THE ECONOMIC JOURNAL

The Economic Journal, **131** (*January*), 1–32 DOI: 10.1093/ej/ueaa100
Advance Access Publication Date: 25 August 2020

THE EFFECT OF FERTILITY ON MOTHERS' LABOUR SUPPLY OVER THE LAST TWO CENTURIES*

Daniel Aaronson, Rajeev Dehejia, Andrew Jordan, Cristian Pop-Eleches, Cyrus Samii and Karl Schulze

Using a compiled data set of 441 censuses and surveys from between 1787 and 2015, representing 103 countries and 51.4 million mothers, we find that: (*i*) the effect of fertility on labour supply is typically indistinguishable from zero at low levels of development and large and negative at higher levels of development, (*ii*) the negative gradient is stable across historical and contemporary data, and (*iii*) the results are robust to identification strategies, model specification, and data construction and scaling. Our results are consistent with changes in the sectoral and occupational structure of female jobs and a standard labour–leisure model.

The relationship between fertility and female labour supply is widely studied in economics. For example, the link between family size and mothers' work decisions has helped explain household time allocation and the evolution of women's labour supply, particularly among rapidly growing countries in the second half of the twentieth century (e.g., Angrist and Evans, 1998; Cristia, 2008). Development economists relate the fertility–work relationship to the demographic transition and study its implications on economic growth (Bloom *et al.*, 2001). Yet despite the centrality of these issues in the social sciences, the existing evidence is fragmentary and, as we discuss below, seemingly contradictory.

Our contribution is to provide unified evidence on whether the relationship between fertility and labour supply has evolved over time and with the process of economic development. Using data spanning not only a broad cross-section of countries at various stages of development but historical examples from currently developed countries dating back to the late eighteenth century, we show a strikingly consistent albeit evolving relationship between fertility and mothers' labour supply. To provide consistent estimates over time and space, we use two common instrumental variables strategies: (*i*) twin births introduced by Rosenzweig and Wolpin (1980), and (*ii*) the gender composition of the first two children (Angrist and Evans, 1998). We implement these estimators using four large databases of censuses and surveys: the International Integrated Public Use Micro Sample (IPUMS), the US IPUMS, the North Atlantic Population Project, and the Demographic and Health Surveys. Together, the data cover 441 country–years, and 51.4 million mothers, stretching from 1787 to 2015 and, consequently, a large span of economic development.

A natural starting point in thinking about the fertility–labour supply relationship is Angrist and Evans (1998). Based on US IPUMS data from 1980 and 1990, Angrist and Evans document

* Corresponding author: Rajeev Dehejia, Wagner School of Public Service, New York University, 295 Lafayette Street, 2nd Floor, New York, NY 10012, USA. Email: rajeev@dehejia.net

This paper was received on 1 July 2019 and accepted on 30 July 2020. The Editor was Barbara Petrongolo.

The data and codes for this paper are available on the Journal website. They were checked for their ability to reproduce the results presented in the paper.

We thank seminar participants at the NBER Summer Institute, SOLE, Williams, Cornell, the University of Washington, the University of Illinois Chicago, the University of Connecticut, and GREQAM and especially Quamrul Ashraf and Leah Boustan for invaluable suggestions. The views expressed in this article are not necessarily those of the Federal Reserve Bank of Chicago or the Federal Reserve System.

a negative effect of fertility on female labour supply using both gender mix and twin births as instruments for subsequent children, a result also established by Bronars and Grogger (1994).[1] Alternative instruments that rely on childless mothers undergoing infertility treatments in the USA and Denmark (Cristia, 2008; Lundborg *et al.*, 2017) or natural experiments like the introduction of birth control pills (Bailey, 2013) or changes in abortion legislation (Bloom *et al.*, 2009) similarly conclude that children have a negative effect on their mother's labour supply or earnings. That the results are consistent across instruments is notable since each IV uses a somewhat different subpopulation of compliers to estimate a local average treatment effect (LATE), and therefore is suggestive of wide external validity (see also Dehejia *et al*., 2020).

However, we show that the negative relationship between fertility and mother's work behaviour holds only for countries at a later stage of economic development. At a lower level of income, including the USA and Western European countries prior to WWII, there is no causal relationship between fertility and mothers' labour supply. The lack of a negative impact at low levels of development aligns with Aguero and Marks' (2008; 2011) studies of childless mothers undergoing infertility treatments in 32 developing countries, Godefroy's (2017) analysis of changes to women's legal rights in Nigeria, and Heath (2017) who finds an economically small effect of fertility on women working using non-experimental evidence from urban Ghana. Strikingly, combining US historical censuses with data from a broad set of contemporary developing countries, we find that the negative gradient of the fertility–labour supply effect with respect to economic development is remarkably consistent across time and space. That is, women in the USA at the turn of the twentieth century make the same labour supply decision in response to additional children as women in developing countries today. We show that the negative gradient is robust to a wide range of data, sampling, and specification issues, including alternative instruments, development benchmarks, sample specification criteria, conditioning covariates including those highlighted by Bhalotra and Clarke (2016), additional measures of mother's labour supply, and a variety of other adjustments to make our data historically consistent.

That said, our main results come with important qualifications, some of which we can address with additional assumptions or subsets of data and some of which we cannot. First, there are significant measurement concerns about female LFP in historical and modern developing country data. As we explain in detail below, our results are robust to excluding historical data and to using developing country samples where female labour participation is externally validated by the International Labour Organization (ILO), the most reliable outside source. Second, the complier population varies from developed to developing countries and over the two-hundred year span of our data, as does the base rate of women's LFP. We can address this heterogeneity, in part, by weighting our results to a constant complier covariate profile or scaling by the complier outcome mean. Our results are robust to both methods, although each comes with assumptions. Third, exact dates of birth and complete birth histories are available only for a subset of our data. We show that our results are similar in this subset of the data. Fourth, our main results are based primarily on LFP rather than the intensive margin of hours worked; we present results on hours below, although they are based on much more limited samples.

There are two important issues our data do not allow us to consider. First, by construction, the twins and same gender instruments cannot be applied to the birth of first children. Indeed, we are only aware of two research strategies that focus on the effects of first children. The first

[1] Bhalotra and Clarke (2016) and Clarke (2018) provide useful summaries of the validity of various fertility instruments and the broader empirical literature.

uses longitudinal data in event studies of first birth (e.g., Angelov *et al.*, 2016; Kleven *et al.*, 2019a). These studies find large negative labour supply effects in several developed countries, though this strategy has not, to our knowledge, been applied in a developing country. The second approach to first births relies on the random success of *in vitro* fertilisation (IVF), which is not classified in any of our data sets. That said, the contrast between Aguero and Marks' (2008; 2011) IVF-based finding of a zero effect in developing countries and Cristia's (2008) and Lundborg *et al.*'s (2017) large negative effect in developed countries is tellingly consistent with the patterns in our data. Moreover, we show a similar pattern, albeit with a monotonically declining magnitude, across all family size parities beyond one child, at least suggestive that the negative gradient is a general result. Second, our data are cross-sectional and therefore only allow identification of the short-run effect of fertility. As noted in Adda *et al.* (2017) among others, the life-cycle response is often attenuated compared to the short-run effect, and late-in-life (rather than early) shocks are more likely to have lasting impacts on fertility.

The empirical regularities we describe are consistent with a standard labour–leisure model augmented to include a desire for children. As wages increase during the process of development, households face an increased time cost of fertility but also experience increased income. With a standard constant elasticity of substitution utility function, the former effect dominates as countries develop, creating a negative gradient (Online Appendix A provides a sketch of the model).

Indeed, in exploring the mechanism behind our result, we document that the substitution effect falls from zero to negative and is economically important as real GDP per capita increases. We argue that the declining substitution effect arises from changes in the sectoral and occupational structure of female jobs, as in Schultz (1991) and Goldin (1995). As economies evolve, women's labour market opportunities transition from agricultural and self-employment to urban wage work. The latter tends to be less compatible with raising children and causes some movement out of the labour force. In support of this channel, we show that the negative gradient is steeper among mothers with young children that work in non-professional and non-agricultural wage-earning occupations (e.g., urban wage work). Moreover, a growing literature documents a causal relationship between access to childcare or early education and the propensity of mothers to work (e.g., Baker *et al.*, 2008; Havnes and Mogstad, 2011), a finding that is consistent with leaving the workforce when labour market opportunities become less compatible with child rearing. We cannot rule out that the income effect from rising wages could also be playing a role in the negative gradient but the evidence is at best mixed. Other explanations, most notably the widespread adoption of modern contraceptives and shifting social norms about female work (Goldin, 1977; Boustan and Collins, 2014) could also be compatible with our results. While we can find little evidence consistent with these alternative mechanisms, our data do not allow us to rule them out.

Our main empirical findings have important implications both for understanding the historical evolution of women's labour supply and the relationship between the demographic transition and the process of economic development. As Goldin (1995) documents in her comprehensive study of women's work in the twentieth century, women's labour supply follows a U-shape over the process of economic growth, first declining before eventually increasing (see also Mammen and Paxson, 2000). Our results suggest that declining fertility may have contributed to the upswing in women's labour supply in much of the developed world during the second half of the century. Moreover, family policies (Olivetti and Petrongolo, 2017) and childcare costs (Del Boca, 2015) likely played a role. At the other end of the economic development spectrum,

our results suggest that the demographic transition to smaller families probably does not have immediate implications for women's labour supply and growth. This in turn reinforces a claim in the demographic transition literature (Bloom *et al.*, 2001) that family planning policies are unlikely to enhance growth through a labour supply channel, although such policies could still be desirable for other reasons.

The article is organised as follows. Section 1 explains the empirical strategy, followed in Section 2 by a description of the data. Section 3 presents our main findings. Section 4 discusses potential channels, and presents a series of robustness checks, for our results. Section 5 concludes.

1. Empirical Strategy

Our empirical analysis adopts the standard approach of exploiting twin births and gender composition as sources of exogenous variation in the number of children to identify the causal effect of an additional child on the labour force activity of women (Rosenzweig and Wolpin, 1980; Bronars and Grogger, 1994; Angrist and Evans, 1998; Black *et al.*, 2005; Caceres-Delpiano, 2006; Vere, 2011). In particular, for twin births, we consider a first stage regression of the form:

$$z_{ijt} = \gamma S_{ijt} + \boldsymbol{w}'_{ijt}\rho + \pi_{jt} + \mu_{ijt}, \tag{1}$$

where z_{ijt} is an indicator of whether mother i in country j at time t had a third child, the instrument S_{ijt} is an indicator for whether the second and third child are the same age (twins), $\boldsymbol{w}_{ijt}$ is a $k \times 1$ vector of demographic characteristics that typically include the current age of the mother, her age at first birth, and an indicator for the gender of the first child, and π_{jt} are country–year fixed effects. γ measures the empirical proportion of mothers with at least two children who would not have had a third child in the absence of a multiple second birth.

LATE among mothers with multiple children is identified from a second stage regression:

$$y_{ijt} = \beta z_{ijt} + \boldsymbol{w}'_{ijt}\alpha + \theta_{jt} + \varepsilon_{ijt}, \tag{2}$$

where y_{ijt} is a measure of labour supply for mother i in country j at time t and β is the IV estimate of the pooled labour supply response to the birth of twins for women with at least one prior child.[2] Our baseline twin estimates condition on one child prior to the singleton or twin so that all mothers have at least two children, as in Angrist and Evans (1998). This restriction provides a family-size-consistent comparison so that both the same-gender and twins IV study the effect of a family growing from two to three children.

While twins are a widely-used source of variation for studying childbearing on mothers' labour supply, it is by no means the only strategy in the literature. Perhaps the leading alternative exploits preferences for mixed gender families (Angrist and Evans, 1998). Angrist and Evans estimate a first-stage regression like equation (1) but, for S_{ijt}, substitute twin births for an indicator of whether the first two children of woman i are of the same gender (boy–boy or girl–girl). Again, the sample is restricted to women with at least two children and γ measures the likelihood that a mother with two same gendered children is likely to have additional children relative to a mother with a boy and a girl.

[2] We also aggregate the results in a procedure that is analogous to a hierarchical Bayesian model with a flat prior. To identify the gradient, we use a local polynomial smoother with a bandwidth of $1,500, where each country–year point estimate is weighted by its precision. That has no impact on our inferences.

Both twins and same gender children have been criticised as valid instruments on the grounds of omitted variables biases. Twin births may be more likely among healthier and wealthier mothers and can consequently vary over time and across geographic location (Rosenzweig and Wolpin, 2000; Hoekstra *et al.*, 2007; Bhalotra and Clarke, 2016; Clarke, 2018). Rosenzweig and Zhang (2009) also argue that twin siblings may be cheaper to raise, leading to a violation of the exclusion restriction. While the same gender instrument has proven quite robust for the USA and other developed countries (Butikofer, 2011), there are many reasons to be cautious in samples of developing countries (Schultz, 2008). Among other factors, households may practice either sex selection or selective neglect of children based on gender (Ebenstein, 2010; Jayachandran and Pande, 2017).

We adopt the broad view of Angrist *et al.* (2010) that the sources of variation used in various IV strategies are different and, therefore, so are the biases. As such, each IV provides a specification check of the other. Besides the basic LATE estimates underlying the multiple instrument methodology of Angrist *et al.* (2010), we also report: (*i*) a third instrument introduced by Klemp and Weisdorf (2019), which relies on exogenous variation in the timing of first births, (*ii*) twin results at alternative family parities, (*iii*) estimates that control for education and health measures to the greatest extent possible, including height and body mass index that have been highlighted as key determinants of twin births (Bhalotra and Clarke, 2016), and (*iv*) estimates by same gender versus mixed gender twins.[3] All these specification checks (see Online Appendix B for details) are consistent with a declining labour supply gradient over development when they can be implemented across the GDP distribution.

The literature analyses a number of measures of y_{ijt}, including whether the mother worked, the number of hours worked, and the labour income earned. These measures are sometimes defined over the previous year or at the time of the survey. In order to include as wide a variety of consistent data across time and countries as possible, we typically focus on the labour force participation (LFP) of mothers at the time of a census or survey. When LFP is unavailable, especially in pre-WWII censuses, we derive LFP based on whether the woman has a stated occupation. Online Appendix B discusses the robustness of the results to several alternative labour market measures, including mismeasurement of occupation-based LFP (Goldin, 1990).

In concordance with much of the literature (especially Angrist and Evans, 1998), our standard sample contains women aged 21 to 35 with at least two children, all of whom are 17 or younger. We exclude families where a child's age or gender or mother's age is imputed. We also drop mothers who gave birth before age 15, who live in group quarters, or whose first child is a multiple birth. It is worth emphasising that the restrictions on mother's (21–35) and child's (under 18) age may allay concerns about miscounting children that have moved out of the household.[4] We also experiment with even younger mother and child age cut-offs, which additionally provide some inference about difference in the labour supply response to younger and older offspring. Further sample statistics, single sample estimates, as well as results when these restrictions are relaxed, are provided in the Online Appendix tables.

We present our results stratified by time, country, level of development, or some combination. The prototypical plot stratifies countries-years into seven real GDP per capita bins (in 1990 US

[3] Monozygotic (MZ) twinning is believed to be less susceptible to environmental factors. Hoekstra *et al.* (2007) provides an excellent survey of the medical literature. Since we cannot identify MZ versus dizygotic (DZ) twins in our data, we take advantage of the fact that MZ twins are always the same gender, whereas DZ twins share genes like other non-twin siblings and therefore are 50% likely to be the same gender.

[4] As a robustness check, we also use information about complete fertility when it is available.

dollars): under $2,500, $2,500–5,000, $5,000–7,500, $7,500–10,000, $10,000–15,000, $15,000–20,000, and over $20,000. To be concrete, in this example, all country–years where real GDP per capita are, say, under $2,500 in 1990 US dollars are pooled together for the purpose of estimating equations (1) and (2). Similarly, countries with real GDP per capita between $2,500 and $5,000 are also pooled together for estimation, and so on. The plots report weighted estimates of γ and β, and their associated 95% confidence interval based on country–year clustered standard errors, for each bin.[5]

2. Data

We estimate the statistical model using four large databases of country censuses and surveys.

2.1. *US Census, 1860–2010*

The USA is the only country for which historical microdata over a long stretch of time is *regularly* available. We use the 1% samples from the 1860, 1870, 1950, and 1970 censuses; the 5% samples from the 1960, 1980, 1990, and 2000 censuses; the 2010 American Community Survey (ACS) five-year sample, which combines the 1% ACS samples for 2008 to 2012; and the 100% population counts from the 1880, 1900, 1910, 1920, 1930, and 1940 censuses.[6] Besides additional precision, the full count censuses allow us to stratify the sample (e.g., by states) to potentially take advantage of more detailed cross-sectional variation.

IPUMS harmonises the US census samples to provide comparable definitions of variables over time. However, there are unavoidable changes to some of our key measures. For example, the 1940 census is the first to introduce years of completed schooling and earnings; therefore, when we show results invoking education or earnings, we exclude US data prior to 1940. Perhaps most important, the 1940 census shifted our labour supply measure from an indicator of reporting any 'gainful occupation' to the modern labour force definition of working or looking for work in a specific reference week. Fortunately, there does not appear to be a measurable difference in our results between these definitions in 1940 when both measures are available. Nevertheless, there is concern that women's occupations (Goldin, 1990) as well as fertility (Moehling, 2002) could be systemically under- or over-reported, especially in US census samples for 1910 and earlier. We present a number of robustness checks meant to isolate these mismeasurement issues in Online Appendix B, and in Subsection 3.5 we present results that exclude historical data.

For Puerto Rico, we use the 5% census samples from 1980, 1990, and 2000 and the 2010 Community Survey, which combines the 1% samples for 2008 to 2012. Censuses prior to 1980 are missing labour force data or reliable information about real GDP per capita.

[5] Household weights are supplied by the various surveys or censuses, normalised by the number of mothers in the final regression sample.

[6] For information on the IPUMS samples, see Steven Ruggles, J. Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek, *Integrated Public Use Microdata Series: Version 5.0* [Machine-readable database], Minneapolis: University of Minnesota, 2010. The 100% counts were generously provided to us by the University of Minnesota Population Center via the data collection efforts of ancestry.com. Those files have been cleaned and harmonised by IPUMS. The 1890 US census is unavailable and US censuses prior to 1860 do not contain labour force information for women. In some figures, we also report single-year estimates from the 1880 10%, 1900 and 1930 5%, as well as the 1910, 1920, and 1940 1% random IPUMS samples.

2.2. *IPUMS International Censuses, 1960–2015*

IPUMS harmonises censuses from around the world, yielding measures of our key variables that are roughly comparable across countries and time. We use data from 212 of the 301 non-US country–year censuses between 1960 and 2015 that were posted at the IPUMS-I website as of May 2017. Censuses are excluded if mother–child links or labour force status is unavailable (83 censuses) or age is defined by ranges rather than single-years (6 censuses).[7]

2.3. *North Atlantic Population Project (NAPP), 1787–1911*

The North Atlantic Population Project (NAPP) provides 18 censuses from Canada, Denmark, Germany, Great Britain, Norway, and Sweden between 1787 and 1911. As with IPUMS, these data are made available by the Minnesota Population Center.[8] For most samples, NAPP generates family interrelationship linkages. However, in a few cases (Canada for 1871 and 1881 and Germany in 1819) such linkages are not available. In those cases, we use similar rules developed to link mothers and children in the US full count census. Also, consistent with the pre-1940 US censuses, labour force activity is based on whether women report an occupation rather than the modern definition of working or seeking work within a specific reference period, and education is unavailable.

2.4. *Demographic and Health Surveys (DHS), 1990–2014*

We supplement the censuses with the Demographic and Health Surveys (DHS).[9] From the initial set of 254 country–year surveys, spanning six waves from the mid-1980s onward, we exclude samples missing age of mother, marital status of mother, current work status, whether the mother works for cash, birth history, and comparable real GDP per capita. These restrictions force us to drop the first wave of the DHS, leaving 692,923 mothers in 192 country–years.

The DHS includes a number of questions that are especially valuable for testing the robustness of our census results. First, detailed health information allows us to control for characteristics that may be related to a mother's likelihood of twinning (Bhalotra and Clarke, 2016). Second, we can use an indicator of whether children are in fact twins to test the accuracy of our coding of census twins.[10] To keep the DHS results comparable to the censuses, our baseline DHS estimates

[7] Similar to the USA, the international linking variables use relationships, age, marital status, fertility, and proximity in the household to create mother–child links. Sobek and Kennedy (2009) compute that these linking variables have a 98% match rate with direct reports of family relationships. However, we are not able to compute linkages that do not include relevant household information on relationship and surname similarity. Unfortunately, this affects some censuses from Canada and the UK. Although the 1971 to 2006 Irish censuses use age ranges for adults, they do not for children under 20 (so we literally include Irish twins!).

[8] See Minnesota Population Center (2015), North Atlantic Population Project: Complete Count Microdata, Version 2.2 [Machine-readable database], Minneapolis: Minnesota Population Center.

[9] For additional information about the DHS files see ICF International (2015). The data is based on extracts from DHS Individual Recode files. See http://dhsprogram.com/Data/.

[10] Online Appendix Figure A1 illustrates the high degree of correspondence between twinning rates when we define twins using 'real' multiple births and those imputed for children sharing the same birth-year. The DHS has a number of labour force variables but none that directly compare to those in the censuses. We chose to use an indicator of whether the mother is currently working since it is most correlated with the IPUMS labour force measures (see Online Appendix Figure A2).

identify twins based on the census year-of-birth criterion and consider only living children who reside with the mother.

2.5. *Real GDP per Capita*

Real GDP per capita (in US$1,990) is collected from the Maddison Project.[11] To reduce measurement error, we smooth each GDP series by a seven year moving average centred on the survey year. We are able to match 441 country–years to the Maddison data, leaving 51,449,770 mothers aged 21 to 35 with at least two children in our baseline sample.[12]

When we split the 1930 and 1940 full population US censuses into the 48 states and DC, we bin those samples by state-specific 1929 or 1940 income-per-capita.[13] The income data are converted into 1990 dollars using the Consumer Price Index.

2.6. *Summary Statistics*

Table 1 provides summary statistics separately for the US and non-US samples and by real GDP per capita bins. Although the first bin (less than $2,500 GDP per capita) is dominated by DHS samples, most bins have a large number of mothers for both US and non-US samples. Online Appendix Table A1 provides additional descriptive statistics and estimates by individual country–year data sets.

3. Results

3.1. *OLS Estimates*

We begin with estimates from OLS regressions of the labour supply indicator on the indicator for a third child and the controls described above. These results do not have a clear causal interpretation, but they are useful for establishing key data patterns. In Figure 1, we plot the coefficients for the USA, the non-US countries, and the combined world sample (labelled 'All'), binned into the seven ranges of real GDP per capita reported on the x-axis ($0–2,500, $2,500–5,000, etc.). Point estimates and country–year clustered standard errors are provided in Table 2. The three samples exhibit a similar pattern. At low levels of real GDP per capita, the OLS estimate of the effect of children on mother's labour supply is negative and statistically significant at the 5% level but economically small in magnitude (e.g., −0.022 (0.005) in the lowest GDP bin). As real GDP per capita increases, the effect becomes more negative, ultimately flattening out between −0.15 and −0.25 beyond real GDP per capita of $15,000.

Figure 2 plots the US-only OLS results over time. Circles represent IPUMS samples and diamonds represent full population counts. These estimates start out negative, albeit relatively

[11] See The Maddison-Project (2013).

[12] In a few minor cases, we were not able to match a country to a specific year but still left the census in our sample because we did not believe it would have impacted their placement in a real GDP per capita bin. Specifically, the censuses of Denmark in 1787 and 1801 are matched to real GDP per capita data for Denmark in 1820 and Norway in 1801 is matched to data for Norway in 1820. Excluding these country–years has no impact on our results. More importantly, the Maddison data ends in 2010 and therefore censuses or surveys thereafter are assigned their most recently available real GDP per capita data.

[13] http://www2.census.gov/library/publications/1975/compendia/hist_stats_colonial-1970/hist_stats_colonial-1970p1-chF.pdf.

Table 1. *Sample Summary Statistics by Real GDP/Capita Bin.*

	Mothers	Samples	In labor force	3 or more children	2nd child is multiple birth	First 2 children are same gender	Children in household	Mother's age at survey	Mother's age at first birth
USA									
0−2,500	32,531	2	5.12%	62.47%	0.74%	49.48%	3.27	29.02	21.04
2,500−5,000	5,530,793	2	6.30%	62.47%	0.89%	50.27%	3.28	29.10	21.06
5,000−7,500	12,899,725	3	8.68%	55.75%	0.81%	50.37%	3.10	29.29	21.15
7,500−10,000	4,724,927	2	10.68%	47.07%	0.87%	50.50%	2.88	29.48	20.94
10,000−15,000	470,378	1	22.85%	55.09%	1.70%	50.38%	2.99	29.30	21.40
15,000−20,000	692,165	2	44.95%	40.85%	1.31%	50.48%	2.62	29.62	21.03
20,000−35,000	1,312,550	3	62.90%	36.64%	1.46%	50.58%	2.50	30.28	21.85
Non-USA									
0−2,500	9,676,791	213	43.33%	57.20%	1.28%	50.22%	3.06	29.07	20.66
2,500−5,000	7,617,815	103	36.14%	50.66%	1.05%	50.34%	2.96	29.82	21.19
5,000−7,500	4,192,823	52	36.77%	45.95%	1.22%	50.39%	2.77	29.43	20.46
7,500−10,000	2,184,583	20	34.95%	43.88%	1.25%	50.65%	2.69	29.54	20.66
10,000−15,000	614,503	19	37.90%	36.34%	1.19%	50.57%	2.61	29.99	21.63
15,000−20,000	415,161	10	56.06%	30.65%	1.19%	50.53%	2.41	30.73	22.61
20,000−35,000	1,085,025	9	73.66%	28.99%	1.44%	50.58%	2.38	31.23	24.00

Notes: This table displays summary statistics for the baseline sample of mothers by real GDP/capita bins. The sample consists of all two-child mothers aged 21 to 35 that were at least 15 when they had their first child, their oldest child is younger than 18, they do not live in group quarters, their first child is not a multiple birth, and mother and child have no imputations on age and gender. A twin is defined as the second and third birth being the same age. The samples directly correspond to those used in Table 2 and Figures 1, 3, and 6.

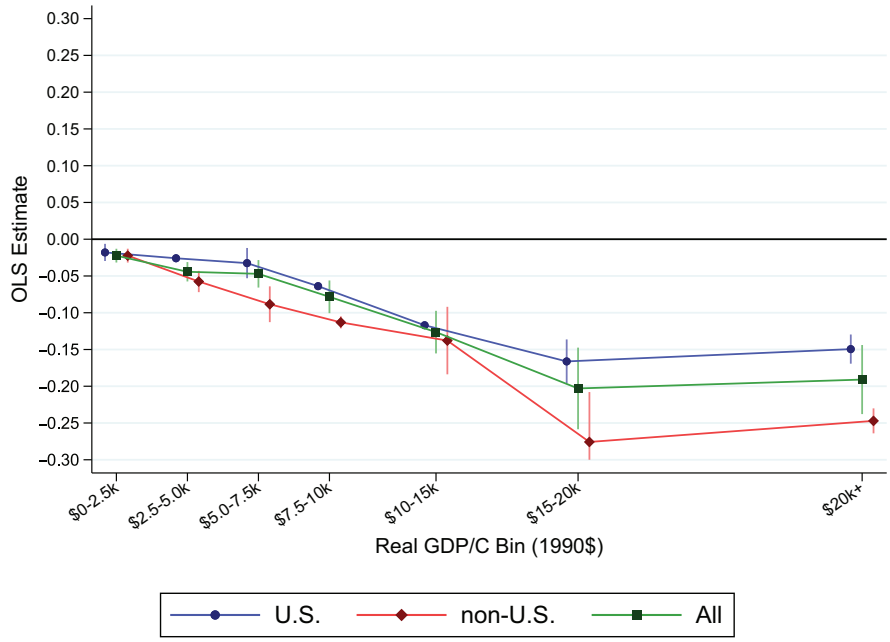

Fig. 1. *OLS, by Real GDP/Capita.*
Notes: This figure displays OLS estimates of the relationship between having a third birth and mothers' labor force participation using the baseline sample of mothers in each GDP/capita bin. Matching OLS estimates for US and non-US samples are reported in Table 2. Regressions control for mother's age, age at first birth, gender of first child, and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. Ninety-five per cent confidence intervals are displayed but may not always be visible at the scale of the figure.

small (e.g., −0.011 (0.004) in 1860 and −0.008 (0.0004) in 1910), decrease from 1910 to 1980, at which point the magnitude is −0.177 (0.001), and flatten thereafter.

Online Appendix Figure A3 plots the OLS estimates by real GDP per capita separately by time periods (pre-1900, 1900–1949, 1950–1989, and 1990+). Years prior to 1950 combine US census and NAPP data. Years thereafter include all four of our databases. The same general pattern appears *within* time periods.[14] The effect of fertility on labour supply tends to be small at low levels of GDP per capita but increases as GDP per capita rises.

3.2. *Twins IV*

Figure 3, Panel (a) shows the first-stage effect, γ in equation (1), of a twin birth on our fertility measure, the probability of having three or more children. For the US, non-US, and combined world samples, there is a positive and concave pattern, with the first-stage increasing with higher real GDP per capita up to $15,000 or so and flattening thereafter. Note that the regression specification controls for the mother's age, but does not, indeed cannot, control for the number of children or target fertility. Therefore, the positive gradient over real GDP per capita reflects

[14] Relative to Figure 1, we combined some real GDP per capita bins because of small sample sizes within these tight time windows.

Table 2. *Baseline Estimates by Real GDP/Capita Bin.*

	Mothers	Samples	LFP	OLS	Twin FS	Twin 2S	Same-Gender FS	Same-Gender 2S
USA: 0−2,500	32,531	2	5.12%	−0.018***	0.345***	0.119***	0.015*	−0.068
				(0.006)	(0.018)	(0.005)	(0.007)	(0.162)
USA: 2,500−5,000	5,530,793	2	6.30%	−0.026***	0.363***	0.022***	0.009***	0.064***
				(0.002)	(0.013)	(0.008)	(0.000)	(0.013)
USA: 5,000−7,500	12,899,725	3	8.68%	−0.033***	0.451***	0.012	0.014***	0.032***
				(0.010)	(0.017)	(0.010)	(0.002)	(0.004)
USA: 7,500−10,000	4,724,927	2	10.68%	−0.064***	0.540***	−0.017***	0.021***	0.072***
				(0.001)	(0.002)	(0.001)	(0.000)	(0.002)
USA: 10,000−15,000	470,378	1	22.85%	−0.117***	0.452***	−0.033***	0.035***	−0.084**
				(0.001)	(0.002)	(0.010)	(0.001)	(0.034)
USA: 15,000−20,000	692,165	2	44.95%	−0.166***	0.575***	−0.059***	0.049***	−0.121***
				(0.015)	(0.065)	(0.018)	(0.006)	(0.007)
USA: 20,000−35,000	1,312,550	3	62.90%	−0.149***	0.636***	−0.070***	0.049***	−0.121***
				(0.010)	(0.007)	(0.008)	(0.001)	(0.008)
Non-USA: 0−2,500	9,676,791	213	43.33%	−0.022***	0.411***	−0.005	0.028***	−0.046**
				(0.005)	(0.018)	(0.009)	(0.007)	(0.019)
Non-USA: 2,500−5,000	7,617,815	103	36.14%	−0.058***	0.473***	−0.014	0.030***	−0.018
				(0.007)	(0.036)	(0.011)	(0.007)	(0.012)
Non-USA: 5,000−7,500	4,192,823	52	36.77%	−0.088***	0.545***	−0.003	0.035***	−0.037***
				(0.012)	(0.020)	(0.015)	(0.002)	(0.013)
Non-USA: 7,500−10,000	2,184,583	20	34.95%	−0.113***	0.548***	−0.033***	0.032***	−0.001
				(0.004)	(0.023)	(0.011)	(0.001)	(0.029)
Non-USA: 10,000−15,000	614,503	19	37.90%	−0.138***	0.604***	−0.089***	0.035***	−0.061*
				(0.023)	(0.064)	(0.016)	(0.004)	(0.035)
Non-USA: 15,000−20,000	415,161	10	56.06%	−0.276***	0.719***	−0.127***	0.042***	−0.205***
				(0.035)	(0.038)	(0.036)	(0.002)	(0.020)
Non-USA: 20,000−35,000	1,085,025	9	73.66%	−0.247***	0.706***	−0.105***	0.038***	−0.173***
				(0.009)	(0.003)	(0.003)	(0.001)	(0.019)

Notes: This table displays OLS, same gender and twin first stage (FS) and second stage (2S) IV estimates of the effect of a third birth on mother's labor force participation using the baseline sample of mothers described in the text and Table 1. Regressions control for mother's age, age at first birth, gender of first child (and second child for same gender IV), and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. These estimates are plotted in Figures 1, 3, and 6. Statistical significant at the one, five, and ten percent levels are denoted by ***, **, and *.

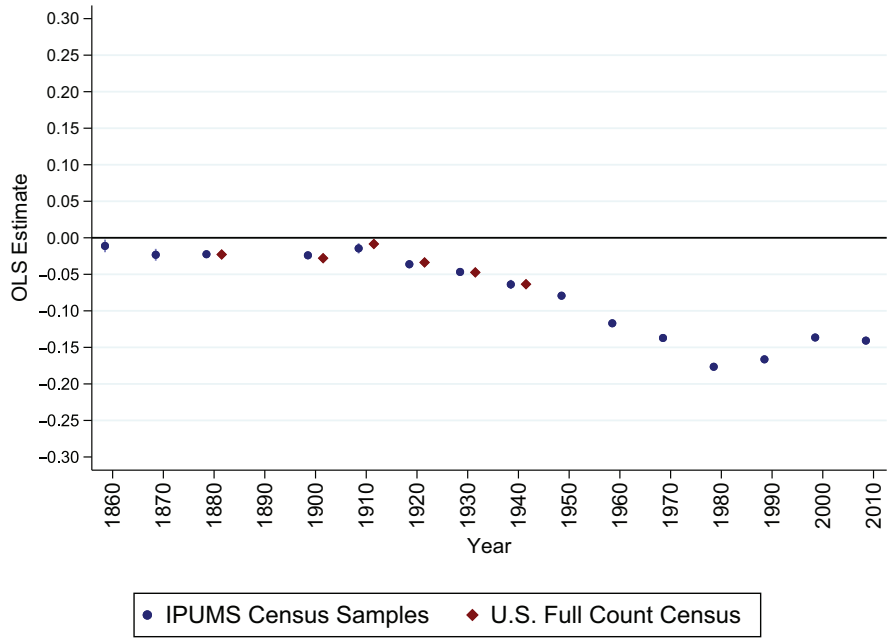

Fig. 2. *OLS, USA, by Time.*

Notes: This figure displays OLS estimates of the relationship between having a third birth and mothers' labor force participation, binned by census year. It uses the baseline sample of mothers for the USA only. Regressions control for mother's age, age at first birth, and gender of first child. Standard errors are robust to heteroskedasticity. Ninety-five per cent confidence intervals are displayed but may not always be visible at the scale of the figure. The estimates from this figure are reported in Online Appendix Table A1.

the negative impact of income on target fertility and hence the heightened impact of a twin birth on continued fertility relative to a non-twin birth.[15] In all cases, the instrument easily passes all standard statistical thresholds of first-stage relevance, including among countries with low real GDP per capita and high fertility rates.[16]

Figure 3, Panel (b) (and Table 2) plots β, the instrumental variables effect of fertility on mother's labour supply. In the world sample, β is mostly statistically indistinguishable from zero among countries with real GDP per capita of \$7,500 or less. Subsequently, β begins to decline and eventually flattens out between -0.05 and -0.10 at real GDP per capita of around \$15,000 and higher. The results for the US and non-US samples are similar in that there is a notable negative gradient with respect to real GDP per capita. For example, above \$20,000, the US estimate is -0.070 (0.008)[17] while the non-US estimate is -0.105 (0.003). The US (non-US) estimate implies that an extra child is associated with a decrease in a mother's labour supply of

[15] The first stage coefficient, γ, is $E\{z = 1|S = 1, w\} - E\{z = 1|S = 0, w\}$. Mechanically, $E\{z = 1|S = 1, w\} = 1$ because of the definition of twins. This means that if, for example, $\gamma = 0.6$, then $E\{z = 1|S = 0, w\} = 0.4$, implying that 40% of mothers would have a third child if their second child is a singleton. The increasing coefficient over real GDP per capita means having a third child after a singleton second child is declining with development. The reversal of this pattern at real GDP per capita of \$10–15,000 in the USA represents the Baby Boom.

[16] The smallest first stage *F*-statistic for the results displayed in Figure 3 is 170 for the non-US-only results for countries between \$2,500–5,000 GDP per capita.

[17] By comparison, Angrist and Evans (1998) report a twin IV estimate of -0.079 for the 1980 US census. Vere (2011) estimates twin IV coefficients for a third child of -0.086, -0.095, and -0.078 for 1980, 1990, and 2000, respectively.

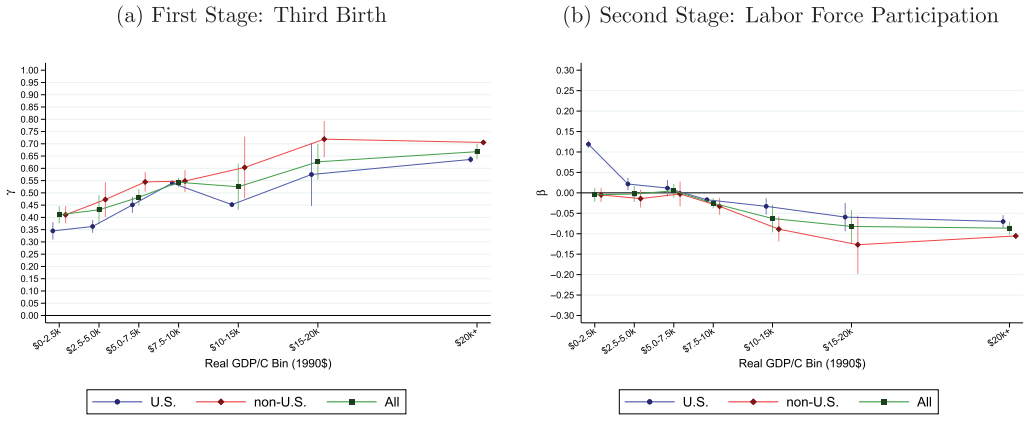

Fig. 3. *Twin IV, by Real GDP/Capita.*

Notes: This figure displays twin IV estimates using the baseline sample of mothers for each each real GDP/capita bin. Panel (a) shows the first-stage estimates of the relationship between twins and having a third birth. Panel (b) shows the second-stage estimates of the relationship between having a third birth and mothers' labor force participation. These estimates are also reported in Table 2. Regressions control for mother's age, age at first birth, gender of first child, and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. Ninety-five per cent confidence intervals based on robust standard errors clustered at the country–year level are displayed but may not always be visible at the scale of the figure.

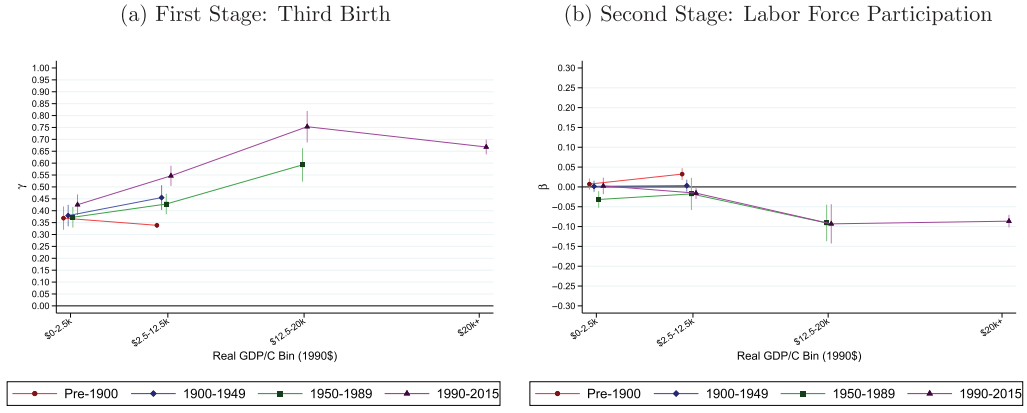

Fig. 4. *Twin IV, by Time and Real GDP/Capita Bin.*

Notes: This figure presents twin IV estimates stratified by year. Regressions control for mother's age, age at first birth, gender of first child, and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. Ninety-five per cent confidence intervals based on robust standard errors clustered at the country–year level are displayed but may not always be visible at the scale of the figure.

around 11 (14)%, relative to an average base rate of 62.9 (73.6) percentage points (e.g., $-0.070/0.629 = -0.111$).

In Figure 4, we show the results by time window. This gives us a sense of how much of the pattern we observe is due to differences in development instead of secular changes across time.

The central message of this figure is that the results are very consistent across time periods at similar levels of GDP per capita.[18] We think it is particularly notable that the declining β appears prior to the wide-spread availability of modern fertility treatments like IVF in wealthy countries and after modern census questions on LFP and fertility were introduced in 1940. We further address these potential issues below.

3.3. *Are There Positive Labour Supply Effects Among the Lowest Income Countries?*

One surprising finding is that at low real GDP per capita levels, we sometimes estimate a positive labour supply response to childbearing. This result is particularly evident in the pre-WWI USA (displayed in Online Appendix Figure A4), but also periodically appears, although not always statistically significantly so, for some low-income, post-1990 countries. The positive US results are not statistically different from zero for the early census samples (1860, 1870) but are for the full population counts of 1880 and 1910.

While these positive results are not artefacts in the statistical sense, it is worth noting that the underlying rates of LFP for US women are very low at this time in history (e.g., 6.2% and 10.0% for 1880 and 1910 mothers, respectively). As such, a positive effect could reflect that low-income mothers are more likely to work after having children, for example because subsistence food and shelter are necessary, whereas childcare might be cheaply available.

To gain further insight into the low real GDP sample results, we split the US's 1930 and 1940 full population counts by state of residence and pool states into income-per-capita estimation bins (matching what we did with countries in previous figures). Figure 5 shows the now familiar upward sloping pattern to the first stage results by real income per capita. In the second stage, we see that the effect of fertility on labour supply is in general statistically indistinguishable from zero at low-income levels in 1930 and 1940 and overlaps with the low-income post-1990 non-US results (shown in the line with squares). But we also find a small positive effect from the lowest income states in 1930, seemingly corroborating the positive estimates from a lower income USA prior to WWI.[19] These findings are directionally consistent with Godefroy (2017) and Heath (2017).

3.4. *Same Gender IV*

Next, we discuss results, displayed in Figure 6 and Table 2, which use the same gender instrument. Like the twins IV, we estimate a positive gradient to the first stage with respect to real GDP per capita, although the interpretation of this pattern is different than for twins. In particular, the same-gender first-stage picks up the increased probability that a mother opts to have more than two children based on the gender mix of her children (rather than picking up the proportion of mothers with incremental fertility when the twin instrument is zero, i.e., for non-twin births).[20]

[18] In Online Appendix Figures A4 and A7, we present US twin and same gender results by census decade. The pattern is broadly similar to the previous figure. The magnitude of the first stage is increasing over time, and the second-stage IV results begin to exhibit a pronounced negative gradient, particularly post-WWII. The same pattern arises within data sets (Online Appendix Figure A5 and Figure A8) and within geographic regions of the world (Online Appendix Figure A6 and Figure A9, although again with much noisier estimates for the same gender instrument).

[19] For the 1930 census, the states in that lowest bin ($2,000–3,000) are: Alabama, Arkansas, Georgia, Mississippi, North Carolina, North Dakota, New Mexico, South Carolina, and Tennessee.

[20] We find that the first stage of the same gender instrument is overall weaker than the twin instrument but passes the usual tests of relevance in binned samples, such as those in Figure 6. The one case with a weak first stage is the US estimates with GDP less than $2,500, which is based on the 1860 and 1870 US censuses. See Bisbee *et al.* (2017) for more details.

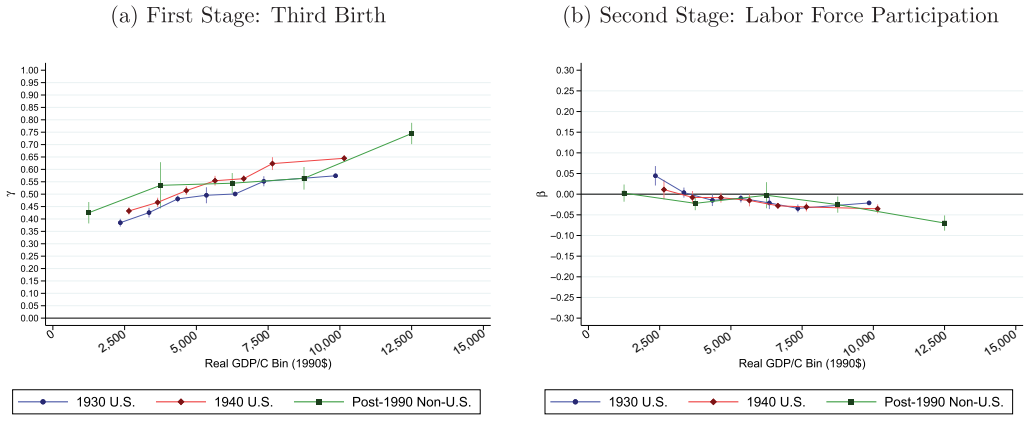

Fig. 5. *Twin IV by 1930 and 1940 US State Compared to Modern Non-US Countries.*
Notes: This figure displays twin IV estimates from the 1930 and 1940 full count censuses, binned by state real income per capita. For comparison, we also plot the post-1990 non-US estimates over the same real GDP/capita range. Income/capita for US states is taken from the US Census Bureau (see http://www2.census.gov/library/publications/1975/compendia/hist_stats_colonial-1970/hist_stats_colonial-1970p1-chF.pdf). Regressions control for mother's age, age at first birth, gender of first child, and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. Ninety-five per cent confidence intervals based on robust standard errors clustered at the country–year level are displayed but may not always be visible at the scale of the figure.

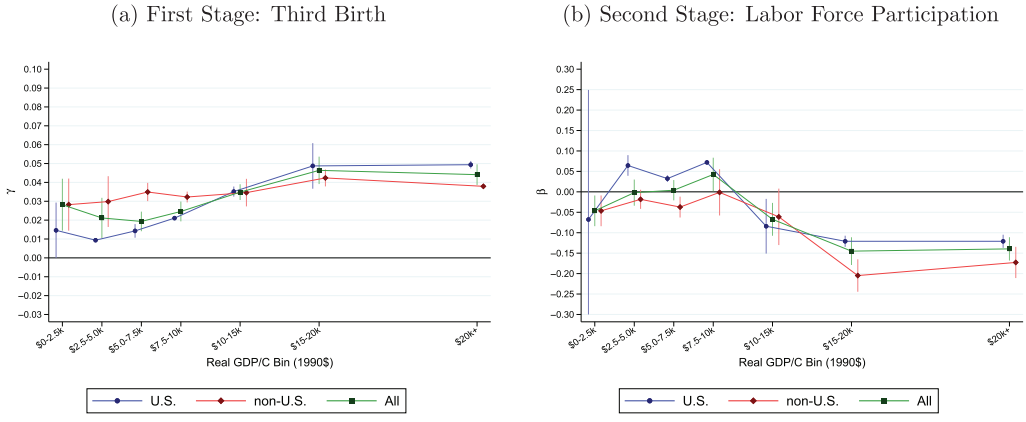

Fig. 6. *Same Gender IV, by Real GDP/Capita.*
Notes: This figure displays same gender IV estimates using the baseline sample of mothers for each each real GDP/capita bin. Analogous to Figure 3, Panel (a) shows the first-stage estimates of the relationship between same gender children and having a third birth and Panel (b) shows the second-stage estimates of the relationship between having a third birth and mothers' labor force status. These estimates are also reported in Table 2. Regressions control for mother's age, age at first birth, gender of first two children, and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. Ninety-five per cent confidence intervals based on robust standard errors clustered at the country–year level are displayed but may not always be visible at the scale of the figure.

Most importantly, we again see a negative gradient on the second stage IV estimates, from a close-to-zero effect among low GDP countries to a negative and statistically significant effect at higher real GDP per capita that flattens at around $15,000. As with the twins estimates, the negative estimates appear in the USA post-WWII (Online Appendix Figure A7).[21]

Our main intention is to highlight the similar shapes of the labour supply effect across the development cycle, despite using instruments that exploit different sources of variation. Indeed, when we combine all possible instrument variation into a singled pooled estimator, as in Angrist *et al.* (2010), our weighted average twin and same gender IV results also, unsurprisingly, shows the same strong negative gradient. That said, the magnitude of the same gender IV result is larger than the twin IV result at the high GDP per capita bins. For example, at the $20,000 and above bin, the twin estimate is −0.070 (0.008) for the US sample and −0.105 (0.003) for the non-US sample. By comparison, the same gender estimates are −0.121 (0.008) for the US sample and −0.173 (0.019) for the non-US sample. Since this is a LATE, this disparity suggests a greater effect of fertility on labour supply for those women encouraged to have an incremental child based either on son preference or the taste for a gender mix compared to those induced to higher fertility by a twin birth.

3.5. *Measurement Concerns with Female Labour Force Participation*

There are significant concerns with how female labour participation is measured in pre-1940 US censuses and modern developing country surveys and censuses, especially relative to measurement in modern developed country censuses. With regard to the historical USA, pre-1940 censuses use an occupation-based measure of LFP and introduce a number of miscodings highlighted in Goldin (1990). For developing countries, women's work may not be as clearly defined in informal settings, home production, and agriculture.

To address these concerns, Figure 7 compares our baseline estimates to results that exclude two sets of potentially mismeasured data. First, we throw out pre-1940 censuses and non-US pre-1950 data.[22] Second, we exclude IPUMS and DHS samples where our measure of female LFP fails to adequately match female LFP that was independently validated by the ILO (see International Labour Organization, 2019). We identify 177 countryyears where the ILO estimate of female LFP for 25 to 34 year-olds is within 4.8 percentage points (the median difference) of a comparable IPUMS or DHS estimate.[23]

The key patterns are the same as our baseline results: the negative effect of fertility on female LFP starts out small and becomes more negative over the process of development for both twin instrument (panel A) and the same gender instrument (panel B). One difference is that the two lowest GDP bins for the twins instrument (and the first and third bins for the same gender instrument) have statistically significant negative effects. However, the magnitude of the negative effect in the lowest GDP bins is small both relative to female LFP (56.9%) in these samples and

[21] Like the twins estimates, we also find systematic evidence of a positive fertility–labour supply effect at low levels of income, which are statistically significant for the 1900, 1930, and 1940 US censuses (see Online Appendix Figure A7).

[22] In Online Appendix B, we also partially recode the miscoded occupations following Goldin (1990). Those results are also similar to the baseline estimates.

[23] Since our surveys do not always align with ILO's periodicity, when necessary we extrapolate or linearly interpolate between ILO estimates (up to a maximum of four years) to obtain an estimate of female LFP in the IPUMS and DHS years. In the end, we are able to match 355 of our 441 country–years to the ILO, with all samples based on 1950 or later—that is, historical data is excluded. The 177 country–years that we use in this exercise represent the best half of the country–year matches.

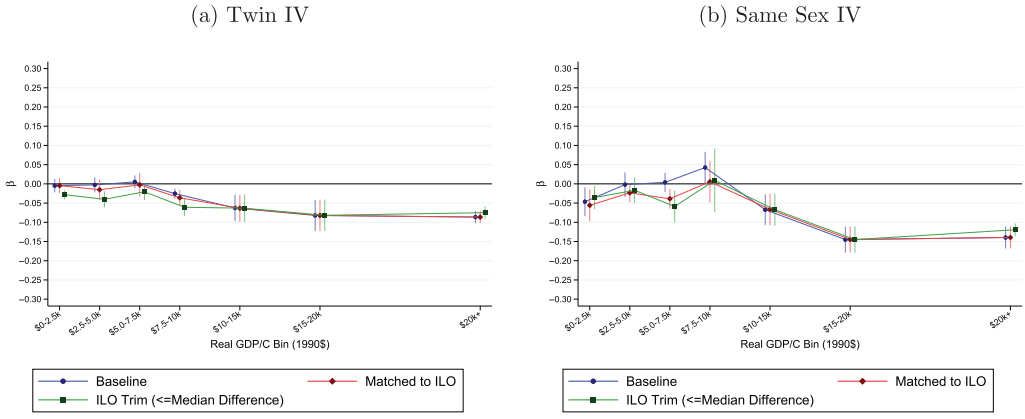

Fig. 7. *Second Stage Estimates Matched to ILO Statistics.*

Notes: This figure compares our baseline second-stage twin and same sex results to results that restrict to surveys that match well to ILO female labor supply statistics. Regressions control for mother's age, age at first birth, gender of first child, and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. Ninety-five per cent confidence intervals based on robust standard errors clustered at the country–year level are displayed but may not always be visible at the scale of the figure.

to the point estimates at higher levels of GDP. The results are similar when we retain the best third or best two-thirds of ILO matches. In the former case, the ILO LFP rates are nearly identical to the IPUMS/DHS LFP rates.

Finally, in principle, we would like to analyse the effect of fertility on hours worked and participation separately. However, this would require instruments for the intensive and extensive margins. Nonetheless, as an exploratory analysis, Online Appendix Figure A10 plots twin and same gender instrumental variables results for the number of hours worked per week where those out of the labour force are coded as working zero hours. We include all country–years that contain a measure of hours worked, which unfortunately limits us to only 56 censuses – eight from the USA (1940–2010) and 48 from the International IPUMS (the DHS and NAPP do not contain hours worked per week). We continue to find a negative gradient to labour supply, with the difference between hours worked among mothers in low-income and high-income countries being about 1.3 for the twins instrument and 4.3 hours for same gender instrument. As a benchmark, all mothers work, on average, just under 23 hours per week in countries with real GDP per capita above $20,000, suggesting a roughly 4 to 18% average decline in hours as a result of an additional child, conditional on working.

4. Channels

This section explores some of the potential mechanisms that account for the remarkably robust negative income gradient of mother's labour supply response to children.

4.1. *Accounting for a Changing Complier Participation*

A key challenge in interpreting our results is that the complier population is likely to change across our data. The group of women induced to have more than two children because of an

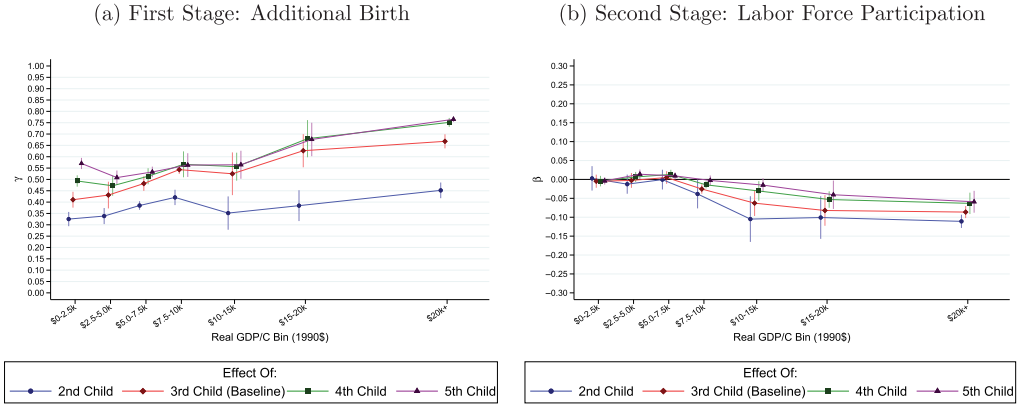

Fig. 8. *Twin IV Estimates at Different Family Sizes.*

Notes: This figure displays twin IV estimates by the size of the family when the twins were born. For example, the line labeled '2nd child' includes mothers with at least one child and where twins are the first and second child born. The line labeled '3rd child' is our baseline. Regressions control for mother's age, age at first birth, gender of first child, and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. Ninety-five per cent confidence intervals based on robust standard errors clustered at the country–year level are displayed but may not always be visible at the scale of the figure.

initial twin birth in a context where most women have more than three children (e.g., in a developing country or in historical data from developed countries) is presumably different from the women encouraged to have more than two children by a twin birth in a low fertility context. It is important to acknowledge that we cannot directly address this issue, at least without a stronger set of assumptions.

An indirect approach to capture variation from different sets of mothers is to condition on different family size parities (Angrist *et al.*, 2010). For example, one might expect that mothers with a large number of previous children would be less likely to adjust their labour supply in response to unexpected incremental fertility (for example, because of low incremental childcare costs for higher births). Indeed, as shown in Figure 8, we observe a stronger first stage effect for the sample that conditions on more children, especially at higher income levels. In the second stage, we see a notably, although not always statistically significantly, more negative effect in high-income countries for women starting with one child. The pattern of results is similar regardless of how many children are in the household when the twins are born. In all non-zero family size circumstances (up to three initial children), we continue to find no effect among low-income countries and an increasingly larger negative effect among higher income countries, flattening out around $20,000 per capita.[24] The continued robustness of the negative gradient to family parity suggests that the key patterns in our results may not solely be driven by changes in the complier population, although as noted above this is at best indirect evidence.[25]

[24] Additionally, we restrict the DHS sample to mothers whose report their ideal number of children as less than three (or four) and obtain nearly identical point estimates. This test loosely addresses concern that the parities we consider would not be binding and, consequently, have no labour supply effect in high-fertility, low-income countries.

[25] Unfortunately, by construction, the twin and same gender instruments are unable to identify the labour supply effect from an unexpected first child. Causal evidence on the impact of first births sometimes uses childless mothers undergoing IVF treatments. Interestingly, Cristia (2008) and Lundborg *et al.* (2017) find large negative labour supply responses to

More direct evidence requires stronger assumptions. Using the approach suggested by Angrist and Fernandez-Val (2013) (see Bisbee *et al.*, 2017 for a related application), we can adjust for changes in the complier population by reweighting our IV estimate to a constant complier profile. This adjustment assumes a constant treatment effect conditional on a covariate profile, in other words that heterogeneity in IV causal effects is driven by observable changes in complier characteristics. We use Abadie's (2003) kappa function to recover the covariate profile of compliers in a target year. We then compute covariate-specific IV treatments in other years, and reweight these to match the twin IV complier covariate profile in 1980's USA. Specifically, given a $k \times 1$ vector of covariates that have been de-meaned by the means of the target complier population, $\tilde{w}_{ijt}$, we augment the standard 2SLS framework by estimating the following second-stage equation and reporting estimates of β:[26]

$$y_{ijt} = \beta z_{ijt} + z_{ijt}\tilde{w}'_{ijt}\delta + \tilde{w}'_{ijt}\alpha + \theta_{jt} + \varepsilon_{ijt}. \tag{3}$$

The procedure in (3) involves estimating $k+1$ corresponding first-stage equations for $\{z_{ijt},\ z_{ijt}\tilde{w}'_{ijt}\}$ using $\{S_{ijt},\ S_{ijt}\tilde{w}'_{ijt}\}$ as instruments. Reweighting by age and education bins significantly impact the first stage at low levels of GDP per capita, but there are no significant changes in the IV estimates (see Figure 9).

Together, these results lead us to postulate—albeit with significant qualifications on the available evidence—that the key patterns in our results are not driven by changes in the complier population over the process of development.

4.2. *Accounting for Changing Base Rates of Labour Force Participation*

A related possibility is that the negative gradient is driven by the changes in the base rate of LFP. A lower base rate of LFP would imply less scope for a negative fertility effect on labour supply. This mechanically limits the scale of any average causal effect of fertility. We can account for this possibility by rescaling estimates to the relevant base rate (as in Angrist *et al.*, 2013). The rescaling relies on the assumption that effects tend to be monotonic in the population under study. That is, write the average effect in population s as

$$\beta_s = E_s[Y_1 - Y_0]\,,$$

where Y_1 and Y_0 are potential labour outcomes (with support $\{0, 1\}$) under the condition of three or more children and less than three children, respectively. Effect monotonicity implies $Y_1 \leq Y_0$, which also means:

$$E_s[Y_1 - Y_0 | Y_0 = 0] = 0.$$

This further implies that:

$$\beta_s = E_s[Y_1 - Y_0 | Y_0 = 1]\, E_s[Y_0]\,,$$

successful IVF treatment in the US and Denmark, respectively. By contrast, Aguero and Marks (2008; 2011) find no impact among 32 developing countries. While, we cannot replicate these findings with our data, the patterns seem to further validate a negative labour supply gradient across all family parities. See also Angelov *et al.* (2016), Kuziemko *et al.* (2018), Kleven *et al.* (2019a; 2019b) for an event study approach in developed countries. For comparable results across family size parities, see Bronars and Grogger (1994), Angrist and Evans (1998), Cruces and Galiani (2007), Maurin and Moschion (2009), Vere (2011), and Lundborg *et al.* (2017).

[26] See Propositions 1 and 2 of Bisbee *et al.* (2017). To be as flexible as possible, we discretise our baseline covariates into dummy variables of mother's age (3-year bins), age at first birth (3-year bins), first child gender, and education (<8, 8–11, 12–15, 16+ years of schooling where applicable). Note that we include country–year fixed effects as usual.

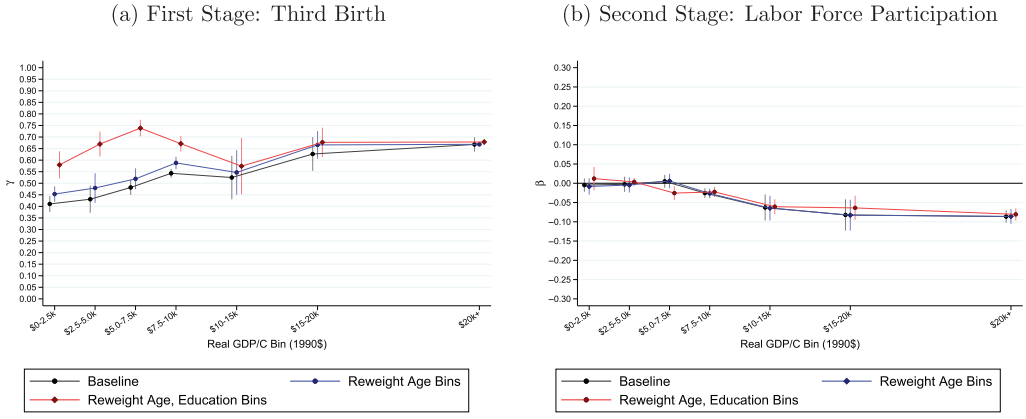

Fig. 9. *Reweight Covariates to 1980 US Compliers, Twin IV.*
Notes: This figure adjusts for changes in the twin IV complier population by reweighting IV estimates to the USA's 1980 complier profile (see Angrist and Fernandez-Val, 2013 and Bisbee *et al.*, 2017). The sample is restricted to the set of mothers who report education. Regressions control for mother's age, age at first birth, gender of first child, and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. Ninety-five per cent confidence intervals based on robust standard errors clustered at the country–year level are displayed but may not always be visible at the scale of the figure.

in which case the average effect of having three or more children *among those for which there can be an effect* is given by:

$$\beta_s^r = E_s[Y_1 - Y_0 | Y_0 = 1] = \frac{\beta_s}{E_s[Y_0]}.$$

Comparing trends in β_s versus β_s^r allows us to assess the influence of base participation rates.[27]

Given that we are estimating complier LATEs via IV, the populations indexed by s correspond to the compliers in our various country years. As such, the relevant base rate, $E_s[Y_0]$, corresponds to the LFP rate among compliers with instrument values equal to 0. We compute these complier-specific rates using the IV approach of Angrist *et al.* (2013).[28]

Figure 10 shows the rescaled baseline twins estimates (rescaled estimates for the ILO-restricted sample are shown in Online Appendix Figure A11). For the USA, the rescaling results in a substantial flattening past $7,500 per capita. For the non-US populations, the rescaled estimates are consistent (taking into account the uncertainty in the estimates) with a flattening after $10,000 per capita. However, a negative gradient is still evident over lower levels of income. This indicates

[27] This rescaling recovers a meaningful effect in populations for which the monotonicity assumption is reasonable. Rescaling would not be valid in country–years, such as those described in Subsection 3.3, where we estimate statistically significant positive fertility effects. Our figures are based on samples that include positive estimates, except for the pre-1920 US which shows the most consistently positive results. If we apply our rescaling strategy to country–year samples for which we observe either negative or (statistically indistinguishable from) zero fertility effects, we still recover a comparable negative gradient, although, unsurprisingly, labour supply responses at all real GDP per capita levels become more negative.

[28] Specifically, we stack the two-stage estimation used in Angrist *et al.* (2013) to calculate the complier-control mean with our baseline two-stage least squares regression to get the covariance between the base rate and the labour supply effect.

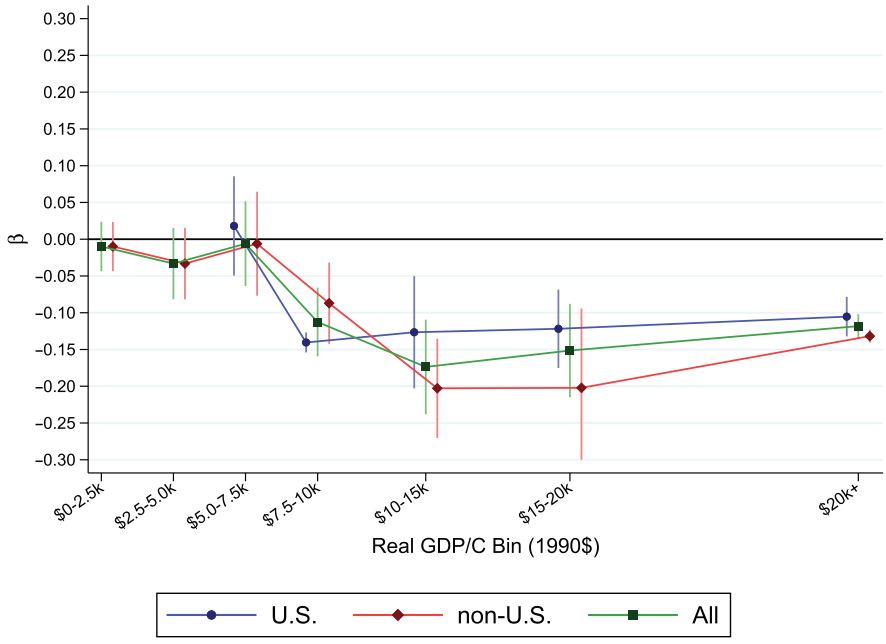

Fig. 10. *Second Stage Twin IV Estimates, Rescaled by the Complier-Control Outcome Mean.*
Notes: This figure rescales the baseline, second-stage twin IV estimates by the complier-control mean of mothers' labor force status. The calculation of the complier-control mean follows the IV methodology of Angrist *et al.* (2013). To get standard errors, unscaled coefficients and the complier-control mean are calculated in a seemingly unrelated regression framework and the standard errors of the ratio of the unscaled estimate to the control mean are calculated via the delta method. We exclude US samples prior to 1920 since these surveys often exhibit strongly positive labour supply responses. Regressions control for mother's age, age at first birth, gender of first child, and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. Ninety-five per cent confidence intervals based on robust standard errors clustered at the country–year level are displayed but may not always be visible at the scale of the figure.

that the decline in the labour supply effect of an additional child is not solely driven by increases in the base rate of mother's LFP and motivates further analysis into the channel driving the negative gradient, particular over income levels under $10,000 per capita. The analyses below examine results both with and without the base-rate rescaling.

It is worth noting that this procedure does not adjust for changing selectivity into the complier population, which the literature (e.g., Olivetti and Petrongolo, 2008) suggests is likely to be occurring.

4.3. *Changes to the Income and Substitution Effect Across Stages of Development*

We believe much of the remaining negative gradient is due to a declining substitution effect, in combination with a mostly unchanging income effect, resulting from increasing wages for women during the process of economic development.

We identify the substitution effect primarily through changes in job opportunities. This exercise is motivated by previous work that documents a U-shape of female employment with development

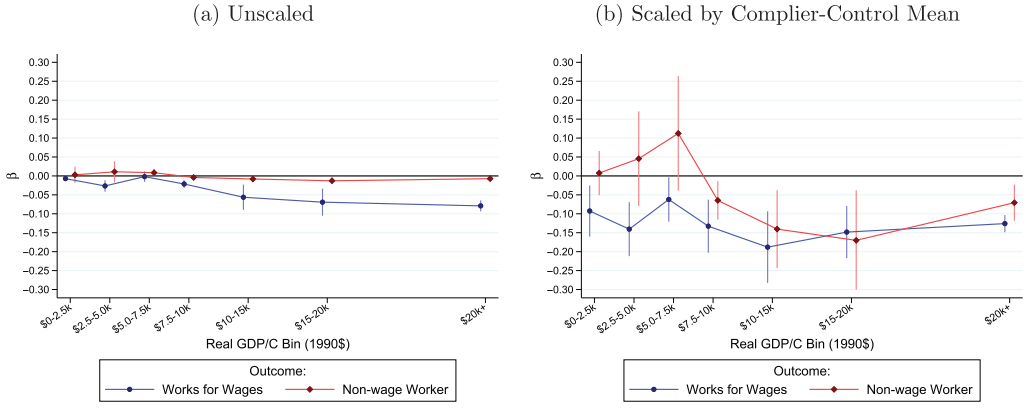

Fig. 11. *Twin IV Estimates by Class of Worker.*

Notes: This figure displays second-stage twin IV estimates, unscaled (panel A) and scaled by the complier-control mean (panel B). The outcome for the line with circles is an indicator of whether the mother works for wages. The outcome for the line with diamonds is an indicator of whether a mother works but not for wages. The sample is restricted to the set of mothers with nonmissing data on wage work and held constant across panels. We exclude US samples prior to 1920 since these surveys often exhibit strongly positive labour supply responses. The calculation of the complier-control mean follows the IV methodology of Angrist *et al.* (2013). To get standard errors, unscaled coefficients and the complier-control mean are calculated in a seemingly unrelated regression framework and the standard errors of the ratio of the unscaled estimate to the control mean are calculated via the delta method. Regressions control for mother's age, age at first birth, gender of first child, and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. Ninety-five per cent confidence intervals based on robust standard errors clustered at the country–year level are displayed but may not always be visible at the scale of the figure.

in the USA and across countries (Schultz, 1991; Goldin, 1995; Mammen and Paxson, 2000). Schultz (1991) shows that the U-shape is not observed within sector. Rather, it is explained by changes in the sectoral composition of the female labour force. Specifically, women are less likely to participate in unpaid family work (mostly in agriculture) and self-employment and more likely to be paid a wage in the formal sector in the later stages of the development process. In addition, we have reason to believe that the types of jobs that women have over time might change in a way that is less suitable to raising children. For example, in rural, agricultural societies, women can work on family farms while simultaneously taking care of children, but the transition to formal urban wage employment is less compatible with providing care at home (Jaffe and Azumi, 1960; McCabe and Rosenzweig, 1976; Kupinsky, 1977; Goldin, 1995; Galor and Weil, 1996; Edwards and Field-Hendrey, 2002; Szulga, 2014).

Given that consistent information on occupations and sectors across our many samples is limited, we rely on two coarse indicators of job type that can be consistently measured in almost all of our data. First, we try to capture the distinction between urban/rural and formal/informal occupations by changing the outcome to be whether women work for a wage or work but are unpaid. These results, unscaled (left) and scaled (right), are presented in Figure 11 (results for the ILO-restricted sample are presented in Figure A12). The unscaled results show that the changing relationship between fertility and labour supply is driven by women who work for wages. The scaled results are consistent with this finding, in that the effect is greater for wage workers than

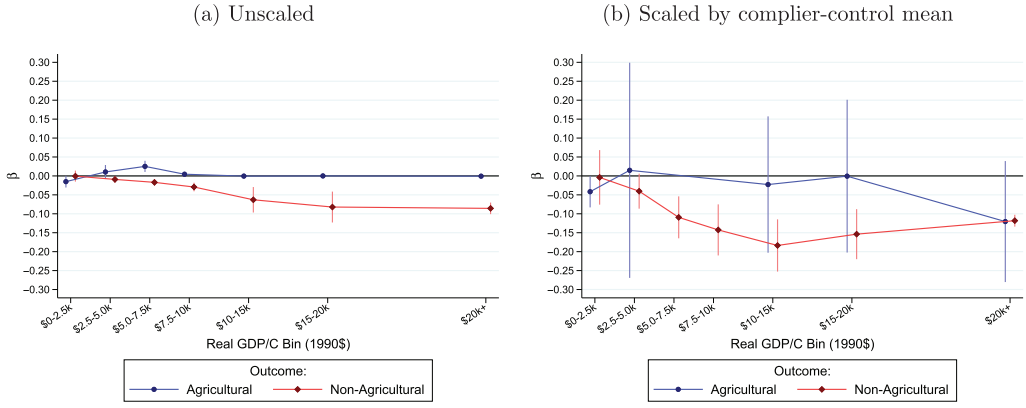

Fig. 12. *Twin IV Estimates by Agricultural Occupation.*
Notes: This figure displays second-stage twin IV estimates unscaled (panel A) and scaled by the complier-control mean (panel B). The outcome for the line with circles is an indicator of whether the mother works in agriculture (defined as a farm laborer, tenant, manager, or owner). The outcome for the line with diamonds is an indicator of whether a mother works but not in agriculture. The sample is restricted to the set of mothers with nonmissing data on occupation and held constant across panels. We exclude US samples prior to 1920 since these surveys often exhibit strongly positive labor supply responses. The calculation of the complier-control mean follows the IV methodology of Angrist *et al.* (2013). To get standard errors, unscaled coefficients and the complier-control mean are calculated in a seemingly unrelated regression framework and the standard errors of the ratio of the unscaled estimate to the control mean are calculated via the delta method. Regressions control for mother's age, age at first birth, gender of first child, and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. Ninety-five per cent confidence intervals based on robust standard errors clustered at the country–year level are displayed but may not always be visible at the scale of the figure. The point estimates for 5–7.5k and 7.5–10k in the agricultural subsample of panel B are not displayed because the denominator is very small, and the point estimate does not fit on the figure.

non-wage workers, so that the gradient is driven by changes in the sectoral composition of the labour force toward wage workers.

A second proxy of sectoral shifts is whether women work in agricultural or non-agricultural sectors (Figure 12 for the main sample, and Online Appendix Figure A13 for the ILO-restricted sample). Although the scaled results presented in the right plot are unfortunately noisy for agricultural labour, the labour supply response of women in non-agricultural sectors becomes clearly more negative as real GDP per capita rises. We also observe in Figure 13 (Online Appendix Figure A14 for the ILO-restricted sample) that fertility has almost no differential effect across the development cycle on female labour supply in professional occupations, despite the fact that these occupations tend to have higher wages.[29] Instead, the changing gradient seems to be driven entirely by women who work in non-professional occupations, suggesting either that education and professional status are poor proxies for the substitution effect or that the opportunity

[29] Professional occupations are defined somewhat differently across data sources. For the USA, we define professionals as Professional, Technical, or Managers/Officials/Proprietors. This definition corresponds to 1950 occupation codes 0–99 and 200–290. In all non-US sources, we define professionals as close as possible to the US For IPUMS-I, we use the International Standard Classification of Occupations (ISCO) occupation codes. For the NAPP, we use the Historical ISCO codes, except for 1911 Canada where we use 1950 US occupation codes. We dropped the 1851 and 1881 UK censuses due to difficulty convincingly identifying professionals.

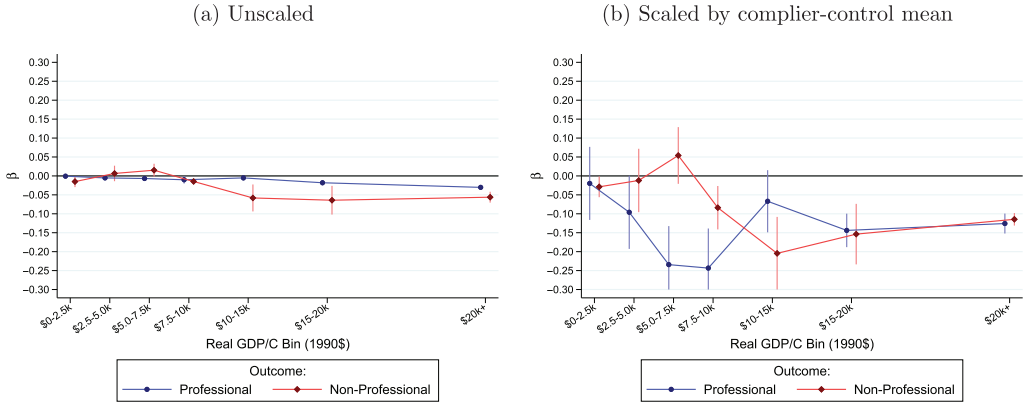

Fig. 13. *Twin IV Estimates by Professional Occupation.*
Notes: This figure displays second-stage twin IV estimates unscaled (panel A) and scaled by the complier-control mean (panel B). The outcome for the line with circles is an indicator of whether the mother works in a professional occupation. The outcome for the line with diamonds is an indicator of whether a mother works but not in a professional occupation. The sample is restricted to the set of mothers with nonmissing data on occupation and held constant across panels. We exclude US samples prior to 1920 since these surveys often exhibit strongly positive labor supply responses. The calculation of the complier-control mean follows the IV methodology of Angrist *et al.* (2013). To get standard errors, unscaled coefficients and the complier-control mean are calculated in a seemingly unrelated regression framework and the standard errors of the ratio of the unscaled estimate to the control mean are calculated via the delta method. Regressions control for mother's age, age at first birth, gender of first child, and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. Ninety-five per cent confidence intervals based on robust standard errors clustered at the country–year level are displayed but may not always be visible at the scale of the figure.

differences they capture are small in comparison to the sectoral shifts out of agricultural and non-wage work. This is consistent with an implication of the model laid out in Online Appendix A, which predicts that the negative gradient will be sharper among lower-skilled women.[30]

By contrast, we believe that the income effect of rising wages on fertility is likely small and invariant to the stage of development, as in Jones and Tertilt (2008), although the evidence is admittedly somewhat mixed. We further investigate the relevance of income effects in two ways. First, we examine the husband's labour supply response to children using the same twin IV estimator. A long literature, tracing back to classic models of fertility such as Becker (1960) and Willis (1973), uses the husband's labour supply response as a proxy of the income effect, since the substitution effect is likely to be smaller for men, who typically spend less time rearing children than women. In Figure 14, we return to the unscaled estimates and show that the husband's labour supply response is economically indistinguishable from zero and invariant to the level of real GDP per capita. Second, we examine the 1940 to 2010 US censuses, which contain hourly wages of husbands, to measure the differential labour supply response of married women throughout

[30] The fertility response literature has long used a woman's education to proxy for the type of jobs and wages available to her. While Gronau (1986) documents several results finding education is correlated with a fertility response, this correlation appears to reverse once Angrist and Evans (1998) apply instrumental variables. While we are able to replicate their results, we find that this education gradient is sensitive to instrument and the sample used. Overall, we find no strong heterogeneity by education (Online Appendix Figure A22).

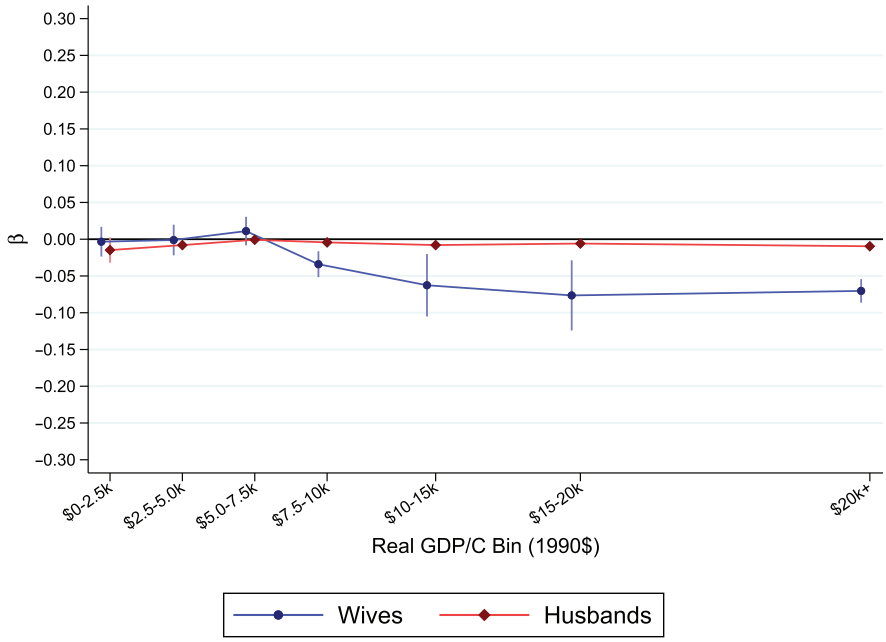

Fig. 14. *Twin IV Estimates for Fathers.*

Notes: This figure displays second-stage twin IV estimates for fathers living in the same household as mothers. The line with circles shows our baseline mother labour supply estimates, restricted to those where the father also lives in the same household. The line with diamonds shows the analaogous estimate for fathers. Regressions control for mother's age, age at first birth, gender of first child, and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. Ninety-five per cent confidence intervals based on robust standard errors clustered at the country–year level are displayed but may not always be visible at the scale of the figure.

the hourly wage distribution of their spouse. Although we continue to see a negative gradient over time, there are some, not always statistically significant, cross-sectional differences among women based on their husband's wage (see Online Appendix Figure A15). In particular, the negative gradient appears to be more pronounced among women whose spouses are high wage earners. Thus, we cannot rule out that the income effect could be playing a (smaller) role in the negative gradient as well.

4.4. *Childcare Costs*

A key factor driving the relationship between mother's labour supply and children is the time cost of raising kids (e.g., see equation (A.7) in Online Appendix A). One simple indication that childcare costs could be a relevant channel is visible in Figure 15, which stratifies the samples by six year age bins of the oldest child (similar results by the age of youngest child are presented in Online Appendix Figure A16). Regardless of kids' ages, we find a negative gradient, with the labour supply elasticity declining at real GDP per capita around $7,000 to $15,000. However, the gradient is monotonically sharper for families with younger children who

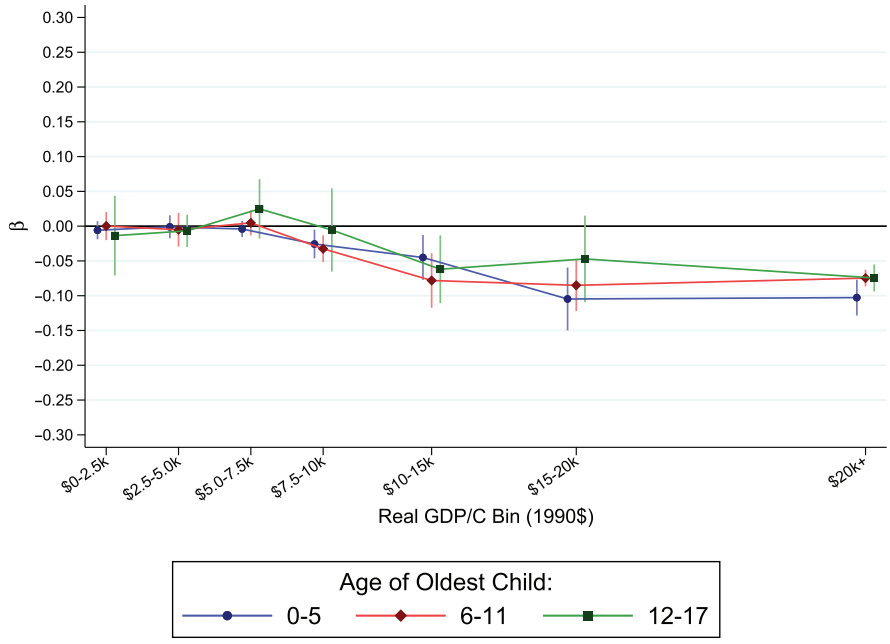

Fig. 15. *Twin IV Estimates by the Age of the Oldest Child.*
Notes: This figure displays second-stage twin IV estimates, stratified by the age of the oldest child in the household. Regressions control for mother's age, age at first birth, gender of first child, and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. Ninety-five per cent confidence intervals based on robust standard errors clustered at the country–year level are displayed but may not always be visible at the scale of the figure.

typically require more care, and especially among mothers in non-professional occupations with younger children (Table 3).[31] In particular, among mothers with a child under six, the impact of a child on working in a non-professional occupation falls by −0.067 (0.010) in countries with real GDP per capita above $10,000 relative to countries below $10,000.[32] By comparison, the non-professional gradient falls to −0.053 (0.011) and −0.020 (0.021) for mothers with a youngest child between six to 11 and 12 to 17. Strikingly, the labour supply gradient among professional occupations is invariant to the age of the youngest child. These results are at least suggestive that non-professional mothers, who are most exposed to sectoral shifts over the development cycle, may also be least likely to be able to pay for childcare costs through formal wage work.

[31] There is a monotonic relationship between age of children and time spent on childcare. For example, in the US Time Use Survey, 21- to 35-year-old women with two children at home where one was under six spent 2.9 hours per day, on average, on childcare (plus an additional 2.5 hours per day on other household activities). By comparison, when the youngest child is six to 11 or 12 to 17, mothers spend 1.8 and 1.3 hours per day, respectively, on childcare. For the subset of mothers who are not working, childcare takes up 6.8 (youngest child under 6), 5.4 (6 to 11), and 4.7 (12 to 17) hours per day.

[32] For exposition and due to sample size concerns that arise when dividing samples too finely, country–years in Table 3 are sorted into two real GDP per capita bins: above and below $10,000. The bottom row, labeled 'gradient,' is the difference.

Table 3. *Estimates by Mother's Professional Status by the Age of Youngest Child.*

	Mom occupation is professional			*Mom occupation is non-professional*		
	0 to 5	6 to 11	12 to 17	0 to 5	6 to 11	12 to 17
≤ 10k	−0.007***	−0.005***	−0.006**	0.004	−0.008	−0.007
	(0.002)	(0.002)	(0.003)	(0.006)	(0.005)	(0.015)
> 10k	−0.025***	−0.015***	−0.024***	−0.064***	−0.061***	−0.027*
	(0.004)	(0.005)	(0.006)	(0.008)	(0.009)	(0.015)
Gradient	−0.019***	−0.009*	−0.018***	−0.067***	−0.053***	−0.020
	(0.004)	(0.005)	(0.006)	(0.010)	(0.011)	(0.021)

Notes: This table displays second-stage, twin IV estimates of the effect of a third birth on the occupational status of mothers using the baseline sample who also report occupational status. The samples are stratified by the age of the youngest child (0–5, 6–11, and 12–17). 'Gradient' refers to the difference between row 2 (countries with real GDP per capita of at least 10,000 in 1990$) and row 1 (countries with real GDP per capita under 10,000 in 1990$). Regressions control for mother's age, age at first birth, gender of first child (and second child for same gender IV), and country–year fixed effects. Country–year weights are normalised to the number of mothers in a survey. Robust standard errors are clustered at the country–year level. See footnote 27 in the text for a description of the definition of professional and non-professional occupations in each data source. Statistical significant at the one, five, and ten percent levels are denoted by ***, **, and *.

Ideally, we would test the importance of childcare costs using exogenous variation across countries or over time. Unfortunately, we are not aware of such variation that spans our data. There is, however, a growing literature that uses quasi-experimental variation in access to childcare or early education to study mother's labour supply in individual countries, including the USA (Cascio, 2009; Fitzpatrick, 2012; Herbst, 2017), Argentina (Berlinski and Galiani, 2007), Canada (Baker *et al.*, 2008) and Norway (Havnes and Mogstad, 2011).[33] Summarising this literature, Morrisey (2017) concludes that the availability of childcare and early education generally increases the labour supply of mothers, although there is some response heterogeneity across countries. We view this literature as at least consistent with the possibility that the negative labour supply gradient may be amplified if childcare costs increase because jobs become less conducive to child rearing, and, if so, this dynamic could be stronger among lower wage mothers with less flexibility to provide childcare to young children (Blau and Winkler, 2019).

4.5. *Other Explanations and Robustness*

The evidence from the USA shows that mothers' labour supply response to children likely fell in the decades immediately after WWII.[34] This is a period in which at least two important developments may have impacted female LFP: the introduction and wide-spread usage of modern contraceptives and shifts in the social norms of female work.

To explore the importance of birth control pills, we exploit differences in the timing in which US states allow access to the pill among 18- to 21-year-olds (Bailey *et al.*, 2012). Using mothers in the 1970 and 1980 censuses and a difference-in-difference design, we could not find evidence that access to birth control impacted the labour supply decisions of mothers with either of our main instruments. Combined with a robust cross-sectional negative mother labour supply gradient

[33] To take one example, Herbst (2017) is based on the WWII-era US Lanham Act that provided childcare services to working mothers with children under 12. State variation of funding offered a natural experiment in a period when we find the aggregate labour supply response of mothers to additional children was close to 0. Herbst reports that additional Lanham Act childcare funding raised mother's LFP.

[34] The evidence from other countries for which we have data spanning the development cycle (Canada, France, Ireland, the United Kingdom) suggests a similar pattern (see Online Appendix Table A1).

over the last couple of decades, when much of the world has access to oral contraceptives, we do not see support for changing access to birth control as an important explanation of our main findings.

We looked at two exercises for evidence on the role of changing social norms. Our first attempt borrows an idea from the important work of Goldin (1977), who traced persistent differences in black–white female LFP to different social norms about female work by race that arose during slavery. Boustan and Collins (2014) further show that this disparity persisted into the mid-twentieth century through the intergenerational transmission of work norms between mothers and daughters. Following them, we looked for differences in the labour supply gradient in the USA over time by race. We find that the gradients for whites and blacks follow the same general pattern, with the black labour supply gradient enduring a steeper decline in the 1950 and 1960 censuses (Online Appendix Figure A17). While interesting in its own right, the lack of any economic or statistical difference in the pre-WWII period when the labour supply effect of children is zero indicates that race-specific social norms about female work cannot explain the increasing costliness of a second child over development, at least in the USA.

Secondly, we looked more directly at female work norms using a question from the General Social Survey (GSS): 'Do you approve or disapprove of a married woman earning money in business or industry if she has a husband capable of supporting her?' We show (Online Appendix Figure A18) that the negative gradient across real GDP/capita is similar in economic magnitude for the bottom, middle, and top terciles of state-census years ranked by the share of respondents who do not approve of married women working outside the home within each year. That is, there is a declining labour supply elasticity between 1970 and 1980 that flattens out thereafter for each of the three 'women work norm' tercile samples. Consequently, although these tests are limited to the US experience, we see no compelling evidence to claim that evolving social norms influence our main results during this narrow time period.

We perform a wide range of robustness checks, examining the consequence of omitted variables bias, alternative benchmarks of development, and a variety of variable definition, specification, and sampling considerations. In particular, we examine the robustness of our instrumental variables strategy by trying an alternative instrument (time to first birth; see Klemp and Weisdorf, 2019), by including additional controls suggested by Bhalotra and Clarke (2016), and by splitting the results by same gender versus different gender twins. We present results that look at alternative development benchmarks on the x-axis, including average female wages (for the 1940 to 2010 US censuses), and the average education level of women. Finally, we examine the robustness of our results to the choice of sample and specification.

The full set of results are described in detail in Online Appendix B. Among these, one result to highlight is the robustness of our results to the use of a more precise date of birth when available. In order to maximise data coverage, our main results define twins as being born in the same calendar year. However, for a subset of our data we also observe the month or quarter of birth, allowing us to rule out so-called 'Irish twins'. The key patterns in our result are robust to the choice of sample and the more precise definition of twinning (Online Appendix Figure A19).

5. Conclusion

In her classic monograph of the evolution of women's work in the USA, Goldin (1995) documents a U-shaped evolution of women's labour supply over the twentieth century. At the same time,

she notes the paucity of historical causal evidence on the link between fertility and labour supply. A parallel literature in development economics has investigated the implications of evolving patterns of fertility in developing countries on economic growth (and implicitly labour supply). While there have been many notable and pioneering studies on the effect of fertility on labour supply in developing countries, they naturally tend to focus on single countries or non-causal evidence.

Using a twin birth and same gender of the first two children as instruments for incremental fertility, this article links these two literatures by examining causal evidence on the evolution of the response of labour supply to additional children across a wide swath of countries in the world and over 200 years of history. Our article has two robust findings. First, the effect of fertility on labour supply is small, indeed typically indistinguishable from zero, at low levels of income and both negative and substantially larger at higher levels of income. Second, the magnitude of these effects is remarkably consistent across the contemporary cross-section of countries and the historical time series of individual countries, as well as across demographic and education groups.

The results are consistent with an increased time cost of looking after children, which seems to arise from changes in the sectoral and occupational structure of female jobs, in particular the rise of formal, non-professional, and non-agricultural wage work that flourishes with development. We also show that the negative gradient is steeper among mothers with young children that work in non-professional occupations and argue that access to childcare subsidies may attenuate the negative gradient, suggesting that the affordability of childcare costs may play a key role in declining LFP during the development cycle.

It is important to note that our findings are also consistent with and complementary to other explanations. Over the two hundred years-plus that we examine, there have been significant shifts in social norms regarding both work and fertility, parenting styles, and wide-spread adoption of modern contraceptives, among other plausible changes in the response of mother's work to fertility (Mammen and Paxson, 2000). While we have provided indirect evidence from the USA against some of these mechanisms, our data does not allow us to fully disentangle these plausible channels.

In discussing the evolution of female LFP in the USA, Goldin (1990) notes that ' . . . women on farms and in cities were active participants [in labour] when the home and workplace were unified, and their participation likely declined as the marketplace widened and the specialisation of tasks was enlarged.' In examining the relationship between labour supply and fertility over the process of development, we arrive at a parallel conclusion. The declining female labour supply response to fertility is especially strong in wage work that is likely the least compatible with concurrent childcare.

We see three implications of our results. First, in thinking about the U-shaped pattern of LFP that has been widely documented in the economic history literature, our results suggest that decreases in fertility play an explanatory role. That is, as fertility rates have declined over the latter half of the twentieth century, the responsiveness of labour supply to fertility has increased, contributing to increases in female LFP. Second, among developing countries, our results however suggest that changes in fertility (such as those documented in Chatterjee and Vogl, 2018) tend not to have a large impact on LFP, arguing against fertility-reduction policies specifically motivated by women's LFP and its contribution to growth. Third, our results provide an interesting example of the external validity of a diverse and seemingly different set of results on fertility and labour supply across the development, labour, and economic history literature.

Federal Reserve Bank of Chicago, USA
New York University, USA
University of Chicago, USA
Columbia University, USA
New York University, USA
Princeton University, USA

Additional Supporting Information may be found in the online version of this article:

Online Appendix
Replication Package

References

Abadie, A. (2003). 'Semiparametric instrumental variables estimation of treatment response models', *Journal of Econometrics*, vol. 133(2), pp. 231–63.
Adda, J., Dustman, C. and Stevens, K. (2017). 'The career cost of children', *Journal of Political Economy*, vol. 125(2), pp. 293–337.
Agüero, J. and Marks, M. (2008). 'Motherhood and female labor force participation: evidence from infertility shocks', *American Economic Review*, vol. 98(2), pp. 500–4.
Agüero, J. and Marks, M. (2011). 'Motherhood and female labor supply in the developing world: evidence from infertility shocks', *Journal of Human Resources*, vol. 46(4), pp. 800–26.
Angelov, N., Johansson, P. and Lindahl, E. (2016). 'Parenthood and the gender gap in pay', *Journal of Labor Economics*, vol. 34(3), pp. 545–79.
Angrist, J. and Evans, W. (1998). 'Children and their parents' labor supply: evidence from exogenous variation in family size', *American Economic Review*, vol. 88(3), pp. 450–77.
Angrist, J. and Fernández-Val, I. (2013). 'ExtrapoLATE-ing: external validity and overidentification in the LATE framework', in (D. Acemoglu, M. Arellano and E. Dekel, eds.), *Advances in Economics and Econometrics: Tenth World Congress*, Econometric Society Monographs, pp. 401–34, New York: Cambridge University Press.
Angrist, J., Lavy, V. and Schlosser, A. (2010). 'Multiple experiments for the causal link between the quantity and quality of children', *Journal of Labor Economics*, vol. 28(4), pp. 773–824.
Angrist, J., Pathak, P. and Walters, C. (2013). 'Explaining charter school effectiveness', *American Economic Journal: Applied Economics*, vol. 5(4), pp. 1–27.
Baker, M., Gruber, J. and Milligan, K. (2008). 'Universal child-care, maternal labor supply, and family well-being', *Journal of Political Economy*, vol. 116, pp. 709–45.
Bailey, M., Hershbein, B. and Miller, A. (2012). 'The opt-in revolution? Contraception and the gender gap in wages', *American Economic Journal: Applied Economics*, vol. 4(3), pp. 225–54.
Bailey, M. (2013). 'Fifty years of family planning: new evidence on the effects of increasing access to contraception', *Brookings Papers on Economic Activity*, vol. 46(1), pp. 341–409.
Becker, G. (1960). 'An economic analysis of fertility', in *Demographic and Economic Change in Developed Countries*, Universities-National Bureau of Economic Research Conference Series 11, pp. 209–40, Princeton, NJ: Columbia Universities Press.
Berlinski, S. and Galiani, S. (2007). 'The effect of a large expansion of pre-primary school facilities on preschool attendance and maternal employment', *Labour Economics*, vol. 14, pp. 665–80.
Bisbee, J., Dehejia, R., Pop-Eleches, C. and Samii, C. (2017). 'Local instruments, global extrapolation: external validity of the labor supply-fertility local average treatment effect', *Journal of Labor Economics*, vol. 35(S1), pp. 99–147.
Bhalotra, S. and Clarke, D. (2016). 'The twin instrument', Working Paper, IZA.
Black, S., Devereux, P. and Salvanes, K. (2005). 'The more the merrier? The effect of family composition on children's education', *Quarterly Journal of Economics*, vol. 120(2), pp. 669–700.
Blau, F.D. and Winkler, A.E. (2019). 'Women, work, and family', in (S. Averett, L. Argys and S. Hoffman, eds.), *The Oxford Handbook of Women and the Economy*, pp. 395–424, New York: Oxford University Press.
Bloom, D., Canning, D. and Sevilla, J. (2001). 'Economic growth and demographic transition', NBER Working Paper no. 8685.
Bloom, D., Canning, D., Fink, G. and Finlay, J. (2009). 'Fertility, female labor force participation, and the demographic dividend', *Journal of Economic Growth*, vol. 14(2), pp. 79–101.
Boustan, L.P. and Collins, W. (2014). 'The origin and persistence of black–white differences in women's labor force participation', in (L.P. Boustan, C. Frydman and R.A. Margo, eds.), *Human Capital in History: The American Record*, pp. 205–40, Chicago: University of Chicago Press.
Bronars, S. and Grogger, J. (1994). 'The economic consequences of unwed motherhood: using twin births as a natural experiment', *American Economic Review*, vol. 84(5), pp. 1141–56.

Butikofer, A. (2011). 'Sibling sex composition and the cost of children', Working Paper.
Caceres-Delpiano, J. (2006). 'The impact of family size on investment in child quality', *Journal of Human Resources*, vol. 41(4), pp. 738–54.
Cascio, E. (2009). 'Maternal labor supply and the introduction of kindergartens into American public schools', *Journal of Human Resources*, vol. 44, pp. 140–70.
Chatterjee, S. and Vogl, T. (2018). 'Escaping Malthus: economic growth and fertility change in the developing world', *American Economic Review*, vol. 108(6), pp. 1440–67.
Clarke, D. (2018). 'Children and their parents: a review of fertility and causality', *Journal of Economic Surveys*, vol. 32(2), pp. 518–40.
Cristia, J. (2008). 'The effect of a first child on female labor supply: evidence from women seeking fertility services', *Journal of Human Resources*, vol. 43(3), pp. 487–510.
Cruces, G. and Galiani, S. (2007). 'Fertility and female labor supply in Latin America: new causal evidence', *Labour Economics*, vol. 14(3), pp. 565–73.
Dehejia, R., Pop-Eleches, C. and Samii, C. (2020). 'From local to global: external validity in a fertility natural experiment', *Journal of Business & Economic Statistics*, vol. 39(1).
Del Boca, D. (2015). 'Child care arrangements and labor supply', Working Paper, Inter-American Development Bank.
Ebenstein, A. (2010). 'The "Missing Girls" of China and the unintended consequences of the one child policy', *Journal of Human Resources*, vol. 45(1), pp. 87–115.
Edwards, L. and Field-Hendry, E. (2002). 'Home-based work and women's labor force decisions', *Journal of Labor Economics*, vol. 20(1), pp. 170–200.
Fitzpatrick, M. (2012). 'Revising our thinking about the relationship between maternal labor supply and preschool', *Journal of Human Resources*, vol. 47, pp. 583–612.
Galor, O. and Weil, D. (1996). 'The gender gap, fertility, and growth', *American Economic Review*, vol. 86(3), pp. 374–87.
Godefroy, R. (2017). 'How women's rights affect fertility: evidence from Nigeria', Economic Journal, vol. 129(3), pp. 1247–80.
Goldin, C. (1977). 'Female labor force participation: the origin of black and white differences, 1870–1880', *Journal of Economic History*, vol. 37(1), pp. 87–108.
Goldin, C. (1995). 'The U-shaped female labor force function in economic development and economic history', in (T.P. Schultz, ed.), *Investment in Women's Human Capital and Economic Development*, pp. 61–90, Chicago: University of Chicago Press.
Goldin, C. (1990). *Understanding the Gender Gap: An Economic History of American Women*, New York: Oxford University Press.
Gronau, R. (1986). 'Home production—a survey', in (O. Ashenfelter and R. Layard, eds.), *Handbook of Labor Economics*, pp. 273–304, Amsterdam: North-Holland.
Havnes, T. and Mogstad, M, (2011). 'Money for nothing? Universal child care and maternal employment', *Journal of Public Economics*, vol. 95, pp. 1455–65.
Heath, R. (2017). 'Fertility at work: Children and women's labor market outcomes in Urban Ghana', *Journal of Development Economics*, vol. 126, pp. 190–214.
Herbst, C. (2017). 'Universal child care, maternal employment, and children's long-run outcomes: evidence from the U.S. Lanham Act of 1940', *Journal of Labor Economics*, vol. 35(2), pp. 519–64.
Hoekstra, C., Zhao, Z.Z., Lambalk, C., Willemsen, G., Martin, N., Boomsma, D. and Montgomery, G. (2007). 'Dizygotic twinning', *Human Reproduction Update*, vol. 14(1), pp. 1–11.
ICF International. (2015). "Demographic and health surveys (various) [Datasets]", Calverton, Maryland: ICF International [Distributor].
International Labour Organization. (2019). https://www.ilo.org/ilostat, retrieved February 8, 2019.
Jaffe, A.J. and Azumi, K. (1960). 'The birth rate and cottage industries in under-developed countries', *Economic Development and Cultural Change*, vol. 9, pp. 52–63.
Jayachandran, S. and Pande, R. (2017). 'Why are Indian children so short? The role of birth order and son preference', *American Economic Review*, vol. 107(9), pp. 2600–29.
Jones, L. and Tertilt, M. (2008). 'An economic history of fertility in the United States: 1826–1960', in (P. Rupert, ed.), *Frontiers of Family Economics*, pp. 165–230, Bingley, UK: Emerald Publishing.
Klemp, M. and Weisdorf, J. (2019). 'Fecundity, fertility and the formation of human capital', Economic Journal, vol. 129(618), pp. 925–60.
Kleven, H., Landais, C. and Søgaard, J.E. (2019a). 'Children and gender inequality: evidence from Denmark', *American Economic Journal: Applied Economics*, vol. 11(4), pp. 181–209.
Kleven, H., Landais, C., Posch, J., Steinhauer, A. and Zweimüller, J. (2019b). 'Child penalties across countries: evidence and explanations', NBER Working Paper 25524.
Kupinsky, S. (1977). *The Fertility of Working Women: A Synthesis of International Research*, New York: Praeger Publishers.
Kuziemko, I., Pan, J., Shen, J. and Washington, E. (2018). 'The mommy effect: do women anticipate the employment effects of motherhood?', NBER Working Paper 24740.

Lundborg, P., Plug, E. and Rasmussen, A.W. (2017). 'Can women have children and a career? IV evidence from IVF treatments', *American Economic Review*, vol. 107(6), pp. 1611–37.
Mammen, K. and Paxson, C. (2000). 'Women's work and economic development', *Journal of Economic Perspectives*, vol. 14(4), pp. 141–64.
Maurin, E. and Moschion, J. (2009). 'The social multiplier and labor market participation of mothers', *American Economic Journal: Applied Economics*, vol. 1(1), pp. 251–72.
McCabe, J. and Rosenzweig, M.R. (1976). 'Female labor force participation, occupational choice, and fertility in developing countries', *Journal of Development Economics*, vol. 3, pp. 141–60.
Minnesota Population Center. (2015). North Atlantic Population Project: Complete Count Microdata, Version 2.2 [Machine-readable database], Minneapolis: Minnesota Population Center.
Moehling, C. (2002). 'Broken homes: the 'missing' children of the 1910 census', *Journal of Interdisciplinary History*, vol. 33(2), pp. 205–33.
Morrissey, T. (2017). 'Child care and parent labor force participation: a review of the research literature', *Review of Economics of the Household*, vol. 15, pp. 1–24.
Olivetti, C. and Petrongolo, B. (2008). 'Unequal pay or unequal employment? A cross-country analysis of gender gaps', *Journal of Labor Economics*, vol. 26(4), pp. 621–54.
Olivetti, C. and Petrongolo, B. (2017). 'The economic consequences of family policies: lessons from a century of legislation in high-income countries', *Journal of Economic Perspectives*, vol. 31(1), pp. 205–30.
Rosenzweig, M. and Wolpin, K. (1980). 'Testing the quantity-quality fertility model: the use of twins as a natural experiment', *Econometrica*, vol. 48(1), pp. 227–40.
Rosenzweig, M. and Wolpin, K. (2000). 'Natural 'natural experiments' in economics', *Journal of Economic Literature*, vol. 38(4), pp. 827–74.
Rosenzweig, M. and Zhang, J. (2009). 'Do population control policies induce more human capital investment? Twin, birth weight and China's 'One-Child' policy', *The Review of Economic Studies*, vol. 76(3), pp. 1149–74.
Ruggles, S., Alexander, J.T., Genadek, K., Goeken, R., Schroeder, M.B. and Sobek, M. (2010), 'Integrated Public Use Microdata Series: version 5.0 [Machine-readable database]'. Minneapolis: University of Minnesota.
Schultz, T.P. (1991). 'International differences in labor force participation in families and firms', Working Paper, Yale Economic Growth Center Discussion Paper No. 634.
Schultz, T.P. (2008). 'Population policies, fertility, women's human capital, and child quality', in (P.T. Schultz and J.A. Strauss, eds.), *Handbook of Development Economics, Volume Four*, pp. 3249–303, Amsterdam: Elsevier Science B.V.
Sobek, M. and Kennedy, S. (2009). 'The development of family interrelationship measures for international census data', Working Paper, University of Minnesota Population Center.
Szulga, R. (2014). 'A dynamic model of female labor force participation rate and human capital investment', *Journal of Economic Development*, vol. 39(3), pp. 81–114.
The Maddison-Project. (2013). https://www.rug.nl/ggdc/historicaldevelopment/maddison/, 2013 version.
Vere, J. (2011). 'Fertility and parents' labour supply: new evidence from US Census Data', *Oxford Economic Papers*, vol. 63(2), pp. 211–31.
Willis, R. (1973). 'A new approach to the economic theory of fertility behavior', *Journal of Political Economy*, vol. 81, pp. S14–64.