THE TWIN INSTRUMENT: FERTILITY AND HUMAN CAPITAL INVESTMENT

Sonia Bhalotra
University of Essex

Damian Clarke
Universidad de Santiago de Chile

Abstract
Twin births are often used as an instrument to address selection of women into fertility. However, recent work shows selection of women into twin birth such that, while OLS estimates 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 to estimate the trade-off between fertility and children's human capital, we demonstrate the nature and size of the bias in the twin-IV estimator and estimate bounds on the true parameter. (JEL: J12, J13, C13, D13, I12)

1. Introduction

Following Becker (1960), fertility has been modeled jointly with each of investments in children, and women's labour force participation. On account of the average tendency for negative selection into high fertility, linear least squares estimates of the association of fertility with children's human capital, or with women's employment tend to be downward biased (that is, to be "too negative" or to over-estimate the trade-off). Since the pioneering work of Rosenzweig and Wolpin (1980a,b), a considerable literature has attempted to address this by using twins to instrument fertility. The premise is that

The editor in charge of this paper was Imran Rasul.

Acknowledgments: We are grateful to four anonymous referees for very helpful suggestions and to the editor Imran Rasul for providing excellent suggestions and very clear guidance. We are also grateful to Paul Devereux, James Fenske, Judith Hall, Christian Hansen, Martin Karlsson, Toru Kitagawa, Magne Mogstad, Rohini Pande, Adam Rosen, Paul Schulz, Margaret Stevens, Atheendar Venkataramani and Marcos Vera-Hernandez, along with various seminar audiences and discussants for helpful comments. Bhalotra acknowledges partial funding from the ESRC Centre Grant ES/L009153/1 awarded to the Research Centre for Micro-Social Change at ISER, Essex. Clarke acknowledges financial support from the Institute for Research in Market Imperfections and Public Policy, MIPP, ICM IS130002, Ministerio de Economía, Chile. 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 and considerably extended version of Part 2 of the working paper.

E-mail: srbhal@essex.ac.uk (Bhalotra); damian.clarke@usach.cl (Clarke)

Journal of the European Economic Association 2019 00(0):1–50 DOI: 10.1093/jeea/jvz058

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, we presented new population-level evidence that challenges this premise (Bhalotra and Clarke 2019). 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 the outcome of interest (child human capital or women's labour force participation), then twin-instrumented estimates of the relationship between fertility and these outcomes will tend to be upwards biased (i.e., to move towards a null-estimate when there is an underlying trade-off).[1]

This is pertinent as it could resolve the ambiguity of the available evidence on these relationships. Some recent studies using the twin instrument reject the presence of a quantity-quality (QQ) trade-off (Black et al. 2005; Cáceres-Delpiano 2006; Angrist et al. 2010; Åslund and Grönqvist 2010; Fitzsimons and Malde 2014), challenging a long-standing theoretical prior of Becker (1960), Becker and Lewis (1973) and Becker and Tomes (1976).[2] 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 1980b; Bronars and Grogger 1994; Jacobsen et al. 1999; Vere 2011). We argue that, in principle, addressing the omission of maternal health related variables could adjust for the upwards 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 in linear and non-linear models. Since covariate adjustment is necessarily partial, we proceed to estimate bounds on the IV estimates. Given that the first stage (twins predicting fertility) is

1. A recent study proposes a formal test of instrument invalidity (Kitagawa 2015). Albeit with restricted controls, we will show that this test rejects the twin instrument in our data.

2. More recent studies that produce a trade-off once a linearity assumption is relaxed are discussed in what follows and in the following section. The QQ trade-off refers to the inverse relationship between fertility and human capital investment or attainment in a family, posited by Becker and co-authors. Unless we refer to the original "QQ" model of Becker and coauthors (where "child quality" refers to human capital), we will henceforth refer to this as a trade-off between fertility and human capital investment or attainment, rather than use the word "quality".

powerful, we follow Conley et al. (2012) in estimating bounds on the premise that twin births are plausibly if not strictly exogenous. We also estimate bounds under the different assumptions of Nevo and Rosen (2012), using twin births as an "imperfect instrumental variable". Both of these bounding procedures are based on linear IV models, but we document how they compare with the non-parametric "Monotone IV" bounds of Manski and Pepper (2000), which are based on weaker assumptions.

We provide estimates for the United States 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 (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 versus 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. Yet the twin-IV estimates replicate, in our samples, the finding in a number of recent studies that there is no discernible trade-off. However, we find that 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 trade-off.

For instance, in samples with at least three births, an additional child is associated with lower human capital outcomes for the first two births: in linear IV models with the most complete set of controls, this is estimated as −0.05 s.d. for years of education in developing countries (0.16 fewer years of education), and −0.06 s.d. for an index of child health in the United States, and in the sample with at least two births it is −0.10 s.d. for grade progression in the United States (or 0.38 fewer grades progressed). If instead we consider estimates that pool twin-instrumented fertility movements at parities 2, 3 and 4, these estimates are, respectively, −0.04 s.d. (0.12 years of education in developing countries), −0.01 s.d. for the child health index in the United States, and −0.08 s.d. for education in the US sample (0.3 grades progressed).[3,4]

3. We will discuss the relevance of each of the indicators of human capital chosen. In particular, we show evidence using the National Longitudinal Survey of Youth (NLSY) that grade progression in the United States is (i) a function of parental investment measured as reading with the child, and (ii) that it predicts completed educational attainment.

4. Since conception of monozygotic (MZ) twins is thought to be quasi-random even if conception of dizygotic (DZ) twins is not (Farbmacher et al. 2016), we subjected our argument to a harsher test by using only same-sex twins to construct the twin instrument, same-sex twins being more likely to be MZ. We still observe the QQ trade-off diverging from zero and becoming significant when controls for maternal health

The bounds we estimate also confirm the presence of a trade-off at certain parities. Bound end points for the pooled samples based on "plausible exogeneity methods" are −0.04 to −0.05 s.d. for education in developing countries (0.12–0.15 fewer years of education), −0.01 to −0.02 for a health index in the United States, and −0.077 to −0.094 s.d. for education in the United States (0.3–0.36 fewer grades progressed). These values all refer to the lower and upper bounds themselves, rather than to the confidence intervals of the bounds. We place these effect sizes in perspective in Section 4.2.4.[5]

Observe that IV point estimates suggest that the trade-off is not smaller in the United States than in developing countries. This is important given that the recent studies, cited previously, arguing there is no trade-off are set in richer countries, and one may be tempted to conclude that the trade-off may exist 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 *confidence intervals* on both the IV estimates and the partially identified bounds are correspondingly wider. For example, the 95% confidence intervals on the aforementioned bounds estimates span from −0.012 to −0.068 in the developing country sample, and from 0.046 to −0.078 (US health) or 0.012 to −0.18 (US education). We also note here that we will later show that the trade-off is clearer in the United States among families in which the mother does not have a college education.

Although our focus is on our innovation—on invalidity of the twin instrument deriving from healthier women being more likely to give birth to twins—we incorporate in the analysis other recent innovations in the literature that bear upon inference or interpretation in twin-IV studies. We investigate whether our argument is robust to these other innovations. We check whether introducing controls changes the complier group and LATE identified in the linear IV model. We look at two potential issues here. First, following Angrist et al. (2010) and Angrist and Imbens (1995) who describe the weighting functions over parities induced by instrumental variation, we investigate how the weighting function changes conditional on controls for maternal health and education—and we find it does not significantly change. Second, we consider what heterogeneity in the estimates (as in Brinch et al. 2017 for instance) implies for the coefficient movements in 2SLS estimates that arise from conditioning on additional covariates. We explain that the weighting on different households will depend positively on the rate of twinning (provided that the proportion of twins is less than 0.5). Given that we provide evidence that positively selected groups (women with higher education or better health) (a) have weaker (less negative) QQ trade-offs, and (b) higher rates of twinning, the inclusion of additional covariates gives more weight to groups with smaller QQ trade-offs, and so our argument likely holds a fortiori.

are included. This is what we would expect since our argument pertains to the role of maternal health not in determining conception but instead in ensuring survival of twin conceptions to birth.

5. We investigate if using the sex-mix instrument of Angrist et al. (2010) together with the twin instrument, as in Chesher and Rosen (2013, 2018), tightens the linear bounds but find that, as it is quite weak, it does not.

We follow Brinch et al. (2017) and Mogstad and Wiswall (2016) in estimating non-linear models that allow the trade-off to vary with the parity at which twins are born. These authors made the point that once the common linearity assumption is relaxed, a trade-off emerges, at least at some parities and for some households. We confirm their results on new data samples. We extend their specification to include controls for maternal characteristics and interactions between parity and every maternal characteristic. Controlling for characteristics additively increases the trade-off (makes the coefficient more negative), and further allowing these controls to have different effects at different parities makes point estimates even more negative. Thus our argument continues to hold in this much richer model.

In an extension of the main analysis, we investigate the twin-IV critique of Rosenzweig and Zhang (2009), which draws attention to the fact that twins have lower birth endowments than their siblings, which may lead to reinforcing or compensating parental investments.[6] We assess parental investment responses to twins on our data samples (incorporating the NLSY for the United States) and consider how any nonzero responses influence interpretation of our results. Since we find that some parental responses vary systematically with the mother's education (reading with the child does, breastfeeding responses are similar across education groups), we repeat the main analysis to investigate bias adjustment for mothers with versus without college-education. We conclude that we can more readily sign the bias in cases where parents do not compensate and, for the indicators and data samples we analyse, this is for less educated mothers.[7]

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, whereas many countries including Italy and Canada have incentivized it, often with non-linear rules.[8] Moreover, advocates of policies encouraging smaller families rest their case on larger families investing less in each child, limiting human capital accumulation and living standards (Galor and Weil 2000; Moav 2005). Although we do

6. The literature often uses birth weight as an indicator of the birth endowment. We show that the impact of close birth spacing between twins on always-taker families and on complier families will interact with our IV estimates in the same way as the lower birth weight of twins.

7. We find that parents in developing countries and the United States reinforce endowments in their breastfeeding behaviour. In the United States, for which data on this additional investment are available, we find that college-educated women (only) compensate the endowment when reading with the child.

8. 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.

not directly analyse women's labour supply as an outcome, the analysis here suggests that it is likely to have a stronger relationship with fertility than existing twin-IV estimates suggest.

The rest of this paper unfolds as follows. In Section 2, we formalize the bias in the OLS and twin-IV estimators in the linear and non-linear models, and describe the partial identification methodologies for estimating the fertility–human capital trade-off using twins. In Section 3 we describe the microdata, the measures of human capital analysed, the construction of parity-specific and pooled samples, and the empirical strategy. Section 4 presents all results, including estimates of the linear, non-linear and partial identification methods proposed, and extensions including the role of parental behaviour, and heterogeneity by mother's education. Section 5 concludes.

2. Methodology—Biases and Bounds

2.1. Estimating the Quantity–Quality Trade-Off with Twins

A long-standing theoretical result in the literature on human capital formation is the existence of a 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 (Blake 1989; Hanushek 1992; Galor 2012). This empirical regularity is also evident in the data sets we analyse from the United States and developing countries (see Online Appendix Figures A.1 and A.2).

However, as discussed in the previous section, empirical evidence of a fertility–human capital trade-off is ambiguous. Early work including Hanushek (1992) and Rosenzweig and Wolpin (1980b) documented significant negative effects of additional births within a family on average child educational outcomes. Since then, research has found estimates of no significant relationship (Black et al. 2005; Cáceres-Delpiano 2006; Angrist et al. 2010; Åslund and Grönqvist 2010; Fitzsimons and Malde 2014), a significantly positive relationship (Qian 2009) and a significantly negative relationship (Grawe 2008; Lee 2008; Ponczek and Souza 2012; Bougma et al. 2015; Mogstad and Wiswall 2016; Brinch et al. 2017), see the review in Clarke (2018).[9] The two studies of Brinch et al. (2017) and Mogstad and Wiswall (2016) show 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. A trade-off has emerged also in studies that, rather than

9. Black et al. (2005) show that OLS estimates of the trade-off in their Norwegian sample are sensitive to controlling for birth order. We find this is not the case in our data samples.

use instrumental variables, use quasi-experimental variation in either the quantity or "quality" of children.[10]

In this section we formalize the biases that concern us. We start with the standard linear model and we then develop the non-linear case.[11]

2.1.1. Linear Fertility Models. **OLS**. Analyses of the fertility–human capital trade-off aim to arrive at a causal estimate of β_1 in the following model:

$$y_{ij} = \beta_0 + \beta_1 fertility_j + \mathbf{X}\boldsymbol{\beta_x} + \varepsilon_{ij}. \quad (1)$$

Here, y_{ij} is a measure of human capital attainment of or investment in child i in family j, and $fertility_j$ is the number of children in family j. The vector $\mathbf{X}$ includes relevant (exogenous) family and child level covariates. As discussed in Section 3, we consider two measures of human capital: a standardized measure of education, and a measure of health. A fertility–human capital trade-off implies that $\beta_1 < 0$. As has been extensively discussed in a previous literature, estimation of β_1 using OLS will result in biased coefficients given that child human capital and fertility are jointly determined (Becker and Lewis 1973; Becker and Tomes 1976), and relevant parental behaviours and attributes that influence both fertility decisions and investments in children's human capital are unobserved, and relegated to the error term (Qian 2009). The direction of the OLS bias is determined by the sign on the conditional correlation between $fertility_j$ and the relevant omitted factors in ε_{ij}. If mothers with weaker unobserved preferences for investing in children (or other unobserved constraints to investing in child human capital) have more children, OLS estimates will overstate the magnitude of the true trade-off.

IV. Following the seminal work of Rosenzweig and Wolpin (1980b), fertility has been instrumented with the incidence of twin births on the premise that they constitute an exogenous shock to family size. The first stage linear projection can be written as

$$fertility_j = \pi_0 + \pi_1 twin_j + \mathbf{X}\boldsymbol{\pi_x} + \nu_{ij}, \quad (2)$$

10. For instance, Bhalotra et al. (2018) show that the sharp improvement in newborn health (quality) following the introduction of antibiotics led to lower fertility and, Bhalotra and Venkataramani (2015) verify that the children born in that era had higher education and income. Similarly, Ager et al. (2018) find that vaccinations that reduced child mortality led to lower fertility. Bailey et al. (2019) show that access to family planning led to higher quality births. Anukriti et al. (2016) show that the introduction of technology facilitating sex-selective abortion that (selectively) reduced the quantity of girls born led to higher investments in surviving girls.

11. The twin instrument has also been used to estimate effects of childbearing on women's labour force participation with varying results (Rosenzweig and Wolpin 1980a; Angrist and Evans 1998; Jacobsen et al. 1999), and to estimate the consequences of out of wedlock births on marriage market outcomes, poverty and welfare receipt (Bronars and Grogger 1994). The discussion here applies to these cases as well. In this paper we focus nearly exclusively on the internal validity of twins estimates (IV consistency). In recent work, Aaronson et al. (2017) and Bisbee et al. (2017) examine the external validity of IV estimates of the relationship between fertility and female labour supply using either twins or the sex-mix of the first two births as instruments. Still, as observed in Bhalotra and Clarke (2019), our estimates suggest considerable heterogeneity by country income levels.

where $twin_j$ is an indicator for whether the nth birth in family j is a twin birth. $E[(1, twin_j, \mathbf{X})' v_{ij}] = 0$ by construction. As described further in Section 3, 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 (the control group) or two siblings (the treatment group) at the nth birth. Comparing these allows us to estimate causal impacts of the additional birth on the child outcome. The twins themselves are excluded from the estimation sample.[12] If twins are a valid instrument, the parameter β_1 can be consistently estimated by IV where equation (1) is the second stage. In particular, for validity we require $\text{Cov}(twin_j, \varepsilon_{ij}) = 0$, or that the occurrence of twins is uncorrelated with unobserved factors conditional on observables $\mathbf{X}$. We turn to this in what follows, and discuss additional threats to the exclusion restriction related to parental behaviour in Section 4.2.5.

Bhalotra and Clarke (2019) provide evidence that omitted variables indicating maternal health and relegated to ε_{ij} are correlated with twinning, and in Section 4.1 we document this for the data used in this paper. If healthier mothers invest more in their pregnancies (e.g., by averting smoking before birth) and 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. Provided that $\pi_1 > 0$, implying that twinning increases fertility, (something consistently observed in data), positive selection of mothers of twins implies:

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

where $\hat{\beta}_1^{IV}$ signifies that β_1 was estimated using the twin instrument in equation (2).

We can partition the stochastic error term from equation (1) into a vector of observable measures of mother's health capital (H), socioeconomic variables ($\mathbf{S}$), and all other unobserved components, as $\varepsilon_{ij} = \mathbf{H} + \mathbf{S} + \varepsilon_{ij}^*$. Assuming a positive (or zero) covariance between the three components of the error term,[13] the step-by-step removal of twin selection predictors will result in the estimated coefficient becoming continually closer to the true parameter. Thus, provided any changes in first stage estimates across

12. This takes care of the concern that since twins tend to be born with weaker endowments (e.g., birth weight), they will tend to have systematically different quality outcomes. Using data from the United States, Almond et al. (2005) document that twins have substantially lower birth weight, lower APGAR scores, higher use of assisted ventilation at birth and lower gestation period than singletons. We document similar endowment differences in our data samples (Online Appendix Figures A.3 and A.4).

13. Given that the covariance between elements of $\mathbf{S}$ and $\mathbf{H}$ is found to be positive, and given that the covariance between each of these and other unobserved variables that positively affect child human capital 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.

specifications are of second order importance[14]:

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

The coefficients $\hat{\beta}_1^{IV,H}$ and $\hat{\beta}_1^{IV,S+H}$ refer to coefficients in a model augmented to control for observable health capital **H**, and then also observable socioeconomic status **S**. Since, as discussed further in Section 4.1, all determinants of twin birth are virtually impossible to account for, the twin-IV will over-estimate β_1, under-estimating the magnitude of the fertility–human capital trade-off if $\beta_1 < 0$, although addition of predictors of twins as controls will lead to the IV estimate approaching the true value from above.

Although inequality 3 suggests that the inclusion of additional covariates will cause the IV estimate to approach the parameter of interest, the addition of controls in a 2SLS model implies that the composition of the LATE estimate will change. In particular, the LATE estimate is based on a weighted average of covariate specific LATEs and variation in fertility induced by the instrument at different parities (Angrist and Imbens 1995). So, under heterogeneity, the inclusion of additional covariates may change the IV estimate, independent of any selection into twinning. In Section 4.2.2 we consider this explicitly, documenting that coefficient movements observed owe to twin selection in this case.

Pooled IV Models. In the following section we motivate and describe the use of a series of samples, varying in the parity at which the potential twin birth occurs. As IV estimates are considerably less precise than OLS estimates, we follow Angrist et al. (2010) and Mogstad and Wiswall (2012) in presenting parity-pooled estimates in certain cases to gain power. Consider pooling the 2+, 3+ and 4+ samples, which are first born children in families with at least two births, first and second born children in families with at least 3 births, and first to third born children in families with at least four births. In this case, fertility is instrumented with the variables $twin^*_{2j}$, $twin^*_{3j}$ and $twin^*_{4j}$, which are defined as

$$twin^*_{cj} = \begin{cases} 0, & \text{if } fertility_j < c \\ twin_{cj} - \hat{E}[twin_{cj}|X_j, fertility_j \geq c], & \text{if } fertility_j \geq c \end{cases} \tag{4}$$

for each $c \in 2, 3, 4$. Replacing $twin$ from equation (2) with $twin^*$ resolves the problem that in the pooled sample, the twin instrument will be missing for certain parity groups. For example, for observations in the 2+ group who are in a sibling group of two, the twin at third birth and twin at fourth birth variables will not be

14. These inequalities come about given that conditional correlations between $twin_j$ and the stochastic error term are reduced by the inclusion of relevant omitted variables. However, the inclusion of additional covariates will also impact the first stage estimates. Thus, for these inequalities in limits to hold, we require changes in first stage estimates to be second order. It is useful to note that later in the paper, we typically observe first stage estimates to be weakly *increasing* with the addition of controls, which only serves to reinforce these inequalities.

defined, given that these births have not occurred. $twin_j^*$ is a valid instrument under the same twin exogeneity assumptions as in the linear IV models (Mogstad and Wiswall 2012). For each observation, $twin_{cj}^*$ is formed by assigning a value of 0 if the family has not reached a particular fertility threshold, and for those who have, generated as $twin_{cj} - \hat{E}[twin_{cj}|X_j, c_j \geq c]$, where $\hat{E}[twin_{cj}|X_j, c_j \geq c]$ is a non-parametric estimate of the conditional mean of the probability of twin birth in the non-missing subsample. The validity of this procedure is demonstrated in Angrist et al. (2010) and Mogstad and Wiswall (2012, 2016).

2.1.2. 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 investment. Recent theoretical (Aizer and Cunha 2012) and empirical (Rosenzweig and Zhang 2009; Bagger et al. 2013; Mogstad and Wiswall 2016; Brinch et al. 2017) papers relax this assumption. Among other things, this allows for reinforcing or compensating behaviours in parental investment choices (Almond and Mazumder 2013), something we explicitly consider later. This implies allowing the coefficient β_1 to vary across children in the family.

In this paper we focus initially on linear IV models given that it allows us to document our bias argument in a setting used in the majority of papers estimating the fertility–human-capital trade-off. However, when using data for which we have sufficient power to split instruments, we estimate non-linear marginal fertility models as in Brinch et al. (2017) and Mogstad and Wiswall (2016). Following Mogstad and Wiswall (2016), this consists of the following 2SLS procedure (illustrated for the two-plus sample):

$$\begin{aligned} fertility_{sj} &= \lambda_{s2} twin_{2j} + \lambda_{s3} twin_{3j}^* + \lambda_{s4} twin_{4j}^* + \lambda_{s5} twin_{5j}^* \\ &\quad + \mathbf{X}\lambda_{Xs} + \nu_{sj}, \quad \text{for } s = 2, \ldots, 5 \end{aligned} \tag{5}$$

$$\begin{aligned} y_{ij} &= \alpha_0 + \alpha_1 fertility_{2j} + \alpha_2 fertility_{3j} + \alpha_3 fertility_{4j} + \alpha_4 fertility_{5j} \\ &\quad + \mathbf{X}\alpha_X + \eta_{ij}, \end{aligned} \tag{6}$$

where $fertility_{sj}$ is an indicator variable taking the value of 1 if $fertility_j \geq s$, (5) are a series of first stages for each fertility indicator, and (6) 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 the $twin_{cj}^*$ instruments defined in equation (4). We also follow Mogstad and Wiswall (2016) in considering family sizes up to 6 children. Although the specifications (5) and (6) are for the two-plus sample, we 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 in this paper is in examining the impact of positively selected twin births, we estimate the previous 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 equations (5) and (6).

2.2. 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 fertility–human capital trade-off is not possible, see Bhalotra and Clarke (2019) for a complete discussion. However, under additional assumptions relating to the failure of the IV exclusion restriction, or correlations between the instrument, the endogenous variable, and unobservables we can bound the fertility–human capital trade-off. As there appears to be no alternative instrument for fertility that has not been critiqued, we investigate different procedures for bounding the trade-off using the twin instrument. Two of these are recent bounding procedures for linear IV models. We also briefly consider non-parametric monotone IV bounds that are comparable to non-linear models but that tend to deliver wider bounds.

First we describe our use of the Nevo and Rosen (2012) "Imperfect IV" procedure. This is ideally suited to the current context because it suggests that if twins are positively selected, 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 previously.[15] If we are willing to additionally assume that the twin instrument is "less endogenous" than fertility (Nevo and Rosen's assumption 4), we can tighten the bounds by forming a compound instrument based on the endogenous fertility variable, and the imperfect twin instrument. This instrument, $(V = \sigma_{fertility} Twin_j - \sigma_{Twin} fertility_j)$, where σ refers to the standard deviation, can provide tighter bounds on the β_1 parameter where $\hat{\beta}_{IV}^{V} \leq \beta_1 \leq \hat{\beta}_{IV}^{twin}$, suggesting end points for a series of IV bounds on the parameter β_1.

The parameters $\hat{\beta}_{IV}^{V}$ and $\hat{\beta}_{IV}^{twin}$ give end points of the bounds. To conduct inference we follow the Adaptive Inequality Selection (AIS) procedure described in Chernozhukov et al. (2013). This accounts for the fact that there are potentially multiple moment conditions giving upper and lower bounds, and uncertainty related to each moment restriction must be accounted for when generating confidence intervals. Using the algorithm described in Chernozhukov et al. (2015, p. 27), we consistently

15. 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, this assumption 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 *fertility* in equation (1), z refers to *twin* in equation (2), 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 $\hat{\beta}_{IV}^{twin} \leq \beta_1 \leq \hat{\beta}_{OLS}$.

report both the end points on the Nevo and Rosen bounds, as well as their 95% confidence intervals. We have programmed this estimator for Stata (Clarke and Matta 2018).

The Nevo and Rosen (2012) procedure is straightforward and relies on fairly weak assumptions. In particular, the only assumptions we require are that (i) there is negative selection into fertility. This a common 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); (ii) twins are positively selected, which is shown using a variety of data sources and measures of maternal health in Bhalotra and Clarke (2019), (iii) twin births are positively associated with fertility, which we show in the first stage regressions in what follows, and (iv) there is less selection into twin birth than into fertility, which seems reasonable.

The upper bound in the case of Nevo and Rosen is given by the original twin IV estimate $\hat{\beta}_{IV}^{twin}$. 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 fertility–human capital model for this paper):

$$y_{ij} = \delta_0 + \delta_1 fertility_j + \gamma twin_j + \mathbf{X}\boldsymbol{\delta}_{\boldsymbol{x}} + \vartheta_{ij}. \tag{7}$$

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. 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 $\gamma \in [0, \infty]$". They state that their procedure is hence an extension of the Manski and Pepper procedure.

The approach in Conley et al. is 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 in what follows). In Section 4.1, we provide evidence that leads us to suspect that γ will not equal zero. Specifically, γ is the degree to which twin mothers are healthier than non-twin mothers multiplied by the effect of (unobserved) maternal health on child quality. If one or other of these conditional correlations is equal to zero, IV estimates will not be inconsistent.[16]

16. Section 4.1 only shows that twin mothers are healthier than mothers of singletons. To complement this, we also show in what follows a series of positive associations of maternal health with both investments in children and child quality outcomes.

Conley et al. (2012) show that bounds for the IV parameter β_1 from equation (1) 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 to capture uncertainty over γ, and the second involves assuming a distribution to capture the uncertainty in γ. When we present bounds based on the LTZ approach, we document bounds under a range of assumed distributions. Once again, in the case of Conley et al. (2012) we present both bound end points (in the case of the UCI approach) or midpoints (in the case of the LTZ approach), as well as the 95% confidence intervals on these partially identified bounds. In each case, we follow the inference procedures documented to have asymptotic coverage of at least 95% in Conley et al. (2012), which are implemented following Clarke and Matta (2018).

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, whereas in Nevo and Rosen (2012) we only need to provide the sign.[17] The advantage of the Conley et al. procedure that makes it worthwhile despite its stronger assumptions is that it potentially returns tighter bounds at both the upper and lower end, whereas Nevo and Rosen retains the original IV upper bound and only tightens the lower bound using information from the original OLS estimates.

Both of these bounds procedures are based upon the linear IV model that has been routinely used in the twin QQ literature. However, as described in Section 2.1.2, we will relax the linear assumption and investigate the potential impact of positive twin selection in non-linear models such as in Mogstad and Wiswall (2016). Manski and Pepper (2000) describe a non-parametric procedure allowing us to estimate bounds on *average* treatment effects of a family moving from any fertility level s to any other fertility level t, such that $s > t$. This bounds procedure is based on a Monotone IV assumption, which is suitable in our case, requiring that investments in children are weakly higher among women with twin births than those with singleton births, given positive selection of women into twin births. However, these bound are typically estimated invoking additional assumptions—either "Monotone Treatment Selection" (MTS) or "Monotone Treatment Response" (MTR). In our case these additional assumptions are not justified, given that they assume a weakly increasing relationship between child outcomes and fertility reductions, while our goal is precisely to estimate, rather than assume, this relationship. Invoked alone, Monotone IV bounds are typically wide, as they are based on a similar logic to the "worst case" (or "no assumptions") bounds in Manski (1989). That these non-parametric bounds are uninformative in this context has been documented in other settings, for example, Brinch et al. (2017). We

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

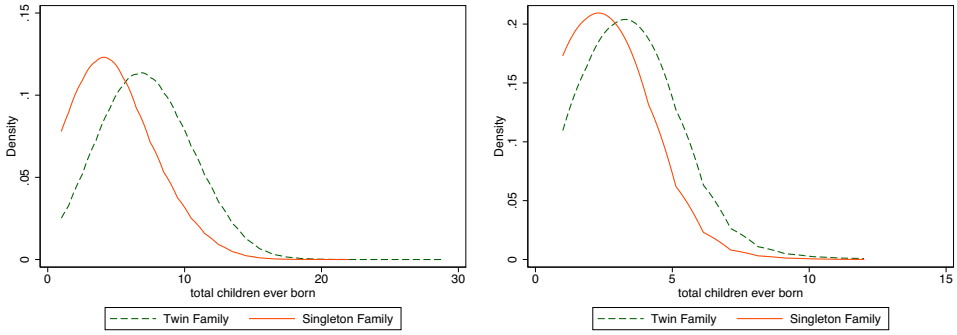
nevertheless present these bounds along with their full derivation in Online Appendix D, in order to assess the relative benefits of the updated bounding procedures described in this section. In this Online Appendix we also present joint MTS–MTR bounds for completeness, though note that these bounds impose that a fertility–human capital trade-off exists, and so are only informative for the lower-bound, rather than the existence of the trade-off per se.

The bounds from Manski and Pepper (2000), Conley et al. (2012) and Nevo and Rosen (2012) all relax the IV assumption of a strict exclusion restriction. The lack of point identification comes with the benefit of allowing that there is positive selection of twin birth. An alternative set of IV bounds has been described in Chesher and Rosen (2013, 2018) for discrete IV models, where set, rather than point, identification owes to a complete non-parametric specification of the underlying model, additionally leading to the potential to estimate an ATE rather than a LATE. However, in these cases, typically some form of independence is assumed for the IV (including weaker-than-standard independence assumptions, such as conditional mean independence or conditional quantile independence). Given that our interest in bounds in this setting is principally in accounting for failure of IV independence assumptions, we do not estimate these bounds. An example of generalized IV bounds based on the twin instrument and an independence assumption can be found in Chesher and Rosen (2018).

3. Data and Estimation

We shall consistently estimate OLS and twin-IV estimates employing microdata from the United States and a sample of 68 developing countries. In order to estimate specification 1, augmented with health and SES controls, we require information on sibling-linked births, measures of child human capital, 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 for the years 2004–2014, and the DHS for 68 countries, which have been applied over 20 years using a broadly similar design. To examine parental investments, we use the NLSY.

The NHIS is an ongoing household survey released yearly in the United States from 1957. The survey has been updated every 10 or 15 years, and as such, we focus on a particular time period (2004–2014) which ensures we can generate all variables in a consistent way. It is representative of the population of (non-institutionalized) US residents and has a household file, a family file, a sample adult and a sample child file. The household and family files give basic demographic information on each household member and other basic demographic outcomes, and the sample adult and child files provide more in depth health information on one particular adult and child in the family. Our sample is generated using all mothers who are included in the sample adult file, and so for whom health measures are recorded, and this is merged with demographic information and information on all child outcomes from the family and household files. We do not use the sample child files given that we require consistent measures of outcomes for all children in the family.



(a) Developing Countries (b) United States

FIGURE 1. Twins shift the fertility distribution outwards. 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.

The DHS implements standardized questionnaires across countries (see Online Appendix Table A.1 for the full list of countries in our microdata) on nationally representative samples of reproductive-age women (15–49 years). It includes a household questionnaire recording basic information on each household member including children (such as their sex, education and relationship to the household head), and a women's questionnaire to eligible women in the household. The latter is more extensive, recording, among other things, each woman's full fertility history and health and socioeconomic characteristics including her height, BMI, and education. Our data consist of a merged file of all reproductive age women from the women's questionnaire with information on their children's educational attainment from the household questionnaire.

Online Appendix Table A.2 provides summary statistics for the DHS and NHIS data. Fertility and maternal characteristics are described at the level of the mother, whereas child education and health outcomes are described at the level of the child. Twin births make up 1.98% of all births in the DHS sample, and 2.57% in the NHIS sample. 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 in what follows.

Auxiliary Tests on Parental Investment Responses to Endowments. For the extension we will discuss that investigates parental behaviour we require data on parental investments in different children at similar points of their children's lives. For this, we use the National Longitudinal Survey of Youth (NLSY79) Children and Young Adults survey, which registers information on all children born to the original NLSY79 female respondents, and gathers a much wider range of information on parent-child

interactions. We generate a panel dataset of children based on each biennial survey through 1988–2012. We will use two measures of investment—whether the mother reads frequently with the child between ages 6–9 years, and whether the mother breastfed the child, as both of these are consistently recorded for each child at comparable stages of the child's life. For the developing country sample, we use the duration of breastfeeding in months, available for all children for whom breastfeeding is completed. These data are analysed in Section 4.2.5.

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 human capital (proxies for parental investment) are the first-born children. Following Black et al. (2005), we similarly construct a 3+ sample that 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 Online Appendix Figure A.5), increasing family size raises the chances of having a twin birth. In the DHS sample, 42% of all children are in one 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.[18]

Measures of Human Capital, and Relevance. A measure of child human capital available in both data sets is educational attainment. Our final estimation sample consists of children aged 6–18 when surveyed, selected to represent children who have begun their education but still reside in the same household as their mother. Ideally we would observe completed education but, to our knowledge, no large datasets are

18. The pooled sample will provide a weighted average of the parity-specific estimates 2+, 3+ and 4+. It will not include the trade-off induced by twins at 5th or 6th or higher order births because we do not instrument these, similar to other papers based on the twin IV. Note however that, by design, many of the subjects of a 5+ family (the children born before birth 5 in a family with at least 5 births) will be in the samples with twins at birth orders up to order 4, for example, recall that the 4+ sample contains children born before birth 4 in a family with at least 4 births, so this will include the first 3 children of 5+ families. The difference is that we are not adding higher order instruments (e.g., twins at birth 5).

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 fertility–human capital 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.

Since children age 6–18 are 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, whereas in the NHIS, it consists of children with the same month and year of birth. Thus coefficients are expressed in standard deviations. The NHIS also provides a subjectively assessed binary indicator of child health (excellent or not), which we model as an additional indicator of child human capital.[19]

We now consider the relevance of the measures of child human capital that we use in the poor and richer country settings. While schooling rates have been increasing globally, of 163 developing countries, only 47 have achieved universal primary education (UNESCO 2005) and in two-thirds of sub-Saharan African countries, more than 30% of students who start primary school are expected to drop out. Credit constraints have been identified as a factor (UNESCO 2011), and these tend to tighten as the number of dependent children increases.

In the United States, although all children stay in school until the legislated minimum school leaving age, grade retention is (i) common and (ii) a significant marker of educational progress. It is estimated that over 2.4 million (5%–10%) students are retained every year in the United States. Rising through the 25 years up to 2003, this was estimated to cost over 13 billion dollars per year just to pay for the extra year of schooling (i.e., ignoring its long run costs) (Anderson et al. 2002). Retention rates are higher among boys, ethnic minorities and children of less educated parents (Warren et al. 2014). Passage in 2002 of the No Child Left Behind Act in the United States, with its emphasis on mastery of minimum grade-level competencies as a condition for promotion, has renewed discussion of grade retention in public policy making. Systematic reviews examining research over almost a century conclude that grade retention is one of the strongest predictors of high school dropout, and is associated with lower earnings in adulthood (Jimerson 1999, 2001; Jimerson et al. 2002; Lavy et al. 2012; Manacorda 2012).

Using the National Longitudinal Survey of Youth for 1986–2014, and following students from ages 7 up to their education completed at the age of 25 or older, we investigated ourselves the associations of grade retention with (i) final attainment and (ii) parental investment in children. We find that, at every age, children who are behind

19. While we would also like to analyse 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 human capital for children born *prior to* twins who will have already been fully exposed to infant mortality risk by the time the twins were born.

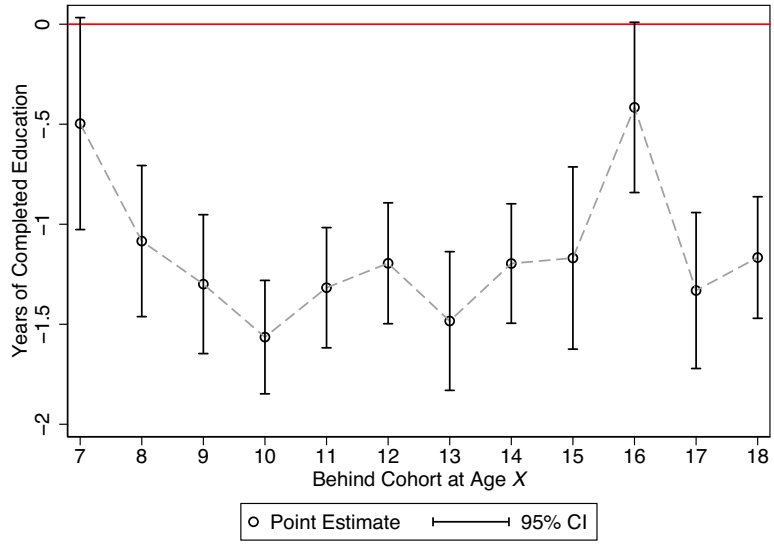


FIGURE 2. School completion rates and lifetime educational accumulation. Each point estimate and 95% confidence interval displays the coefficient from a separate regression of an individual's eventual completed education on whether the individual was behind his or her cohort at the age indicated on the horizontal axis. All data from the NLSY79 child and young adult panel are used, covering individuals who were born to NLSY79 women, and who were children or young adults at some point between 1986 and 2014. In each survey wave, the individual's current age and education is reported, and at the end of the panel survey we observe their eventual completed education. Regressions are only estimated on observations who are at least 25 years old in the final wave, to ensure that education is approximately completed. Similar results are observed if we condition on being 30 years old in the final wave. In each case, an individual is behind their cohort if their grade accumulation is at least 1 year less than their age minus 6. Each coefficient and point estimate comes from a separate regression, given that NLSY survey waves occur every 2 years, and so we do not observe the same sample of respondents at each age.

their cohort in school years end up with significantly lower completed schooling, for example, final attainment is 1.5 years lower for children who are behind their birth cohort at age 10 and it is 0.5 years lower for children who are behind at ages 7 or 16 (Figure 2). This alleviates concerns that our school for age measure represents only a temporal effect, for example, capturing red-shirting or other strategic parental behaviour. In the Online Appendix (Online Appendix Figure A.6), using the same data sample, we provide complementary evidence showing that children of parents who read frequently with their children during the ages of 6–9 years exhibit a lower likelihood of being behind their cohort in the future, at ages 10–18.

The third measure of child quality we use is a subjective measure of child health. Beyond its intrinsic value, the long term health and socio-economic payoff to improved child health is estimated to be large (Almond and Currie 2011). In particular, Case et al. (2002) and references therein demonstrate that a similar self-reported measure of health predicts mortality and morbidity in the US population. Full variable definitions are provided in Online Appendix B.

Birth Order. The set-up we have just described will yield estimates of the human capital outcomes for the first born child in the 2+ sample, and estimates that average over the human capital of the first and second born child in the 3+ sample, and so on. Like Angrist et al. (2010), we could also show results separating the first and second born child in the 3+ sample, which we do not do just to conserve power and because identifying birth order specific effects of family size is not closely related to our purpose. However, in every specification in the paper, OLS and IV, we control for birth order fixed effects on the right hand side. These allow for direct impacts of birth order of the pre-twins on their education. Black et al. (2005) find that the trade-off (the coefficient on fertility) is eliminated in OLS when controlling for birth order effects, but this is not the case in our data samples.[20] A different consideration of birth order pertains to the order at which twin births impact completed fertility. This involves estimating non-linear IV models that allow the trade-off to vary with birth order (parity) of the twin. We shall do this, following Brinch et al. (2017), for example. We further allow for interactions between parity and controls for the health and education of the mother.

4. Results

4.1. Twin Births and Maternal Condition

In Bhalotra and Clarke (2019) 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 (positively) selected 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 (2019) 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. The controls were described previously and are listed in notes to the tables. 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 (2019) we use the universe of births in the United States, between 2010 and

20. In Online Appendix Tables A.4 and A.5 (OLS) and A.6–A.8 (IV) we show the coefficients on the birth order fixed effects in the OLS and IV models. Birth order has significant direct effects on education but it does not eliminate (or attenuate) the estimated trade-off.

TABLE 1. Probability of giving birth to twins USA (NHIS).

	All	Time	
		1982–1990	1991–2013
Mother's education (years)	0.010**	0.020	0.009*
	(0.005)	(0.019)	(0.005)
Mother's height (in.)	0.002	0.007	0.001
	(0.004)	(0.013)	(0.004)
Mother's BMI	0.001**	0.004**	0.001**
	(0.001)	(0.002)	(0.001)
Smoked prior to birth	−0.042*	−0.217***	−0.027
	(0.021)	(0.084)	(0.022)
Observations	103,249	6,160	96,374
Pseudo *R*-squared	0.015	0.075	0.015

Notes: This table presents probit 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. Average marginal effects are reported. Height is measured in inches and BMI is weight in kg divided by height in meters squared. Standard errors are included in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

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.[21] In Bhalotra and Clarke (2019), we show with NLSY data that time-varying indicators of maternal health influence the chances of twin birth conditional on woman fixed effects that purge any genetic influences on twinning (and also purge the effects of mother's education). In that earlier paper we also report estimates using the much larger vital statistics database for the United States that includes several additional measures of maternal health. Although that evidence is more compelling, here our purpose is primarily to show that the results hold on the data that are analysed in this paper.

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). In Bhalotra and Clarke (2019), we show that the health indicators are significant conditional and unconditional on education. These results hold even in the 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)).[22]

21. 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.

22. 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.

TABLE 2. Probability of giving birth to twins (developing countries by income and time period).

	All	Income		Time		Prenatal
		Low inc	Middle inc	1990–2013	1972–1989	
Mother's age	0.098***	0.100***	0.096***	0.105***	0.071***	0.099***
	(0.004)	(0.005)	(0.007)	(0.005)	(0.009)	(0.004)
Mother's age squared	−0.001***	−0.001***	−0.001***	−0.001***	−0.001***	−0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age at first birth	−0.005***	−0.009***	−0.000	−0.005***	−0.006***	−0.005***
	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
Mother's education (years)	0.002***	0.002**	0.001	0.002**	0.002	0.002**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
Mother's height (cm)	0.008***	0.008***	0.008***	0.008***	0.007***	0.008***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Mother's BMI	0.006***	0.008***	0.005***	0.006***	0.008***	0.006***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Prenatal care (doctor)						0.046**
						(0.022)
Prenatal care (nurse)						0.042*
						(0.022)
Prenatal care (none)						−0.001
						(0.028)
Observations	2,210,676	1,379,640	830,716	1,536,262	674,414	2,206,009
Pseudo *R*-squared	0.034	0.037	0.030	0.034	0.031	0.034

Notes: This table presents results for the developing country sample splitting by pre- and post-1990, and by country income level. 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 using a probit model. Average marginal effects are reported. Height is measured in cm and BMI is weight in kg divided by height in meters 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. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

We also investigated whether the source of positive selection of twins 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 standardized 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

TABLE 3. Maternal health and child investments/outcomes (NHIS).

	No health insurance	Health limits	Education *z*-score	Excellent health
Mother's education (years)	−0.076***	−0.009***	0.019***	0.052***
	(0.002)	(0.003)	(0.002)	(0.002)
Mother's height (in.)	−0.007**	−0.012***	0.005***	0.020***
	(0.003)	(0.003)	(0.002)	(0.002)
Mother's BMI	0.000	0.002***	0.000	−0.004***
	(0.000)	(0.000)	(0.000)	(0.000)
Smoked prior to birth	0.101***	0.183***	−0.046***	−0.153***
	(0.016)	(0.015)	(0.009)	(0.011)
Observations	103,502	103,502	74,777	103,502
R-squared/pseudo *R*-squared	0.072	0.021	0.019	0.025

Notes: Regressions are presented of child investments or child outcomes on a number of maternal characteristics. Dependent variables are indicated in column headings. 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 insurance, health limits and excellent health are binary variables, and models are estimated as probit 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 meters squared. Standard errors are included in parentheses. $^{**}p < 0.05$; $^{***}p < 0.01$.

having health limitations or not having insurance (columns (1) and (2)), and correlated with positive measures of human capital outcomes (education and self-informed health status; columns (3) and (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 on the child's education.[23]

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. Note that previous twin-IV studies cited earlier often control for parental characteristics including age and education, but not for maternal health (see Online Appendix Table A.3, which reviews previous twin-IV studies). In Bhalotra and Clarke (2019) we show that numerous different measures of maternal health have large and statistically significant impacts on the probability of twin birth (in many different samples) even after conditioning upon maternal education and age.

23. 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 19.

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	0.084***	0.024***	0.004***	0.085***	0.030***	0.003*
	(0.002)	(0.005)	(0.001)	(0.002)	(0.004)	(0.001)
Mother's age squared	−0.001***	−0.001***	−0.000***	−0.001***	−0.001***	−0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age at first birth	−0.042***	0.070***	0.009***	−0.039***	0.061***	0.008***
	(0.001)	(0.002)	(0.000)	(0.001)	(0.001)	(0.000)
Mother's education (years)	−0.113***	0.267***	0.079***	−0.100***	0.220***	0.075***
	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)
Mother's height (cm)	−0.005***	0.020***	0.005***	−0.004***	0.016***	0.005***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
Mother's BMI	−0.046***	0.076***	0.023***	−0.041***	0.060***	0.021***
	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)
Prenatal care (doctor)				−0.766***	1.217***	0.066***
				(0.016)	(0.031)	(0.008)
Prenatal care (nurse)				−0.153***	0.092***	0.060***
				(0.016)	(0.030)	(0.008)
Prenatal care (none)				1.027***	−3.715***	−0.438***
				(0.020)	(0.035)	(0.010)
Observations	750,213	616,448	1,289,528	750,211	616,446	1,285,129
R-squared/pseudo *R*-squared	0.253	0.334	0.138	0.278	0.383	0.145

Notes: 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. Specifications with binary outcome variables (home birth) are estimated using a probit model, and average marginal effects are reported. Other models are estimated using OLS. Home birth and antenatal visits are recorded only for children aged 0–4 at the time of the survey, and the standardized education score is recorded only for children aged 6–18 (of school age). Additional notes are available in Table 2. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

4.2. The Trade-off: Impacts of Fertility on Child Human Capital

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 2.1, 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 upwards by virtue of twins being born to healthier mothers.

4.2.1. 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, child birth order, 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. The available controls for mother's health are height, BMI and cluster-level health service availability in the developing country sample, and BMI and a self-reported assessment of own health on a Likert scale in the US sample. In both samples, the control for socioeconomic status is years of education of the mother (see Table A.2 for summary statistics of these variables) and in the developing country sample we also control for the wealth quintile of the family. Given that our interest is in capturing mother-level variation that predicts twinning, in the models with the richest controls (labelled +S&H in Table 5) we additionally include quadratic interactions between continuous measures of maternal health (height and height squared, and BMI and BMI squared) and categorical variables for maternal education. We note that this sequence implies that the health and socioeconomic interaction terms will all "load onto" the last column, when in fact they "belong" in controls for health as much as in controls for socioeconomic indicators. We chose this ordering as it makes our estimates of the coefficient change associated with maternal health covariates more conservative, though we note that, in general these interaction terms have much less impact than first-order controls. 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 6.1% and 8.4% of a standard deviation. In the United States they are between 1.1% and 2.4% for education and between 0.3% and 1.6% 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 1–1.2 times as important as observables for these estimate of the fertility–human capital trade-off to be entirely driven by selection into fertility. The corresponding ratio in the United States varies from between 1 and 3. In developing countries, the estimated education-fertility trade-off is decreasing in the birth order at which twins (the additional child) occur, that is, it is largest in the 2+ sample and smallest in the 4+ sample. In the United States, 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 (Online Appendix Tables A.4 and A.5).

4.2.2. IV Estimates with the Twin Instrument. IV estimates using the twin instrument are in Tables 6 (DHS) and 7 (United States), 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 Online Appendix Tables A.6–A.8.

TABLE 5. OLS estimates of the fertility–human capital trade-off: Developing country and the United States.

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***	−0.129***	−0.084***	−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	0.134	0.196	0.093	0.122	0.193	0.081	0.115	0.191
Altonji et al. Ratio			1.243			1.131			1.044
Panel B: US Results									
Dependent variable = school z-score									
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									
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

Notes: OLS regressions of 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 3). Base controls consist of fixed effects for child's age and year of birth, child gender, mother's age at birth, birth order, and a cubic for mother's age at time of survey. For the US 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 its square, and for DHS, mother's height and its square, and coverage of prenatal care at the level of the survey cluster. For US data, we include controls for mother's self-assessed health on a Likert scale. In the final columns (+S&H), quadratic health variables are interacted with year of education indicators. Standard errors are clustered by mother. $^{**}p < 0.05$; $^{***}p < 0.01$.

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***	0.844***	0.842***	0.829***	0.838***	0.840***	0.860***	0.866***	0.867***
	(0.029)	(0.028)	(0.028)	(0.025)	(0.025)	(0.025)	(0.026)	(0.026)	(0.026)
Kleibergen–Paap rk statistic	826.32	876.23	931.50	1066.33	1124.84	1158.90	1119.75	1096.09	1145.77
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	−0.018	−0.012	−0.028	−0.041**	−0.048**	−0.027	−0.038*	−0.036*
	(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		0.005	0.414		0.007	0.032		0.108	0.348

Notes: Panels A and B present coefficients and standard errors for the first and second stages of twin instrumented impacts of fertility on schooling outcomes. The 2+, 3+, and 4+ samples are defined in the estimation sample section of the paper (Section 3). Base controls consist of child age, mother's age at birth, and birth order 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 their square, and coverage of prenatal care at the level of the survey cluster. In the final columns (+S&H), quadratic health variables are interacted with year of education indicators. In each case the sample is made up of all children aged between 6 and 18 years from families in the DHS who fulfil 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)									
Twins	0.698***	0.701***	0.704***	0.739***	0.739***	0.742***	0.801***	0.804***	0.818***
	(0.026)	(0.026)	(0.026)	(0.047)	(0.047)	(0.047)	(0.079)	(0.079)	(0.078)
Kleibergen–Paap rk statistic	704.66	710.97	737.35	247.09	250.27	251.38	101.80	102.35	108.62
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*	−0.011	−0.014	−0.017	−0.125	−0.134	−0.142
	(0.061)	(0.061)	(0.060)	(0.067)	(0.067)	(0.067)	(0.151)	(0.151)	(0.148)
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		0.546	0.390		0.233	0.115
Dependent variable = excellent health									
Fertility	0.009	0.027	0.026	−0.036	−0.058*	−0.057*	0.032	−0.025	−0.031
	(0.025)	(0.021)	(0.021)	(0.039)	(0.032)	(0.032)	(0.060)	(0.053)	(0.051)
Observations	70,277	70,277	70,277	53,393	53,393	53,393	24,358	24,358	24,358
Coefficient difference		0.164	0.213		0.340	0.365		0.123	0.090

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), birth order FEs, 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 and 18 years from families in the NHIS who fulfil 2+ to 4+ requirements for schooling variables, and for children aged between 1 and 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 Online Appendix Table A.2. Standard errors are clustered by mother. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

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 0.7–0.8 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 fertility 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 (2) 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 (Black et al. 2005; Cáceres-Delpiano 2006; Åslund and Grönqvist 2010; Angrist et al. 2010; Fitzsimons and Malde 2014), we find no significant 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 4.1 that twin mothers are positively selected on health (and education), we see that upon introducing controls for maternal selectors of twinning, a fertility–human capital 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 4.8% (if mother's education is also included). In the 4+ sample, the corresponding figures are 2.7%, 3.8% and 3.6%.[24]

Although 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 Online 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,

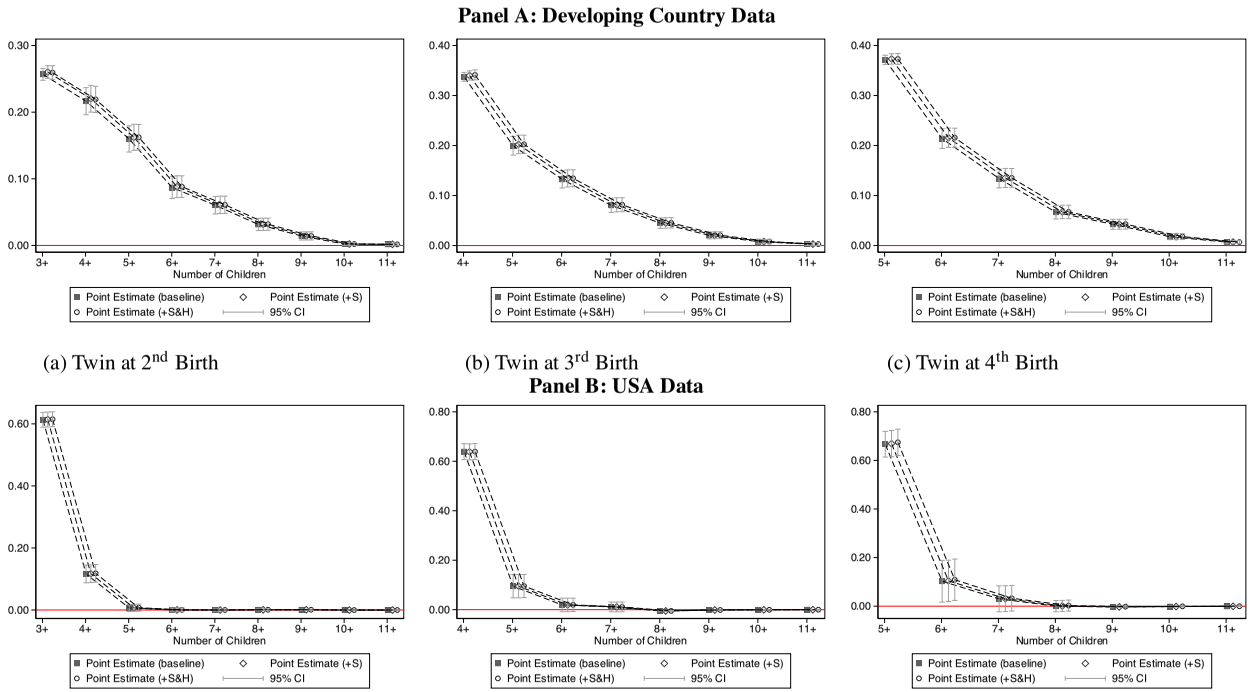
24. We experimented with adding education first, and then health. For example, in the case of the 3+ sample in DHS regressions, the baseline estimate of the trade-off is 2.8%. When we control for just maternal education, this moves to is 3.8%. When we additionally control for maternal health, it rises to 4.8%. Crudely, this suggests that maternal education alone accounts for 47% of the movement, whereas the health indicators explain the remaining 53%. This calculation is: $(0.0375–0.0284)/(0.0476–0.0284) = 0.472$.

3 or 4 leading to an additional 0.7–0.8 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 fertility–human capital 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 US samples range between about 21,000 and 70,000 individuals whereas 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 recognized that twin-IV estimates are often not precise, and indeed, note that in this case IV confidence intervals entirely overlap with the OLS confidence intervals meaning that formal statements of their relative magnitudes cannot be made, something we do not observe in the large developing country sample. So it is quite striking that we find any significant trade-off for education and health. Overall, partial bias adjustment reveals a statistically significant 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).

IV Estimates: Who are the Compliers? We have argued that maternal health is an omitted variable in the 2SLS model with the twin instrument and that, as a result, the progressive removal of mother-level twin selection characteristics from the unobserved term in IV models will render them progressively more negative. However, as discussed in Section 2.1.1, introducing controls changes the complier group and LATE identified in the linear IV model. There are two potential issues here, (i) non-linearity—the inclusion of controls may change the weighting across parities and (ii) heterogeneous effects—inclusion of controls may change the weighting on households with different characteristics (e.g., high vs. low maternal health).

First, inclusion of additional covariates may change the weights assigned to fertility change occurring at different parities. Angrist et al. (2010) show that the estimated impact is a weighted average of the variation in fertility at each birth induced by the twin birth. They demonstrate that the weights correspond to the first stage impact of twin birth at birth k on an indicator of fertility exceeding each particular parity j, denoted $d_{ji} \equiv 1[fertility_i \geq j] \; \forall j \in (k+1), \ldots, 11$. We replicate Angrist et al. (2010)'s weighting estimates in Figure 3, regressing each d_{ji} on the twin birth indicator, including identical controls as in each IV model. Our innovation here is that as well as estimating these parity-specific weights for the baseline model, we document how these weights change when additionally adding socioeconomic and health covariates. We observe, as also found by Angrist et al. (2010), that twin births have the largest impact on fertility at parities close to the parity of twin birth, but that twins also impact the likelihood of exceeding higher birth parities, and this is much more so in the



(a) Twin at 2nd Birth (b) Twin at 3rd Birth (c) Twin at 4th Birth

(d) Twin at 2nd Birth (e) Twin at 3rd Birth (f) Twin at 4th Birth

FIGURE 3. First stage impact of twinning on fertility, and average causal response weights. Each panel documents the impact of a twin at a particular birth order on the likelihood of exceeding a particular family size (indicated on the x-axis). These "average causal response" function give the weights placed on each parity in LATEs estimated using twins at a particular birth order as an IV. In each panel, point estimates and 95% confidence intervals are plotted in three cases, corresponding to the impact of twins on fertility conditional on baseline controls, baseline plus health controls, or baseline, plus health, plus socioeconomic controls. Each of the three estimates refer to the same fertility outcomes documented on the x-axis, however have been slightly shifted for ease of visualization.

developing country sample where access to contraceptives is lower.[25] Importantly, we note that the weights are very stable to adding additional controls. This suggests that heterogeneity in fertility impacts owing to inclusion of the additional covariates does not explain the change in the IV estimates that we see with these additional covariates.

Second, parameter heterogeneity in LATEs for different groups of mothers may explain the coefficient movements we observe, for instance, Brinch et al. (2017) document considerable heterogeneity in effects of fertility on children's education. In a fully saturated model, 2SLS estimates correspond to a weighted average of covariate specific LATEs, with the weights assigned to each group determined by the conditional variance of the first stage fitted value at each point of support of the covariates (Angrist and Pischke 2009). Thus it is possible that the more negative coefficients we observe in our 2SLS estimates after we add controls for socioeconomic and health characteristics arise because larger weights are assigned to groups with more negative LATEs. However, as shown in Angrist and Pischke, the covariate specific weights will be larger for groups with more instrumental variation. In our case, given that the twin instrument is a binary variable, the variance of the instrument will be increasing in the proportion of twins.[26] Our evidence suggests that twins are positively selected on maternal characteristics. Since the proportion of twins is significantly lower than 0.5 in any population group, the weight given to the LATE of any group will be increasing in the rate of twin births, so positively selected groups (healthier and more educated mothers) will be given larger weights in the 2SLS estimates conditional on these controls. We posit that these positively selected groups will have *smaller* fertility–human capital trade-offs (and we document this later in the paper using maternal education). In this case, adding health and socioeconomic controls will give larger weights to mothers with smaller trade-offs. Based on this line of reasoning, if anything, heterogeneity in LATEs will cause our estimates to be move upwards and closer to zero when adding controls, so our finding that the coefficients become more negative are unlikely to owe to heterogeneity, and our IV bounds argument holds a fortiori.

4.2.3. 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

25. For the United States, we estimate that twins at second birth increase the chances of exceeding a family size of 3 by 60%, and the chances of exceeding a family size of 4 by 10%, and have no impact on the chances of exceeding family sizes of 5 or larger. In developing countries, a twin birth at order 2 increases the chances of exceeding family sizes of 3 up to 9.

26. This is provided that the proportion of twins does not exceed 0.5. Note that the variance of a binary variable is $p(1-p)$, where p is the proportion of observations for which the variable is equal to 1. This value is increasing between $p = 0$ and $p = 0.5$, and then decreasing up until $p = 1$.

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) and Mogstad and Wiswall (2016). Models of this type loosen the assumption of 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. The restricted (linear) and unrestricted (non-linear) IV models are in Table 8, and we report results by the same parity samples as the main IV results in Table 6.

In Table 8 we observe, firstly, that as described in Table 6, the linear specifications are universally lower (more negative), and often become statistically significant when partially controlling for the selection of twins. The only difference in these linear results and those reported earlier in that we now restrict the sample to families with 6 or fewer children in line with results in Mogstad and Wiswall (2016), which involves a loss of between 5% and 18% of the sample depending on the parity sample used. Descriptive statistics on family size in each parity group are in Online Appendix Figure A.7. Panel B reports non-linear estimates, which confirm the results 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, whereas the impact of moving from two to three siblings is large and negative.

However, what is most pertinent to the present analysis is that the coefficients in the non-linear model are nearly universally more negative 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 some cases become statistically significant. Thus, our finding that the twin-IV estimator tends to underestimate the magnitude of the causal effect of fertility on child human capital holds in the linear and non-linear specifications. Departing from any previous work, we investigated a further generalization that is pertinent to our purpose. In addition to allowing for differential effects of additional children at each level of family size, we allow these differential effects of family size to vary with the health and SES of the mother. It seems plausible that the impact of maternal health could be quite different when considering a change of family size from one to two children versus a change in family size from three to four children. The final column of each panel in Table 8 shows estimates allowing these interactions. The trade-off parameter is not significantly changed and our main result—that controlling for mother's health and SES sharpens the trade-off—holds a fortiori.

4.2.4. IV Effect Sizes in Perspective. Since the fertility–human capital 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. Using summary statistics from Table A.2, we can convert standardized estimates into years of education. Our results (in the linear model) imply that an additional birth in a family is associated with between 0.04 (2+) and 0.15 (3+) fewer years of completed education (developing countries) or 0.06 to 0.5 fewer grades progressed (United States). In a widely cited

TABLE 8. Linear and nonlinear IV estimates for marginal effects with and without full twin controls.

	Two-plus			Three-plus			Four-plus		
	Baseline	+S&H	Interactions	Baseline	+S&H	Interactions	Baseline	+S&H	Interactions
Panel A: Linear estimates of marginal effects									
Number of children	0.034	0.014	0.091	−0.015	−0.052**	−0.062*	0.009	−0.024	−0.019
	(0.036)	(0.034)	(0.055)	(0.030)	(0.026)	(0.035)	(0.037)	(0.031)	(0.037)
Panel B: Unrestricted estimates of marginal effects									
Siblings ≥ 2	0.222***	0.193**	0.182**						
	(0.081)	(0.087)	(0.085)						
Siblings ≥ 3	−0.094**	−0.115**	−0.153	−0.016	−0.070*	−0.076**			
	(0.045)	(0.051)	(0.106)	(0.044)	(0.039)	(0.035)			
Siblings ≥ 4	−0.036	−0.039	−0.058	−0.039	−0.041	−0.035	0.013	−0.031	−0.022
	(0.077)	(0.083)	(0.093)	(0.062)	(0.055)	(0.088)	(0.062)	(0.061)	(0.046)
Siblings ≥ 5	0.051	0.057	0.024	0.030	0.030	0.025	0.028	0.026	−0.002
	(0.063)	(0.064)	(0.089)	(0.058)	(0.049)	(0.060)	(0.064)	(0.060)	(0.055)
Observations	239,898	239,898	239,898	350,155	350,155	350,155	329,230	329,230	329,230
Joint significance (p-value)	0.000	0.000	0.022	0.843	0.109	0.080	0.790	0.870	0.879

Notes: Each column and panel present a separate regression using DHS data. Siblings ≥ 2 refer to the marginal effect of moving from 1 to 2 siblings, Siblings ≥ 3 refer to moving from 2 to 3 siblings, and so forth. Each model includes maternal age, country, survey year, birth order, 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 re-estimated. The p-value for "Joint Significance" refers to the null hypothesis that each parameter is zero. This is implemented using a χ^2 test based on the estimated coefficients, and block-bootstrap variance–covariance matrix. Low p-values provide evidence against joint insignificance of the Sibling indicators. First stage regressions are displayed in Online Appendix Table A.9. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

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, whereas 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 United States, our estimates are of similar magnitude to estimates of the effect of an additional 1,000 grams of birth weight over the normal birth weight 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% and 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.

4.2.5. Parental Investment Behaviour as a Response to Twinning. A twin birth is used to instrument fertility because it creates an unexpected increment to family size. To recapitulate, our critique has centred upon mothers of twins being selected, in particular, mothers of twins are systematically healthier. In this section, we consider whether it matters for the current analysis that the twin births themselves are less healthy. Twins are lower birth weight than singleton births and closely spaced.[27] Rosenzweig and Zhang (2009) made the observation that if parents reinforce the initial endowment (as they find in China), then they will tend to allocate resources away from the low-endowment twins to other children in the family and, in this case, we will tend to under-estimate the QQ trade-off (i.e., our IV subjects, the pre-twins, may gain higher education than in the absence of reinforcement). The same arguments hold for close birth spacing of twins as for low birth weight of twins. Close spacing may make investments in twins more costly, as suggested by Rosenzweig and Zhang (2009) and, thereby, encourage parents to shift investments to pre-twin children.[28]

27. Refer to Online Figures A.3–A.4. Our data suggest birth weight differentials of 697 g in the developing country sample, and 885 g in the United States.

28. It is interesting to note that this birth-spacing effect may flow not only from compliers but additionally from always-takers (individuals who would have had an additional birth if they had not had twins). To

If, as Rosenzweig and Zhang (2009) suggest, parents reinforce child endowments, their argument provides a reason independent of our argument for why twin-IV estimates will tend to be upwards biased. This said, if the lower endowments of twins were the only challenge to IV, then adjusting for maternal health would not produce the coefficient shifts that we document.[29]

There is nevertheless the potential concern that parents may compensate rather than reinforce the child endowment. There is some evidence of this, for instance among more educated mothers Conley (2008) and Hsin (2012). If parents compensate, and if this is more so among relatively educated and healthy mothers who are more likely to have twins, then this may act to offset the bias that we highlight (the positive selection of twin births). If this were the case, it is no longer clear that we can sign the bias on IV estimates. To investigate this concern, we examine parental investments in a child in response to their endowment, and the endowment of their subsequent siblings. We extend this analysis to look at heterogeneity in parental responses to endowments by the education of the mother. As we do find heterogeneity, we repeat our earlier investigation of how OLS and twin-IV estimates vary upon adjusting for maternal characteristics that predict twin birth, but now we separate the sample into more versus less educated mothers. We similarly allow heterogeneity by an indicator of the health of the mother.

Tables 9 and 10 present estimates of parental investment responses to initial endowments, measured by birth weight. We also present responses to whether or not the child is a twin. As twins have lower birth weight, we show the twin coefficient conditional and unconditional on birth weight, expecting that the conditional case will capture other factors like the close spacing of twins. We use a panel of mothers from the DHS and, since this is not feasible with the NHIS, for the United States, we use the NLSY. The indicator for parental investment in the developing country sample is the duration of breastfeeding in months, available for each child born to DHS mothers for whom breastfeeding is complete. For the US sample, we model two investments, both binary, indicating whether or not the mother breastfeeds and reads frequently with her child. We consistently condition on mother fixed effects to purge effects of the mother's (time invariant) health and socioeconomic status on their investment behaviour and

see this, note that there are no never-takers from the twin-instrument (twinning must always shift fertility for those with the instrument switched on), and so from Imbens and Wooldridge (2007) we can write the numerator of the LATE term as

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] =$$
$$[E(Y_i|Z_i = 1, complier)\pi_c + E(Y_i|Z_i = 1, always-takers)\pi_a]$$
$$-[E(Y_i|Z_i = 0, complier)\pi_c + E(Y_i|Z_i = 0, always-takers)\pi_a],$$

where π_c and π_a refer to the proportions of compliers and always-takers respectively. If the exclusion restriction holds, $E(Y_i|Z_i = 1, always\text{-}takers) = E(Y_i|Z_i = 0, always\text{-}takers)$, and so the IV estimand is local to compliers. However, if birth spacing due to twins impacts investments in other children, this equality may not hold. If the effect of birth spacing on investments in other children is similar for both compliers and always-takers, the always-takers will additionally contribute to the inconsistency in IV estimates, in the same way that the compliers do.

29. Notice that our argument refers to cross-mother differences (that influence the chances of twin vs. singleton birth) and highlight within-family re-allocations (between twin and singleton siblings).

TABLE 9. Parental investment responses to the child endowment (DHS).

Dependent variable:	All observations			<Mean education			≥Mean education		
Duration of breast-feeding	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Birth weight in kg	0.445***		0.386***	0.330***		0.278***	0.510***		0.445***
	(0.056)		(0.057)	(0.093)		(0.095)	(0.071)		(0.071)
Twin		−1.903***	−1.627***		−1.761***	−1.571***		−2.022***	−1.696***
		(0.295)	(0.298)		(0.507)	(0.516)		(0.359)	(0.361)
Observations	65,815	65,815	65,815	22,361	22,361	22,361	43,454	43,454	43,454
R-squared	0.8133	0.8133	0.8136	0.8179	0.8179	0.8181	0.8098	0.8097	0.8101
Mean breastfeeding	10.795	10.795	10.795	11.643	11.643	11.643	10.359	10.359	10.359

Notes: Regressions of duration of breast-feeding in months on each child's birth weight include maternal fixed effects as well as fixed effects for the child's age, mother's age at birth, and birth order of the child. The sample consists of all DHS observations for which birth weight is recorded, and whose mothers report that breastfeeding is no longer ongoing. Birth weight is recorded in grams and rescaled to kilograms for ease of presentation. DHS record breastfeeding duration for all children born within 5 years of the date of the survey. Results are split by the mother's education level, compared to the average of all women in her country and survey wave (irrespective of whether the woman is in the breastfeeding sample). In the table footer, mean breastfeeding refers to the average duration of breastfeeding of each estimation sample in months. $^{***}p < 0.01$.

TABLE 10. Parental investment responses to the child endowment (NLSY child and young adult survey).

	All observations			No college			Some college		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Dependent variable = child breastfed									
Birth weight in kg	0.060***		0.060***	0.061***		0.064***	0.057***		0.050**
	(0.012)		(0.012)	(0.015)		(0.015)	(0.021)		(0.021)
Twin		−0.069	−0.010		−0.008	0.051		−0.157	−0.104
		(0.052)	(0.054)		(0.034)	(0.038)		(0.118)	(0.119)
Observations	8,605	8,605	8,605	5,640	5,640	5,640	2,965	2,965	2,965
Mean breastfeeding	0.438	0.438	0.438	0.361	0.361	0.361	0.584	0.584	0.584
Panel B: Dependent variable = mother reads to child frequently									
Birth weight in kg	−0.023		−0.024	0.015		0.006	−0.070***		−0.066**
	(0.016)		(0.016)	(0.019)		(0.020)	(0.027)		(0.027)
Twin		−0.001	−0.025		−0.156	−0.150		0.128	0.066
		(0.078)	(0.079)		(0.097)	(0.099)		(0.119)	(0.121)
Observations	6,636	6,636	6,636	3,949	3,949	3,949	2,687	2,687	2,687
Mean reading	0.477	0.477	0.477	0.424	0.424	0.424	0.554	0.554	0.554
Panel C: Dependent variable = mother reads to previous child frequently									
Birth weight in kg	0.005		−0.001	0.039		0.027	−0.027		−0.027
	(0.024)		(0.025)	(0.030)		(0.031)	(0.036)		(0.039)
Twin		−0.066	−0.066		−0.166**	−0.140		0.028	−0.000
		(0.076)	(0.080)		(0.082)	(0.086)		(0.114)	(0.125)
Observations	2,907	2,907	2,907	1,894	1,894	1,894	1,013	1,013	1,013
Mean reading	0.429	0.429	0.429	0.386	0.386	0.386	0.508	0.508	0.508

Notes: Each regression includes maternal fixed effects and FEs for the mother's age at birth, and birth order of the child. The sample consists of all NLSY child observations for which mothers report having breastfed their child (top panel), and report the frequency of reading with their child (bottom panels). Birth weight is recorded in ounces and rescaled to kilograms for ease of presentation. The frequency of reading is recoded from a categorical question with responses, "never", "several times a year", "several times a month", "once a week", "about 3 times a week", and "everyday". Frequently reading is an indicator variable coded as one if the response is "about 3 times a week" or "everyday". All responses are for when children are aged between 6 and 9 years of age. Results are split by the whether the mother has completed any college education, or not. The final panel regresses mother's reading to previous children on the birth weight/twin status of each child, and so is only defined for children with younger siblings. $^{**}p < 0.05$; $^{***}p < 0.01$.

their twinning chances. We also control for fixed effects for the age of the mother, the age of the child, birth order, and child sex.

In the full sample, we observe that breastfeeding is longer for higher birth weight children (increasing by 0.3−0.4 months per additional kilogram of birth weight), in line with reinforcing behaviour. And, in line with this, unconditional on birth weight, twins are breastfed for around 1.9 months less than singletons. In the DHS sample, where twins weigh on average 697 grams less than singletons, this translates to a birth weight impact of twins of $0.445 \times 0.697 = 0.31$ months less breastfeeding. *Conditional on birth weight*, twins are still breastfed for around 1.6 months less than singletons, consistent with their close spacing that makes it harder to breastfeed both twins.

In panel A of Table 10 estimates from the NLSY similarly show that American mothers' breastfeeding responses are to reinforce the birth weight endowment. The coefficient on the twin indicator has a sign consistent with reinforcement but it not statistically significant in this sample. Note, though, that the NLSY measure of breastfeeding is binary, and so we cannot rule out that the twin indicator influences the duration of breastfeeding conditional on breastfeeding. In the developing country and American samples, the results for breastfeeding behaviour are broadly similar for women with educational attainment below and above their country-level average (columns (4)–(9)).

Panel B of Table 10 investigates using the NLSY whether the mother read to the child frequently during the ages 6–9 years. Overall, and for women with no college, we can detect no significant response of reading to the child endowment but, for women with college, we see evidence of *compensating* behaviour, consistent with Hsin (2012). The estimates in panels A and B describe parental investments in a child responding to the endowment of that child. However, the typical twin IV experiment consists of examining the impact of a twin birth on their older siblings. We investigate this in panel C, using reading in the NLSY and limiting the estimation sample to children who have younger siblings.[30] We find no significant impact associated with birth weight. However, among women without a college education, we find that mothers with twins read less to children born before the twins (column (5)). This seems plausible as maternal time may be so stretched after having twins that investments may fall for *all* children in the family. It demonstrates that, in contrast to the common assumption in the literature, reinforcing behaviour may not imply that siblings born prior to twins receive more resources when twins are born, and we expect this will vary with the type of investment we consider.

Overall, the results in this section suggest that our running argument that twin IV estimates act as an upper bound on the impact of fertility on child outcomes will hold in the DHS sample, and likely in the US sample for women with no college provided that reinforcing behaviours dominate any direct impact of twin births (as suggested by coefficient movements discussed in the following paragraph). However, for women

30. We cannot conduct this test using the developing country data because we do not observe investments for different siblings.

TABLE 11. Maternal education, the twin instrument and the fertility–human capital trade-off.

	OLS			IV		
	Base	+H	+S&H	Base	+H	+S&H
Panel A: Developing country results						
Fertility (less educated)	−0.099***	−0.085***	−0.068***	−0.034**	−0.042***	−0.043***
	(0.002)	(0.002)	(0.002)	(0.017)	(0.016)	(0.015)
Observations	366,799	366,799	366,799	366,799	366,799	366,799
Fertility (more educated)	−0.094***	−0.084***	−0.055***	−0.024	−0.031	−0.029
	(0.002)	(0.002)	(0.002)	(0.019)	(0.019)	(0.018)
Observations	247,258	247,258	247,258	247,258	247,258	247,258
Panel B: US results						
Dependent variable = school z-score						
Fertility (less educated)	−0.030***	−0.029***	−0.022**	−0.015	−0.015	−0.022
	(0.010)	(0.010)	(0.010)	(0.075)	(0.074)	(0.074)
Observations	41,733	41,733	41,733	41,733	41,733	41,733
Fertility (more educated)	−0.021***	−0.022***	−0.022***	−0.094	−0.098	−0.098
	(0.007)	(0.007)	(0.007)	(0.059)	(0.059)	(0.059)
Observations	46,553	46,553	46,553	46,553	46,553	46,553
Dependent variable = excellent health						
Fertility (less educated)	−0.009***	−0.006**	−0.005*	−0.053	−0.051*	−0.051*
	(0.004)	(0.003)	(0.003)	(0.038)	(0.031)	(0.031)
Observations	49,050	49,050	49,050	49,050	49,050	49,050
Fertility (more educated)	0.007*	−0.001	−0.001	0.018	0.012	0.012
	(0.004)	(0.003)	(0.003)	(0.025)	(0.022)	(0.022)
Observations	54,070	54,070	54,070	54,070	54,070	54,070

Notes: OLS and IV results are shown for the pooled 2+, 3+ and 4+ samples, splitting samples by the educational level of each mother. In the case of IV estimates, fertility is instrumented using the twin instruments with pooling procedure described in Angrist et al. (2010) and refinement discussed in Mogstad and Wiswall (2012). In the developing country sample, less and more educated refers to mothers with education, respectively, below and above the country-level mean (calculated in each survey) given heterogeneity in educational attainment by countries. In the case of the United States, less educated refers to mothers with high school education or less, and more educated refers to mothers with college education or higher. All other details follow OLS and IV estimates in Tables 5–7. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

with college in the United States, our observation that one of two studied investments compensates birth weight may mean that we cannot necessarily sign the bounds. To examine this, we present IV estimates separating the more and less educated samples in Table 11, using the same educational splits as in Tables 9 and 10. Splitting the sample challenges statistical power and the IV estimates are imprecise, particularly for the smaller US sample, so we now present pooled estimates (described in the Section 2). We discuss pooled estimates in the full sample in the following section (Figure 4). We also present OLS estimates as they are more precise, and they provide an alternative metric indicating the nature of the bias through covariate adjustment.

In the developing country sample, we observe that the inclusion of covariates tends to move OLS point estimates upwards (towards zero), shift IV point estimates downwards (away from zero), and to produce a clearly negative relationship in the less educated sample. These results are in line with our main argument of positive selection of mothers of twins. In the US data, the less educated sample behaves broadly like the developing country sample. However in the sample of more educated mothers, the inclusion of controls typically does not increase OLS estimates, or lead to the emergence of statistically significant trade-offs in conditional IV models.[31] We obtain broadly similar results splitting by an indicator of maternal health rather than education, see Online Appendix Table A.10. In summary, our findings suggest that it can be instructive not only to allow for heterogeneity in treatment effects by maternal characteristics that influence investments in children, but to model and account for parental responses to birth endowments over and above allowing for positive selection into twin births (the latter being the issue that we emphasize because it has not been rigorously assessed before).

4.3. Bounding the QQ Trade-off

The adjusted twin-IV results will not provide consistent estimates of β_1 as there are almost certainly omitted indicators of maternal health.[32] 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. Figure 4 presents the range of the alternative linear bounds estimates discussed in Section 2.2, along with the corresponding OLS and IV estimates with base controls, health, and health and socioeconomic controls. In each case we present the bound end points (or point estimates) along with the 95% confidence intervals associated with each parameter/partially identified set. We present results for the 2+, 3+ and 4+ samples, as well as for the pooled sample using the methodology discussed in Section 2.1.1. Figure 4 is based on bound estimates for the developing country DHS data that are sufficiently large.[33]

As discussed previously, OLS estimates tend to be more negative than the true estimate and IV estimates, more positive. The inclusion of controls for maternal health in the IV models makes the point estimates more negative, and/or increases their precision. Nevo and Rosen bounds are based on the premise of opposite directions

31. Comparing these results with those in Tables 6 (DHS) and 7 (United States), note that the parity-specific results are typically only significant at one parity, so that pooling across parities as we do here, we lose significance. The coefficient movements are nevertheless illustrative for the current purpose.

32. 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.

33. The NHIS data contain only 21,000–70,000 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 confidence intervals for bounds, we further challenge statistical power. Still, we do discuss bounds for the American NHIS sample, albeit imprecise, in the following section.

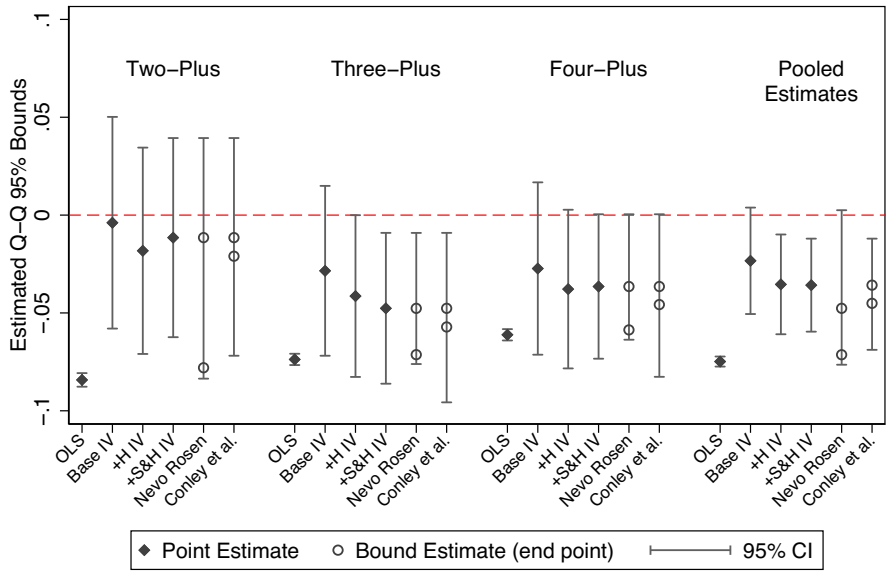


FIGURE 4. Parameter and bound estimates of the Q–Q trade-off. Each set of estimates refer to the 95% confidence intervals on parameter bounds of the impact of fertility on child education. Two-plus, three-plus and four-plus refer to parity specific groups, and pooled estimates refer to these samples pooled following the procedure described in Angrist et al. (2010) and refinement discussed in Mogstad and Wiswall (2012). 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. Confidence intervals on Nevo and Rosen bounds are estimated following Chernozhukov et al. (2013). Upper and lower end points of the interval estimates are plotted with hollow circles.

of selectivity into fertility and twinning, and are approximately bounded between the IV estimate as the upper bound and the OLS estimate as the lower bound. Where multiple potential upper or lower bound candidates exist, inference is based on adaptive inequality selection. Conley et al. IV bounds can lead to tightening of the upper bound (as we discuss in what follows and document in Figure 5). The estimates in Figure 4 are based on the UCI approach in Conley et al., with priors allowing for the exclusion restriction to fall anywhere between 0 (in which case the IV is valid) and 0.008, in which case being born to a twin mother implies a direct benefit of 0.8% of a standard deviation in educational outcomes *beyond* the impact mediated by fertility. A discussion of the calculation of these priors is provided in Bhalotra and Clarke (2016), though we document robustness to alternative priors in the following paragraphs. We also note that this is providing quite a wide range for the failure of the exclusion restriction, as 0.8% of a standard deviation is around 20% of the total estimated impact of *fertility* on school outcomes in the largest IV estimate (the 3+ sample).

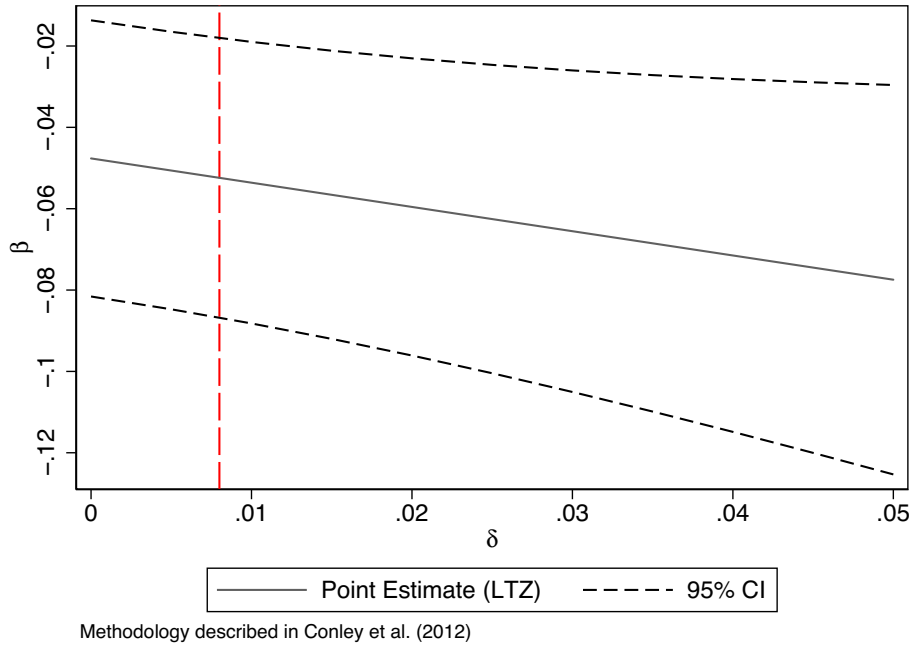


FIGURE 5. Plausibly exogenous bounds: school z-score. Confidence intervals and point estimates are calculated according to Conley et al. (2012) using DHS data for the 3+ sample. Estimates reflect a range of priors regarding the validity of the exclusion restriction required to consistently estimate $\hat{\beta}_1$ using twinning in a 2SLS framework. The local to zero (LTZ) approach treats the uncertainty surrounding γ, the coefficient on the instrument when included in the structural equation, as being normally distributed, with the mean and variance of a $U(0, \delta)$ variable. The vertical dashed line indicates the point at which priors are comparable to those in Figure 4.

The informativeness of the bounds is evaluated against the criteria laid out by Hotz et al. (1997): first do the bounds enable us to determine if the effect is negative or positive, second can we reject the point estimates of linear IV, and third do our bounds allow us to reject the OLS estimate. In general, for the 3+, 4+ and pooled 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 in the case of 2+ and pooled estimates 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).

Once again, using summary statistics from Table A.2, we can convert standardized estimates from these bounds into years of education. The estimated pooled effect of fertility shocks on education from Conley et al.'s bounds is between around 4 and 5% of a standard deviation. Using the standard deviation in the sample of 3.15 years, this implies an average effect of around 0.12−0.15 years of education per additional sibling at the age of 11 years (the average age in the sample). In the case of the US estimates, where the s.d. of years of education is 3.85 years, the average estimated

effect based on the pooled bounds estimates is 8%–9% of a standard deviation in grade retention, which equates to a marginal effect of 0.3–0.36 years of education. On average the likelihood of the child being in excellent health in the American sample falls by 1%–2% according to the pooled bounds estimates. Overall, these are quite large effects relative to the marginal effects of different policy interventions considered in the literature (see Section 4.2.4).

Although Nevo-Rosen bounds are based on simple direction-of-selection arguments, Conley et al. bounds are based on priors over the failure of the exclusion restriction. In Figure 4 we present bounds based on quite a broad prior, so we investigated sensitivity to the choice of prior in Figure 5. We follow Conley et al. (2012) in plotting bounds estimates under their "LTZ" assumptions, for a series of priors imposing γ (the measure of the violation of the exclusion restriction) to be normally distributed, with the mean and variance of a $U(0, \delta)$ variable. On the horizontal axis, we denote each value of δ considered, and on the vertical axis, the resulting confidence intervals and bound mid-point estimates. Figure 5 provides these estimates for the DHS 3+ sample, other parity specific estimates and pooled estimates are discussed in robustness checks in what follows. In general, note that regardless of priors over the violation of the exclusion restriction, these bounds are informative, though they do widen considerably when priors that imply very strong violations of the exclusion restriction are entertained.

So far we have discussed bounds on parameters in linear models. The cost of presenting linear bounds is that they may hide considerable heterogeneity within an average parameter estimate. The benefit of the linear bounds, on the other hand, is that they tend to be tighter than earlier non-parametric versions of IV bounds, such as the Monotone IV bounds of Manski and Pepper (2000) *if* the underlying assumptions are correct. We estimated Monotone IV bounds, but given that such bounds impose very weak assumptions, namely a bound on outcome variables and monotonicity in IV, the resulting bounds are too wide to allow any firm conclusions. In theory, these bounds provide a partially identified comparison to the non-linear models discussed in Section 4.2.3, however, as observed in other implementations of "worst-case" non-parametric IV bounds such as those presented in Brinch et al. (2017), the bounds are non-informative. We provide precise values of these non-parametric bounds in Online Appendix D. We additionally document the MTS–MTR bounds, though as discussed in the Methods section and at more length in Online Appendix D, these bounds require us to make sign restrictions on the fertility–human capital trade-off such that we would have to assume that it does exist, which would defeat the purpose of the current analysis. In this case, we are able to considerably tighten estimated lower bounds (upper bounds are 0 by construction), and these lower bounds on the non-parametric ATE are generally comparable in magnitude to lower bounds estimated on the LATE documented previously. As we document in Online Appendix D, lower bounds range from slightly less than zero to around −0.2 standard deviations for child educational outcomes. What's more, as is the case with the non-linear estimates that follow Mogstad and Wiswall (2016), the MTS–MTR lower bounds become more negative as fertility shifts occur at higher parities.

4.4. Robustness and Extensions

We undertake a number of robustness and extensions to the methods discussed in the previous sections, results of which are presented in the Online Appendix. A recent study proposes a formal test of instrument invalidity (Kitagawa 2015). Using this test on the 2+ sample for the DHS data, we reject the validity of the twin instrument, see Online Appendix Figure A.8 and Table A.12. However this test is sensitive to curse of dimensionality considerations, and so to implement it we had to simplify the specification of controls.[34] We do not report results for the NHIS data because the sample is too small to obtain informative confidence intervals.

There is evidence to suggest that conception of monozygotic (MZ) twins is quasi-random even if conception of dizygotic (DZ) twins is not (Farbmacher et al. 2016). Since our argument pertains to the role of maternal health in ensuring survival of twin conceptions to birth, this distinction may not be relevant. We nevertheless subjected our argument to a harsher test by using only same sex twins to construct the twin instrument, as same-sex twins are more likely to be MZ (since our data do not identify MZ vs. DZ). Even with this restriction, we observe a similar pattern, of estimates of the fertility–human capital trade-off diverging from zero and becoming significant when controls for maternal health are included. See results for the DHS in Online Appendix Table A.13 and for the NHIS in Online Appendix Table A.14.

In Online Appendix Figure A.9 we present a series of bounds for the US sample, similar to those plotted for the DHS sample in Figure 4. The smaller sample results in considerably wider 95% confidence intervals. Additionally, given that IV confidence intervals often contain OLS parameters, the Nevo-Rosen bounds cross or flip in certain cases (suggesting incorrect identifying assumptions), and in these cases of bound crossing they are suppressed from the figure. We document the full set of linear bounds, OLS, and IV estimates along with their confidence intervals in Table A.10. This includes the few cases of crossing bounds in US data. Here we additionally provide a preferred set of Conley et al. (2012) bounds based on the LTZ approach, as described in Figure 5. Identical robustness tests of Conley et al. bounds are provided for the US 3+ sample in Figure A.10, and for alternative samples and pooled estimates in Online Appendix Figures A.11 (developing countries) and A.12 (United States).

Finally, we present bounds estimates using the twin instrument together with the sex-mix instrument (an indicator for whether each of the first 2, 3 or 4 children in a family were of the same sex, for the 2+, 3+ and 4+ families, respectively). This strategy has been implemented in Angrist et al. (2010) and Chesher and Rosen (2013, 2018). We present bounds for the DHS and the smaller NHIS samples as Online Appendix Figures A.13–A.14.[35] As documented in Chesher and Rosen (2013), the sex

34. 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.

35. Note that in the case of Nevo and Rosen's bounds with multiple instruments we follow their Proposition 5, which suggests using the minima of all IV estimates where each instrument is used separately, rather than a 2SLS model where all instruments are entered together.

mix instrument is considerably weaker than the twin instrument, and so its inclusion does little to tighten bounds.

5. Conclusion and Discussion

This paper demonstrates that twin-IV estimates of the fertility–human capital trade-off tend to be biased upwards (towards zero) 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 relationship down by about 2% 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 large as −9% of a standard deviation, though end point estimates of the upper bound are typically closer to −1% to −3%.

We conclude that additional unexpected births do have quantitatively important effects on their siblings' educational outcomes. The estimated −4% to −5% of a standard deviation impact in developing countries is equivalent to 0.12–0.15 fewer years in the classroom, and estimates of approximately −8% of a standard deviation in the United States implies 0.3 fewer 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. A recent survey of national governments suggests that fertility was perceived as too high in 50% of developing countries, with this figure rising to 86% among the least developed countries (United Nations 2010).

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% in some countries, with a mean of 27%. Unwanted fertility is not unique to poorer countries. For instance, despite access to contraceptive methods, 21% of all pregnancies in 2011 in the United States ended in elective abortion (Guttmacher Institute 2016). Moreover, there is a strong trend in IVF use, and up to 40% of IVF successes result in multiple births to women who wanted one child (Kulkarni et al. 2013), creating a growing set of unwanted children. This might exacerbate impacts of additional births on investments in preceding births.

References

Aaronson, Daniel, Rajeev Dehejia, Andrew Jordan, Cristian Pop-Eleches, Cyrus Samii, and Karl Schulze (2017). "The Effect of Fertility on Mothers' Labor Supply over the Last Two Centuries." IZA Discussion Paper 10559, Institute for the Study of Labor (IZA), Bonn, Germany.

Ager, Philipp, Casper Worm Hansen, and Peter Sandholt Jensen (2018). "Fertility and Early-Life Mortality: Evidence from Smallpox Vaccination in Sweden." *Journal of the European Economic Association*, 16, 487–521.

Aizer, Anna and Flávio Cunha (2012). "The Production of Human Capital: Endowments, Investments and Fertility." NBER Working Paper 18429, National Bureau of Economic Research, Inc., Cambridge, MA.

Almond, Douglas (2006). "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, 672–712.

Almond, Douglas, Kenneth Y. Chay, and David S. Lee (2005). "The Costs of Low Birth Weight." *Quarterly Journal of Economics*, 120, 1031–1083.

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

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

Almond, Douglas and Bhashkar Mazumder (2013). "Fetal Origins and Parental Responses." *Annual Review of Economics*, 5, 37–56.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber (2005). "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy*, 113, 151–184.

Anderson, Gabrielle E., Angela D. Whipple, and Shane R. Jimerson (2002). *Grade Retention: Achievement and Mental Health Outcomes*. National Association of School Psychologists, http://www.wrightslaw.com/info/fape.grade.retention.nasp.pdf.

Angrist, Joshua, Victor Lavy, and Analia Schlosser (2010). "Multiple Experiments for the Causal Link between the Quantity and Quality of Children." *Journal of Labor Economics*, 28, 773–824.

Angrist, Joshua D and William N. Evans (1998). "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *American Economic Review*, 88(3), 450–477.

Angrist, Joshua D. and Guido W. Imbens (1995). "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association*, 90, 431–442.

Angrist, Joshua D. and Jörn-Steffen Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton.

Anukriti, S, Sonia R. Bhalotra, and Hiu Tam (2016). "On the Quantity and Quality of Girls: New Evidence on Abortion, Fertility, and Parental Investments." IZA Discussion Paper 10271, Institute for the Study of Labor (IZA), Bonn, Germany.

Åslund, Olof and Hans Grönqvist (2010). "Family Size and Child Outcomes: Is There Really No Trade-off?" *Labour Economics*, 17, 130–139.

Bagger, Jesper, Javier A. Birchenall, Hani Mansour, and Sergio Urzúa (2013). "Education, Birth Order, and Family Size." NBER Working Paper 19111, National Bureau of Economic Research, Inc., Cambridge, MA.

Bailey, Martha J., Olga Malkova, and Zoë M. McLaren (2019). "Does Access to Family Planning Increase Children's Opportunities? Evidence from the War on Poverty and the Early Years of Title X." *Journal of Human Resources*, 54, 825–856.

Baird, Sarah, Joan Hamory Hicks, Michael Kremer, and Edward Miguel (2016). "Worms at Work: Long run Impacts of a Child Health Investment." *Quarterly Journal of Economics*, 131, 1637–1680.

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

Becker, Gary S. (1960). "An Economic Analysis of Fertility." In *Demographic and Economic Change in Developed Countries*, NBER Chapters. National Bureau of Economic Research, Inc., Cambridge, MA, pp. 209–240.

Becker, Gary S. and H. Gregg Lewis (1973). "On the Interaction between the Quantity and Quality of Children." *Journal of Political Economy*, 81, S279–S288.

Becker, Gary S. and Nigel Tomes (1976). "Child Endowments and the Quantity and Quality of Children." *Journal of Political Economy*, 84, S143–S162.

Bhalotra, Sonia and Damian Clarke (2019). "Twin Birth and Maternal Condition." *Review of Economics and Statistics*, 101, 853–865.

Bhalotra, Sonia and Tom Cochrane (2010). "Where Have All the Young Girls Gone? Identification of Sex Selection in India." IZA Discussion Paper 5381, Institute for the Study of Labor (IZA), Bonn, Germany.

Bhalotra, Sonia R. and Atheendar Venkataramani (2015). "Shadows of the Captain of the Men of Death: Early Life Health Interventions, Human Capital Investments, and Institutions." Available at SSRN: https://ssrn.com/abstract=1940725 or http://dx.doi.org/10.2139/ssrn.1940725.

Bhalotra, Sonia R. and Damian Clarke (2016). "The Twin Instrument." IZA Discussion Paper 10405, Institute for the Study of Labor (IZA), Bonn, Germany.

Bhalotra, Sonia R. and Atheendar Venkataramani (2013). "Cognitive Development and Infectious Disease: Gender Differences in Investments and Outcomes." IZA Discussion Paper 7833, Institute for the Study of Labor (IZA), Bonn, Germany.

Bhalotra, Sonia R., Atheendar Venkataramani, and Selma Walther (2018). "Fertility and Labor Market Responses to Reductions in Mortality." IZA Discussion Paper 11716, Institute for the Study of Labor (IZA), Bonn, Germany.

Bisbee, James, Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii (2017). "Local Instruments, Global Extrapolation: External Validity of the Labor Supply–Fertility Local Average Treatment Effect." *Journal of Labor Economics*, 35, S99–S147.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes (2005). "The More the Merrier? The Effect of Family Size and Birth Order on Children's Education." *Quarterly Journal of Economics*, 120, 669–700.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes (2016). "Does Grief Transfer across Generations? Bereavements during Pregnancy and Child Outcomes." *American Economic Journal: Applied Economics*, 8, 193–223.

Blake, Judith (1989). *Family Size and Achievement*. University of California Press, Berkeley.

Bougma, Moussa, Thomas K. LeGrand, and Jean-François Kobiané (2015). "Fertility Decline and Child Schooling in Urban Settings of Burkina Faso." *Demography*, 52, 281–313.

Brinch, Christian N., Magne Mogstad, and Matthew Wiswall (2017). "Beyond LATE with a Discrete Instrument." *Journal of Political Economy*, 125, 985–1039.

Bronars, Stephen G. and Jeff Grogger (1994). "The Economic Consequences of Unwed Motherhood: Using Twin Births as a Natural Experiment." *American Economic Review*, 84(5), 1141–1156.

Cáceres-Delpiano, Julio (2006). "The Impacts of Family Size on Investment in Child Quality." *Journal of Human Resources*, 41, 738–754.

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

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

Chernozhukov, Victor, Wooyoung Kim, Sokbae Lee, and Adam M. Rosen (2015). "Implementing intersection bounds in Stata." *Stata Journal*, 15, 21–44.

Chernozhukov, Victor, Sokbae Lee, and Adam M. Rosen (2013). "Intersection Bounds: Estimation and Inference." *Econometrica*, 81, 667–737.

Chesher, Andrew and Adam Rosen (2018). "Generalized Instrumental Variable Models, Methods, and Applications." CeMMAP Working Paper CWP43/18, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, London.

Chesher, Andrew and Adam M. Rosen (2013). "What Do Instrumental Variable Models Deliver with Discrete Dependent Variables?" *American Economic Review*, 103(3), 557–562.

Clarke, Damian (2018). "Children and Their Parents: A Review of Fertility and Causality." *Journal of Economic Surveys*, 32, 518–540.

Clarke, Damian and Benjamín Matta (2018). "Practical Considerations for Questionable IVs." *Stata Journal*, 18, 663–691.

Conley, Dalton (2008). "Bringing Sibling Differences in: Enlarging Our Understanding of the Transmission of Advantage in Families." In *Social Class: How Does it Work*, edited by A. Lareau and D. Conley. Russell Sage Foundation, New York, NY, pp. 179–200.

Conley, Timothy G., Christian B. Hansen, and Peter E. Rossi (2012). "Plausibly Exogenous." *The Review of Economics and Statistics*, 94, 260–272.
Currie, Janet and Enrico Moretti (2003). "Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings." *Quarterly Journal of Economics*, 118, 1495–1532.
De Tray, Dennis N. (1973). "Child Quality and the Demand for Children." *Journal of Political Economy*, 81, S70–S95.
Farbmacher, Helmut, Raphael Guber, and Johan Vikström (2016). "Increasing the Credibility of the Twin Birth Instrument." Working Paper Series 2016:10, IFAU—Institute for Evaluation of Labour Market and Education Policy, Swedish Ministry of Employment, Uppsala.
Fitzsimons, Emla and Bansi Malde (2014). "Empirically Probing the Quantity-Quality Model." *Journal of Population Economics*, 27, 33–68.
Galor, Oded (2012). "The Demographic Transition: Causes and Consequences." *Cliometrica, Journal of Historical Economics and Econometric History*, 6, 1–28.
Galor, Oded and David N. Weil (2000). "Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond." *American Economic Review*, 90(4), 806–828.
Grawe, Nathan D. (2008). "The Quality–Quantity Trade-off in Fertility across Parent Earnings Levels: A Test for Credit Market Failure." *Review of Economics of the Household*, 6, 29–45.
Guttmacher Institute (2016). "Induced Abortion in the United States." Fact Sheet, Guttmacher Institute.
Hanushek, Eric A. (1992). "The Trade-Off between Child Quantity and Quality." *Journal of Political Economy*, 100, 84–117.
Heckman, James, Rodrigo Pinto, and Peter Savelyev (2013). "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review*, 103(6), 2052–86.
Hotz, V. Joseph, Charles H. Mullin, and Seth G. Sanders (1997). "Bounding Causal Effects Using Data From a Contaminated Natural Experiment: Analysis the Effects of Teenage Chilbearing." *Review of Economic Studies*, 64, 575–603.
Hsin, Amy (2012). "Is Biology Destiny? Birth Weight and Differential Parental Treatment." *Demography*, 49, 1385–1405.
Imbens, Guido and Jeffrey Wooldridge (2007). "What's New in Econometrics. Instrumental Variables with Treatment Effect Heterogeneity: Local Average Treatment Effects." NBER Methods Lectures. https://www.nber.org/WNE/lect_5_late_fig.pdf, retrieved 1 February 2019.
Jacobsen, Joyce P., James Wishart Pearce, III, and Joshua L. Rosenbloom (1999). "The Effects of Childbearing on Married Women's Labor Supply and Earnings: Using Twin Births as a Natural Experiment." *Journal of Human Resources*, 34, 449–474.
Jensen, Robert (2010). "The (Perceived) Returns to Education and the Demand for Schooling." *Quarterly Journal of Economics*, 125, 515–548.
Jimerson, Shane R. (1999). "On the Failure of Failure: Examining the Association Between Early Grade Retention and Education and Employment Outcomes During Late Adolescence." *Journal of School Psychology*, 37, 243–272.
Jimerson, Shane R. (2001). "Meta-Analysis of Grade Retention Research: Implications for Practice in the 21st Century." *School Psychology Review*, 30, 313–330.
Jimerson, Shane R., Phillip Ferguson, Angela D. Whipple, Gabrielle E. Anderson, and Michael J. Dalton (2002). "Exploring the Association Between Grade Retention and Dropout: A Longitudinal Study Examining Socio-Emotional, Behavioral, and Achievement Characteristics of Retained Students." *The California School Psychologist*, 7, 51–62.
Kahn, Robert S., Barry Zuckerman, Howard Bauchner, Charles J. Homer, and Paul H. Wise (2002). "Women's Health After Pregnancy and Child Outcomes at Age 3 Years: A Prospective Cohort Study." *American Journal of Public Health*, 92, 1312–1318.
Kitagawa, Toru (2015). "A Test for Instrument Validity." *Econometrica*, 83, 2043–2063.

Kulkarni, Aniket D., Denise J. Jamieson, Howard W. Jr. Jones, Dmitry M. Kissin, Maria F. Gallo, Maurizio Macaluso, and Eli Y. Adashi (2013). "Fertility Treatments and Multiple Births in the United States." *New England Journal of Medicine*, 369, 2218–2225.

Lavy, Victor, M. Daniele Paserman, and Analia Schlosser (2012). "Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom." *Economic Journal*, 122, 208–237.

Lee, Jungmin (2008). "Sibling Size and Investment in Children's Education: An Asian Instrument." *Journal of Population Economics*, 21, 855–875.

Lleras-Muney, Andrea and Frank Lichtenberg (2005). "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, 671–696.

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

Manski, Charles F. (1989). "Anatomy of the Selection Problem." *Journal of Human Resources*, 24, 343–360.

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

Mazumder, Bhashkar and Zach Seeskin (2015). "Breakfast Skipping, Extreme Commutes and the Sex Composition at Birth." *Biodemography and Social Biology*, 61, 187–208.

Moav, Omer (2005). "Cheap Children and the Persistence of Poverty." *Economic Journal*, 115, 88–110.

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

Mogstad, Magne and Matthew Wiswall (2012). "Instrumental Variables Estimation with Partially Missing Instruments." *Economics Letters*, 114, 186–189.

Nevo, Aviv and Adam M. Rosen (2012). "Identification with Imperfect Instruments." *Review of Economics and Statistics*, 94, 659–671.

OECD (2002). *Taxing Wages: 2001 Edition*. Organization for Economic Cooperation and Development, Paris.

Ponczek, Vladimir and André Portela Souza (2012). "New Evidence of the Causal Effect of Family Size on Child Quality in a Developing Country." *Journal of Human Resources*, 47, 64–106.

Qian, Nancy (2009). "Quantity-Quality and the One Child Policy: The Only-Child Disadvantage in School Enrollment in Rural China." NBER Working Paper 14973, National Bureau of Economic Research, Inc., Cambridge, MA.

Rosenzweig, Mark R. and Kenneth I. Wolpin (1980a). "Life-Cycle Labor Supply and Fertility: Causal Inferences from Household Models." *Journal of Political Economy*, 88, 328–348.

Rosenzweig, Mark R. and Kenneth I. Wolpin (1980b). "Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment." *Econometrica*, 48, 227–40.

Rosenzweig, Mark R. and Kenneth I. Wolpin (2000). "Natural "Natural Experiments" in Economics." *Journal of Economic Literature*, 38, 827–874.

Rosenzweig, Mark R. and Junsen Zhang (2009). "Do Population Control Policies Induce More Human Capital Investment? Twins, Birth Weight and China's One-Child Policy." *Review of Economic Studies*, 76, 1149–1174.

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

Schultz, T. Paul (2004). "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics*, 74, 199–250.

Uggla, Caroline and Ruth Mace (2016). "Parental Investment in Child Health in Sub-Saharan Africa: A Cross-national Study of Health-seeking Behaviour." *Royal Society Open Science*, 3, 150460.

UNESCO (2005). "Education for All Global Monitoring Report 2006: Literacy for Life." United Nations Educational, Scientific and Cultural Organization, Paris.

UNESCO (2011). "Education for All Global Monitoring Report 2011: The Hidden Crisis: Armed Conflict and Education." United Nations Educational, Scientific and Cultural Organization, Paris.

United Nations (2010). "World Population Policies 2009." Tech. rep., Department of Economic and Social Affairs: Population Division.
Vere, James (2011). "Fertility and Parents' Labour Supply: New Evidence from US Census Data." *Oxford Economic Papers*, 63, 211–231.
Warren, John Robert, Emily Hoffman, and Megan Andrew (2014). "Patterns and Trends in Grade Retention Rates in the United States, 1995–2010." *Educational Researcher*, 43, 433–443.
Willis, Robert J. (1973). "A New Approach to the Economic Theory of Fertility Behavior." *Journal of Political Economy*, 81, S14–S64.
Young, Alwyn (2018). "Consistency without Inference: Instrumental Variables in Practical Application." Working paper, London School of Economics.

Supplementary Data

Supplementary data are available at *JEEA* online.