**Appendix**

Children are identified by their relationship with the head of household. Any relationship outside of “child of head of household" is excluded because other relationships lack accurate family data necessary for analysis including sibling information and mother fertility history. Mother’s number of children is collected for women between the ages of 15 and 50.

Because sibling information is necessary to identify legality of birth, I exclude households where the mother's[[1]](#footnote-1) reported number of children does not match the number of children in the household. I am able to match 82.96 per cent of children born after 1963 with mothers in the sample. Children whose mothers are identified as born before 1950 and children born before 1963 are excluded from the analysis because mother fertility data is not reliable.

It is possible that individuals’ *hukou* status changed between having children and the 2000 census. If a woman gave birth in one province then changed *hukou* location or status, she may have faced different birth planning regulations than measured here. During the 1980s and early 1990s, however, the *hukou* system was highly regulated and opportunities to change one’s status were few,[[2]](#footnote-2) suggesting measurement error of sorting families by *hukou*-based policy should be minimal.

**Imputation**

The biggest limitation of the census sample is the reliance on reported births. Families with out-of-plan births have a large incentive to hide children from census enumerators, as discussed by many demographers in China. Reporting an out-of-plan birth may result in fines. In order to overcome this shortfall, models presented in this paper replicate the full census model with data imputed to simulate the un-enumerated individuals. The following section outlines the process of imputation.

I use standard techniques for estimating missing individuals in the 2000 census using cohort analysis from the 2010 census using the base formula used by the US Census Bureau, where missing individuals are estimated by:[[3]](#footnote-3)

$$P\_{t}=P\_{t+1}-B\_{t,t+1}+D\_{t,t+1}-M\_{t,t+1}$$

where $P\_{t}$ is the population in 2000, $P\_{t+1}$ is the population in 2010, $B\_{t,t+1}$ is the number of births between 2000 and 2010, $D\_{t,t+1}$ is the number of deaths between 2000 and 2010 and $M\_{t,t+1}$ is the net migration during the interval from 2000 to 2010. Because population information is reported by five-year age cohorts and I am estimating backwards, I ignore births, since any births since 2010 would not appear in the 2000 census. Population and migration rates come from the county-level census data in 2000 and 2010. Death rates were estimated using gendered cohort mortality rates reported in Yu et. al (2015). Net migration rates are estimated based on the 2000 and 2010 census by municipality. I then adjust the formula accordingly to:

$$P\_{2000, 0-4|G}=P\_{2010, 10-14|G}+P\_{2000, 0-4|G}\*D\_{0-4|G}-M$$

Where each of the data points is gendered (given gender *G*) and based on a given age cohort, $D\_{C|G}$ is the estimated death rate given gender *G* and age cohort as reported by Ye et al (2015)*,* and *M* is a migration rate estimated for the given county. The total population of girls aged 0-4 in 2000 is equal to the total population of girls in 2010 aged 10-14 plus any population who may have died between 2000 and 2010 minus net migration. Using this formula, I estimated the proportion of missing individuals by gendered age cohort by subtracting the reported population in the 2000 census given age, gender, in a given county from the estimated population resulting from the above equation. Because these are estimates, assign each county-gender-age group a proportion of missing by total reported.

I then created observations based on these estimates to add to the existing census sample. For example, if a county had 100 children in the census sample and I found that 5% of girls were missing in the 0-4 age group, five observations were added in this county. I do this for all age groups under 20. Given the dependence on mother’s information, the original census sample restricted to children living with their parents faces significant limitations for individuals over 20, so the imputation models only include individuals aged 1-20. The average missingness for males is 13.28 per cent and the average missingness for females is 14.22 per cent in the total census.

This data allows me to identify the estimated number of individuals missing in a given county, their gender, and age group (five-year cohort). I assume a uniform distribution of the missing observations across age and randomly assign individuals an age within their age cohort.[[4]](#footnote-4) I used multiple imputation (mi command in Stata 15) to create the imputed data. Imputations were monotone, because missingness was constant across variables for each of the added observations: all added observations were missing mother’s *hukou* and migration status, house price, mother’s education level, rank age, and mother’s date of birth. While one could assign *hukou* status based on location, with women living in city districts being assigned urban and in counties rural, but this ignores the impact migration has on the mixing of the population across space. Mother’s *hukou* and migration status were estimated using a logit regression, house price, mother’s education level, and rank age used ordered logit given that the variables are ranked categorical variables. Mother’s date of birth is imputed as a linear regression. Best practice suggests that anywhere between five and twenty imputations are needed to converge on a mean. I completed both five and ten imputations for robustness. The results from the ten imputations model are reported in the analysis.

Stata 15’s mi estimate command was used to complete the models. This command runs the base model over each of the ten imputations of missing data and presents more accurate results combined across the ten imputations according to Rubin’s rules. Given the number of categorical variables, the imputed data generated a small number of cases (less than 0.5%) that were perfectly predicted. I used the esampvaryok command to estimate the models, allowing these perfectly predicted cases to be dropped from the analysis. Because the number of cases is so small, the subsequent bias is minimal.

**Table A1 Full Logit Model Results for Interaction Models**

|  |  |  |
| --- | --- | --- |
| VARIABLES | Rural | Urban |
|  |  |  |
| Migrant Mother | -1.36\*\*\* | -0.684\*\* |
|  | (0.194) | (0.224) |
| Out-of-plan | -3.66\*\*\* | -4.484\*\*\* |
|  | (0.182) | (0.439) |
| Female | -0.129\*\*\* | -0.109\*\* |
|  | (0.030) | (0.051) |
| Out-of-Plan X Female | -0.158\*\*\* | -0.056 |
|  | (0.029) | (0.057) |
|  |  |  |
| Age | 0.254\*\* | 0.398\*\*\* |
|  | (0.096) | (0.039) |
| Age Squared | -0.018\*\*\* | -0.016\*\*\* |
|  | (0.005) | (0.002) |
| Birth Order | -0.226\*\*\* | -0.231\*\* |
|  | (0.065) | (0.109) |
| House Price | 0.025 | 0.061\*\*\* |
|  | (0.034) | (0.014) |
| Constant | -35.99 | 26.565 |
|  | (26.58) | (42.371) |
|  |  |  |
| Observations | 2,574,206 | 466,656 |

*Note.* Robust standard errors clustered at the province in parentheses. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Policy Controls include indicator variables for the policy (OCP, OHCP, TCP) children were born under.

Mother Controls include mother’s year of birth and education level.

1. Mother is identified as either the female head of household or female spouse of head of household. [↑](#footnote-ref-1)
2. Chan and Buckingham 2008; Li, Yi and Zhang 2011 [↑](#footnote-ref-2)
3. Population projection estimate methodology can be found at https://www.census.gov/programs-surveys/popproj.html [↑](#footnote-ref-3)
4. For example, an individual identified as belonging to the 0-4 years old category is randomly assigned to an age between 0 and 4 with equal probability. [↑](#footnote-ref-4)