				-0.121***	-0.053**	-0.050**	-0.046*
				(0.025)	(0.025)	(0.025)	(0.025)
Birth Order 5				-0.095**	0.002	0.012	0.018
				(0.046)	(0.045)	(0.045)	(0.045)
Birth Order 6				-0.194**	-0.065	-0.057	-0.043
				(0.083)	(0.081)	(0.081)	(0.081)
Birth Order 7				-0.236	-0.079	-0.062	-0.047
				(0.157)	(0.157)	(0.156)	(0.158)
Birth Order 8				0.012	0.191	0.196	0.220
				(0.498)	(0.497)	(0.495)	(0.495)
Birth Order 9				-0.460***	-0.259**	-0.250**	-0.207
				(0.107)	(0.115)	(0.123)	(0.133)
Birth Order $\geq$ 10				-0.421***	-0.181***	-0.184***	-0.148**
				(0.054)	(0.056)	(0.054)	(0.068)
Observations	163,931	163,931	163,931	163,931	163,931	163,931	163,931

Table A5: Full Output on Health and Socioeconomic Controls from IV Estimates (Developing Countries)

Dependent Variable	2+		3+		4+	
School Z-Score	+H	+S&H	+H	+S&H	+H	+S&H
Fertility	-0.015	-0.012	-0.041*	-0.046**	-0.038*	-0.037**
	(0.027)	(0.026)	(0.021)	(0.020)	(0.021)	(0.019)
Mother's Height	0.009***	0.003***	0.009***	0.003***	0.008***	0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mother's BMI	0.026***	0.012***	0.027***	0.013***	0.028***	0.014***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Doctor Availability	0.203***	-0.029	0.189***	-0.056***	0.194***	-0.048***
	(0.024)	(0.019)	(0.020)	(0.016)	(0.020)	(0.017)
Nurse Availability	0.103***	0.113***	0.122***	0.114***	0.156***	0.120***
	(0.013)	(0.012)	(0.013)	(0.012)	(0.014)	(0.013)
No Prenatal Care Available	-0.457***	-0.259***	-0.483***	-0.302***	-0.523***	-0.364***
	(0.022)	(0.019)	(0.019)	(0.017)	(0.019)	(0.017)
Poorest Quintile		-0.275***		-0.265***		-0.246***
		(0.014)		(0.011)		(0.011)
Quintile 2		-0.114***		-0.114***		-0.088***
		(0.011)		(0.010)		(0.010)
Quintile 3		-0.037***		-0.030***		0.002
		(0.011)		(0.010)		(0.010)
Quintile 4		0.026**		0.058***		0.116***
		(0.010)		(0.010)		(0.010)
Richest Quintile		0.155***		0.229***		0.327***
		(0.012)		(0.011)		(0.012)
Observations	259,958	259,958	395,687	395,687	409,576	409,576
R-Squared	0.075	0.153	0.078	0.158	0.071	0.154

Notes: Full output is presented from IV regressions displayed in Table 6 on health and socioeconomic controls from models denoted "+H" (adding health controls) and "+S&H" (adding health and socioeconomic controls). Additionally, fixed effects for years of education of the mother are included in regressions though are not displayed in the interests of space. These fixed effects show a positive gradient with higher education associated with additional child education. Full notes are available in Table 6.

Table A6: Full Output on Health and Socioeconomic Controls from IV Estimates (USA Education)

Dependent Variable	2+		3+		4+	
School Z-Score	+H	+S&H	+H	+S&H	+H	+S&H
Fertility	-0.099	-0.101*	-0.015	-0.017	-0.134	-0.142
	(0.061)	(0.060)	(0.067)	(0.067)	(0.152)	(0.149)
Excellent Health	0.139	0.131	-0.047	-0.027	0.325	0.353
	(0.181)	(0.178)	(0.228)	(0.230)	(0.602)	(0.606)
Very good Health	0.141	0.134	-0.049	-0.028	0.293	0.322
	(0.181)	(0.178)	(0.228)	(0.229)	(0.602)	(0.606)
Good Health	0.080	0.086	-0.102	-0.066	0.247	0.289
	(0.181)	(0.178)	(0.228)	(0.230)	(0.602)	(0.606)
Fair Health	0.006	0.024	-0.186	-0.140	0.200	0.249
	(0.181)	(0.179)	(0.229)	(0.230)	(0.602)	(0.606)
Poor Health	-0.098	-0.070	-0.293	-0.232	-0.020	0.047
	(0.186)	(0.183)	(0.235)	(0.236)	(0.609)	(0.612)
Mother's Height	0.079	0.061	0.187	0.168	0.123	0.128
	(0.102)	(0.102)	(0.139)	(0.138)	(0.239)	(0.240)
Mother's Height Squared	-0.001	-0.000	-0.001	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Smoked Prior to Pregnancy	-0.047***	-0.041***	-0.051**	-0.046**	-0.055	-0.051
	(0.015)	(0.015)	(0.020)	(0.020)	(0.040)	(0.041)
No Response to Smoking	0.046*	0.041	0.062*	0.052	0.094*	0.079
	(0.026)	(0.025)	(0.034)	(0.033)	(0.055)	(0.054)
Observations	61,267	61,267	47,308	47,308	21,352	21,352
R-Squared	0.000	0.003	0.003	0.008	-0.005	-0.004

Notes: Full output is presented from IV regressions displayed in Table 7 on health and socioeconomic controls from models denoted "+H" (adding health controls) and "+S&H" (adding health and socioeconomic controls). Additionally, fixed effects for years of education of the mother are included in regressions though are not displayed in the interests of space. These fixed effects show a positive gradient with higher education associated with additional child education. Full notes are available in Table 7.

Table A7: Full Output on Health and Socioeconomic Controls from IV Estimates (USA Health)

Dependent Variable	2+		3+		4+	
Excellent Health	+H	+S&H	+H	+S&H	+H	+S&H
Fertility	0.027	0.026	-0.058*	-0.057*	-0.025	-0.031
	(0.021)	(0.021)	(0.032)	(0.032)	(0.053)	(0.052)
Excellent Health	0.501***	0.499***	0.451***	0.455***	0.089	0.090
	(0.090)	(0.090)	(0.136)	(0.136)	(0.134)	(0.128)
Very good Health	-0.022	-0.023	-0.076	-0.071	-0.435***	-0.434***
	(0.090)	(0.090)	(0.136)	(0.136)	(0.134)	(0.128)
Good Health	-0.112	-0.107	-0.172	-0.164	-0.547***	-0.541***
	(0.090)	(0.090)	(0.136)	(0.136)	(0.134)	(0.128)
Fair Health	-0.096	-0.087	-0.146	-0.136	-0.492***	-0.485***
	(0.090)	(0.090)	(0.137)	(0.136)	(0.134)	(0.128)
Poor Health	-0.097	-0.085	-0.132	-0.119	-0.598***	-0.588***
	(0.092)	(0.091)	(0.139)	(0.138)	(0.138)	(0.132)
Mother's Height	-0.018	-0.024	-0.001	-0.003	0.013	0.022
	(0.046)	(0.046)	(0.068)	(0.068)	(0.120)	(0.121)
Mother's Height Squared	0.000	0.000	0.000	0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Smoked Prior to Pregnancy	0.016**	0.019***	0.008	0.011	0.008	0.010
	(0.007)	(0.007)	(0.010)	(0.010)	(0.019)	(0.019)
No Response to Smoking	0.001	-0.001	-0.004	-0.005	-0.025	-0.027
	(0.011)	(0.011)	(0.016)	(0.016)	(0.027)	(0.027)
Observations	70,277	70,277	53,393	53,393	24,358	24,358
R-Squared	0.295	0.298	0.295	0.296	0.304	0.306

Notes: Full output is presented from IV regressions displayed in Table 7 on health and socioeconomic controls from models denoted "+H" (adding health controls) and "+S&H" (adding health and socioeconomic controls). Additionally, fixed effects for years of education of the mother are included in regressions though are not displayed in the interests of space. These fixed effects show a positive gradient with higher education associated with additional child education. Full notes are available in Table 7.

Table A8: Developing Country IV Estimates Using Same Sex Twins Only

	2+			3+			4+		
	Base	+H	+S&H	Base	+H	+S&H	Base	+H	+S&H
Panel A: First Stage									
Dependent Variable = Fertility									
Same Sex Twins	0.703***	0.713***	0.717***	0.687***	0.709***	0.713***	0.773***	0.776***	0.783***
	(0.034)	(0.034)	(0.033)	(0.031)	(0.030)	(0.030)	(0.033)	(0.034)	(0.034)
Kleibergen-Paap rk statistic	419.61	440.97	475.44	506.94	547.91	561.61	552.39	517.49	544.88
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	0.007	-0.006	-0.007	-0.039	-0.065*	-0.072**	0.013	0.006	-0.000
	(0.046)	(0.045)	(0.043)	(0.036)	(0.033)	(0.031)	(0.035)	(0.032)	(0.028)
Observations	259,954	259,954	259,954	395,693	395,693	395,693	409,573	409,573	409,573
Coefficient Difference		0.102	0.313		0.001	0.022		0.522	0.399

Refer to notes to table 6. This table follows identical specifications, however now only same sex twins are used as an instrument instead of all twins. In the DHS, 64.1% of twin pairs are of the same gender. Standard errors are clustered by mother.*p<0.1; **p<0.05; ***p<0.01

Table A9: US IV Estimates Using Same Sex Twins Only

	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)									
Same Sex Twins	0.715***	0.717***	0.718***	0.767***	0.767***	0.770***	0.840***	0.845***	0.853***
	(0.031)	(0.031)	(0.031)	(0.055)	(0.054)	(0.054)	(0.112)	(0.112)	(0.111)
Kleibergen-Paap rk statistic	522.05	526.26	541.38	196.87	199.35	202.38	55.86	56.51	59.56
p-value of rk statistic	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Dependent Variable = Fertility (Excellent Health Second Stage)									
Same Sex Twins	0.752***	0.754***	0.755***	0.780***	0.781***	0.783***	0.823***	0.831***	0.841***
	(0.030)	(0.030)	(0.030)	(0.051)	(0.050)	(0.050)	(0.105)	(0.105)	(0.104)
Kleibergen-Paap rk statistic	630.23	637.62	654.36	235.37	239.03	243.67	61.01	62.54	66.02
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	-0.063	-0.063	-0.065	-0.018	-0.022	-0.026	0.096	0.089	0.084
	(0.061)	(0.060)	(0.060)	(0.082)	(0.082)	(0.083)	(0.119)	(0.116)	(0.115)
Observations	61,267	61,267	61,267	47,308	47,308	47,308	21,352	21,352	21,352
Coefficient Difference		0.963	0.772		0.421	0.272		0.536	0.436
Dependent Variable = Excellent Health									
Fertility	0.003	0.032	0.031	-0.020	-0.062*	-0.061*	0.074	-0.001	-0.004
	(0.030)	(0.025)	(0.025)	(0.046)	(0.037)	(0.037)	(0.067)	(0.055)	(0.054)
Observations	70,277	70,277	70,277	53,393	53,393	53,393	24,358	24,358	24,358
Coefficient Difference		0.056	0.070		0.133	0.143		0.083	0.073

NOTES: Refer to notes in table 7. This table follows identical specifications, however now only same sex twins are used as an instrument instead of all twins. In the NHIS, 66.0% of twin pairs are of the same gender. Standard errors are clustered by mother. *p<0.1; **p<0.05; ***p<0.01

Table A10: Results for Kitagawa (2015) Tests with Controls (DHS)

	Baseline	Socioeconomic	Socioeconomic plus Health
Kitagawa Test Statistic	14.559	15.963	16.558
Instrumental Validity (p-value)	0.028	0.224	0.462
Coefficient (IV model)	-0.013	-0.032	-0.042
	(0.073)	(0.068)	(0.068)
Observations	251,831	251,831	251,831

Notes: Results are presented for the Kitagawa (2015) test of instrumental validity. This test exists for a binary endogenous variable, and as such rather than estimate a model with fertility as the endogenous variable, we estimate a model with the binary variable "greater than 2 births" as the endogenous variable. The instrument considered is twinning at birth order 2. The estimation results of a typical IV model using this specification are presented and indicated as "IV model". Instrumental validity can not be proven, but can be disproven, with low *p*-values being evidence against instrumental validity. The first row shows the value for the variance weighted test statistic proposed by Kitagawa (2015), and the second row displays the *p*-values associated with the Kitagawa test. Baseline controls consist of mother year of birth fixed effects, continent fixed effects, child sex, and decade of birth fixed effects. Socioeconomic controls add indicators for mother's education (0 years, 1-6 years, 7-11 years, or 12+ years), and Health controls add indicators for overweight or underweight mothers, and whether the majority of births in the mother's region were attended by doctors, nurses or unattended. A trimming constant of 0.07 is used for the instrumental validity test, (as laid out in Kitagawa (2015)), and 500 bootstrap replications are run to determine the p-value.

Table A11: Maternal Characteristics and Child Educational Outcomes

	Z-Scores				Continuous Variables/Indicators			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Educ	Educ	Educ	Educ	Educ	Educ	Educ	Educ
Maternal Education (Z)	0.225***			0.215***				
	(0.002)			(0.002)				
BMI (Z-Score)		0.088***		0.070***				
		(0.001)		(0.001)				
Height (Z-Score)			0.032***	0.019***				
			(0.001)	(0.001)				
Body Mass Index					0.014***		0.015***	0.014***
					(0.000)		(0.000)	(0.000)
Height in Centimetres						0.003***	0.003***	0.003***
						(0.000)	(0.000)	(0.000)
Availability of Doctors								-0.058***
								(0.009)
Availability of Nurses								0.101***
								(0.008)
No Prenatal Care								-0.305***
								(0.010)
R-Squared	0.16	0.13	0.12	0.17	0.17	0.16	0.17	0.17
Observations	1,128,699	1,128,699	1,128,699	1,128,699	1,128,699	1,128,699	1,128,699	1,128,699

The dependent variable in each model is each child's standardized educational attainment compared with his/her cohort. Table headers (Z-Scores and Continuous Variables) refers to the form of the independent variables at the level of the mother. All regressions are estimated by OLS, and cluster standard errors by mother. Each specification includes fixed effects for mother and child age, total fertility, country and year, and family wealth quintile. In columns 5-7 maternal year of education fixed effects are included. Columns 1-4 use standardised variables for education, BMI and height, where Z-scores are constructed comparing each mother to those in her country and survey wave.

Table A12: 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.

B Data Definitions

All outcome and control variables used in principal IV and OLS analyses are described in the following table. As well as variable definitions, units and any functional forms are indicated, which refer to the way variables enter IV or OLS models.

Table A13: Variable Definitions

Variable	Definition
Panel A: DHS Data	
School Z-score	Z-score of years of schooling, standardised relative to country and year of birth cohort.
Male Child	Binary measure, one for boy, zero for girls
Country	Fixed effect for country of survey
Year of Birth	Fixed effect for year of birth
Child's Age	Fixed effect for child's age
Contraceptive Intent	Fixed effect for mother's use of contraceptive methods
Mother's Age	Fixed effect for mother's age at child birth
Mother's Age at First Birth	Inferred from age at survey time and age of child
Mother's Education	Fixed effect for total years of education achieved
Family Wealth	Fixed effect for DHS-assigned wealth quintile. Where not recorded a separate fixed effect for "no wealth quintile" is included
Mother's Height	Measured in centimetres
Mother's BMI	Measured in units (weight in kilograms divided by height in metres squared)
Prenatal Doctor Availability	Proportion of births in the same DHS cluster which received a prenatal check-up from a doctor
Prenatal Nurse Availability	Proportion of births in the same DHS cluster which received a prenatal check-up from a nurse
No Prenatal Care	Proportion of births in the same DHS cluster which received no prenatal check-ups from health professionals
Panel B: NHIS Data	
Education Z-Score	Z-score of grade progression, standardised relative to month and year of birth cohort
Excellent Health	Indicator of whether a child is classified by the family as being in "excellent health" (chosen from a categorical list)
Male Child	Binary measure, one for boy, zero for girls
Survey Year	Fixed effect for year NHIS wave was run
Child Age	Fixed effect for age at interview in months and years
Region	Fixed effect for census bureau region of residence
Mother's Race	Fixed effect for mother's race
Mother's Age	Fixed effect for mother's age in years
Mother's Age at First Birth	Inferred from age at survey time and age of child
Mother's Education	Fixed effects for mother's highest completed year of education
Mother's Health Status	Self-reported based on categorical list
Mother's Height	Mother's Height in Inches
Smoking Status	Binary variable indicating whether the mother smoked prior to pregnancy
Smoking Status Missing	Binary variable indicating no response to the mother's smoking status

C Testing for Equality of Coefficients Between IV Models

When estimating subsequent IV models with the progressive inclusion of controls to capture maternal selection, our point is really that column 1 (“Base”) is not distinguishable from 0, while column 3 (“+S&H”) often is, as this is the important thing in considering the literature and in showing that partial bias adjustment recovers the trade-off. We have nevertheless added a formal test of coefficients between IV models in all IV tables. This is added as a row called “Coefficient Difference” at the bottom of Tables 6 and 7. This computation is not entirely trivial, as these tests must take account of correlations between variance-covariance matrices of each IV regression in the style of seemingly unrelated regression. Thus, we calculate these test statistics by jointly estimating the models with GMM (seemingly unrelated regression is an Feasible Generalised Least Squares technique, and hence not suitable for IV models). To do this we form two equations which are the two models we wish to compare in the following format:

$$
\begin{aligned}
quality_{ij} &= b_0 + b_1 \times quantity_j + \boldsymbol{baseline}'_{ij} \times \boldsymbol{b}_b && \text{(A1)}\\
quality_{ij} &= c_0 + c_1 \times quantity_j + \boldsymbol{baseline}'_{ij} \times \boldsymbol{c}_b + \boldsymbol{health}'_{ij} \times \boldsymbol{c}_h. && \text{(A2)}
\end{aligned}
$$

Our goal is to test the equality of coefficients $b_1 = c_1$. Given that we are using instruments for endogenous quantity (fertility) in each case, we can thus form the following population moment conditions which hold under the null of instrumental validity in each case (ie, replicate the specifications we are estimating in the paper):

$$
\begin{aligned}
twin'_i(quality_{ij} - b_0 - b_1 \times quantity_j - \boldsymbol{baseline}'_{ij} \times \boldsymbol{b}_b) &= 0 && \text{(A3)}\\
twin'_i(quality_{ij} - c_0 - c_1 \times quantity_j - \boldsymbol{baseline}'_{ij} \times \boldsymbol{c}_b - \boldsymbol{health}'_{ij} \times \boldsymbol{c}_h) &= 0. && \text{(A4)}
\end{aligned}
$$

Using the sample analogues of these moments, we can then estimate the parameters $\boldsymbol{b}$ and $\boldsymbol{c}$ via GMM. Denoting the two moments as the 2 element vector $g(\widehat{bc})$, we then estimate the parameters $\widehat{b}$ and $\widehat{c}$ using the GMM objective function $J(\widehat{bc}) = ng(\widehat{bc})'Wg(\widehat{bc})$. An unadjusted weight matrix is used which assumes that the moment conditions are independent, which replicates all parameters and standard errors from the original IV model, but now the estimates can be formally tested for equality against one-another using a χ^2 test which also considers the correlation between the observations in the two models when estimating the eventual variance-covariance matrix.

D Loosening the Linear Effect Specification of the Q–Q Trade-off

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. (2017); Mogstad and Wiswall (2016); Bagger et al. (2013) 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.[1]

Using DHS data for which we have sufficient power to split instruments, we re-estimate our regressions following the non-linear marginal fertility models of Brinch et al. (2017); Mogstad and Wiswall (2016), and find that as is the case with the linear models reported in Tables 6 and 7, the inclusion of twin predictors nearly universally increases the size of the estimated QQ trade-off in non-linear models, and in some specifications, the trade-off is statistically significant. Thus the emergence of a trade-off following partial correction for twin non-randomness is not sensitive to functional form and, in particular, holds when the impact of fertility is allowed to vary by parity.

[1] In this paper we focus nearly exclusively on the internal validity of twins estimates (IV consistency). In recent work, Aaronson et al. (2017) examine the external validity of twin instrumented estimates of the impact of fertility on female labour supply.

As laid out in Mogstad and Wiswall (2016), this consists of the following 2SLS procedure (for the two-plus sample):

$$\begin{aligned} quantity_{sj} &= \lambda_{s2} twin_{2j} + \lambda_{s3} twin^*_{3j} + \lambda_{s4} twin^*_{4j} + \lambda_{s5} twin^*_{5j} + \boldsymbol{X}\boldsymbol{\lambda}_{\boldsymbol{X}\boldsymbol{s}} + \nu_{sj}, \qquad \text{for } s = 2,3,4,5 \quad \text{(A5)} \\ quality_{ij} &= \beta_0 + \beta_1 \widehat{quantity}_{2j} + \beta_1 \widehat{quantity}_{3j} + \beta_1 \widehat{quantity}_{4j} + \beta_1 \widehat{quantity}_{5j} + \boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{X}} + \eta_{ij}, \quad \text{(A6)} \end{aligned}$$

where (A5) is a series of first stages for the likelihood effect of moving from the s^{th} to $(s+1)^{th}$ child, and (A6) is the second stage estimate of the effect of an additional child after s births on the human capital of the first born child. As the estimation sample consists of families with at least two births, $twin_{2j}$: a binary variable for a twin at the second birth, is defined for all families. However, when moving to higher birth orders, $twin_{3j}$ is not defined for families with only two births. We thus follow Mogstad and Wiswall (2016) in replacing higher-order twin birth indicators with:

$$twin^*_{cj} = \begin{cases} 0, & \text{if } c_j < c \\ twin_{cj} - \hat{E}[twin_{cj}|X_j, c_j \geq c], & \text{if } c_j \geq c \end{cases}$$

where, as described in Mogstad and Wiswall (2016) $\hat{E}[twin_{cj}|X_j, c_j \geq c]$ is a non-parametric estimate of the conditional mean of the probability of twin birth in the non-missing subsample. We similarly follow Mogstad and Wiswall (2016) in considering family sizes up to 6 children. The above specification (A5 and A6) is estimated for the two-plus sample, however we also estimate analogous specifications for the three-plus sample, and four-plus sample, where in each case we only consider the marginal impacts of fertility at birth orders greater than the birth orders of the children included in the estimation sample.

As our interest is in examining the impact of non-random twin births, we estimate the above specifications in two circumstances: the first, following exactly the procedure laid out in Mogstad and Wiswall (2016) where twins are assumed to be exogenous, and the second where we additionally control for observable health and socioeconomic predictors of twins in A5 and A6. These results are presented and discussed in Section 5.2.3 of the paper.

Table A14: First Stages for Non-Linear IV Estimates

Instrument	twin_{2j}	twin^*_{3j}	twin^*_{4j}	twin^*_{5j}
Panel A: Two Plus Sample				
Siblings $\geq$ 2	0.296***	0.213***	0.121***	0.032***
	(0.005)	(0.012)	(0.010)	(0.007)
Siblings $\geq$ 3	-0.011	0.429***	0.189***	0.077***
	(0.008)	(0.007)	(0.013)	(0.010)
Siblings $\geq$ 4	-0.014*	-0.012	0.525***	0.174***
	(0.008)	(0.017)	(0.008)	(0.017)
Siblings $\geq$ 5	-0.001	-0.023	-0.009	0.653***
	(0.009)	(0.021)	(0.034)	(0.012)
Panel B: Three Plus Sample				
Siblings $\geq$ 3		0.393***	0.199***	0.075***
		(0.004)	(0.009)	(0.007)
Siblings $\geq$ 4		0.014	0.518***	0.186***
		(0.009)	(0.006)	(0.011)
Siblings $\geq$ 5		-0.007	0.008	0.645***
		(0.011)	(0.020)	(0.008)
Panel C: Four Plus Sample				
Siblings $\geq$ 4			0.480***	0.190***
			(0.004)	(0.009)
Siblings $\geq$ 5			0.009	0.634***
			(0.012)	(0.005)

Each row reports the first stage estimate of the number of children on twin births from the IV regressions displayed in table 8. In each case we report the first stages for the baseline specification of the Non-Linear IV, although results are quantitatively similar in the case of the +S&H specification. Standard errors are clustered by family(three plus and four plus samples), or robust to heteroscedasticity when only one child from each family is included in the regressions (two plus sample).

E Estimating Values for γ

We propose a number of methods of arriving to a non-arbitrary prior regarding γ in the Conley et al. (2012) method, where γ is the violation of the exclusion restriction when using twins as an instrument in the QQ model. From equation 4, 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. In this appendix we first discuss a proposed manner to causally estimate γ, and then present a number of consistency checks based on the data used in the QQ models of the paper which support estimated values of γ

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:

$$\gamma = \left.\frac{\partial quality_{ij}}{\partial twin_j}\right|_X = \left.\frac{\partial quality_{ij}}{\partial health_j}\right|_X \times \left.\frac{\partial health_j}{\partial twin_j}\right|_X = \phi_q \times \phi_t. \tag{A7}$$

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[2], or indeed use the analytical distribution of $\widehat{\gamma}$ directly in the bounds estimate of β_1.[3] We provide a summary of the assumptions underlying our bounds estimates and evidence in their favour in appendix E.4 and a full description of the resampling process in appendix E.5.

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 following 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, 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. We observe that bounds

[2] If, for example, we determine that γ is normally distributed, estimation then proceeds by imposing the prior distribution for γ as: $\gamma \sim \mathcal{N}(\widehat{\mu}_\gamma, \widehat{\sigma}^2_\gamma)$.

[3] 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 estimates are based on the entire empirical distribution, and we proceed using Conley et al. (2012)'s suggested simulation method.

estimates of this type are necessarily case specific (see, for instance, the examples provided in the Conley et al. paper) so, although our approach is of general interest in suggesting a process for bounding when violation of the exclusion restriction is small, the estimates produced here are only representative of the cases examined.

E.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 World 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.[4] 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 pneumonia 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 Figure A12).[5]

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

$$quality^{m+1}_{stc} = \alpha + \phi^q_1(Post_t \times basePneumonia^m_s) + \theta_{rs} + \eta_{rt} + \varphi \mathbf{X}^m_{st} + \lambda_{rc} + (\theta_s \times t) + \varepsilon_{stc} \qquad \text{(A8)}$$

where ϕ^q_1 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 $basePneumonia^m_s$ 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 sulfonamides) interacted with the indicator post.

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 twin non-randomness associations we present in the 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 A8. In other words, we calculate

$$\phi^t_1 = \overline{bP}_{stwin_j=1}|_X - \overline{bP}_{stwin_j=0}|_X = \left.\frac{\partial bP_s}{\partial twin_j}\right|_X .$$

[4] Results from (Bhalotra and Venkataramani, 2014) show that on a range of outcomes, scarring dominates selection – ie Sulfa exposure improves all socioeconomic outcomes. This suggests that selection due to survival of weaker births is small, and that the arrival of Sulfa drugs is appropriately viewed as a positive health shock.

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

In view of our findings related to twin selection, 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:

$$\phi_1^q \times \phi_1^t = \frac{\partial quality_{ij}}{\partial bP_s} \times \left.\frac{\partial bP_s}{\partial twin_j}\right|_X = \left.\frac{\partial quality_{ij}}{\partial twin_j}\right|_X = \gamma_{US}. \tag{A9}$$

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

E.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:

$$quality_{ites}^{m+1} = \alpha + \phi_2^q war_{te}^m + \alpha_t + \theta_e + \lambda_s + \mu_e t + u_{ites} \tag{A10}$$

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:

$$\phi_2^q \times \phi_2^t = \frac{\partial quality_{ij}}{\partial war_s} \times \left.\frac{\partial war_s}{\partial twin_j}\right|_X = \left.\frac{\partial quality_{ij}}{\partial twin_j}\right|_X = \gamma_{Nigeria}. \tag{A11}$$

E.3 Estimated Values for γ in US and Nigeria

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 A8 (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.[6] This estimate is the quantity ϕ_1^q in equations A8 and A9. In the second column,

[6] 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.

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 twin non-randomness tests of this paper, adding a further case of twin births being a function of health conditions.

Following equation A9, in column 3 we interact $\widehat{\phi}^q$ and $\widehat{\phi}^t$ to form a consistent estimate for γ of 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 A13a, overlaid with an analytical normal distribution with the same mean and variance. When comparing our estimate of γ to IV estimates discussed in section 5.2, we see that the direct effect of having a (healthier) twin mother on child quality (the violation) is considerably smaller than the point estimates of the effect of fertility on child outcomes (the parameter of interest). 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 A15. The first column presents an estimate of ϕ_2^q from equation A10. 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).[7] 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 positive estimate for γ of 0.4% of a standard deviation in education, not dissimilar to the value estimated using a different shock to maternal health in early twentieth century America. The bootstrapped distribution of γ based on 100 replications is displayed in Figure A13b (bootstrap variance 0.0022).

E.4 Assumptions and Evidence Underlying the Calculation of Plausibly Exogenous Bounds

The calculation of bounds using Conley et al. (2012)'s plausibly exogenous methodology relies on a number of assumptions relating to the exclusion restriction. We provide a full list of these assumptions, their precedence (be it from Conley et al. (2012)'s methodology or our extension to estimating γ and its empirical distribution), and supporting evidence for each.

1. There exists prior information that implies γ (the violation of the exclusion restriction) is near 0 but perhaps not exactly 0. Precedence: Conley et al. (2012), p. 262. Evidence: Tables 1-2 of our paper documents that twins occur more frequently to healthy women. This renders the exclusion restriction on which the twin-IV rests invalid if, in addition, earlier children of healthy women are higher quality children. Nevertheless, it is unlikely that the violation of the exclusion restriction is very large given that maternal health is only a small part of the production function of child quality.

2. The prior which is assumed for $\gamma \sim F$ is correct. Precedence: Conley et al. (2012), p. 265. Evidence: Refer to subpoints below.

    (a) The average value of γ *for a particular context* can be estimated using a single maternal health shock as a mediator to examine both elements of the violation of the exclusion restriction (twin non-randomness and the effect of maternal health on child quality). Precedence: This paper, equation A7. Evidence: the particular maternal health shock examined is a common factor in both effects. In the simplest case, if we scale a maternal health shock by a fixed parameter (for example

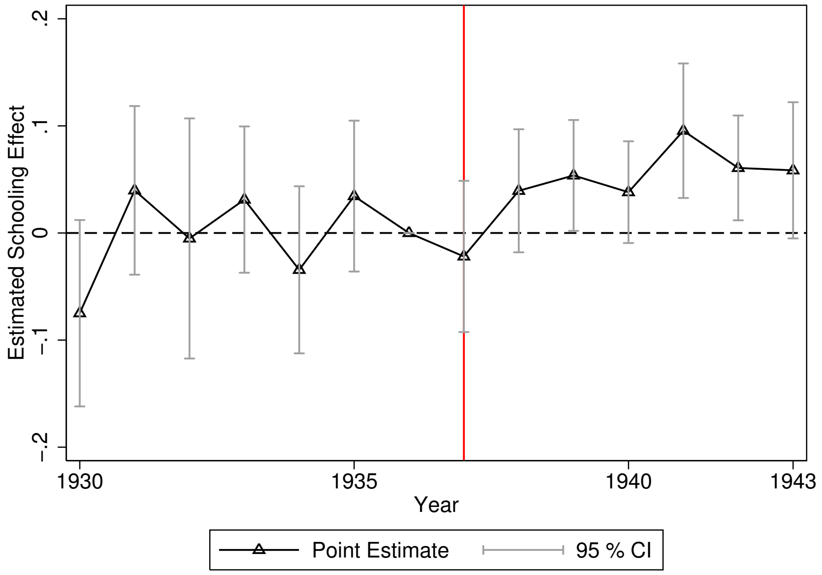
[7] This is not directly relevant here but, again, notice that the second-generation effect is smaller than the impact on the first generation, which is 0.6 years of education (Akresh et al., 2016).

Table A15: 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		

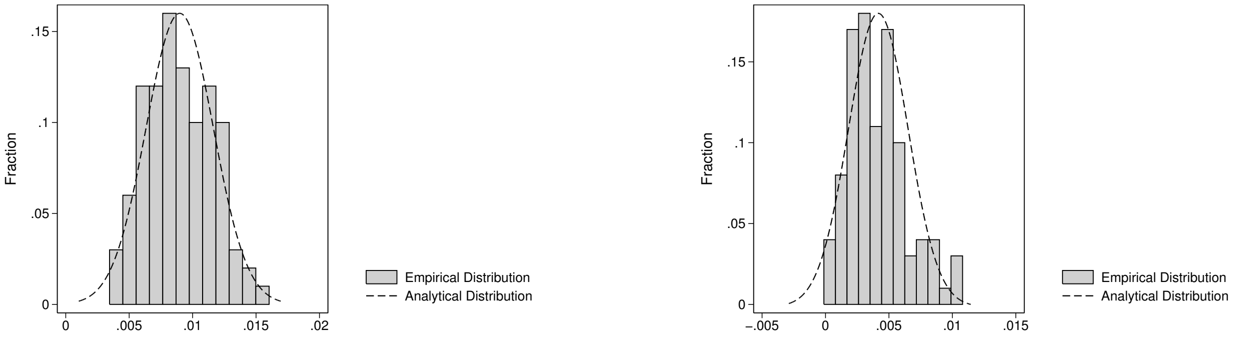
Notes: Regression results for panel A use the 5% sample of 1980 US census data and follow the specifications in Bhalotra and Venkataramani (2014). Regression results from panel B are based on all Nigerian DHS data in which children can be linked to their mothers. Specifications and samples are identical 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 this Appendix.

Figure A12: Test of Parallel Trends of Second Generation Sulfa Effects for γ



Note to Figure A12: Graph replicates specification (A8), 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 A13: Bootstrap Estimates of $\hat{\gamma}$



(a) United States (Sulfonamide)

(b) Nigeria (Biafra Civil War)

Notes to Figure A13: 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 3.2). 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 A15.

considering the effect of being exposed to a 1% reduction in rates of pneumonia or the effect of being exposed to a 10% reduction in pneumonia) these scale effects will be perfectly canceled out in the numerator and denominator of equation A7. To the degree that a large or small health shock impacts maternal health and rates of twinning by a similarly large or small amount, the particular mediator used will produce an identical value for γ. This assumption would be violated if different health shock have different *relative* effects on twinning and on child quality, for example a shock which is particularly important for child quality but not for twinning. We return to this point in the caveat below.

(b) The true distribution of γ around its mean can be approximated by a resampling algorithm. Precedence: Conley et al. (2012), p. 265. Evidence: Conley et al. (2012) demonstrate that a simulation-based estimate for the confidence intervals of β can be generated based on resampling of the underlying distribution of interest. In this paper we propose the use of an analytical distribution. This follows if we view our sample of data as the population, and resample from the population, as is typical in bootstrap methods. In both cases (USA and Nigeria) our resampling is based on a representative sample of the full population of mothers, leading to a valid bootstrapped distribution.

Caveat: If the above assumptions are not met, particularly assumption 2 or any of its parts, our estimate of the bounds on β will no longer be correct. However, as Conley et al. (2012) point out:

> "It [this method] will produce valid frequentist inference under the assumption that the prior is correct and will provide robustness relative to the conventional approach (which assumes $\gamma \equiv 0$) even when incorrect."

In the case that the above assumptions are *not* correct, we provide a full set of bounds over a wider range of values in Figures 2 and 3, to determine the robustness of bounds estimates to (even non-conservative) changes in assumptions of γ.

E.5 Bootstrap Confidence Intervals

The methodology to estimate γ in equations (A9) and (A11) is described in previous sub-sections of this Appendix. 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 $\ln \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$.

Estimates of the full distribution of γ are presented in Figures A13a and A13b. 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.[8]

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 distribution 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 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.

[8] 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.

References

D. Aaronson, R. Dehejia, A. Jordan, C. Pop-Eleches, C. Samii, and K. Schulze. The Effect of Fertility on Mothers' Labor Supply over the Last Two Centuries. IZA Discussion Papers 10559, Institute for the Study of Labor (IZA), Feb. 2017.

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.

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

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. 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. Bagger, J. A. Birchenall, H. Mansour, and S. Urza. Education, Birth Order, and Family Size. NBER Working Papers 19111, National Bureau of Economic Research, Inc, June 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. 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.

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

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.

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.

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

S. Jayachandran, A. Lleras-Muney, and K. V. Smith. Modern Medicine and the 20th-Century Decline in Mortality: Evidence on the Impact of Sulfa Drugs. *American Economic Journal: Applied Economics*, 2(2): 118–46, 2010.

T. Kitagawa. A test for instrument validity. *Econometrica*, 83(5):2043–2063, 2015.

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.

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.

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, 2009.