Chae, Minhee; Meng, Xin; Xue, Sen

Working Paper

Fertility, Son-Preference, and the Reversal of the Gender Gap in Literacy/Numeracy Tests

IZA Discussion Papers, No. 16208

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Chae, Minhee; Meng, Xin; Xue, Sen (2023) : Fertility, Son-Preference, and the Reversal of the Gender Gap in Literacy/Numeracy Tests, IZA Discussion Papers, No. 16208, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/278906

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16208

Fertility, Son-Preference, and the Reversal of the Gender Gap in Literacy/Numeracy Tests

Minhee Chae
Xin Meng
Sen Xue

JUNE 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16208

Fertility, Son-Preference, and the Reversal of the Gender Gap in Literacy/Numeracy Tests

Minhee Chae
Nankai University

Xin Meng
Australian National University and IZA

Sen Xue
Jinan University

JUNE 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

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

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

www.iza.org

ABSTRACT

Fertility, Son-Preference, and the Reversal of the Gender Gap in Literacy/Numeracy Tests*

This study examines the relationship between fertility decline and the reversal/narrowing of the gender gap in literacy/numeracy test scores. Drawing on Becker's Quantity-Quality (Q-Q) trade-off model, we propose that in a society such as China, where son-preference is prevalent, the Q-Q trade-off would be larger for daughters than that for sons. An exogenous reduction of fertility would make girls more likely to live in a single-sex family, which in turn increases the share of human capital investment for girls. We test this empirically. To consider the endogenous nature of the demand for children, we exploit an exogenous variation in fertility due to China's family planning policy. Utilising the policy intensity information collected from hundreds of county gazetteers to construct a novel instrument for fertility, we find that a reduction in one sibling narrows the gender gap in numeracy and literacy test scores by 14.8% and 21.4% of a standard deviation in the rural sample and 4.0% and 6.5% in the urban sample. The pattern is more pronounced in regions with a higher proportion of people who prefer a son over a daughter. We also provide suggestive evidence that the channel of the effect is indeed largely through the increased probability of girls living in single-sex families.

JEL Classification: J13, J16, I24

Keywords: gender gap, family size, test scores

Corresponding author:
Xin Meng
Research School of Economics
College of Business and Economics
Australian National University
Canberra, ACT 2601
Australia

E-mail: Xin.Meng@anu.edu.au

* We are grateful to seminar participants at Paris School of Economics, The European Winter Meeting of the Econometric Society, The CReAM 20th Anniversary Workshop on Topics in Labour Economics, Australian Gender Economics Workshop, The Asian & Australasian Society of Labour Economics meeting, The International Symposium on Contemporary Labor Economics, Australian National University, Peking University, Institute for Economic and Social Research at Jinan University, for their constructive comments.

1 Introduction

Many studies have documented the narrowing, or even a reversal, of the gender educational gap in both developed and developing countries(see, for example, Goldin et al., 2006; Wu and Zhang, 2010; Jones and Ramchand, 2016; Evans et al., 2021; Dao et al., 2021). The causes for this increase in female education discussed in the literature focuses mainly on the improved labour market positions for women (increased female labour force participation and return to education), fuelled by the ***natural*** fertility decline and pro-women technology changes (see, for example, Goldin et al., 2006; Cho, 2007; Chiappori et al., 2009; Asadullah and Chaudhury, 2009; Fortin et al., 2015; Riphahn and Schwientek, 2015; Bossavie and Kanninen, 2018; Dao et al., 2021). In addition, as labour market condition changes, women's more positive expectations of future jobs and earnings also provide a drive for the reversal of the gender educational gap (Goldin et al., 2006; Fortin et al., 2015).[1]

In recent years China also experienced a fast narrowing or even reversal of the gender education gap. Using the census and intercensal population survey data, panel A of Figure 1 shows the average years of schooling by gender for different birth cohorts, beginning with the 1950 and measured for those aged 24-60. The graph shows that the average female education level has increased faster than that of their male counterparts for those born after the early 1970s, and for the late 1980s cohorts, the average years of schooling of females have reached or even exceeded the level of males.[2] What is more, during the same period, adult females' literacy test scores also exceeded that of males and their numeracy test scores are catching up to that of the males (see Panel B of Figure 1).[3] These extraordinary gender education reversal trends pose a puzzle. Since China's market-oriented economic reform in 1979, China's female labour force participation rate has reduced consistently and the gender earnings gap enlarged (see, for example, Maurer-Fazio et al., 2011; Meng, 2012). Thus, the reasoning for the improved market position for women, as discussed above, does not seem to apply.

In this paper, we propose a different story, which links the Quantity-Quality trade-off (Q-Q) model developed by Becker and Lewis (1973) with the Chinese society's deep-rooted son-preference culture together with a ***coercive Family Planning Policy*** (FPP). The general Q-Q model suggests that, with a given household budget, parents invest less in each child as the number of children increases. Thus there exists a negative relationship between fertility and the level of human capital investment in each child, but there is no gender dimension in the original setting of the Q-Q model. However, in a society where son-preference is prevalent, the effect of sibling size on human capital investment may vary between girls and boys. When parents prioritise boys in resource allocation owing to their son-preference, an additional sibling will disadvantage girls more than boys, hence enlarging the gender educational gap. In the situation where fertility is declining coercively,[4] declin-

[1] A recent study proposed a condition under which the reversal of the gender educational gap could occur: if males' ability distribution is flatter than that for females, the increase in education itself would narrow the gender educational gap (Bossavie and Kanninen, 2018).

[2] The narrowing of the gender education gap in China has also been documented in Wu and Zhang (2010); Wu (2012).

[3] Data used for Panel B of Figure 1 are from China Family Project Survey (CFPS) 2010 and 2014.

[4] We distinguish the natural decline of fertility from a coercive FPP induced reduction in fertility in the paper

ing in fertility can increase the probability of girls (and boys) living in single-sex families. This, in itself, increases human capital investment in girls relative to boys compared to the situation when they lived in mixed-sex families. Increased probability of girls living in girls-only households could also benefit girls' educational performance beyond an immediate increase in parental human capital investment. For example, in societies with son-preference, girls often need to do more housework than boys (Edmonds, 2006; Lin and Adserà, 2013; Vu, 2014; Choi and Hwang, 2015). If more girls are living in girls-only households, it will on average reduce girls' family responsibilities and increase their time invested in their own education and health. Moreover, as girls-only households increase, the parental son-preference level may reduce, which can lead to an evolutionary reduction in son-preference in the society. Of course, in a society with strong son-preference, when fertility declined sharply and parents have additional resources to invest on children due to a coercive FPP, a narrowing of the gender gap could also occur in mixed-sex families. This could be due to potentially higher marginal product for education investment for girls than boys as a result of the previous lack of investment in girls' education.

We empirically test this hypothesis. We first examine the relationship between the effect of fertility (sibling size) on the gender gap in education performance. We find that girls in families with more siblings perform much worse than boys in terms of their Chinese and math test scores. To ascertain this relationship is causal, we construct a novel instrument that measures the potential reduction of fertility due to the FPP for each mother in our sample. Specifically, we first calculate the number of years each mother in our sample was exposed to the FPP in her own county during her fertile age (aged 15 to 49). We then generate a FPP intensity index by utilising our hand-collected data, from hundreds of county gazetteers, on each local government's records of contraceptive usage for the period since the year the initial FPP was introduced. To gauge the 'potential reduction' in fertility due to the FPP we also need a counterfactual fertility. For that we use the pre-FPP age-specific average fertility for women in their birth province. Our IV combines these three pieces of information.[5] Our IV closely resembles the identification method utilized in Chen and Fang (2021), but ours is based on more precise measurements of women's actual exposure duration and takes into account the regional variation in the intensity of the FPP implementation. Our IV results show that the fertility reduction induced by the government family planning policies significantly reduced the gender gap in Chinese and math test scores, even after controlling for individuals' education level and health status.

We then test whether this effect of fertility reduction on the gender educational performance gap is related to the variation in son-preference prevalence across different regions. We find a clear heterogeneous effect: regions with higher son-preference exhibit a larger effect.

because the latter is less likely to be subject to self-selection. The situation in a society with son-preference and a natural fertility decline is complicated. This is because in general, educated people are likely to have lower fertility and are also less likely to have son-preference. Thus, in a society with natural decline of fertility and son-preference, the correlation between decline in fertility and the reduction in gender educational gap is likely to be due to the self-selection of educated people reducing fertility and investing equally to sons and daughters. This made it hard to identify whether the narrowing of the gender educational gap is due to the Q-Q trade-off effect or the selection effect.

[5] Our IV will be precisely defined in the Empirical Strategy Section.

Is the effect we observed so far indeed related to the increased probability of girls living in single-sex households due to the forced fertility reduction? Albeit this part of the analysis is more descriptive, we do observe that the probability of girls living in a single-sex household increased as the FPP being implemented. Further, conditional on the same sibling size, girls in single-sex households are doing better academically than their counterparts in mixed-sex households. Some additional evidence are also provided to show that girls in single-sex families are, on average, doing fewer household chores than their counterparts in mixed-sex families while this is not the case for boys.

The contributions of this paper are threefold. First, while studies have discussed the reduction/reversal of the gender gap in schooling for China and elsewhere (Wu and Zhang, 2010; Wu, 2012; Huang et al., 2021; Guo et al., 2022), we are the first to document and analyse the reduction and reversal of the gender gap in adult literacy/numeracy in China. This is important because, relative to the gender gap in schooling years, the gender gap in the quality of education has more fundamental implications to gender equality in the labour market.

Second, we empirically establish that the reduction in sibling size causes a narrowing/reversal of the gender gap in adult test scores by using the novel instrument capturing the unique ***county-level time-varying*** intensity of the Family Planning Policy. Although studies investigating the impact of family size on children's outcomes in China have often used the introduction of the One Child Policy (OCP) in 1979 as a natural experiment (McElroy and Yang, 2000; Li et al., 2008; Rosenzweig and Zhang, 2009; Li et al., 2011; Cameron et al., 2013; Liu, 2014; Qian, 2017; Guo et al., 2018; Huang et al., 2021), China's FPP was introduced in the early 1970s, many years before the final introduction of the OCP, and the main fertility effect of the FPP occurred during this pre-OCP period (Zhang, 2017). Between 1970 and 1979, the total fertility in China dropped from just below 6 to 2.6, while after the introduction of the OCP, it dropped marginally from 2.6 to 2.3 in 1990 and then to 1.64 in 2010. Thus, using pre- and post-OCP comparisons to identify the effect ignores the large variations of the FPP on fertility that took place during the 1970s and before the OCP. Further, the existing studies mainly use the timing and intensity of policy implementation across ***provinces*** while ignoring the large variations within each province (Scharping, 2003). Our identification strategy based on our hand-collected data overcomes these drawbacks and can identify the causal fertility effect more accurately.

Third, we propose a hypothesis that incorporates the Q-Q model for a society with son-preference and with a coercive family planning policy to explain the channels through which a reduction in sibling size can generate the narrowing/reversal of the gender gap in education and academic performance, and provide some evidence to support our hypothesis.

The rest of the paper is organised as follows. Section 2 provides the background of the gender gap in China and the fertility decline over the last decades. Section 3 introduces the data and the main sample for our main analysis. Section 4 describes the empirical strategy with the instrumental variable construction. Section 5 presents the main results and tests the robustness of these results. Section 6 provides the potential channels of our main findings, and Section 7 concludes the paper.

2 Background

2.1 Son-preference

Son-preference refers to the attitude that regards sons being more important and more valuable to the family than daughters (Clark, 2000). son-preference exists in much of East and South Asia and China is one of the societies with a strong son-preference (Jayachandran, 2015).

In China, son-preference can be traced to the influence of the Confucian traditions that stated that only sons can perpetuate family lineage and perform ancestor worships (Milwertz, 1997). The desire to have sons is also based on a traditional custom that sons provide old-age support for their parents (Das Gupta et al., 2003; Ebenstein and Leung, 2010). A daughter is expected to be married out from her birth family to her husband's family, which gives incentives for parents to invest less in a girl's education and more in a boy's education. The cultural norm is described in the Chinese proverb, which goes, "Raising a daughter is ploughing someone else's field" (Jayachandran, 2015). Giving birth to a son, thus, increases a woman's bargaining power in a household (Fan et al., 2018).

The prevalence of son-preference may disadvantage daughters in parental resource allocation. For example, mothers may spend more time with sons (Bo, 2018) and breastfeed sons for longer durations(Jayachandran and Kuziemko, 2005). Sons may receive more childcare time and more vitamin supplementation than daughters (Barcellos et al., 2014). Daughters are less likely to be vaccinated than boys (Borooah, 2004). In short, sons get more resources than daughters when parents prefer sons over daughters. As a consequence, one may expect a male-biased gender gap in education and literacy/numeracy skills in a society with strong son-preference (Cunha et al., 2019; Francesconi and Heckman, 2016).

During the Mao era, the government pursued policies that promoted gender equality in many aspects. The ruling Communist Party of China (CPC) actively denounced the Confucian view that women are subordinate to men and implemented policies that encouraged gender equality in work, earnings, and marriage decisions. In addition, the government, to some extent, replaced family-provided economic security with a social-security in urban areas and collective-security in rural villages (Li and Lavely, 2003; Meng, 2000; Yao and You, 2018; Booth et al., 2019). However, as the market-oriented economic reform took place in the late 1970s and early 1980s, these social/collective protections gradually eroded, more so in rural areas than urban areas. Many responsibilities returned to families and individuals. The most relevant form of these protections for this paper is the erosion of the old-age support system. As families once again became the main source of old-age support, the need for sons increased. In addition, the government no longer denounces the traditional Confucian culture and even began to push the Confucian doctrine in high school curriculum gradually (see, for example, Cantoni et al., 2017). Both the actual change in the economic sphere and the change in attitude towards confucian traditions might have reinvigorated the son-preference culture (Li and Lavely, 2003).

2.2 China's fertility decline

In the late 1950s and early 1960s, China experienced the Great Famine, which caused over 30 million deaths. Soon after, Mao launched the Cultural Revolution, which threw the country into political and economic upheaval. Agricultural productivity was stagnant over this period. The clear memory of the Great Famine and the stagnation of the economy increased the fear that strong population growth might lead to a disastrous outcome. Moreover, high fertility, which exceeded six births per fertile woman throughout the 1960s, concerned the government (Bainster, 1987).[6]

In the early 1970s, a debate over the possibility that China might fall into the Malthusian Population Trap was initiated. Soon after, the Chinese government began a serious family planning campaign in 1971 with the propaganda 'One child isn't too few, two are just fine, and three are too many.' (Zhang, 2017). Moreover, the Leading Group for Family Planning was officially established in 1973 at the central level, and soon after, different provinces began to establish their own Leading Groups and by 1975, all provincial leading groups had begun their role to implement the family planning policy (FPP). At around the same time, the family planning leading groups were also established at the lower level of administration, from prefectures and counties/urban-districts down to villages/urban work-units or residential-committees. One of the initial policies implemented by the Leading Group was to encourage people to get married later, with longer birth spacing, and have fewer children under a new slogan 'Later, Longer and Fewer' (Center for Population Studies, 1986; Peng, 1991; Feeney and Wang, 1993; McElroy and Yang, 2000; Cai, 2010; Ebenstein, 2010; Cameron and Meng, 2014; Whyte et al., 2015; Zhang, 2017).

This family planning campaign was technically voluntary. However, enforcement was coercive, which meant it did not simply rely on persuasion or voluntary compliance (Whyte et al., 2015; Zhang, 2017). In fact, anecdotal historical evidence suggests that many of the coercive enforcement elements, which are commonly known as being implemented during the OCP campaign, can actually be dated from the 'Later, Longer and Fewer' campaign (Whyte et al., 2015). Each local family planning committee kept track of each fertile woman's menstruation, past births, and contraceptive usage. In some rural regions, women who had three children were forced to get a sterilisation operation or insert an intrauterine device to stop further births. In some urban factories, female employees who became pregnant without permission were subjected to regular harassment to get an abortion. Also, above-quota births were sometimes denied household registration; in other words, they would be denied schooling and other government benefits (Whyte et al., 2015). Statistical evidence also shows that the number of birth-control operations, including female sterilisation and intrauterine device insertion in China, began to sharply increase in 1971 and had more than doubled by 1975.[7] As a consequence, the population growth was reduced by half between 1970 and 1976 (Scharping, 2003). Total fertility plummeted from more than 6 births per woman to around 2.8 births in rural areas and from 3 to less than 2 births per woman in urban areas between 1971 and 1978.

Such a significant decline in fertility did not curb the government's determination to further reduce fertility. In December 1977, the State Planning Commission decided on demographic targets,

[6] The government attempted to control fertility in the period but was not successful (Scharping, 2003).

[7] See Figure 1 in Whyte et al. (2015) for more details.

which further tightened previous birth control activities. In January 1978, the fertility control campaign entered a new era where a new policy of 'One is the Best and Two is the Most' and 'Reward Having One-Child and Punish Having Three' was introduced. This was then followed by a compulsory 'One-Child per Couple' policy (OCP) in June 1979 at the second meeting of the fifth People's Congress (Center for Population Studies, 1986; Peng, 1991). The OCP, however, faced strong resistance in rural areas due partly to its deep-rooted son-preference culture. In 1984 a revised version of the policy that allowed rural households to have a second birth if the first birth is a girl was introduced (Scharping, 2003; Zhang, 2017). Later, a second child was generally allowed in many rural areas (Rosenzweig and Zhang, 2009).

The main instrument of enforcement of the OCP is economic and disciplinary sanctions for people giving unauthorised childbirth. In particular, for urban employees, wage deductions have always been the main sanction for violation of birth plans. Until the late 1980s, the majority of provinces deduct monthly wages by 10% as the punishment. There have been non-monetary sanctions as well. For example, those who exceeded the birth quota would not be promoted, the child of an out-of-quota birth would receive no medical insurance, and the family might not receive a housing allocation (Scharping, 2003). In rural areas, a common method is to pose a one-time fine, sometimes as high as several times an average household annual income. In general, the policy has been more forcefully implemented in urban areas and the penalties for unauthorised births are much more severe in urban areas than in rural (Bainster, 1987).

Figure 2 presents the change in total fertility for China as a whole as well as for rural and urban populations separately. The figure clearly shows that China's coercive FPP substantially reduced fertility. However, whether the OCP per se helped reduce fertility much is debatable. After the introduction of the 'Later, Longer, and Fewer' policy, fertility in China was already reduced substantially. The further introduction of OCP, in fact, was not followed by a significant further decline in fertility (Whyte et al., 2015).[8]

3 Data and Sample

3.1 China Family Panel Studies (CFPS)

The main data used in this study are from China Family Panel Studies (CFPS), which is a longitudinal household and individual survey launched in 2010 by the Institute of Social Science Survey at Peking University. As of February 2023, the 2010 baseline study and 2012, 2014, 2016, 2018, and 2020 follow-up studies (waves) had been released.

The sample of the CFPS is drawn from 25 provinces in China, which covers 94.5% of the total Chinese population.[9] It uses the probability-proportionate-to-size sampling with administrative units at province, county/district, and village/neighbourhood/community levels and socioeconomic

[8] According to Zhang (2017) and Garcia (2022), the slight decline in fertility after 1978 might be explained by the rapid economic development.

[9] The excluded provinces/cities/autonomous regions in mainland China are Xinjiang, Xizang, Qinghai, Inner Mongolia, Ningxia, and Hainan.

status as the stratification variables. Households are randomly selected from each sampled community; see Xie and Lu (2015) for a detailed discussion of the sampling design.

The CFPS have two samples of individuals, the adult sample of those aged 16 and above and the child sample of those aged 10-15. The main reason for us to use the CFPS, in addition to its national representativeness and public availability, is that it tests each adult respondent's literacy/numeracy skills in its 2010, 2014 and 2018 waves and these variables are the main dependent variables used in our study. The survey also collects a rich array of individual and household characteristics, including variables that are very important to our study. These variables include one's birth county and hukou registration county at age 3; complete sibling size, the gender composition of their siblings, birth order; parental characteristics, including those parents who were not living with the survey households.[10]

In 2010 the CFPS surveyed 33,598 adult individuals from 14,960 households.[11] Each subsequent survey year, there are certain attritions, though very low, and there are also new households added to the sample due to new households split from the old ones. In 2014 the total number of adults being surveyed was 37,147 and they were from 13,944 households. The numbers in 2018 were 37,944 adult individuals from 14,296 households.

The CFPS is an unbalanced panel survey. In our empirical work, as our main identification strategy comes from respondents' birth cohort and county-of-birth variations, which do not change across different waves, we use one observation from each individual. To fully utilise all the information provided in different years of the survey and to reduce measurement errors, we take the mean values of multiple-year records of variables that are changing over time and are more likely to be subject to measurement errors. These variables include individuals' literacy/numeracy test scores and the complete years of schooling. Variables that do not change over time, such as gender, birth year, ethnicity, parental schooling years, the number of siblings, and birth order, the CFPS survey team has largely harmonised the variables so that they are consistent over the survey years. For these variables, we take the earliest record available.

3.1.1 Main sample

Over the three survey years, we extracted 53,078 adult individuals.[12] We restrict our sample to individuals who were born between 1950 and 1990. The people's Republic of China was established in 1949. The restriction on those born in 1950 and after ensures that all the observations grew up under largely a similar education system. The restriction on cohorts born in 1990 and before is

[10] The birth county and hukou registration county are useful to measure the family planning policy intensity in one's early childhood, which is the main instrument for sibling size in our analysis. Further details are available in Section 4.

[11] These individuals include all adults aged 16 and above who were present at the time of the survey and family members who were elsewhere in the same county at the time of the survey. The adult individuals were interviewed separately. For those who were not present at home at the time of the interview and were not in the same county, their basic information was collected from the family members who were present. However, these individuals are not included in the number quoted here.

[12] The added new observations to the original 2010 adult sample are due to the formation of new families as well as individuals from the children sample moving to the adult sample.

to ensure that, by and large, at the time of the survey, individuals have had complete education (aged 20 in 2010). This restriction excludes 18,549 individuals and leaves us with a total of 34,529 observations (see Appendix A Table A1 for details).

China has had a rural-urban divide policy for a long time and because of these policies, rural and urban areas have experienced different socioeconomic environmental changes over the past decades. In addition, government policies, such as the FPP, were implemented differentially for rural and urban people based on their household registration status (*hukou*) rather than their current residential place.[13] To take this rural-urban divide into account in our analysis, we separately constructed the rural and urban subsamples based on household registration (*hukou*) status in a person's early childhood. The urban sample includes individuals whose household registration at age 3 was urban (non-agricultural) and the rural sample consists of those with rural (agricultural) household registration at age 3. Individuals with missing hukou status at age 3 (2,818 individuals) are excluded from our sample. That left us with a sample of 31,711 individuals, of which 27,003 and 4,708 are rural and urban hukou observations, respectively.

Because our identification strategy requires that we merge the individual sample with their birth county Family Planning Policy variables, a further 1,714 and 416 rural and urban hukou observations with missing birth county information, respectively, are excluded.[14] The next restriction excludes individuals with missing literacy/numeracy test scores (1,210 individuals in total, and 1,124 and 86 of rural and urban hukou, respectively) and with missing full sibling information (3,305 for the total, and 2,912 and 393 for rural and urban sample, respectively). Finally, excluding 450 individuals due to missing values for other control variables, our working samples consist of 20,895 rural individuals and 3,721 urban individuals. Table A1 of Appendix A provides detailed information on our sample construction rules as discussed above.

3.1.2 Literacy/Numeracy test

The literacy/numeracy test scores are used as our main outcome variables. As indicated earlier, CFPS tested each adult respondent's literacy/numeracy skills in 2010, 2014, and 2018. The word test consists of 34 Chinese characters drawn from the language textbooks used in primary and secondary schools. There are 24 math questions, including addition, multiplication, logarithms and trigonometric functions. Both the literacy test and the math test have multiple equivalent forms. When a respondent takes the test for the first time, a form is randomly chosen for the respondents, and at subsequent waves, the computer loads a different form from the one used in the last administration. The details of how these tests were conducted are presented in Appendix B.

[13] The household registration system (*hukou*) in China was initially developed to perform vital registration, limit rural-to-urban migration, and impose effective political and social controls (Potter, 1983). At present, each person in China has an official record (i.e. household registration), recording date of birth, place of birth, present place of residence, and so on. The place at which one is registered becomes the person's official place of residence (Goldstein, 1987; Goldstein and Goldstein, 1990; Potter, 1983; Li and Cooney, 1993)

[14] There are additional 4,625 observations who did not report the birth county, but with the same birth and current province. We assume that their current counties are the same as their birth counties and add an indicator variable in the regression to flag them out.

As there are some changes in the way the literacy/numeracy tests were conducted (See Appendix B for a detailed discussion of the changes), to keep the test score consistent, we first compute the standardised score (z-score) for each test year for each subject and then take the mean z-score for each subject over the different survey years if one attended the test in multiple years. For those who attended the test only once, their scores are from the year the tests were conducted.[15]

An individual final raw score is the number of correctly answered questions. Thus, the maximum score is 34 for literacy and 24 for numeracy. Figure B1 and Figure B2 of Appendix B provide the raw test score distributions for our sample. According to Figure B1, most urban individuals received literacy scores over 20 marks, whereas around 15% of rural individuals have zero scores. Figure B4 in Appendix B presents our sample distribution of the mean z-score variables. Overall, literacy scores are more evenly dispersed than numeracy scores. There are a few spikes in the numeracy score distribution, especially in the urban sample, which accounts for around 35% (bottom right panel). It suggests that numeracy test scores do not have enough variations for the urban sample, so when interpreting numeracy score results, we pay special attention to this problem.

3.1.3 Summary Statistics

Table 1 provides summary statistics by gender and hukou status for the main variables used for our analysis. Men perform better on average in the test score than women in both rural and urban samples. Urban men perform slightly better than women do by 1% of a standard deviation in the literacy test and by 0.05% in the numeracy test. The gap is significantly larger in the rural sample. Rural men perform better than rural women by 36% of a standard deviation on the literacy test and 38% of a standard deviation on the numeracy test. Similar patterns are found in educational attainments. Rural men are 1.87 years more educated than rural women and urban men have completed 0.18 years more education than urban women. The differences are statistically significant at the 1% level for the rural sample but not significant for the urban sample. On the other hand, parental education attainments do not have such a distinct gender difference. The gender differences in parental mean schooling years are less than 0.06 years and not statistically significant at the 10% level.

Women have more siblings than men do on average, particularly more brothers. Rural women have 0.19 more brothers than rural men, and urban women have 0.17 more brothers than urban men, and the differences are both statistically significant at a 1% level. On the other hand, gender differences in the number of sisters are close to zero in both samples. Women are 3% less likely to be a single child for both rural and urban samples. These sibling patterns are to be expected in a society with strict fertility controls and a prevalence of son-preference. Families with a female child are more likely to have additional births until they have a boy, whereas families with a male child are likely to stop giving birth.

Figure 3 summarises the number of siblings across birth cohorts in our sample.[16] Sibship size

[15] The main results are not sensitive to (1) whether we take the average or maximum of individual standardised test scores and (2) whether we use the raw or standardised test scores or not (See the bottom panel of Table 5).

[16] See Figure E2 for the number of sisters and brothers.

started to decrease from around the late 1960s birth cohorts. In the urban sample, the number of siblings already almost reaches the lowest point before the legal introduction of the OCP in 1980 and remains at a similar level after 1980. However, for the rural sample, the sibling size continued to decrease after 1980 to a lesser degree. Overall, this supports the fact that the government's birth control policy has been intensive and effective since the early 1970s, and not only from the OCP implementation.

Figure 4 shows the gender gap in the standardised test scores in our sample across birth cohorts for rural and urban separately. The grey vertical lines indicate 1971 and 1979: the introduction of the first serious family planning campaign and the 'One-Child per Couple' policy, respectively. The figure indicates that the gender gap is narrowed among younger generations. For the urban sample, the gender gaps in both literacy and numeracy scores are reversed for the post-1970 cohorts, whereas for the rural sample, women have made continuous progress in catching up to their male counterparts in both literacy and numeracy test scores, especially after the early 1970s. The early progress for the rural sample was made, to a large extent, because of the introduction of gender equality after the communist party came to power in 1949. The Chinese communist government introduced its first marriage law in 1951 to stipulate that men and women are equal in marriage, family, land rights, education, and labour market participation (Niida, 1964; Yao and You, 2016; Booth et al., 2019). Such gender equality policies had the most obvious effect in rural areas, which could have narrowed the gender gap. To consider this potential confounding factor, we will test the sensitivity of our results for the rural sample by excluding cohorts born before 1965.

3.2 County level data

To identify fertility decline, we collect information at the county level on married-fertile women's birth control rate (BCR) for different years from county/district and prefecture-level gazetteers ('Local Gazetteer' hereafter). According to Almond et al. (2019), local gazetteers are written by local historians and are not used for cadre evaluation, which reduces concerns over misreporting. The BCR is defined as the ratio of the number of married-fertile women using any method for birth control to the total number of married-fertile women in a county/district.

There is a total of 160 counties in our main sample, and we successfully collect the BCR information for 143 counties (88.3%), of which 63% are from county-level gazetteers and 37% from prefecture-level gazetteers. In the remainder of 17 counties, we are unable to find gazetteer information, instead, we used the 1988 2‰ National Sample Survey of Fertility and Contraception to impute missing values. Details of how the variable BCR are reported in the local gazetteers, the 1988 fertility survey, and our imputation method are reported in Appendix C.

There is a total of 160 counties in our main sample, and we successfully collect the BCR information for 143 counties (88.3%), of which 63% are from county-level gazetteers and 37% from prefecture-level gazetteers. In the remainder of 17 counties, we are unable to find gazetteer information. To impute missing values, we used the 1988 2‰ National Sample Survey of Fertility and Contraception. Details of how the variable BCR are reported in the local gazetteers, the 1988

fertility survey, and our imputation method are reported in Appendix C.[17]

In addition to BCR, we also extracted information from the local gazetteers on the initial timing of the establishment of the Family Planning Leading Group (FPLG) at the county level,[18] as well as county-level socio-economic condition variables, including industrial output per capita, agricultural output per capita, mortality rate, and government health expenditure per capita for each year across 1950 and 1990. The variables BCR and the timing of the establishment of the local FPLG are used to construct the IV used in the study, while the socio-economic condition variables are used for sensitivity tests.

4 Empirical Strategy

To establish that fertility affects the gender gap in literacy/numeracy skills, we estimate the following equation:

$$Y_{ict} = \beta_0 + \beta_1 Sib_{ict} + \beta_2 Sib_{ict} * F_{ict} + \beta_3 F_{ict} + X'_{ict}\kappa + W'_{ict}\lambda + \theta_c + \delta_t + \gamma_c I_t + \upsilon_{ict} \quad (1)$$

where Y_{ict} denotes individual standardised literacy/numeracy test scores for individual i born in county c and year t. Sib_{ict} is the total number of siblings individual i has. We interact Sib_{ict} with a female dummy variable F_{ict} to capture the gender heterogenous sibling effects on individual test performance. X_{ict} represents a vector of individual characteristics, including individuals' ethnicity, birth order[19] and its interaction with a female dummy variable, as well as parental years of schooling and a variable capturing the intensity of the FPP during the fertile period of individual i's mother in her residing county .[20] W_{ict} is a vector of other controls related to the sample construction errors. These include a vector of six dummy variables indicating the way the literacy/numeracy test scores were constructed,[21] and a dummy variable indicating whether the individual has a missing birth county variable. θ_c and δ_t capture birth county and birth year fixed effects, respectively, and I_t is county-specific linear time trends.

Our main interest is the coefficient β_2. It captures the differential effect of the increase in the number of siblings on females relative to that on males. The negative coefficient indicates that an additional sibling is associated with a worse performance in the literacy/numeracy tests for females

[17] Given that we have a large number of missing values, we also generate an alternative BCR variable using the 1988 2‰ National Sample Survey of Fertility and Contraception. This alternative variable is then used to construct an alternative instrument. We find that the main estimation results are comparable to those obtained using the BCR information collected from the local gazetteers. For more detailed results, see Appendix C.

[18] There are 21 out of 160 counties that we are unable to obtain timing information from the local gazetteers. For these counties, we instead use their provincial-level timing information.

[19] Given that the number of siblings and birth order are highly correlated, we employ a normalised birth order measure based on the method proposed by (Hatton and Martin, 2010).

[20] This variable varies across individual mothers' birth cohort as well as their birth county. The details as to how the variable is constructed and why it is included in the regression will be given later in the paper.

[21] The six dummy variables are: (1) attended only in 2010, (2) attended only in 2014, (3) attended only in 2018, (4) attended in 2010 and 2014, (5) attended in 2010 and 2018, and (6) attended in 2014 and 2018. The omitted category is individuals that participated in the tests in all three waves.

relative to that for males; and conversely, a reduction in one sibling improves female cognitive performance related to males.

An issue with the OLS estimation of Equation (1) is that the estimates of β_1 and β_2 are generally inconsistent due to the potential omitted variables, such as parental preference and parenting style, which are both correlated with cognitive skills and sibling size.

In the literature, three types of natural experiments are frequently used as instrumental variables (IVs) to mitigate the endogeneity issue of sibling size: the birth of twins (Rosenzweig and Wolpin, 1980; Black et al., 2005; Rosenzweig and Zhang, 2009; Peter et al., 2018), the sex composition of the first two children (due to the preference for variety parents may have more children if the first two have the same sex) (Angrist and Evans, 1998; Lee, 2008) and the family planning policies, such as China's OCP (Liu, 2014; Li and Zhang, 2017; Qian, 2017). While these three types of natural experiments utilise different variations to identify the effect of sibling size, none seems to suit our case. First, we do not have a large enough sample of observations who are twins in the CFPS data we are using.[22] Second, given that Chinese society has a strong son-preference culture, the gender of the first child (or the sex composition of the first two children) may affect not only the potential birth of an additional child but also the intra-household resource allocation among existing children. And hence, it may not be a valid IV (Black et al., 2010). Finally, although many studies use the introduction of China's OCP or its 1980s relaxation (1.5 policy) as policy shocks to identify the effect of sibling size, such policies only affected fertility at the margin, as indicated in the Background Section and also discussed in details in Zhang (2017); Chen and Fang (2021).

Given these issues, in this paper, we use the series of FPP introduced in the early 1970s together with the intensity of the OCP implementation to identify β_1 and β_2 of Equation (1). During the early FPP period, the fertility rate in China declined dramatically. Both the timing of the establishment of the local FPLG and the intensity of the FPP implementation varied significantly across different regions. These variations allow us to better identify the fertility change. We thus utilise the cross-cohort and cross-regional exposure to different intensities of the FPP to construct the instrumental variable as follows:

$$FP^{intens}_{i,c,p,t^m} = \sum_{49}^{a=15} [AFR_{i,p}(a) * I[t^m + a > T_c] * BCR_c(t^m + a)], \tag{2}$$

where FP^{intens}_{i,c,p,t^m} is a measure of the fertility interruption weighted by the number of years and the intensity of exposure to the FPP policy by the mother of the individual i, born in county c of province p, in years t^m. a is the individual i's mother's fertile age (between ages 15 and 49). $AFR_p(a)$ is the province-level age-specific fertility rate in 1969, prior to the enforcement of any effective family planning policy in any province, for a given county.[23] This term provides a measure of counterfactual or 'natural' age-specific fertility. The next term, $I[t^m + a > T_c]$, captures whether

[22] Rosenzweig and Zhang (2009) also points out that given the endowment of twins is lower than singletons, and parents may allocate resources based on children's endowment, the birth of a twin may have a direct impact on the outcome variable, and therefore may not be a valid IV.

[23] Because the county-level age-specific fertility rate is unavailable, we are using the province-level fertility for a given county here. This fertility measure is extracted from Coale and Li (1987), who used 1982 One Thousand Fertility Survey to calculate the age-specific fertility rate.

in a particular year, during her fertile age a, the mother was exposed to the local implementation of the Family Planning Policies, where T_c is the year the Family Planning Leading Group was set up in county c. The third term $BCR_c(t^m + a)$ is the birth control rate over a mother's fertile years for a given county. Intuitively, FP^{intens}_{i,c,p,t^m} can be interpreted as the weighted interrupted fertility due to the FPP.

Our instrumental variable is closely related to the identification strategy used in Chen and Fang (2021). However, there are two important differences. First, our measure of $I[t^m + a > T_c]$ varies across counties and the indicator $I[t^m + a > T_p]$ used in Chen and Fang (2021) only varies across provinces. To compare the difference in the timing variations of the two identification methods, we plot, in Appendix D Figure D1, the level of variations within each introduction year measured in Chen and Fang (2021). The figure shows that there are large differences in the timing of introduction among different counties within a province. The second, and more important difference between our IV and their strategy is we adjust the implementation of the FPP by an intensity factor. The FPP enforcement differs considerably across regions (Peng, 1989, 1990; Whyte et al., 2015; Zhang, 2017). Given this, we use $BCR_c(t^m + a)$ to measure the intensity of the local policy implementation. Therefore, our IV uses more variations than the strategy used in Chen and Fang (2021), and it is a more accurate measure of interrupted fertility due to the FPP. Because of this accuracy, our IV is less likely to be contaminated by other within provinces common confounding factors.

It could be argued that the term $BCR_c(t^m + a)$ in our IV may not only be related to fertility but also to individuals' literacy/numeracy performance due to a potential omitted variable that is related to local governments' abilities to govern. To mitigate this concern we include the term:

$$\sum_{49}^{a=15} BCR_c * I[t^m + a > T_c]$$

directly in Equation (1).

We use "FP^{intens}_{i,c,p,t^m}" and "$FP^{intens}_{i,c,p,t^m} * F_ic$" as the instruments for "Sib_{ict}" and "$Sib_{ict} * F_{ic}$" in Equation (1), respectively.

5 Results

5.1 OLS results

Table 2 reports the results from the OLS estimation of Equation (1) for literacy and numeracy outcomes and for rural and urban samples separately. The top and bottom panels present the results for the literacy and numeracy outcomes, respectively, while the left and right panels report the rural and urban samples, respectively. We include 3 different specifications: 1. only include the siblings, female, their interaction terms, and birth year fixed effects; 2. add birth county fixed effects; and 3. further include all individual and parental characteristics, controls for test year and missing county of birth information indicators, as well as a birth county-specific time trend.

The table shows that there are some different patterns between the results for the rural and the urban samples. In particular, controlling for sibling size and the interaction term between the sibling

and female dummy variable, the coefficient on the female dummy in the rural sample is consistently negative. But this is not the case for the urban sample. Among the observations in the urban sample, the female dummy variable (measuring female single child) is positive for the literacy test score and statistically insignificantly negative for the math test score, controlling for sibling and sibling and female interaction effects.

Looking through Table 2, we find that the coefficient on the sibling variable for the rural sample is either not statistically significant or positive and significant. This suggests that for our rural male sample, sibling numbers either have no role in their ability to achieve high test scores or sometimes even a positive role, ignoring endogeneity concerns. For the urban sample, this is also the case for columns 5 and 6.

The story for our female sample, though, is very different. There is a consistent pattern revealed from Table 2, that is, in all regressions, the coefficient on the interaction term is negative and statistically significant. The negative coefficient indicates that an increase in parental fertility is associated with a reduction in females' literacy/numeracy performance relative to that of males and hence an increase in the gender gap in test scores. Conversely, a reduction in fertility would be associated with an increase in female performance and a reduction in the gender gap in literacy/numeracy test scores.

The magnitude of the coefficient is large for the rural sample. Focusing on columns 3 and 6, we find that an additional sibling is associated with a 7.1% of standard deviation increase in the gender literacy score gap and a 5.8% of standard deviation increase in the gender gap for the numeracy test, albeit the sibling coefficient on their male counterpart is positive for both test scores. The summary statistics presented in Table 1 indicate that the mean difference in the literacy and numeracy z-scores between rural females and males are negative 36% and 38% of a standard deviation, respectively, and our female rural sample on average have 3.07 siblings. Based on our estimated coefficients, it seems that the number of siblings females have alone can explain 61%(=(0.071*3.07)/0.36)) and 47%(=(0.058*3.07)/0.38) of the average gender gap in literacy and numeracy test scores for the rural sample, respectively.

Relative to the rural sample, the magnitudes of the coefficient on the sibling and female interaction term is less than half the size for the urban sample. An additional sibling for a female in the urban sample is associated with an increase of 2.6% of a standard deviation in the gender gap in literacy score and 2.0% for the gender gap in numeracy score. However, as the mean gender gap in these test scores for the urban sample is very small (even zero for the literacy test), at the average number of siblings for urban females (2.17), the implied contribution from fertility can explain full (in the case of numeracy) or even over-compensate (in the case of literacy) the average gender gaps in these test scores. Interestingly, in the case of the literacy test, the coefficient for the female dummy variable for the urban sample is 0.057, positive and statistically significant, suggesting that urban females who grew up in a single-child family performed 5.7% of a standard deviation better than their male counterparts. And it is on top of this positive association, an additional sibling is associated negatively with the urban female literacy test score.[24] The positive female level effect on

[24]In fact, at the mean number of siblings for the female urban sample, the negative value that is associated with a

literacy here could be an indication that despite the prevalence of son-preference in society, when the policy forced upon individuals to have only one child and the child happened to be a girl, parents might have poured their resources into the girl, and as a result, girls in urban China are catching up with boys.

The OLS results suggest that there seems to be a strong Quantity-Quality trade-off effect for females but not for males. Some empirical studies have also found negative sibling effects on quality as measured by the level of education (Conley and Glauber, 2006; Rosenzweig and Zhang, 2009; Li et al., 2008; Lee, 2008). But others have reported non-negative effects of having more siblings on education (Black et al., 2005; Angrist et al., 2005, 2010; Fitzsimons and Malde, 2010). Our finding seems to be closely linked to the son-preference culture. When resources are limited, the needs of daughters are likely to be neglected.

5.2 IV results

As discussed earlier, the OLS estimates of sibling effects on the individual literacy/numeracy z-scores are inconsistent due to potential omitted variable bias. In this sub-section, we use the instrumental variable approach to mitigate the problem.

The results from the first stage regressions explaining sibling size and its interaction terms with a female dummy for the rural and urban samples separately are reported in Table 3. As can be seen in the table, the instrument FP^{intens}_{i,c,p,t^m} is statistically significant and negatively correlated with the number of siblings in both rural and urban samples. This suggests that an individual whose mother, during her fertile years, living in counties where the FPP policy was implemented earlier and with higher intensity is giving birth to fewer children. The instrument $FP^{intens}_{i,c,p,t^m} * F_{ic}$ has the same effect on the interaction term $Sib_{ict} * F_{ic}$. The Kleibergen-Paap Wald rk F statistics for the instruments for the rural and urban samples are 53.13 and 18.98, respectively, indicating that the IVs are strong enough.[25]

Table 4 reports the results from the IV estimation for specification 3 with the full set of controls. The left and right panels present the results for rural and urban samples, respectively. While coefficients on all other control variables are similar to the results obtained from the OLS estimations, the main change occurred in the coefficients for the number of siblings and its interaction term with the female dummy. In particular, the coefficients on the interaction term are much larger in magnitudes. For example, while the coefficient for the rural literacy test obtained from OLS estimation is -0.071, its magnitude increased to -0.214 and for the rural numeracy test, the magnitude of the coefficient increased from -0.058 to -0.148. A similar pattern is also found for the urban sample, where the OLS estimates for literacy and numeracy tests are -0.026 and -0.02, and the IV estimates are -0.065 and -0.04, respectively.

We interpret these IV coefficients as the Local Average Treatment Effects (LATE). In other words, these coefficients are for a group of individuals whose mothers' fertility was reduced by the

sibling is a negative 5.64%=(2.6% * 2.17) of a standard deviation, almost completely offset the advantage the urban women who grew up as a single child has.

[25]The full results of the first stage regressions are available upon request.

coercive FPP, while the control group is those whose mothers had higher fertility only because they were not subject to the FPP. The difference between the OLS and the IV estimations is, therefore, whether the always-takers (those who would reduce fertility even without the FPP) and never-takers (those who would not reduce fertility even with the FPP) are included. Our results suggest the gender bias in quantity-quality trade-off is larger among compliers than among the always-takers and never-takers. This seems to make an intuitive sense if we believe that the always-takers (with or without the FPP always have fewer children) have lower son-preference and hence a smaller gender investment gap relative to compliers, while the never-takers (with or without the FPP never reduce fertility) are less likely to be budget constrained as they are willing to pay the fines for over-quota births and hence will have smaller quantity-quality trade-off relative to the more budget constrained compliers.

Our results suggest that for the rural sample, a reduction in sibling size due to the implementation of the FPP can narrow the gender gap in literacy and numeracy test scores by 21.4% and 14.8% of a SD of these test scores, respectively. For the urban sample, the effect is estimated to be 6.5% and 4.0% of a SD.

5.3 Robustness Tests

In this subsection, we examine whether our above results are sensitive to adding other confounding factors, relaxing some of our assumptions, and using different ways to measure our dependent variables. These results from the IV regression for specification 3 with the literacy z-score as the dependent variable are reported in Table 5, where the left and right panels report the results for rural and urban samples, respectively.[26]

China had experienced socioeconomic environment changes to a great extent over the past decades and individuals in our sample would be exposed to such changes at the different development stages to different extents. If these changes are related to our outcomes, the test scores, and if they are not orthogonal to our instrumental variable, our IV results would be inconsistent. Although in our main estimation, a linear time trend (based on individuals' birth year) for each county is controlled for, in the case of the relationship between the socioeconomic environment changes and our important variables is non-linear, we would still suffer from these problems. Thus, we first examine whether adding the change in the local socio-economic environment over time and across regions affects our estimated results. In particular, we control for the county-level industrial output per capita, agricultural output per capita, local government expenditure on public health per capita, and mortality rate at individuals' birth county and birth year. Columns 1 and 5 in the top panel of Table 5 reports the selected results for the rural and urban samples, respectively. As can be seen that in neither case, the inclusion of these variables affects our estimated results for our main variables of interest.

In addition to other socioeconomic changes, over the past few decades, China implemented two important education policy changes. First, in 1986 it introduced the 9-year Compulsory Schooling Law (CSL), which requires parents to ensure that their school-age children complete up to junior

[26]The results using z-scores for numeracy are similar and are available upon request.

high school education. The implementation was nationwide, but there is a timing difference across different provinces (Liang and Dong, 2019). Second, the Chinese government began to rapidly expand tertiary education in 1999 to achieve mass higher education. As a result, higher education enrollment increased by 47% between 1998 and 1999, and it reached 6.3 million in 2009 from 1.0 million in 1998 (Yeung, 2013; Wan, 2006). The timing of the expansion was the same across all provinces. To avoid the possibility that these education reforms confound our results, we add two variables to capture the cohorts that are subject to these reforms. For the introduction of CSL, we use the timing difference in the policy implementation across provinces to generate a province-cohort varying indicator variable for individuals who were aged below 16 at the time of the introduction of the CSL in their birth provinces.[27] There was no timing variation across provinces in the university expansion in 1999. However, the level of expansion differs across provinces. We use 'the growth rate of the tertiary institutions in one million population across different provinces over the period 2001 and 2017' to interact with a dummy variable indicating whether an individual should enter university in 1999 or after (based on his/her birth cohort) to proxy for the impact of university expansion effect.[28] The results are reported in columns 2 and 6 for the rural and urban samples. Although we find that both education reforms are positively and statistically significantly correlated with the rural literacy z-score, including them in the regression does not change our main results in either sample.

In Section 2.1, we indicated that the gender equality policy pursued during the Mao era significantly narrowed the gender gap in education. As such, it is important to ascertain that our results are not driven by this early gender equality policy, Columns 3 and 8 of the top panel of Table 5 report the results exclude individuals born before 1965, which show that for the rural sample, despite a somewhat decline in the magnitude of the coefficients on the interaction term, our main story is not changed. For the urban sample, excluding these early cohorts slightly enlarged the estimated coefficient.

To maximise our sample size, we made an assumption that those who did not report birth county information and whose current residential province is the same as their birth province were born in their current residential counties (see discussion in section 3.1.1). To test the sensitivity of this assumption, Columns 4 and 9 of the top panel of Table 5 exclude individuals for whom we do not have birth county information. As can be seen that our results are not sensitive to the assumption.

In columns (5) and (10), we further control for years of schooling. Doing so allows us to understand better whether the additional investment towards girls due to the reduction in sibling size is mainly through increases in their years of schooling or also through improving their knowledge over and above the increase in years of schooling. Such a knowledge improvement could be due to girls with fewer siblings would have more time to do their school homework; or there could be a reduced class size which helps to improve children's learning experience; or could simply be that girls in

[27] The detailed timing variation across provinces can be found in Table 2 of Liang and Dong (2019).

[28] To use this growth rate to capture university expansion, we assume that the majority of Chinese students enrol in tertiary institutions within their own province. This seems to be the case. Based on this online article: *https* : *//www.sohu.com/a/*454441108_614566, the rate is around 66%. The variable 'growth rate of tertiary institutions for each province between 2001 and 2017' is calculated based on Table 3 of Borsi et al. (2022).

households with fewer siblings are able to attend better schools. The results for literacy test scores shown in columns (5) and (10) for rural and urban samples, respectively, indicate that the size of the coefficient on the sibling and female interaction terms was reduced by almost half. However, over and above years of schooling effect, we still observe a sizeable gender gap in literacy test scores, suggesting that, indeed, some of the gender gaps in test scores may be due to the improved effort and quality of schooling.

The bottom panel of Table 5 reports results using different measures of literacy score. In our main estimation, we use the individual average of the standardised scores for the three years (2010, 2014, and 2018) as the main outcome variables. In columns 1-5 of the bottom panel of Table 5, we report the results using, instead of the mean z-scores, the maximum z-score of the three survey years, the mean of raw scores, the 2010 z-scores, the 2014 z-scores, and the 2018 z-scores respectively, for the rural sample, while in columns 5-8, the same results for the urban sample. The results presented here suggest that our results are not due to how the dependent variables are measured.

6 Mechanism

The last section established the fact that the decline in fertility caused the reduction of the gender gap in literacy/numeracy test scores. In this section, we investigate the channels.

6.1 Son-preference

Our conceptual framework assumes that the reason the decline in fertility could narrow the gender gap in test scores is probably due, to a large extent, to the combination of the prevalence of son-preference and the coercive FPP. In this sub-section, we try to find a link to the son-preference culture. If son-preference is a mechanism for our effect, we expect to see a negative relationship between the gender gap in the sibling size effect (i.e. the coefficient of the interaction term between sibling size and female) for a region and the region's son-preference intensity.

To this end, we first need to measure the intensity of son-preference prevalence for different regions. We find two surveys with questions indicating one's son-preference. The first is the 2014 CFPS survey. In a group of questions on individuals' opinions, one of them states, 'Women should give birth to at least one son' and the answers are (1) strongly agree, (2) disagree, (3) neutral, (4) agree and (5) strongly agree. We use the proportion of individuals in each province who choose either (4) or (5) as an indication of the level of son-preference in the province.

The second is the Survey of Women's Status in Contemporary China (SWSC) conducted by the Institute of Population Studies of the Chinese Academy of Social Sciences in 1991. SWSC asks the respondents, 'If only one child is allowed, do you prefer a son or a daughter?' and each respondent could choose one answer among (1) a son, (2) a daughter, or (3) indifferent. We use the proportion of women who chose 'a son' over the other two options at the provincial level, by rural and urban separately, to measure the level of son-preference in the province.[29]

[29]The SWSC was conducted in the following provinces: Shanghai, Shanxi, Shandong, Guangdong, Jilin, and

We then re-estimated the IV regression of specification 3 using the rural sample for each province separately. The estimated coefficients of the interaction term between the number of siblings and female dummy variable is plotted against the proportion of individuals who have son-preference in each province from both the CFPS and SWSC survey results. Figures 5 and 6 present these results, respectively. Both figures show that the gender gap in sibling size effect estimated using the CFPS and SWSC data is indeed correlated with the share of people with son-preference at the provincial level. Provinces with the larger gender gap in the sibling size effect also have a higher proportion of people with son-preference.[30]

In summary, we find that the female disadvantage in having additional siblings is more pronounced in regions with a higher prevalence of son-preference.

6.2 Mixed-sex vs. single-sex families

One of the important mechanisms, we conjecture, is that in a society with son-preference, the coercive FPP would aid the narrowing/reversal of the gender gap in literacy/numeracy tests through an increase in the share of girls (and boys) living in single-sex families rather than in the mixed-sex families. To know why a reduction in fertility due to the coercive FPP can increase the probability of girls and boys living in single-sex families, we assume that the probability of a girl and a boy living in a single-sex family is p^n and $(1-p)^n$, respectively; and their probability of living in a mixed-sex family is $1-p^n-(1-p)^n$. It can be seen that as n reduces, both p^n and $(1-p)^n$ will increase.[31] Then, why is it that living in single-sex families would benefit girls more than boys in a society with son-preference? To understand this, we can look at an example of two-children families. In a two-children mixed-sex family with son-preference, the son would receive >50% of the parental education investment, while the girl would receive <50%. However, in a two-children single-sex family, *certeris paribus*, even with son-preference, sons would receive 50% and girls would receive 50%. Thus, relative to mixed-sex families, living in single-sex families would benefit girls' and worsen boys' situation in terms of education investment. In addition, given that, in general, girls in mixed-sex families do more housework than boys, the increase in the share of girls living in single-sex families should also, on average, reduce girls' housework burden.

The issue, though, is whether our data support our conjecture. In this sub-section, we present some evidence. Figure 7 plots the share of women and men in single-sex and mixed-sex families, respectively. The figure shows that before the introduction of the FPP these ratios were rather stable, especially for girls. But since 1971, the ratio of girls living in single-sex families has increased significantly. For boys, the change has been slight, but an increase nevertheless. Table E1 of

Ningxia. It surveys married women and their husbands born between 1926 and 1973. The variations on son-preference within the rural sample across provinces are quite large, varying from 17.7% in Shanghai to over 75% in Guangdong and Ningxia provinces. The variation among urban people across provinces range from 18.4% in Shanghai to 33.8% in Guangdong. These results are available upon request.

[30] Since the CFPS does not include Ningxia province, the figure using SWSC data only comprises five provinces for which we have both CFPS and SWSC data. The sample size for the urban sample by province in both surveys is too small, and we are unable to obtain sensible estimates. The estimated results are available on request from the authors.

[31] Here we are not assuming that there is no sex selection, i.e. we allow p to take any value between 0 and 1.

Appendix E presents the OLS and IV regression results showing that individuals from households with more siblings are more likely to be in the mixed-sex family for both the rural and urban sample. We also estimated the OLS regression to see the non-linear relationship between sibling size and being in a mixed-sex family. The coefficients for each dummy variable indicating the number of siblings together with the 95% confidence intervals are plotted in Figure E1 of Appendix E. The figure shows some evidence that for both rural and urban samples, the probability of living in a mixed-sex family increases sharply between 1 to 3 siblings.

To provide some evidence that the gender gap in test scores is smaller for those from single-sex families than those from mixed-sex families, we estimate Equation 1 for individuals from mixed- and single-sex families, separately. The OLS and IV results for the rural sample are presented in Table 6.[32] The OLS results show that relative to single-sex families, the gender gap in both literacy and numeracy test scores for mixed-sex families is much larger. However, it is unfortunate that our IVs are very weak for the single-sex family sample and we are unable to obtain any statistically significant and sensible results for this group. Nevertheless, the results for the sample of individuals from the mixed-sex family (with strong IVs) are consistent with those obtained from the OLS regressions with somewhat larger coefficients.[33]

But are the results obtained here due to what we conjectured, that is, when these people were young, those (women or men) who lived in the single-sex families received more (less) investment and/or did less (more) housework and hence had more time to invest on their studies relative to their counterparts living in the mixed-sex families? The CFPS data we are using, however, do not provide respondents' retrospective early childhood conditions. To continue onto this line of inquiry, we resort to using a different data set: the China Health and Nutrition Survey (CHNS), to find some evidence. The CHNS is also a longitudinal survey initiated in 1989 and was continued up until 2015.[34] Although the CHNS did not ask about education spending on each child, and hence we are unable to investigate the gender gap in education investment for mixed- and single-sex families, it includes a time-use module. For the wave 1989-2000, we are able to obtain a consistent measure of children's daily time used to help with cooking and laundry. We use children aged 6 to 15 in these survey years (those who were born between 1983 and 1994) to see if the gender gap in doing housework for children from mixed-sex families is larger than that for children from single-sex families. We first confirm that there is a positive correlation between sibling size and whether the child is living in a mixed-sex family (see Table E1 of Appendix E). The OLS results are reported in Table 7.[35] The results confirm that the gender gap in helping out with housework is larger for

[32]The sample size for this analysis is smaller than the main sample because we use the information on the number of sisters/brothers to identify whether one is in a mixed-sex family or not, and the two variables have missing values (See Note 2 in Table 1)

[33]The results for the urban sample are reported in Table E2 of Appendix E. As can be seen there that the results are largely consistent with that for the rural sample.

[34]The survey was conducted by the Carolina Population Center at the University of North Carolina at Chapel Hill together with the National Institute for Nutrition and Health (NINH, former National Institute of Nutrition and Food Safety) at the Chinese Center for Disease Control and Prevention (CCDC).

[35]The CHNS does not publish the detailed county names for their sample and hence we are unable to estimate the IV results here.

children in mixed-sex families than in single-sex families for the rural sample, but the evidence in the urban sample is weak.

The above results provide some suggestive evidence to our conjecture that the coercive FPP may have played a role in increasing the share of girls (and boys) living in single-sex families, which in turn aided the narrowing/reversal of the gender gap in literacy/numeracy scores.

7 Conclusion

In the past 40 years, China has experienced a significant narrowing, even reversal, of the gender gap in literacy/numeracy test scores. This paper explored the potential causes of this phenomenon. Using the Quantity-Quality Tradeoff idea, the paper proposed that in a society where there is a deep-rooted son-preference, parental investment in children would be higher for sons than for daughters, creating a large gender gap in educational outcomes. When such a society enacts a coercive family planning policy, the sudden reduction of fertility could generate the narrowing/reversal of the gender gap in educational outcomes. In mixed-sex families, this occurs probably because the marginal product of education is higher for girls than for boys due to the lacking of investment in girls' education. More importantly, though, the narrowing of the gender gap in education is related to the increased probability of girls (and boys) living in single-sex families.

The paper empirically tested this hypothesis and established that the reduction in sibling size due to the coercive FPP indeed caused the reduction in the gender gap in literacy/numeracy test scores. A reduction of one sibling narrows the gender gap in the literacy and numeracy test scores by 21.4% and 14.8% of a standard deviation for the rural sample, respectively. The effect for the urban sample was estimated to be 6.5% and 4.0% of a standard deviation. Further investigation revealed that the sibling effect on the gender gap in test scores was larger in areas where the son-preference was more prominent. This paper also provided suggestive evidence to support the conjecture that an increasing proportion of girls (boys) was living in single-sex families due to the coercive FPP and that living in single-sex families gave girls advantages in the literacy/numeracy test over their counterparts living in the mixed-sex families, whereas the opposite was true for boys.

References

Almond, D., Li, H., and Zhang, S. (2019). Land Reform and Sex Selection in China. *Journal of Political Economy*, 127(2).

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*, 28(4).

Angrist, J. D. and Evans, W. N. (1998). Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *The American Economic Review*, 88(3).

Angrist, J. D., Lavy, V., and Schlosser, A. (2005). New Evidence on the Causal link between the Quantity and Quality of Children. *NBER Working Paper*, 11835.

Asadullah, M. N. and Chaudhury, N. (2009). Reverse Gender Gap in Schooling in Bangladesh: Insights from Urban and Rural Households. *Journal of Development Studies*, 45(8):1360–1380.

Bainster, J. (1987). *China's Changing Population*. Standford University Press.

Barcellos, S. H., Carvalho, L. S., and Lleras-Muney, A. (2014). Child Gender and Parental Investments in India: Are Boys and Girls Treated Differently? *American Economic Journal: Applied Economics*, 6(1).

Becker, G. S. and Lewis, H. G. (1973). On the Interaction between the Quantity and Quality of Children. *Journal of Political Economy*, 81(2).

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2005). The More the Merrier? The Effect of Family Size and Birth Order on Children's Education. *The Quarterly Journal of Economics*, 120(2).

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2010). Small family, smart family? family size and the iq scores of young men. *Journal of Human Resources*, 45(1):33–58.

Bo, S. (2018). Son preference, children's gender and parents' time allocation: evidence from China. *Applied Economics*, 50(45).

Booth, A., Fan, E., Meng, X., and Zhang, D. (2019). Gender Differences in Willingness to Compete: The Role of Culture and Institutions. *The Economic Journal*, 129.

Borooah, V. K. (2004). Gender bias among children in India in their diet and immunisation against disease. *Social Science & Medicine*, 58(9).

Borsi, M. T., Valerio Mendoza, O. M., and Comim, F. (2022). Measuring the provincial supply of higher education institutions in China. *China Economic Review*, 71(101724).

Bossavie, L. and Kanninen, O. (2018). What Explains the Gender Gap Reversal in Educational Attainment? *World Bank Policy Research Working Paper*, 8308.

Cai, Y. (2010). China's Below-Replacement Fertility: Government Policy or Socioeconomic Development? *Population and Development Review*, 36(3):419–440.

Cameron, L., Erkal, N., Gangadharan, L., and Meng, X. (2013). Little Emperors: Behavioral Impacts of China's One-Child Policy. *Science*, 339(22).

Cameron, L. and Meng, X. (2014). China's One Child Policy. *The New Palgrave Dictionary of Economics*.

Cantoni, D., Chen, Y., Yang, D., Yuchtman, N., and Zhang, J. (2017). Curriculum and ideology. *Journal of Political Economy*, 125(2):419–440.

Center for Population Studies, C. (1986). The major events of China's population activities. In *The China's Population Yearbook (1985)*, pages 1263–1288. Chinese Social Sciences Press, Beijing.

Chen, Y. and Fang, H. (2021). The long-term consequences of China's "Later, Longer, Fewer" campaign in old age. *Journal of Development Economics*, 151.

Chiappori, P.-A., Iyigun, M., and Weiss, Y. (2009). Investment in Schooling and the Marriage Market. *American Economic Review*, 99(5):1689–1713.

Cho, D. (2007). The role of high school performance in explaining women's rising college enrollment. *Economics of Education Review*, 26(4):450–462.

Choi, E. J. and Hwang, J. (2015). Child gender and parental inputs: No more son preference in korea? *The American Economic Review*, 105(5):638–643.

Clark, S. (2000). On Preference and Sex Composition of Children: Evidence from India. *Demography*, 37(1).

Coale, A. J. and Li, C. S. (1987). Basic data on fertility in the provinces of China, 1940-82. *East-West Population inst.*

Conley, D. and Glauber, R. (2006). Parental Educational Investment and Children's Academic Risk: Estimates of the Impact of Sibship Size and Birth Order from Exogenous Variation in Fertility. *Journal of Human Resources*.

Cunha, F., Heckman, J. J., and Schennach, S. M. (2019). Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica*, 78(3).

Dao, N. T., Dávila, J., and Greulich, A. (2021). The education gender gap and the demographic transition in developing countries. *Journal of Population Economics*, 34:431–474.

Das Gupta, M., Zhenghua, J., Bohua, L., Zhenming, X., Chung, W., and Ok, B.-H. (2003). Why is Son preference so persistent in East and South Asia? a cross-country study of China, India and the Republic of Korea. *Journal of Development Studies*.

Ebenstein, A. (2010). The "Missing Girls" of China and the Unintended Consequences of the One Child Policy. *Journal of Human Resources*, 45(1):87–115.

Ebenstein, A. and Leung, S. (2010). Son Preference and Access to Social Insurance: Evidence from China's Rural Pension Program. *Population and Development Review*, 36(1):47–70.

Edmonds, E. V. (2006). Understanding sibling differences in child labor. *Journal of Population Economics*, 19:795–821.

Evans, D. K., Akmal, M., and Jakiela, P. (2021). Gender gaps in education: The long view. *IZA Journal of Development and Migration*, 12(01).

Fan, Y., Yi, J., Yuan, Y., and Zhang, J. (2018). The glorified mothers of sons: Evidence from child sex composition and parental time allocation in rural China. *Journal of Economic Behavior & Organization*, 145.

Feeney, G. and Wang, F. (1993). Parity progression and birth intervals in China: The influence of policy in hastening fertility decline. *Population and Development Review*, 19(1):61–101.

Fitzsimons, E. and Malde, B. (2010). Empirically probing the quantity-quality model. *Journal of Population Economics*, 22.

Fortin, N. M., Oreopoulos, P., and Phipps, S. (2015). Leaving boys behind: Gender disparities in high academic achievement. *Journal of Human Resources*, 50(3):549–579.

Francesconi, M. and Heckman, J. J. (2016). Child Development and Parental Investment: Introduction. *The Economic Journal*, 126.

Garcia, J. L. (2022). Pricing children, curbing daughters: Fertility and sex ratio during China's one-child policy. *Journal of Human Resources*, pages 0820–11118.

Goldin, C., Katz, L. F., and Kuziemko, I. (2006). The Homecoming of American College Women: The Reversal of the College Gender Gap. *Journal of Economic Perspectives*, 20(4):133–156.

Goldstein, A. and Goldstein, S. (1990). China's labor force: The role of gender and residence. *Journal of Women and Gender Studies*, 1:87–118.

Goldstein, S. (1987). orms of mobility and their policy implications: Thailand and China compared. *Social Forces*, 65.

Guo, H., Hu, C., and Ding, X. (2022). Son preference, intrahousehold discrimination, and the gender gap in education in China. *International Review of Economics and Finance*, 79:324–339.

Guo, R., Li, H., Yi, J., and Zhang, J. (2018). Fertility, household structure, and parental labor supply: Evidence from China. *Journal of Comparative Economics*, 46:145–156.

Hatton, T. J. and Martin, R. M. (2010). The effect on stature of poverty, family size, and birth order: British children in the 1930s. *Oxford Economic Papers*, 62(1):157–184.

Huang, W., Lei, X., and Sun, A. (2021). Fertility restrictions and life-cycle outcomes: Evidence from the one-child policy in China. *Review of Economics and Statistics*, 103(4):694–710.

Jayachandran, S. (2015). The Roots of Gender Inequality in Developing Countries. *Annual Review of Economics*, 7.

Jayachandran, S. and Kuziemko, I. (2005). Why Do Mothers Breastfeed Girls Less than Boys? Evidence and Implications for Child Health in India. *Quarterly Journal of Economics*, 126.

Jones, G. W. and Ramchand, D. S. (2016). Closing the gender and socioeconomic gaps in educational attainment: A need to refocus. *Journal of International Development*, 28:953–973.

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

Li, B. and Zhang, H. (2017). Does population control lead to better child quality? Evidence from China's one-child policy enforcement. *Journal of Comparative Economics*, 45.

Li, H., Yi, J., and Zhang, J. (2011). Estimating the Effect of the One-Child Policy on the Sex Ratio Imbalance in China: Identification Based on the Difference-in-Differences. *Demography*, 48.

Li, H., Zhang, J., and Zhu, Y. (2008). The Quantity-Quality Trade-off of Children in a Developing country: Identification using Chinese Twins. *Demography*, 45(1).

Li, J. and Cooney, R. S. (1993). Son Preference and the One Child Policy in China: 1979-1988. *Population Research and Policy Review*, 12(3).

Li, J. and Lavely, W. (2003). Village context, women's status, and son-preference among rural Chinese women. *Rural Sociology*, 68(1):87–106.

Liang, Y. and Dong, Z. (2019). Has education led to secularization? Based on the study of compulsory education law in China. *China Economic Review*, 54.

Lin, T.-c. and Adserà, A. (2013). Son preference and children's housework: The case of india. *Population Research and Policy Review*, 32(4):553–584.

Liu, H. (2014). The quality-quantity trade-off: evidence from the relaxation of China's one-child policy. *Journal of Population Economics*, 27:565–602.

Maurer-Fazio, M., Connelly, R., Chen, L., and Tang, L. (2011). Childcare, eldercare, and labor force participation of married women in urban China, 1982-2000. *Journal of Human Resources*, 46(2):261–294.

McElroy, M. and Yang, D. T. (2000). Carrots and Sticks: Fertility Effects of China's Population Policies. *American Economic Review*, 90(2).

Meng, X. (2000). *Labour Market Reform in China*. Cambridge: Cambridge University Press.

Meng, X. (2012). Labor market outcomes and reforms in China. *Journal of Economic Perspective*, 26(4):75–102.

Milwertz, C. N. (1997). *Accepting Population Control: Urban Chinese Women and the One-child Family Policy*. Curzon Press.

Niida, N. (1964). Land reform and new marriage law in China. *Developing Economies*, 2(1).

Peng, X. (1989). Major determinants of China's fertility transition. *The China Quarterly*, 117:1–37.

Peng, X. (1990). China's population control and the reform in the eighties. *CHINA REPORT*, 30(29):5.

Peng, X. (1991). *Demographic Transition in China*. Oxford: Clarendon Press.

Peter, N., Lundborg, P., Mikkelsen, S., , and Webbink, D. (2018). The effect of a sibling's gender on earnings and family formation. *Labour Economics*, 54.

Potter, S. H. (1983). The position of peasants in modern China's social order. *Modern China*, 9(4):465–499.

Qian, N. (2017). Quantity-Quality and the One Child Policy: The Positive Effect of Family Size on School Enrollment in China. *Gender and Development*.

Riphahn, R. T. and Schwientek, C. (2015). What drives the reversal of the gender education gap? Evidence from Germany. *Applied Economics*, 47(53):5748–5775.

Rosenzweig, M. R. and Wolpin, K. I. (1980). Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment. *Econometrica*, 48(1).

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

Scharping, T. (2003). *Birth control in China 1949-2000: population policy and demographic development*. Routledge.

Vu, T. M. (2014). Are daughters always the losers in the chore war? evidence using household data from vietnam. *The Journal of Development Studies*, 50(4):520–529.

Wan, Y. (2006). Expansion of Chinese Higher Education Since 1998: Its Causes and Outcomes. *Asia Pacific Education Review*, 7(1):19–31.

Whyte, M. K., Feng, W., and Cai, Y. (2015). Challenging Myths About China's One-Child Policy. *The China Journal*, 74.

Wu, X. and Zhang, Z. (2010). Changes in educational inequality in China, 1990-2005: Evidence from the population census data. *Research in the Sociology of Education*, 17:123–152.

Wu, Y. (2012). Gender gap in educational attainment in urban and rural China. *Society (in Chinese)*, 32.

Xie, Y. and Lu, P. (2015). The Sampling Design of the China Family Panel Studies (CFPS). *Chinese Journal of Sociology*, 1(14).

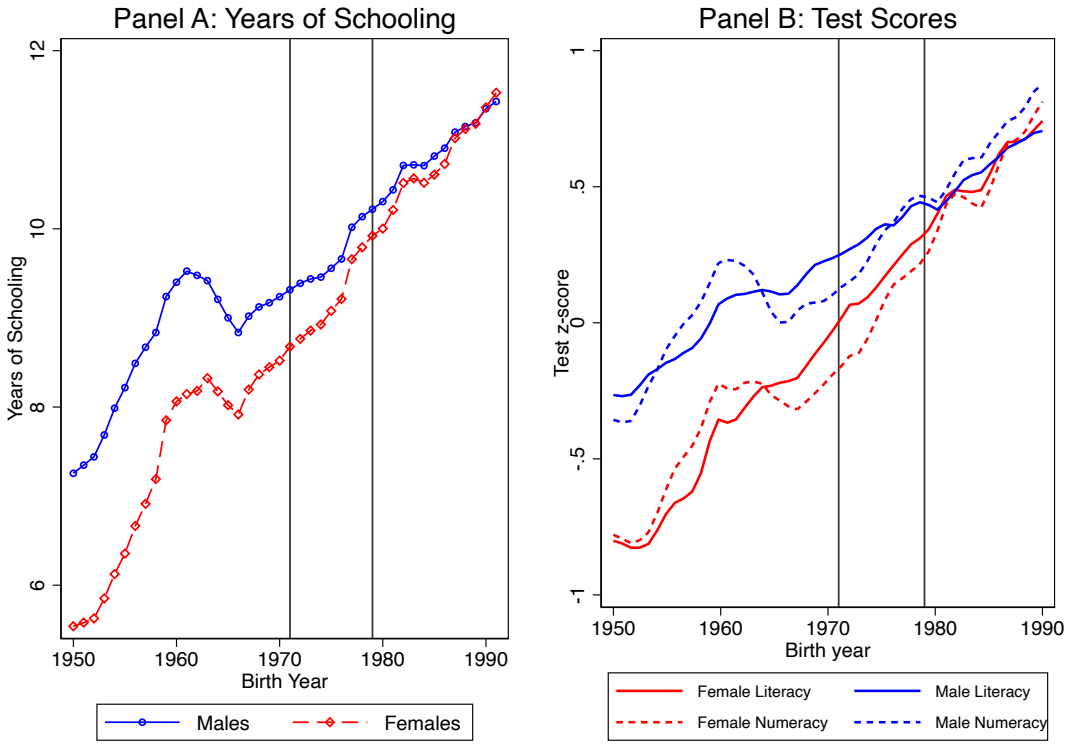
Yao, Y. and You, W. (2016). Half Sky over China: Women's Political Participation and Sex. China Center for Economic Research: Working Paper Series No. E2016009.

Yao, Y. and You, W. (2018). Women's political participation and gender gaps of education in China: 1950-1990. *World Development*, 106:220–237.

Yeung, W.-J. J. (2013). Higher Education Expansion and Social Stratification in China. *Chinese Sociological Review*, 45:54–80.

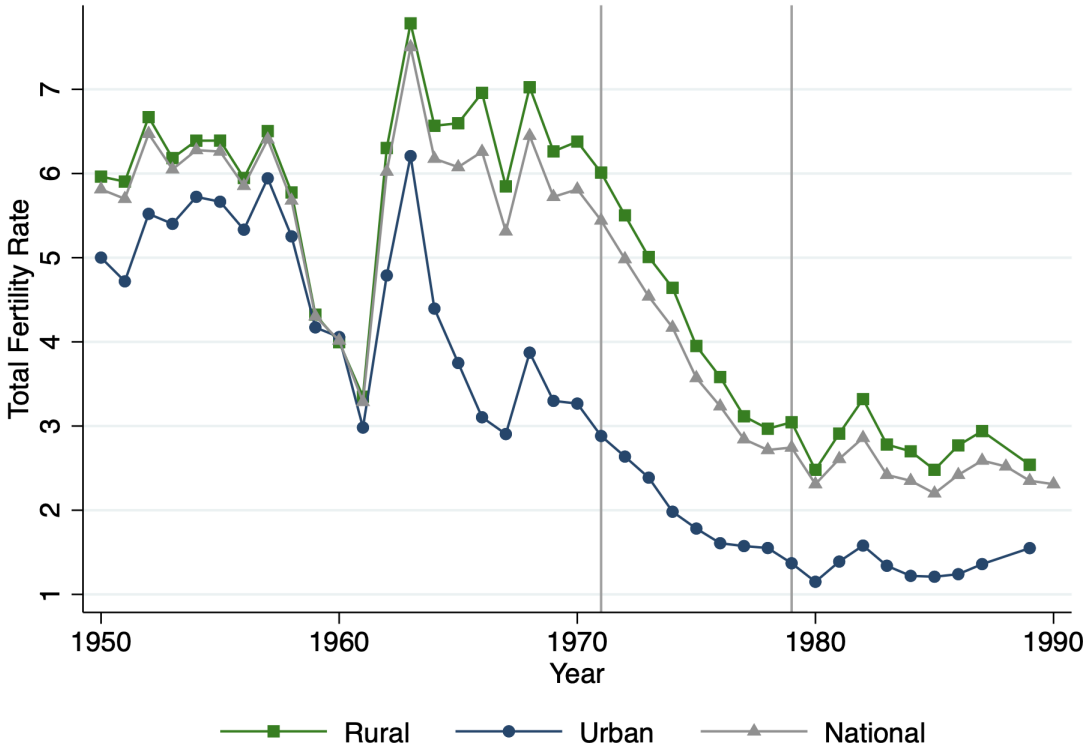
Zhang, J. (2017). The Evolution of China's One-Child Policy and Its Effects on Family Outcomes. *The Journal of Economic Perspectives*, 31(1):141–159.

Figure 1: Gender differences in years of schooling and literacy/numeracy test scores



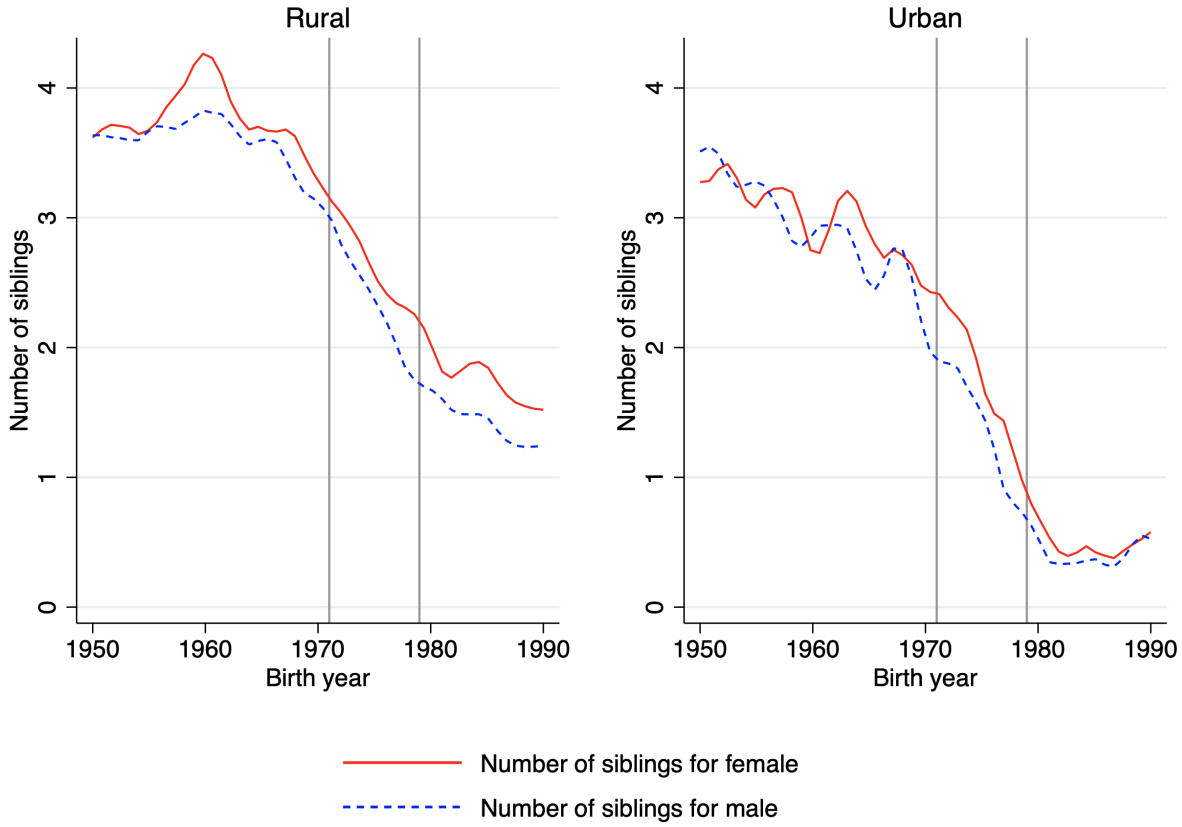
Note: Grey lines correspond to 1971 and 1979, which mark the launch of the campaign 'One child isn't too few, two are just fine, and three are too many' and the introduction of 'One-Child per Couple' policy, respectively

Figure 2: China's Total Fertility: 1950-1990



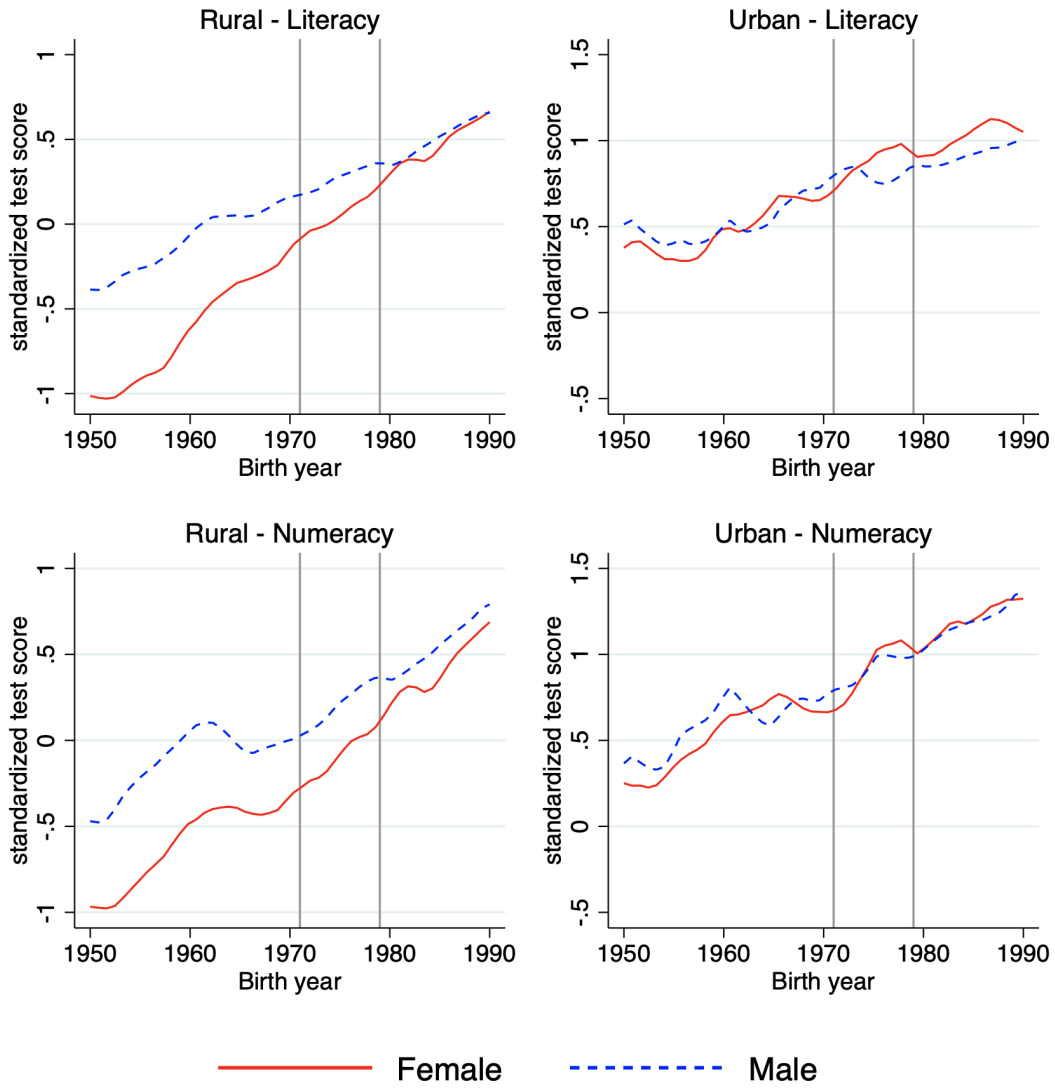
Note: Grey lines correspond to 1971 and 1979, which mark the launch of the campaign 'One child isn't too few, two are just fine, and three are too many' and the introduction of 'One-Child per Couple' policy, respectively

Figure 3: Number of siblings by gender and hukou status



Note: Grey lines correspond to 1971 and 1979, which mark the launch of the campaign 'One child isn't too few, two are just fine, and three are too many' and the introduction of 'One-Child per Couple' policy, respectively; Kernel-weighted local polynomial smoothing with bandwidth 0.8

Figure 4: Gender differences in the standardised cognitive test scores by hukou status



Note: Grey lines correspond to 1971 and 1979, which mark the launch of the campaign 'One child isn't too few, two are just fine, and three are too many' and the introduction of 'One-Child per Couple' policy, respectively; Standardized test scores with zero mean and one standard deviation; Kernel-weighted local polynomial smoothing bandwidth 0.8

Figure 5: Son-preference and sibling effects on gender gap in z-scores, CFPS

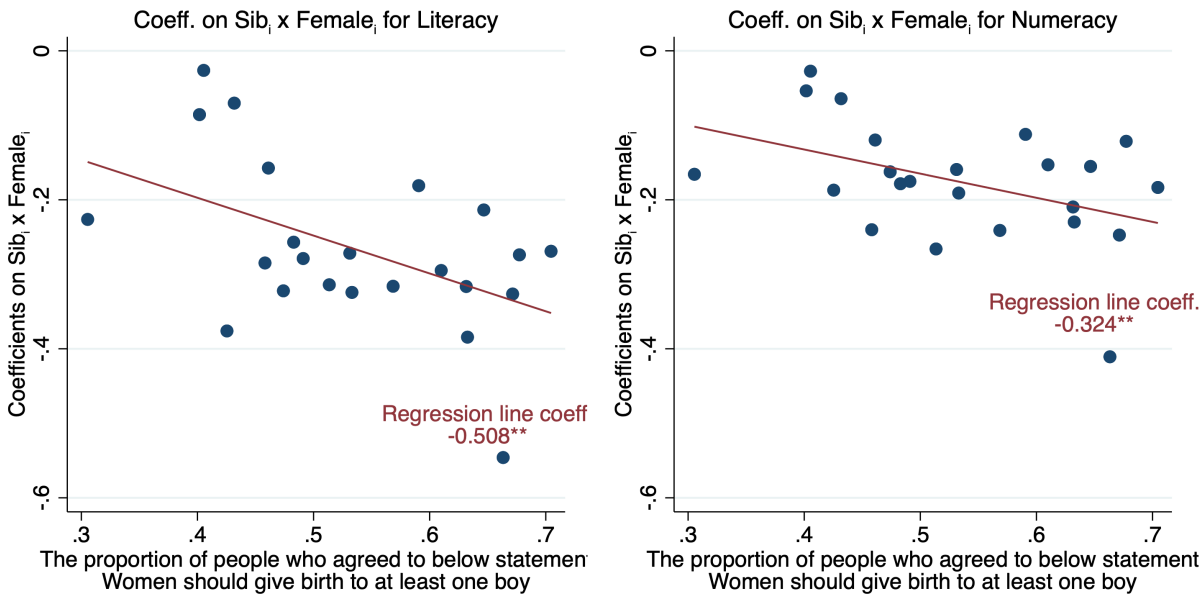


Figure 6: Son-preference and sibling effects on gender gap in z-scores, SWSC

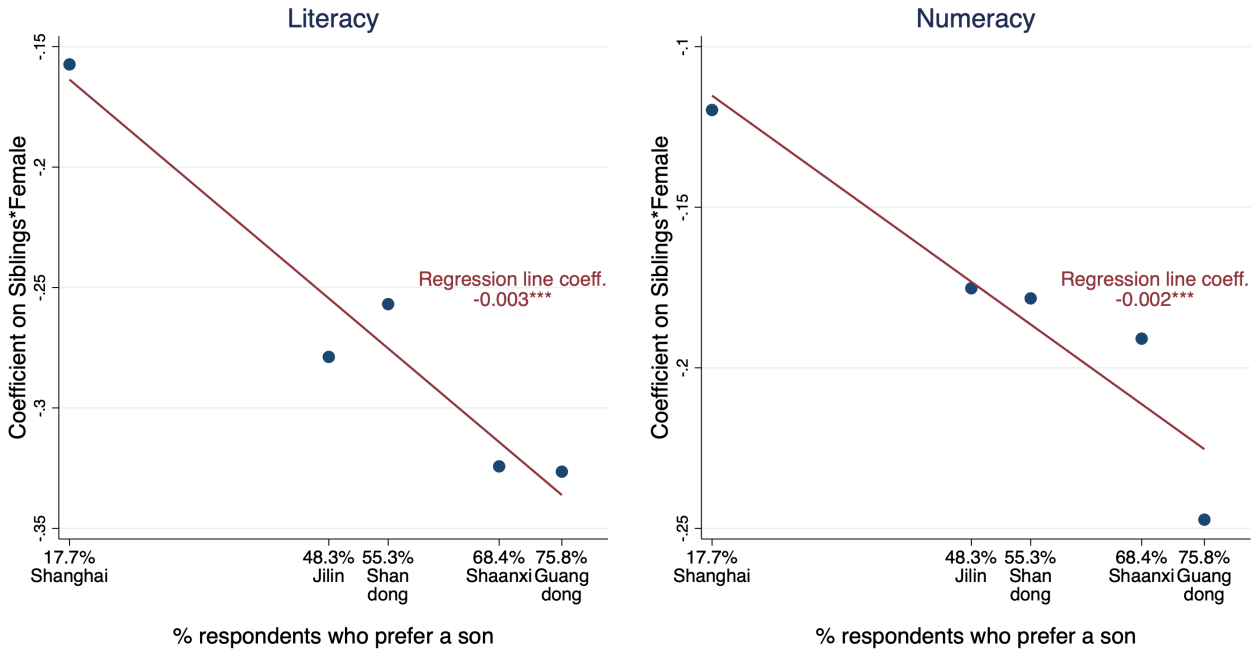


Figure 7: Share of the sample growing up in single-sex vs. mixed-sex families, by gender

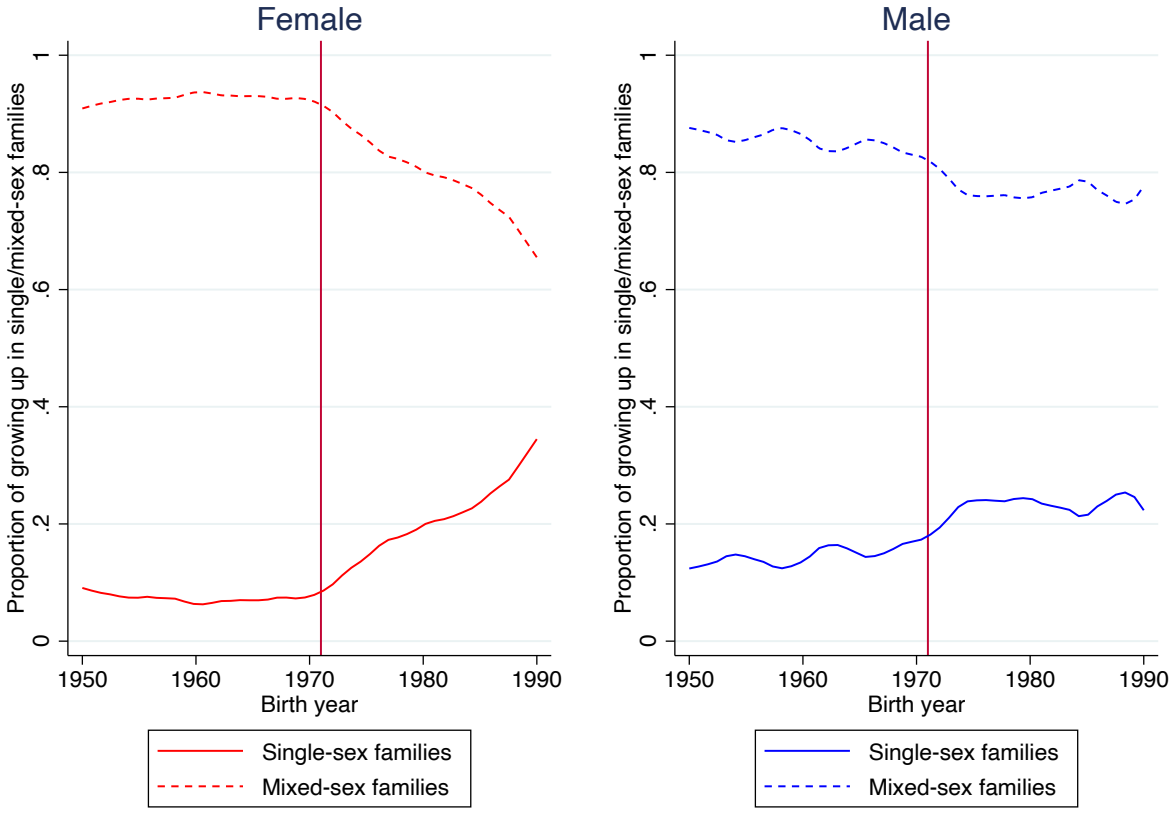


Table 1: Summary statistics

	Rural sample			Urban sample		
	Mean Female	Mean Male	Diff Female-Male	Mean Female	Mean Male	Diff Female-Male
Literacy test raw score (out of 34)	15.17	19.04	-3.87***	24.78	24.71	0.07
	(10.57)	(8.69)		(6.72)	(6.15)	
Numeracy test raw score (out of 24)	8.25	10.74	-2.49***	15.07	15.38	-0.31**
	(6.12)	(5.45)		(4.75)	(4.48)	
Literacy z-score	-0.22	0.14	-0.36***	0.68	0.68	0.01
	(0.98)	(0.80)		(0.62)	(0.56)	
Numeracy z-score	-0.25	0.12	-0.38***	0.77	0.81	-0.05**
	(0.93)	(0.82)		(0.71)	(0.67)	
Age	40.84	40.28	0.56***	41.06	40.46	0.60
	(11.42)	(11.60)		(12.01)	(11.85)	
Years of schooling	6.02	7.89	-1.87***	11.35	11.53	-0.18
	(4.74)	(4.13)		(3.60)	(3.33)	
Number of siblings	3.07	2.79	0.28***	2.17	1.96	0.21***
	(1.85)	(1.87)		(1.90)	(1.81)	
Number of sisters[2)]	1.51	1.53	-0.02	1.06	1.04	0.03
	(1.33)	(1.29)		(1.23)	(1.19)	
Number of brothers[2)]	1.65	1.46	0.19***	1.17	0.99	0.17***
	(1.23)	(1.30)		(1.21)	(1.15)	
One child	0.05	0.09	-0.03***	0.24	0.28	-0.03**
	(0.22)	(0.28)		(0.43)	(0.45)	
Birth order	2.54	2.46	0.08***	2.16	2.11	0.05
	(1.59)	(1.57)		(1.48)	(1.43)	
Parental schooling years	3.07	3.10	-0.03	6.20	6.26	-0.06
	(3.26)	(3.20)		(4.16)	(4.25)	
Current urban hukou registration	0.14	0.16	-0.02***	0.98	0.99	-0.00
	(0.35)	(0.37)		(0.13)	(0.12)	
Han Chinese	0.91	0.90	0.00	0.96	0.96	-0.00
	(0.29)	(0.29)		(0.20)	(0.19)	
Observations	10399	10496		1836	1885	

Note:

1) Years of schooling when one attended the test. For example, for those who attended both in 2010 and 2014, mean years of schooling between 2010 and 2014 is used;

2) The sample size for the number of sisters and the number of brothers is 18321 for rural and 3397 for urban samples due to missing values;

3) Standard deviation in parentheses;

4) * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$;

Table 2: OLS results

Dependent var.	Rural sample			Urban sample		
Literacy test z-score	(1)	(2)	(3)	(4)	(5)	(6)
Siblings	-0.013	0.020***	0.031***	-0.030***	-0.003	0.014
	(0.011)	(0.005)	(0.006)	(0.010)	(0.010)	(0.010)
Siblings × Female	-0.065***	-0.070***	-0.071***	-0.017*	-0.019**	-0.026***
	(0.007)	(0.006)	(0.007)	(0.010)	(0.009)	(0.010)
Female	-0.137***	-0.140***	-0.147***	0.060***	0.055**	0.057**
	(0.029)	(0.027)	(0.028)	(0.022)	(0.022)	(0.022)
Parental schooling years			0.050***			0.030***
			(0.002)			(0.004)
Han Chinese			0.109			0.076
			(0.069)			(0.070)
Birth order index			-0.019***			-0.022
			(0.007)			(0.016)
Birth order index × Female			0.046***			0.035*
			(0.009)			(0.019)
Lifetime BCR			0.001			0.002
			(0.001)			(0.003)
Birth year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Birth county dummies	No	Yes	Yes	No	Yes	Yes
Birth county dummies*time	No	No	Yes	No	No	Yes
Test year indicators	No	No	Yes	No	No	Yes
Missing birth county indicators	No	No	Yes	No	No	Yes
Observations	20895	20895	20895	3721	3721	3721
Adjusted R^2	0.248	0.402	0.433	0.173	0.289	0.332
Joint significance of Sibling and Sibling*Female (P value)	0.000	0.000	0.000	0.000	0.016	0.023
Dependent var.	Rural sample			Urban sample		
Numeracy test z-score	(1)	(2)	(3)	(4)	(5)	(6)
Siblings	-0.021**	0.009*	0.023***	-0.045***	-0.014	0.002
	(0.009)	(0.005)	(0.005)	(0.013)	(0.010)	(0.011)
Siblings × Female	-0.053***	-0.057***	-0.058***	-0.015	-0.016*	-0.020*
	(0.006)	(0.005)	(0.005)	(0.010)	(0.010)	(0.011)
Female	-0.192***	-0.195***	-0.198***	0.005	-0.005	-0.008
	(0.028)	(0.026)	(0.026)	(0.026)	(0.025)	(0.026)
Parental schooling years			0.062***			0.039***
			(0.003)			(0.004)
Han Chinese			0.079			0.091
			(0.066)			(0.092)
Birth order index			-0.015**			-0.007
			(0.006)			(0.019)
Birth order index × Female			0.033***			-0.001
			(0.008)			(0.020)
Lifetime BCR			-0.001			-0.004
			(0.001)			(0.003)
Birth year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Birth county dummies	No	Yes	Yes	No	Yes	Yes
Birth county dummies*time	No	No	Yes	No	No	Yes
Test year indicators	No	No	Yes	No	No	Yes
Missing birth county indicators	No	No	Yes	No	No	Yes
Observations	20895	20895	20895	3721	3721	3721
Adjusted R^2	0.248	0.378	0.427	0.219	0.346	0.390
Joint significance of Sibling and Sibling*Female (P value)	0.000	0.000	0.000	0.000	0.018	0.132

Note:

1) Robust Standard Errors (SEs) are presented in parentheses;

2) SEs are clustered at the county levels;

3) * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$;

Table 3: First stage results

	Rural sample		Urban sample	
	(1)	(2)	(3)	(4)
Dependent var.	Siblings	Siblings*Female	Siblings	Siblings*Female
FPP	-0.484***	-0.031	-0.401***	0.054
	(0.045)	(0.028)	(0.065)	(0.056)
FPP × Female	0.044***	-0.408***	0.008	-0.498***
	(0.009)	(0.014)	(0.018)	(0.023)
Lifetime BCR	0.015	0.011*	0.020	0.003
	(0.010)	(0.006)	(0.014)	(0.011)
All other controls in Equation (1)	Yes	Yes	Yes	Yes
Observations	20895	20895	3721	3721
Kleibergen-Paap Wald rk F statistics	53.127		18.98	

Note:

1) Robust Standard Errors (SEs) are presented in parentheses and clustered at the county level; 2) Controls include birth county fixed effects, birth year fixed effects, birth order index, an interaction of birth order index with female indicator, Han Chinese indicator, parental mean schooling years, county information source indicators, test year indicators, and an interaction between time linear trend variable and birth county fixed effect; 3) * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$;

Table 4: IV results

	Rural sample				Urban sample			
	Literacy		Numeracy		Literacy		Numeracy	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Siblings	0.031***	0.007	0.023***	-0.024	0.014	0.119*	0.002	0.133**
	(0.006)	(0.031)	(0.005)	(0.029)	(0.010)	(0.066)	(0.011)	(0.064)
Siblings × Female	-0.071***	-0.214***	-0.058***	-0.148***	-0.026***	-0.065***	-0.020*	-0.040***
	(0.007)	(0.015)	(0.005)	(0.014)	(0.010)	(0.013)	(0.011)	(0.015)
Female	-0.147***	0.295***	-0.198***	0.084*	0.057**	0.125***	-0.008	0.017
	(0.028)	(0.047)	(0.026)	(0.047)	(0.022)	(0.028)	(0.026)	(0.036)
Parental schooling years	0.050***	0.048***	0.062***	0.060***	0.030***	0.035***	0.039***	0.046***
	(0.002)	(0.003)	(0.003)	(0.003)	(0.004)	(0.005)	(0.004)	(0.005)
Birth order index	-0.019***	-0.025***	-0.015**	-0.017**	-0.022	-0.059**	-0.007	-0.052*
	(0.007)	(0.007)	(0.006)	(0.007)	(0.016)	(0.026)	(0.019)	(0.030)
Birth order index × Female	0.046***	0.061***	0.033***	0.041***	0.035*	0.055***	-0.001	0.015
	(0.009)	(0.010)	(0.008)	(0.008)	(0.019)	(0.019)	(0.020)	(0.021)
Han Chinese	0.109	0.117*	0.079	0.085	0.076	0.091	0.091	0.107
	(0.069)	(0.067)	(0.066)	(0.064)	(0.070)	(0.072)	(0.092)	(0.095)
Lifetime BCR	0.001	-0.005**	-0.001	-0.007***	0.002	0.005	-0.004	0.001
	(0.001)	(0.003)	(0.001)	(0.002)	(0.003)	(0.005)	(0.003)	(0.004)
Birth year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Birth county dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Birth county dummies*time	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Test year indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Missing birth county indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	20895	20895	20895	20895	3721	3721	3721	3721
Kleibergen-Paap Wald F statistics		53.13		53.13		18.98		18.98
P value for Joint significance test								
Sibling and Sibling*Female	0.000	0.000	0.000	0.000	0.023	0.000	0.132	0.000

Note:

1) The outcome variables are test z-scores with zero mean and a standard deviation; 2) Robust Standard Errors (SEs) are presented in parentheses and clustered at the county level; 3) Control variables include the same set of covariates in Table 3; 4) The instruments are the same as those used in Table 3; 5) * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$;

Table 5: Robustness check for numeracy test

Panel A	Rural sample					Urban sample				
Dependent var.	Socio-Econ Condition	Edu reform	Born after 1965	Excl. no B-county	Incl. Eduy	Socio-Econ Condition	Edu reform	Born after 1965	Excl. no B-county	Incl. Eduy
Numeracy test z-score	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Siblings	0.009	0.017	-0.189**	0.004	0.008	0.130*	0.118*	-0.019	0.055	0.084
	(0.030)	(0.031)	(0.085)	(0.033)	(0.022)	(0.068)	(0.066)	(0.126)	(0.081)	(0.056)
Siblings × Female	-0.210***	-0.214***	-0.151***	-0.222***	-0.093***	-0.064***	-0.064***	-0.096**	-0.068***	-0.040***
	(0.015)	(0.015)	(0.025)	(0.016)	(0.009)	(0.015)	(0.013)	(0.038)	(0.018)	(0.011)
Female	0.281***	0.291***	0.239***	0.309***	0.173***	0.122***	0.125***	0.173***	0.165***	0.092***
	(0.048)	(0.047)	(0.078)	(0.053)	(0.028)	(0.030)	(0.028)	(0.044)	(0.038)	(0.026)
Death rate per mil	-0.001					0.006				
	(0.001)					(0.004)				
Log industrial output p.c.	0.007					-0.057*				
	(0.014)					(0.034)				
Log agricultural output p.c.	-0.006					0.011				
	(0.035)					(0.041)				
Log public health expenditure p.c.	-0.011					-0.029				
	(0.019)					(0.025)				
9-year compulsory		0.083**					-0.074			
		(0.038)					(0.065)			
Univ. expansion × HEI growth		0.168***					-0.016			
		(0.050)					(0.061)			
Years of schooling					0.135***					0.101***
					(0.002)					(0.004)
Observations	19966	20895	13242	17842	20895	3509	3721	2153	2409	3721
Kleibergen-Paap Wald F statistics	51.54	51.21	16.41	50.30	52.52	17.85	18.95	5.15	8.53	18.72
P value for Joint significance test										
Sibling and Sibling*Female	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.033	0.000	0.000

Panel B	Rural sample					Urban sample				
Dependent var.	Max score	Raw score	2010 only	2014 only	2018 only	Max score	Raw score	2010 only	2014 only	2018 only
Numeracy test z-scores	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Siblings	-0.014	-0.167	-0.019	0.006	-0.030	0.097	0.878**	0.126**	0.252***	0.116
	(0.030)	(0.190)	(0.031)	(0.034)	(0.031)	(0.067)	(0.424)	(0.060)	(0.089)	(0.135)
Siblings × Female	-0.164***	-0.970***	-0.150***	-0.167***	-0.099***	-0.035**	-0.258***	-0.037**	-0.063***	-0.047*
	(0.015)	(0.093)	(0.016)	(0.015)	(0.017)	(0.017)	(0.099)	(0.016)	(0.022)	(0.026)
Female	0.106**	0.542*	0.057	0.132***	0.005	0.003	0.107	0.018	0.055	-0.043
	(0.050)	(0.311)	(0.052)	(0.051)	(0.064)	(0.040)	(0.240)	(0.037)	(0.057)	(0.072)
Observations	20895	20895	19478	15940	12387	3721	3721	3542	2294	1445
Kleibergen-Paap Wald F statistics	53.13	53.13	54.47	56.57	46.42	18.98	18.98	19.25	13.24	5.49
Sibling and Sibling*Female	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.135

Note:

1) Panel A: Each column checks the robustness of the main IV result in Table 4 by either adding new variables to or restring the samples in a certain way; Columns 1 and 5 add five county-level variables measuring local environmental conditions at one's birth, including death rate, log industrial output per capita, log agricultural output per capita and log local government expenditure on health per capita by year. Columns 2 and 6 add two variables capturing the potential heterogeneous effects of educational reforms. '9-year compulsory' is 1 if one was under 15 (aged between 0 and 14) at the year of the provincial law introduction and 0 otherwise. 'Univ. expansion' is 1 if one started tertiary education after the university expansion in 1999 and 0 otherwise, and 'HEI growth' is the growth rate of the number of universities between 2001 and 2017 at the provincial level provided by Borsi et al. (2022). Columns 3 and 7 include individuals born after 1965 only; Columns 4 and 8 exclude migrants to the survey counties; Columns 5 and 10 add individual years of schooling to the main regression;

2) Panel B: 'Max' means the outcome variable is the individual maximum standardised word test score among the three years (2010, 2014 and 2018); 'Raw' means the outcome variable is the individual average raw word test score; '2010 only' means the outcome variable is 2010 word test standardised score; '2014 only' means the outcome variable is 2014 word test standardised score;

3) The outcome variables are test z-scores with zero mean and a standard deviation;

4) Robust Standard Errors (SEs) are presented in parentheses;

5) SEs are clustered at the county level;

6) Control variables include the same set of covariates in Table 3;

7) The instruments are the same as those used in Table 3;

8) * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$;

Table 6: Gender gap in the mixed- and signle-sex families, test scores – Rural

Dependent variable:	Literacy - OLS		Numeracy - OLS		Literacy - IV		Numeracy - IV	
	Mixed	Single	Mixed	Single	Mixed	Single	Mixed	Single
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Siblings	0.020***	0.026*	0.006	0.016	0.026	-0.771	-0.029	-0.778
	(0.007)	(0.014)	(0.006)	(0.014)	(0.033)	(2.950)	(0.030)	(2.677)
Siblings × Female	-0.056***	0.004	-0.035***	-0.013	-0.174***	-3.316	-0.078***	-3.035
	(0.008)	(0.018)	(0.007)	(0.016)	(0.020)	(7.539)	(0.018)	(6.858)
Female	-0.250***	-0.264***	-0.325***	-0.275***	0.176**	4.549	-0.169**	4.114
	(0.037)	(0.040)	(0.035)	(0.038)	(0.074)	(11.046)	(0.069)	(10.051)
Observations	14713	3608	14713	3608	14713	3608	14713	3608
Kleibergen-Paap Wald rk F stat					58.010	0.073	58.010	0.073

Note:
1) The outcome variables are test z-scores with zero mean and a standard deviation;
2) Robust Standard Errors (SEs) are presented in parentheses;
3) SEs are clustered at the county level;
4) Control variables include the same set of covariates in Table 3;
5) The instruments are the same as those used in Table 3;
6) * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$;

Table 7: Gender gap in the mixed- and single-sex families, housework

	Rural Mixed-sex	Rural Single-sex	Urban Mixed-sex	Urban Single-sex
	(1)	(2)	(3)	(4)
Siblings	-6.092***	-6.296***	-3.669**	-7.509
	(1.239)	(1.676)	(1.750)	(8.953)
Siblings × Female	12.781***	0.516	13.004***	14.870
	(2.796)	(2.681)	(4.108)	(14.105)
Female	2.174	17.530***	9.216	7.301
	(3.200)	(6.040)	(8.987)	(11.907)
Mother's years of schooling	-1.260	-1.573*	-0.740	-2.771
	(0.805)	(0.804)	(1.265)	(1.807)
Birth order	2.359***	1.879**	5.090***	0.213
	(0.690)	(0.754)	(1.625)	(3.380)
Birth order × Female	-5.926***	-2.290	-7.559**	-6.761
	(1.884)	(1.991)	(3.550)	(8.194)
Age fixed effect	Yes	Yes	Yes	Yes
Survey wave effect	Yes	Yes	Yes	Yes
Residential community fixed effect	Yes	Yes	Yes	Yes
Observations	9897	3187	2426	1066
Adjusted R^2	0.082	0.183	0.086	0.133

Note:
1) The outcome variables are housework time (minutes per day);
2) 1989, 1991, 1993, 1997 and 2000 waves of the China Health and Nutrition Survey are used;
3) Robust Standard Errors (SEs) are presented in parentheses;
4) SEs are clustered at the residential community level;
5) * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$;

APPENDIX A Main sample selection

Table A1: Main working sample construction

(Unit: Individual)	**Original obs.**	**Newly added obs.[1)]**		**Working sample**
	2010 [1]	2014 [2]	2018 [3]	[1]+[2]+[3]
Total sample				
Original sample	33598	11529	7951	53078
Born 1950-1990	25241	7480	1808	34529
hukou info. at 3 available	25208	5527	976	31711
Birth county info. available	23525	5141	915	29581
With cognitive test	23516	4341	514	28371
Sibling info. available	23276	1671	119	25066
Other info. available	23032	1480	104	24616
Rural sample				
Born 1950-1990 and hukou info. at 3 available	21171	4948	884	27003
Birth county info. available	19851	4603	835	25289
With cognitive test	19846	3857	462	24165
Sibling info. available	19667	1478	108	21253
Other info. available	19486	1314	95	20895
Urban sample				
Born 1950-1990 and hukou info. at 3 available	4037	579	92	4708
Birth county info. available	3674	538	80	4292
With cognitive test	3670	484	52	4206
Sibling info. available	3609	193	11	3813
Other info. available	3546	166	9	3721

Note:

1) Newly added observations include respondents who turned 16 years old (the minimum age requirement for the adult survey)

APPENDIX B Cognitive ability measures in the CFPS

The China Family Panel Studies (CFPS) in its 2010, 2014, and 2018 waves, tested respondents aged over 10 on their literacy and numeracy ability. The literacy test consists of 34 Chinese characters drawn from the language textbooks used in primary and secondary schools. The test seeks to measure one's knowledge of characters by checking whether one can correctly recognise those characters. The numeracy test has a total of 24 questions, including addition, subtraction, multiplication, division, exponents, logarithms, trigonometric functions, sequence, permutation and combination. Both the literacy and numeracy tests are designed in ascending order in terms of difficulty, i.e. the first question is the simplest one and the last is the most difficult one.

The enumerators were given 8 sets of test papers for both literacy and numeracy with the same difficulty level. At the time of the test, a randomly selected set was given to each respondent.[36] The Chinese characters are presented in cards and the respondent is asked to read them aloud. The math questions are also presented in cards and the respondent is asked to solve and answer.

To conduct the test more efficiently, the test procedure is designed such that not every respondent begins with the easiest question. The respondents were grouped into one of three educational groups - 1) primary school graduation or below; 2) junior high school graduation; and 3) senior high school graduation or above. The tests start with a question suitable for one's education level. In the literacy test, Group 1 starts with the easiest question (i.e., Question 1), Group 2 starts from Question 9, and Group 3 starts from Question 21; whereas for the numeracy test, Group 1 starts with Question 1, Group 2 starts from Question 13, while Group 3 starts from Question 19.

However, the test procedure changed somewhere between 2010 and 2014/2018. In 2010, a respondent with a certain level of education was assumed to be able to answer all the questions for lower education groups correctly. Thus, the final score in 2010 is calculated based on the number of correct answers given for one's own education group plus the total number of questions given to the lower education groups. For example, if one is from Group 3 and in the literacy test he/she was able to correctly read three characters (say, Question 21, Question 22, and Question 23), the final score would be $20 + 3 = 23$. In 2014 and 2018, the procedure was changed. When the first question for the education group could not be correctly answered, the interviewer would next present the first question for the next lower education group. For example, for a respondent in Group 3, in the literacy test, she would be shown Question 21. If she were unable to answer it, her next question would be Question 9, which is the first question for the next lower education group - Group 2. If she answers Question 9 correctly then she would be asked to continue to answer Question 10, 11 and so on so forth until she fails to correctly answer 3 questions in a row. This means that the lowest possible score for a Group 3 respondent is 0 in 2014 and 2018, and 20 in 2010.

To resolve this problem, we opted to use the mean standardised test score for three years as discussed in the main text. Table B1 presents, for our main sample, the distribution of test scores availability by testing years. Figures B1 and B2, below present the raw literacy and numeracy test scores by year and rural/urban samples, respectively. Figure B3 shows the raw test scores distribution by survey year and education groups, where as B4 exhibits our final outcome variables

[36] These test question sets are not publicly available.

— mean literacy and numeracy z-scores.

Table B1: Test score availability by testing years (main sample)

Participated in	Freq.	Percent
2010 only	4,703	19.11%
2014 only	868	3.53%
2018 only	207	0.84%
2010 and 2014	5,213	21.18%
2010 and 2018	1,472	5.90%
2014 and 2018	521	2.01%
2010, 2014 and 2018	11,632	47.25%
Total	24,616	100%

Figure B1: Literacy test raw scores of the main sample

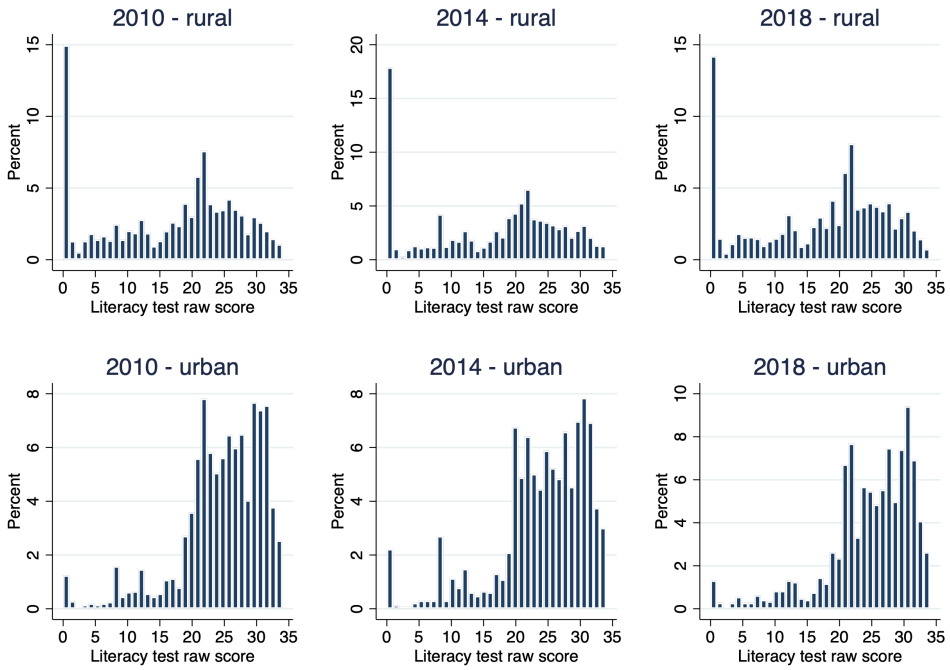


Figure B2: Numeracy test raw scores of the main sample

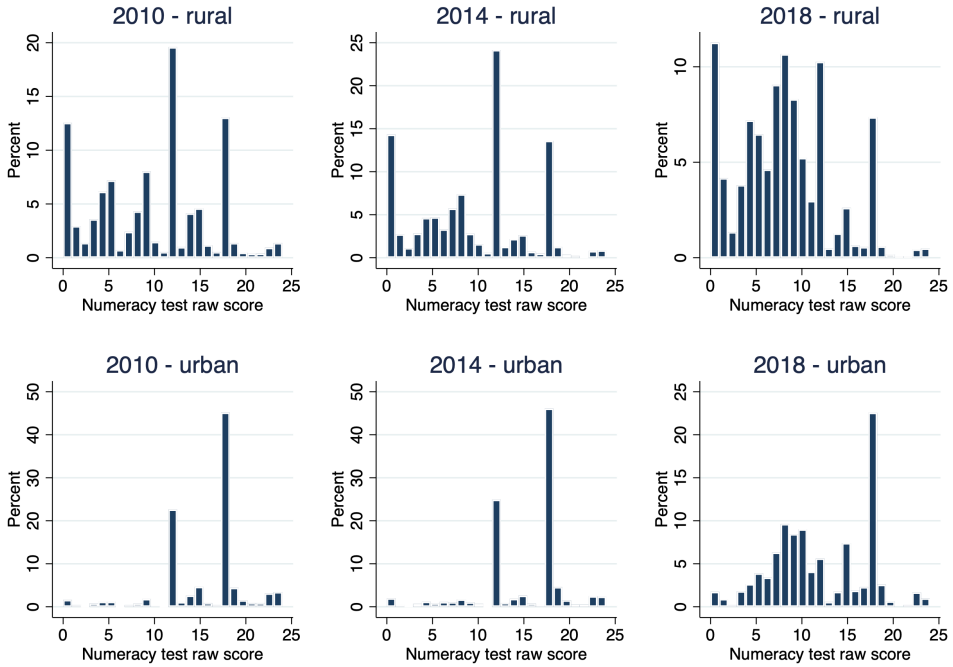


Figure B3: Raw test score distribution by test year and education group

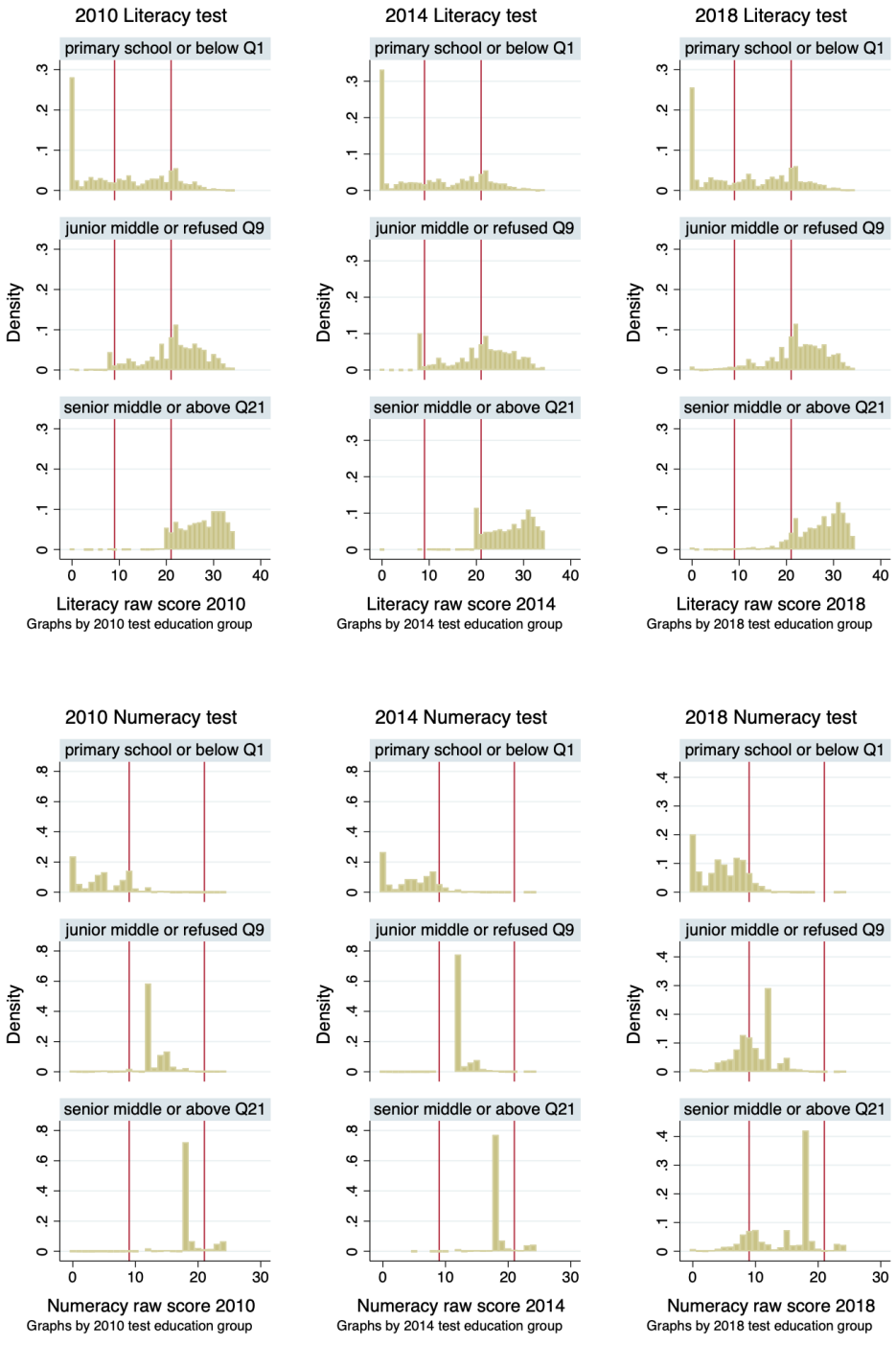
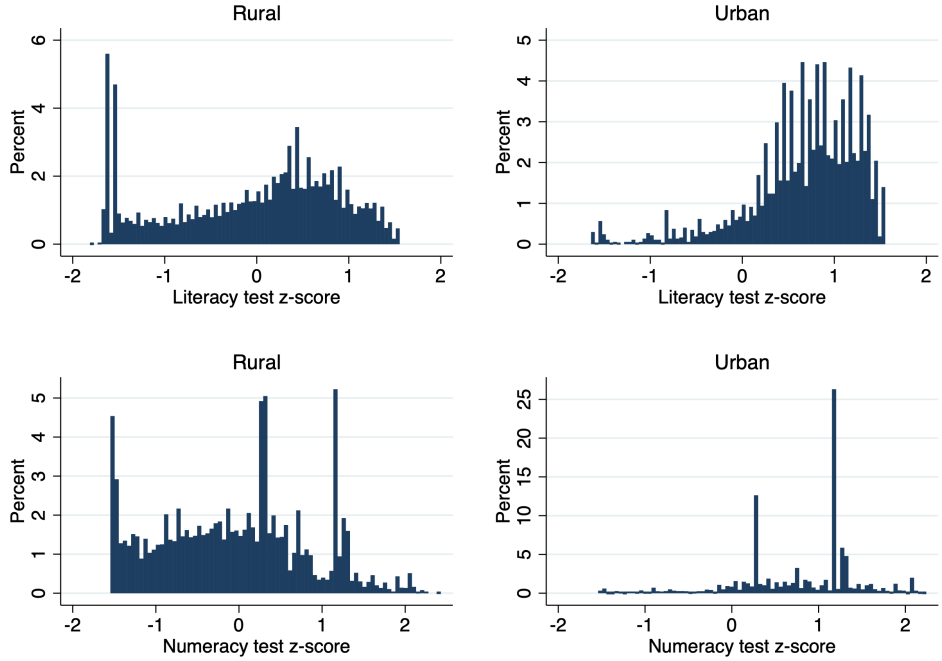


Figure B4: Main outcome variable distributions - Test z-scores



APPENDIX C Birth Control Rate (BCR) data

We collect the government's birth control policy information from each county/district/prefecture gazetteers. Although each county provides information in its own format and covers different periods, there are common variables reported. Figure C1 provides an example of a typical data format. The Birth Control Rate (BCR) is in column (D). It is calculated by dividing the number of fertile and married women in the county who use birth control (C) by the total number of fertile and married women (B). Note that (C) is the number of women who use birth control in the corresponding year, not the number of women who start birth control in that year. For example, column (C-2) indicates 5077 women are sterile in 1981, and the total number of women using contraception in the year is 7613+5077+22198+131=35019, and the BCR is 35019/44065=79.47%.

Figure C2 plots the BCR we collected from the county gazetteers. As can be seen from the figure that there are missing values, especially in the periods before the FPP was introduced. This, however, should have a limited effect on our estimation as our IV construction only use BCR for the years since the FPP was introduced. When BCR information is not available from the county gazetteers for some counties and some years after the FPP was introduced, we looked for information from its neighbouring counties, defined as counties sharing the border, prefecture or provincial gazetteers and statistical yearbooks. The main sources are (1) county or prefecture-level general gazetteers, (2) county or prefecture-level health gazetteers, (3) county or prefecture-level fertility control gazetteers, (4) municipal or provincial population statistical yearbooks, and (5) municipal or provincial fertility control statistical compilations.

We found BCR information for 143 counties out of the total 160 in the main sample, which accounts for 88.3%. Despite our effort to search various sources to collect information on BCR at the county level, we still have a large portion of county-year cells with missing values. To resolve this problem, we use two strategies. First, we use a regression-based imputation method to predict BCR missing values in these county-year cells. Second, instead of using the BCR information from the gazetteers, we use two alternative methods to construct instruments. For the imputation, we estimate the following regression.

$$BCR_{ct} = \beta_0 + \beta_1 SurveyBCR_{pt} + \beta_2 X_{ct} + \theta_c + \gamma_t + \upsilon_{ct} \tag{C1}$$

$$t = 1972, 1973, ..., 1990$$

where BCR_{ct} is the Birth Control Rate collected from local gazetteers/statistical yearbooks for year t in county c; $SurveyBCR_{pt}$ is a large scale fertility survey based measure of BCR for year t in prefecture p.[37] X_{ct} includes four variables. The first is a dummy variable indicating whether the original BCR_{ct} is at the prefecture level; the second is its interaction term with $SurveyBCR_{pt}$; the third variable is a dummy variable capturing whether we use a neighbouring county's BCR; and the last term is its interaction with $SurveyBCR_{pt}$. Adding these indicator variables can help to

[37] We are unable to use information at the county level due to the fact that the survey we are using, despite being a large scale survey, only have limited number of observations at the county-year cell level (11 observations in each cell), whereas at the prefecture level we have on average 105 observations in each prefecture-year cell.

minimise potential measurement errors caused by the unit difference. θ_c is a county fixed effect, and γ_t is a year fixed effect. Using estimated coefficients, $\hat{\beta}$s, from Equation C1, we then predict county-year cells with missing values in BCR_{ct}.

The variable $SurveyBCR_{pt}$ in Equation (C1) are generated from the 'Two-Per-Thousand Fertility and Contraceptive Survey, 1988' conducted by the Chinese State Family Planning Commission in 1988. This survey collected information on each woman's fertility and contraception history among 459,269 ever-married women. The data consists of 544,190 contraception-case observations,[38] including contraception type, starting- and ending-year, and the reason for each contraception.[39] We (1) count the number of fertile women who use contraception by year and prefecture and (2) divide it by the total number of fertile women by year for each prefecture.

The estimated results of Equation (C1) are reported in Table C1. It can be seen that our imputation model (column 5 of Table C1) can explain 75.5% of variations in the actual gazetteer BCR_{ct}.[40] In particular, a one percentage point increase in $SurveyB_{pt}$ is associated with 0.84 percentage point increase in the actual BCR. The goodness of fit can also be shown in Figure C3, where we plot the actual BCR from the gazetteers and that predicted values for the same counties from our model. It shows that the predicted values fairly closely mimic the actual value both in terms of the level of the prediction and the over-time changes in the shape of the curve for the period after the introduction of the FPP. Figure C4 plots the birth control rate by year for the data points that were obtained from the local gazetteers and those of predicted value for missing gazetteer data.

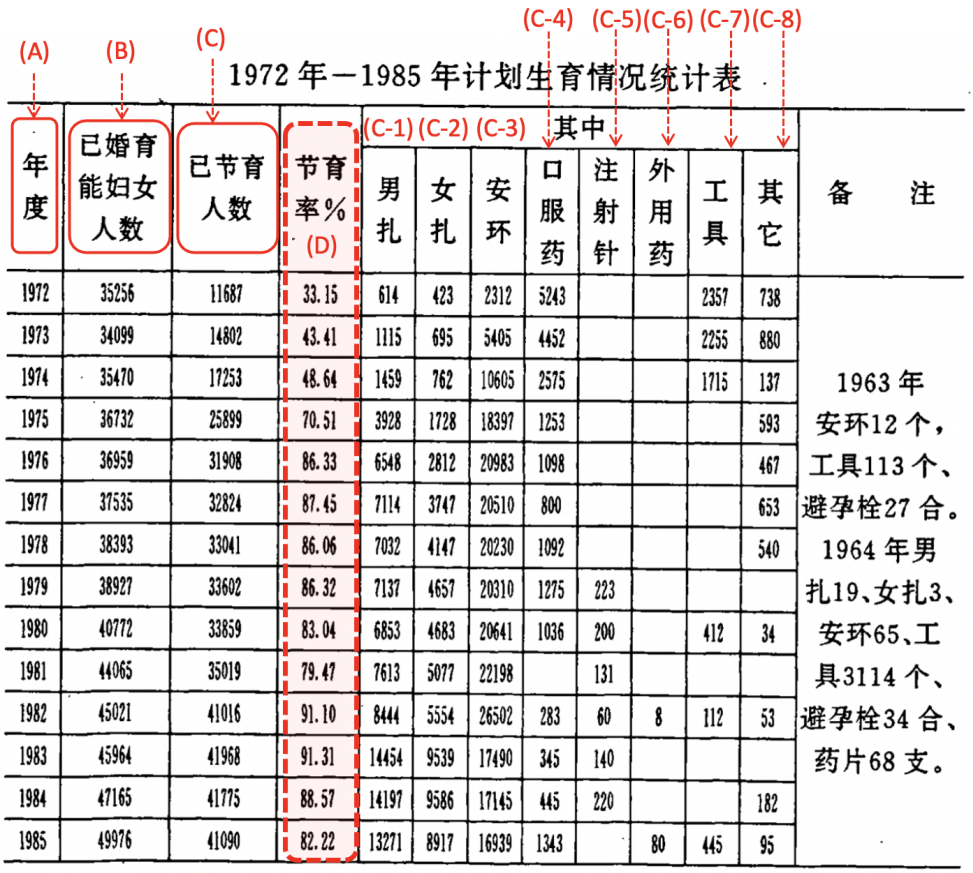
To check whether our main results are sensitive to the imputed values, we use two alternative methods to construct instruments. First, we replace BCR_c in Equation (2) with $SurveyBCR_{pt}$ and perform the baseline IV estimations. Columns 1 to 4 in Table C2 show that the results using the alternative BCR variable are largely similar to the main results in Table 4. Specifically, in the rural sample, the coefficients on the interaction between the number of siblings and the female dummy are -0.222 and -0.160 for literacy and numeracy test scores, respectively. These coefficients are comparable to the values of -0.214 and -0.148 obtained from the main results in Table 4. The other alternative instrument employs the same method as used in Chen and Fang (2021), except that we use T_c (i.e., the initial timing of the establishment of the Family Planning Leading Group at the county level) while they use T_p (the timing information at the provincial level). In other words, we exclude $BCR_c(T_m + a)$ from Equation (2), which means it measures the fertility interruption weighted by the number of years but without considering the intensity of the FPP policy exposure. As shown in Columns 5 to 9 of Table C2, our results remained robust to alternative instruments.

[38] The number of cases used in our calculation of $SurveyBCR_{pt}$ (our sample counties over the period of 1971 to 1988) is 235,948. Note that as the '1988 Fertility Survey' do not have information regarding contraception cases in 1989 and 1990, we assume that for these two years the contraception cases for each county is the same as that in 1988.

[39] The choices for this question are: 1. economic reasons, 2. family issues, 3. work related issues, 4. education related issues, 5. health related issues, 6. government FPP policy, and 7. others. By far, the largest share of people choose the 'government FPP policy', which accounts for 76% of the total cases.

[40] For imputation, we use the actual BCR observations from 1972, as the number of observations is not enough until 1971. The main results are not sensitive to the choice of starting year. The results are upon request from the authors. Another issue is that the survey was conducted in 1988, so we do not have observations for 1989 and 1990. We use 1988 observations for 1989 and 1990 by assuming that the BCR did not significantly change between those years.

Figure C1: Birth Control Rate in a county's gazetteer



1972年—1985年计划生育情况统计表

(A) 年度	(B) 已婚育能妇女人数	(C) 已节育人数	(D) 节育率%	其中 (C-1) 男扎	(C-2) 女扎	(C-3) 安环	(C-4) 口服药	(C-5) 注射针	(C-6) 外用药	(C-7) 工具	(C-8) 其它	备注
1972	35256	11687	33.15	614	423	2312	5243			2357	738	
1973	34099	14802	43.41	1115	695	5405	4452			2255	880	
1974	35470	17253	48.64	1459	762	10605	2575			1715	137	1963年安环12个，工具113个、避孕栓27合。1964年男扎19、女扎3、安环65、工具3114个、避孕栓34合、药片68支。
1975	36732	25899	70.51	3928	1728	18397	1253				593	
1976	36959	31908	86.33	6548	2812	20983	1098				467	
1977	37535	32824	87.45	7114	3747	20510	800				653	
1978	38393	33041	86.06	7032	4147	20230	1092				540	
1979	38927	33602	86.32	7137	4657	20310	1275	223				
1980	40772	33859	83.04	6853	4683	20641	1036	200		412	34	
1981	44065	35019	79.47	7613	5077	22198		131				
1982	45021	41016	91.10	8444	5554	26502	283	60	8	112	53	
1983	45964	41968	91.31	14454	9539	17490	345	140				
1984	47165	41775	88.57	14197	9586	17145	445	220			182	
1985	49976	41090	82.22	13271	8917	16939	1343		80	445	95	

Translated column labels:
(A) Year
(B) Number of married and fertile women
(C) Number of married and fertile women who use birth control
(D) Birth Control Rate (%) $=\frac{(C)}{(B)}\%$
(C-1) controlled by male sterilization
(C-2) controlled by female sterilization
(C-3) controlled by IUD (Intra-Uterine Device)
(C-4) controlled by pills
(C-5) controlled by injections
(C-6) controlled by medicine for external use
(C-7) controlled by condom
(C-8) etc

Figure C2: Birth Control Rate collected from local gazetteers

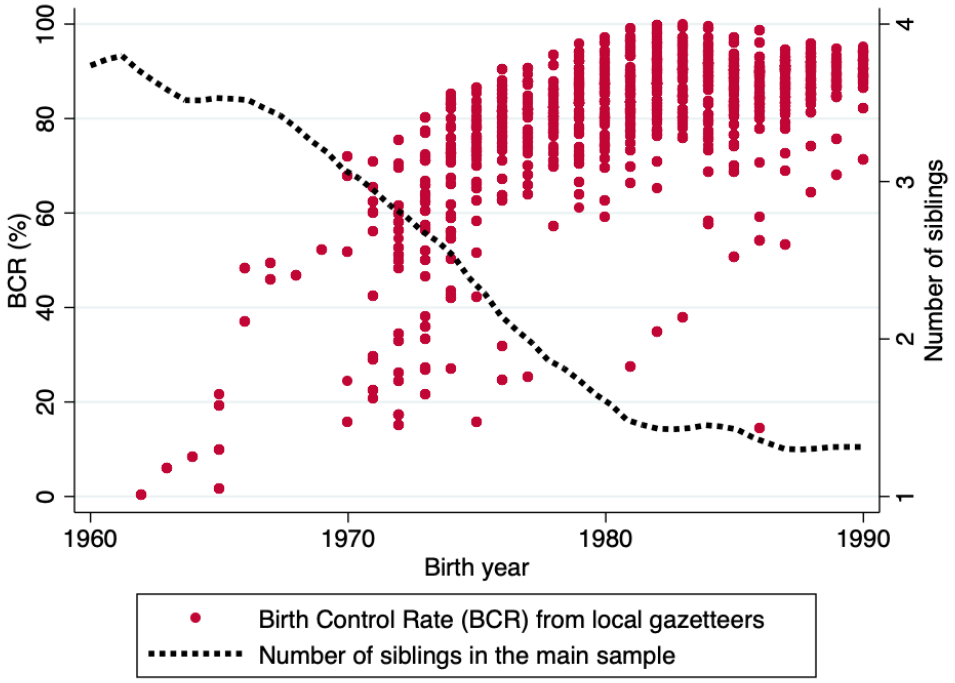


Figure C3: Actual values vs Imputed values of Birth Control Rate for non-missing values

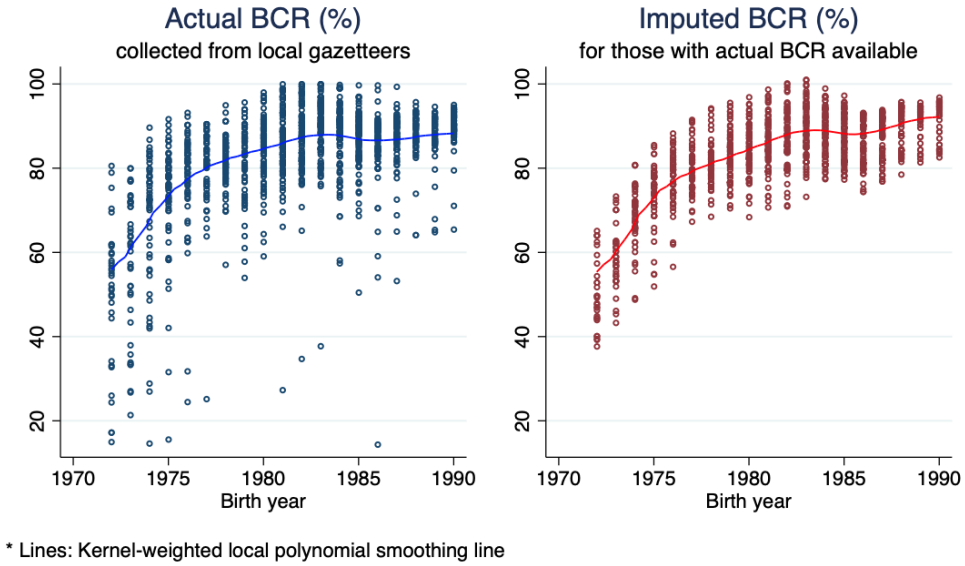


Figure C4: Birth Control Rate (BCR) by year:
Actual values for non-missing cases & imputed values for missing cases

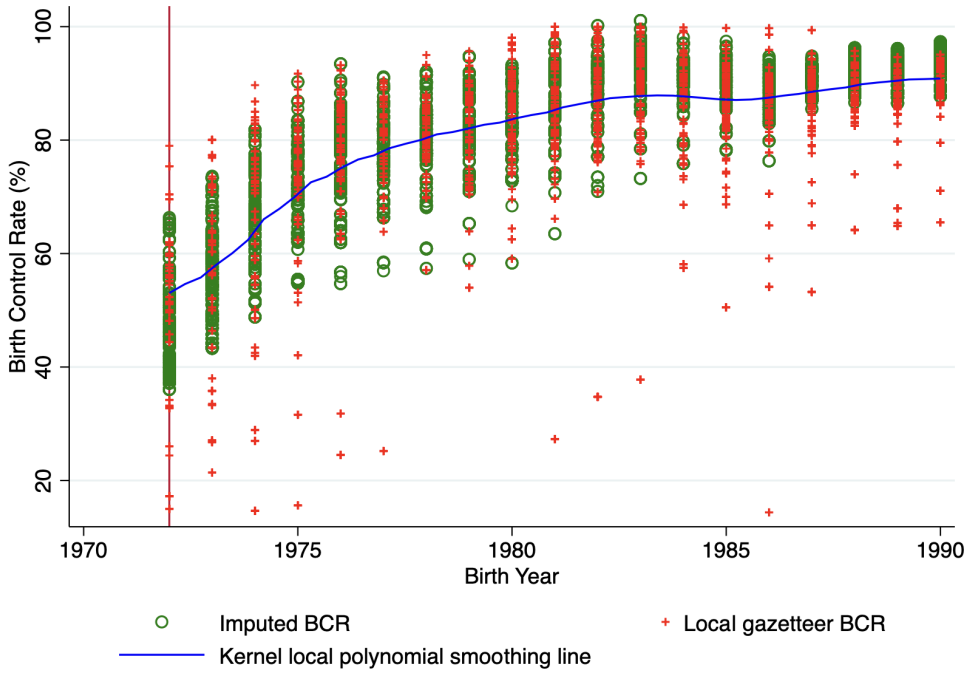


Table C1: Estimation results used for imputation

Dependent var. : Actual gazetteer Birth Control Rate (%) B_{ct}					
	(1)	(2)	(3)	(4)	(5)
$SurveyB_{pt}$	0.310***	0.319***	0.783***	0.783***	0.836***
	(0.013)	(0.012)	(0.044)	(0.044)	(0.045)
$SurveyB_{pt}$ × Neighbouring county data					-0.068**
					(0.029)
$SurveyB_{pt}$ × Prefecture level data					-0.110***
					(0.019)
County fixed effect	No	Yes	Yes	Yes	Yes
Year fixed effect	No	No	Yes	Yes	Yes
Observations	1404	1404	1404	1404	1404
Adjusted R^2	0.296	0.535	0.749	0.749	0.755

Note:
1) Robust Standard Errors (SEs) are presented in parentheses;
2) $SurveyB_{pt}$ is the estimated BCR when all contraception cases are counted;
3) * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$;

Table C2: IV results using alternative instruments

	Instruments constructed using $SurveyBCR$				Instruments constructed without BCR info.			
	Rural sample		Urban sample		Rural sample		Urban sample	
	Literacy	Numeracy	Literacy	Numeracy	Literacy	Numeracy	Literacy	Numeracy
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Siblings	-0.123	-0.160**	0.088	0.075	0.061***	0.055***	0.046	0.116**
	(0.084)	(0.078)	(0.110)	(0.105)	(0.016)	(0.016)	(0.042)	(0.048)
Siblings × Female	-0.222***	-0.160***	-0.065***	-0.036**	-0.209***	-0.137***	-0.066***	-0.041***
	(0.016)	(0.016)	(0.013)	(0.015)	(0.016)	(0.015)	(0.014)	(0.015)
Female	0.344***	0.152**	0.129***	0.017	0.268***	0.037	0.137***	0.023
	(0.059)	(0.060)	(0.029)	(0.039)	(0.049)	(0.050)	(0.030)	(0.036)
Parental schooling years	0.046***	0.058***	0.033***	0.043***	0.049***	0.062***	0.031***	0.045***
	(0.003)	(0.003)	(0.007)	(0.006)	(0.003)	(0.003)	(0.004)	(0.004)
Birth order index	-0.021**	-0.013	-0.052	-0.034	-0.025***	-0.018***	-0.039	-0.047
	(0.008)	(0.008)	(0.037)	(0.038)	(0.007)	(0.007)	(0.024)	(0.030)
Birth order index × Female	0.054***	0.034***	0.052***	0.010	0.064***	0.044***	0.050***	0.014
	(0.011)	(0.011)	(0.019)	(0.019)	(0.009)	(0.008)	(0.018)	(0.021)
Han Chinese	0.123*	0.091	0.087	0.101	0.114*	0.081	0.084	0.105
	(0.069)	(0.065)	(0.070)	(0.090)	(0.067)	(0.065)	(0.068)	(0.093)
Lifetime $SurveyBCR$ (%)	-0.019**	-0.022***	0.003	-0.003				
	(0.008)	(0.007)	(0.008)	(0.007)				
Observations	20895	20895	3721	3721	20895	20895	3721	3721
Kleibergen-Paap Wald F statistics	11.60	11.60	4.219	4.219	197.7	197.7	23.48	23.48

Note:

1) The outcome variables are test z-scores with zero mean and a standard deviation;

2) Robust Standard Errors (SEs) are presented in parentheses;

3) SEs are clustered at the county level;

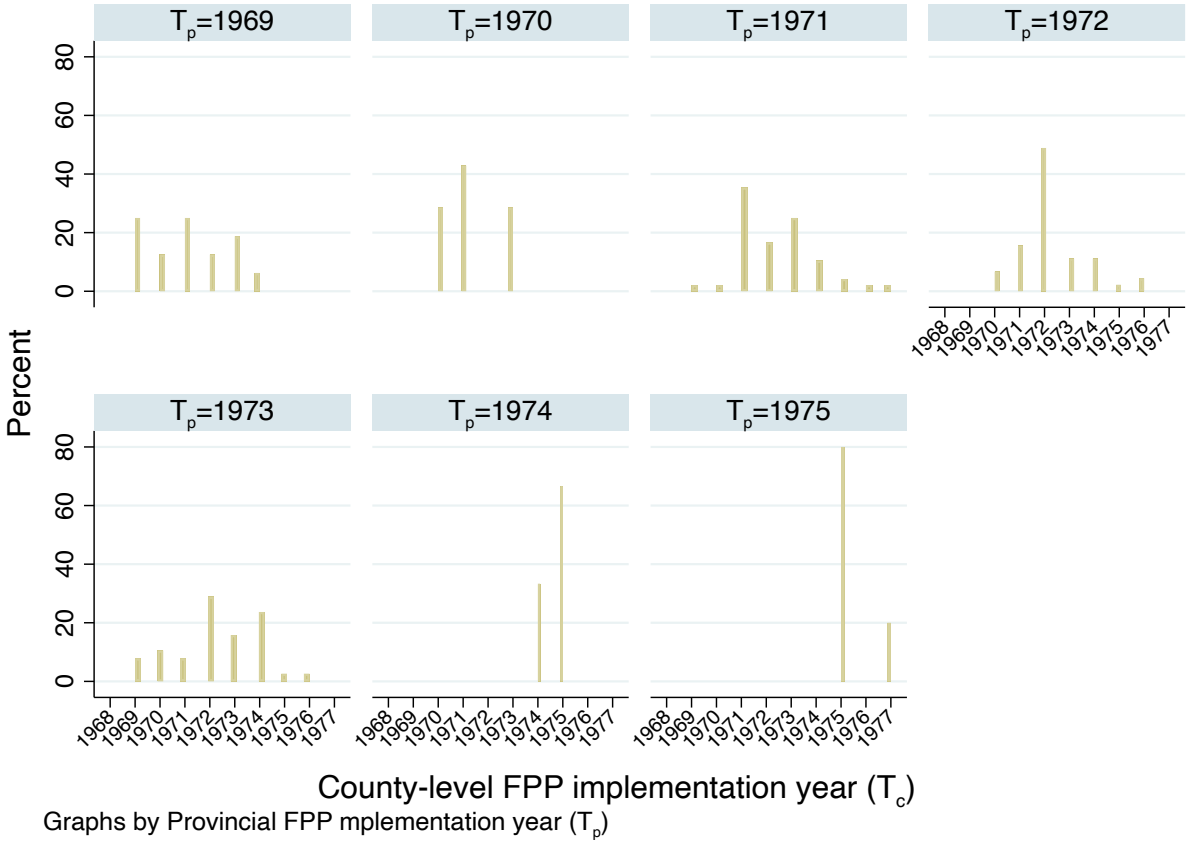
4) Control variables include the same set of covariates in Table 3 except for Lifetime BCR, which is replaced with Lifetime $SurveyBCR$ in columns 1 to 4. Columns 5 to 8 do not include either;

5) For estimations in columns 1 to 4, we replace BCR with $SurveyBCR$ in equation 2 and use this variable and its interaction with a female dummy as instruments. For estimations in columns 5 to 8, we exclude $BCR_c(t_m + a)$ from equation 2 and use the rest and its interaction with a female dummy as instruments. See Appendix C for details;

6) * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$;

APPENDIX D

Figure D1: Comparison of timing difference of initial FPP introduction between across counties and across provinces



APPENDIX E

Table E1: The probability of being in a mixed-sex family

Dependent var.	OLS - Rural	IV - Rural	OLS - Urban	IV - Urban
Mixed-sex dummy	(1)	(2)	(3)	(4)
Siblings	0.091***	0.043***	0.133***	0.217***
	(0.003)	(0.014)	(0.007)	(0.053)
Female	0.058***	0.067***	0.052***	0.041***
	(0.007)	(0.008)	(0.012)	(0.012)
Parental schooling years	0.002	0.001	-0.002	0.003
	(0.001)	(0.001)	(0.002)	(0.004)
Birth order index	0.009***	0.012***	-0.005	-0.037
	(0.003)	(0.003)	(0.011)	(0.025)
Birth order index × Female	-0.006*	-0.010***	-0.018	-0.012
	(0.003)	(0.003)	(0.012)	(0.016)
Han Chinese	-0.006	-0.002	0.036	0.047
	(0.013)	(0.015)	(0.041)	(0.041)
Lifetime BCR (%)	0.005***	0.002	0.003	0.006**
	(0.001)	(0.001)	(0.002)	(0.003)
Observations	18321	18321	3397	3397
Kleibergen-Paap Wald rk F stat		107.682		37.370

Note:

1) The outcome variable is equal to 1 if an individual is in a mixed-sex family and 0 if he/she is in a single-sex family;

2) Robust Standard Errors (SEs) are presented in parentheses;

3) SEs are clustered at the county level;

4) Control variables include the same set of covariates in Table 3 except for Siblings×Female;

5) The instruments are the same as those used in Table 3;

6) * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$;

Figure E1: Comparison of timing difference of initial FPP introduction between across counties and across provinces

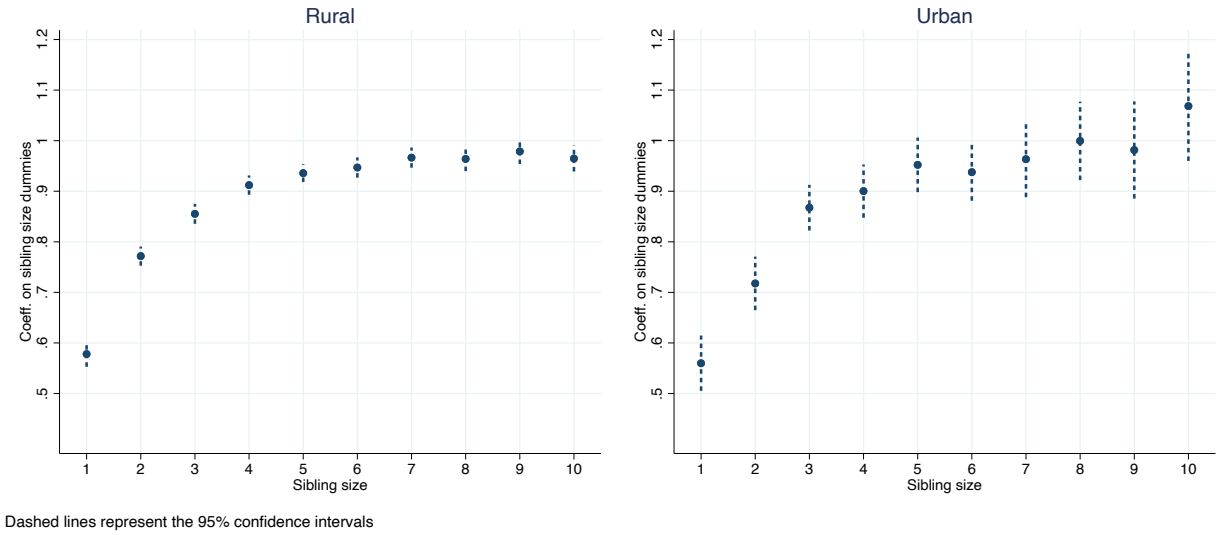


Table E2: Gender gap in the mixed- and single-sex families, test scores – Urban

Dependent variable:	Literacy - OLS		Numeracy - OLS		Literacy - IV		Numeracy - IV	
	Mixed	Single	Mixed	Single	Mixed	Single	Mixed	Single
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Siblings	-0.005	0.027	0.140	0.262*	0.006	-0.024	0.227**	0.170
	(0.016)	(0.027)	(0.092)	(0.146)	(0.016)	(0.020)	(0.111)	(0.165)
Siblings × Female	-0.010	-0.015	-0.056*	-0.033	-0.030*	0.035	-0.019	0.023
	(0.017)	(0.031)	(0.030)	(0.096)	(0.018)	(0.026)	(0.036)	(0.104)
Female	-0.014	0.058**	0.129	0.064	0.014	-0.046	-0.032	-0.043
	(0.056)	(0.027)	(0.093)	(0.056)	(0.056)	(0.032)	(0.116)	(0.061)
Observations	2009	1388	2009	1388	2009	1388	2009	1388
Kleibergen-Paap Wald rk F stat					9.263	2.681	9.263	2.681

Note:

1) The outcome variables are test z-scores with zero mean and a standard deviation;

2) Robust Standard Errors (SEs) are presented in parentheses;

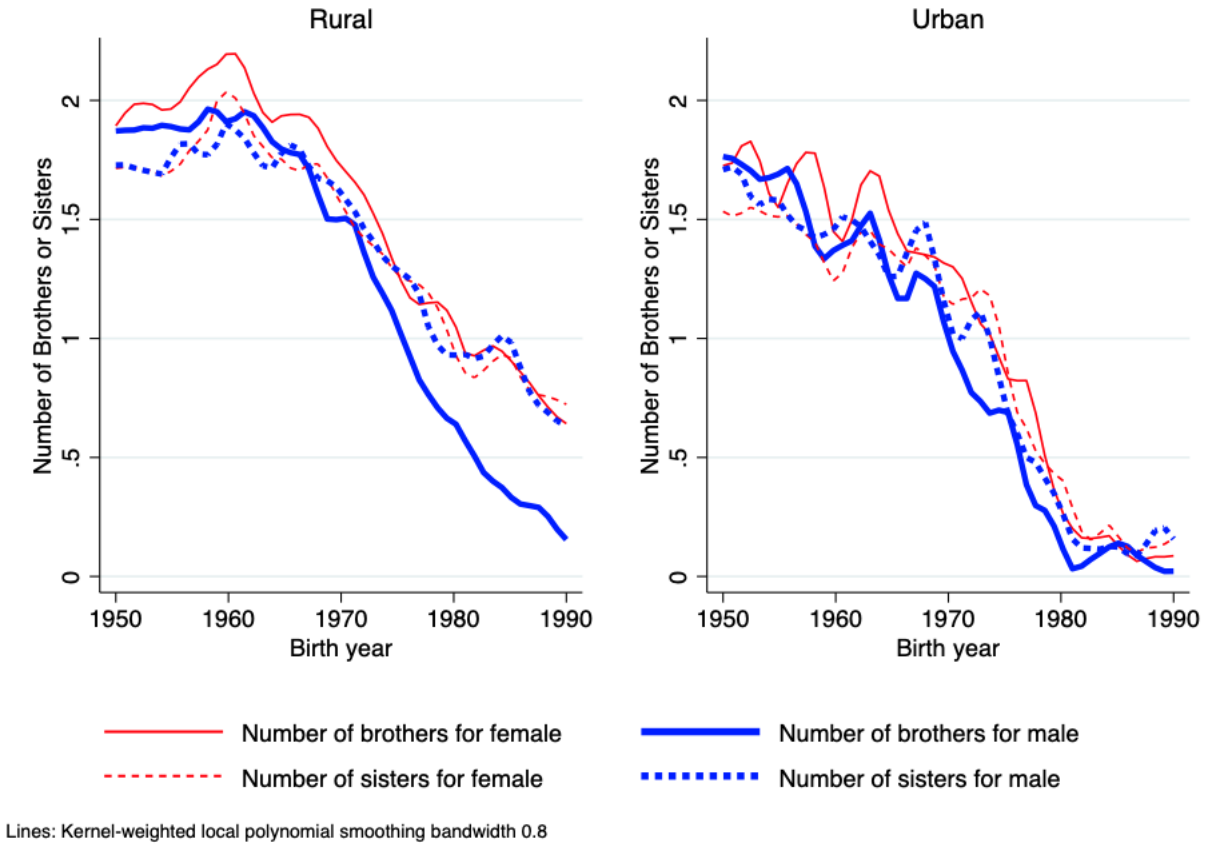
3) SEs are clustered at the county level;

4) Control variables include the same set of covariates in Table 3;

5) The instruments are the same as those used in Table 3;

6) * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$;

Figure E2: Number of brothers/sisters across birth cohorts



Lines: Kernel-weighted local polynomial smoothing bandwidth 0.8