A Appendix

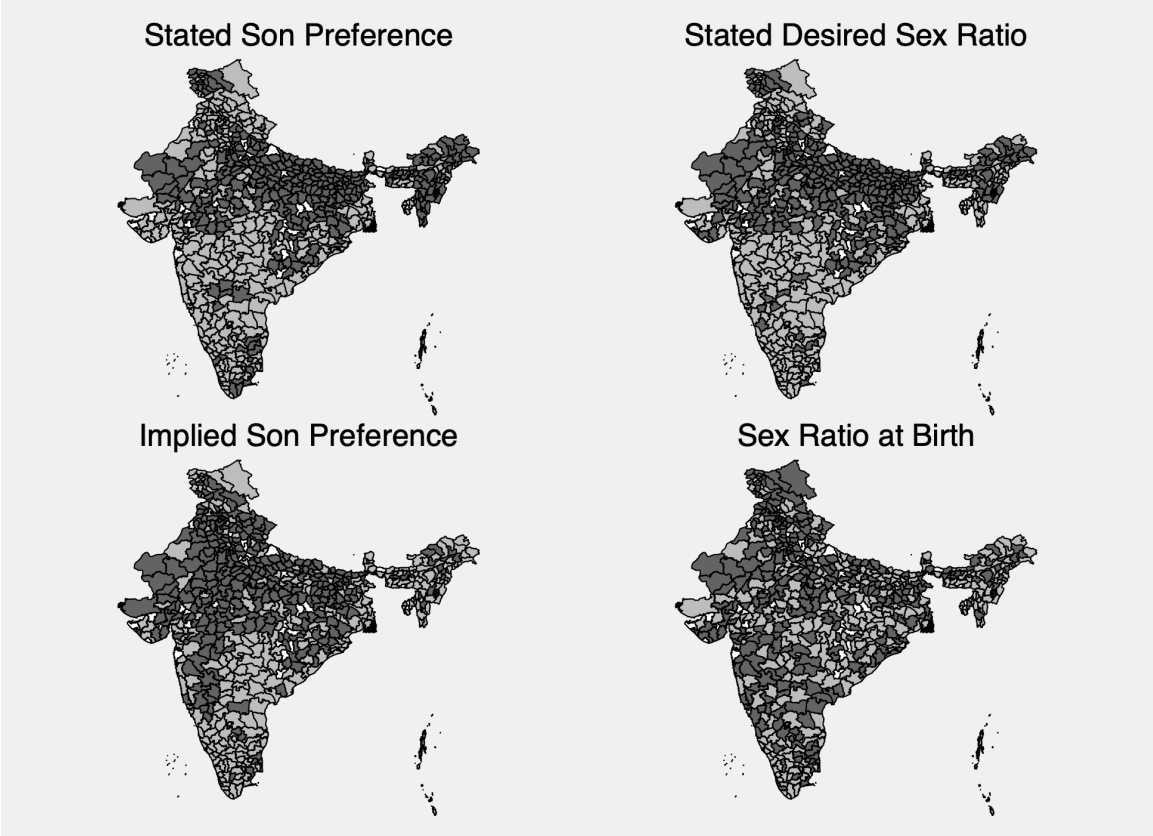

Figure A1: **District level** variation in measures of son preference intensity. The two shades depict the intensity of son preference: darker shade indicates states above the median, and lighter shade indicates states below the median. Measures of son preference are calculated using the 2015/16 Indian NFHS data and a description of how each measure is calculated is provided in Section 3. The sample is restricted to mothers who gave births five years before the survey year.

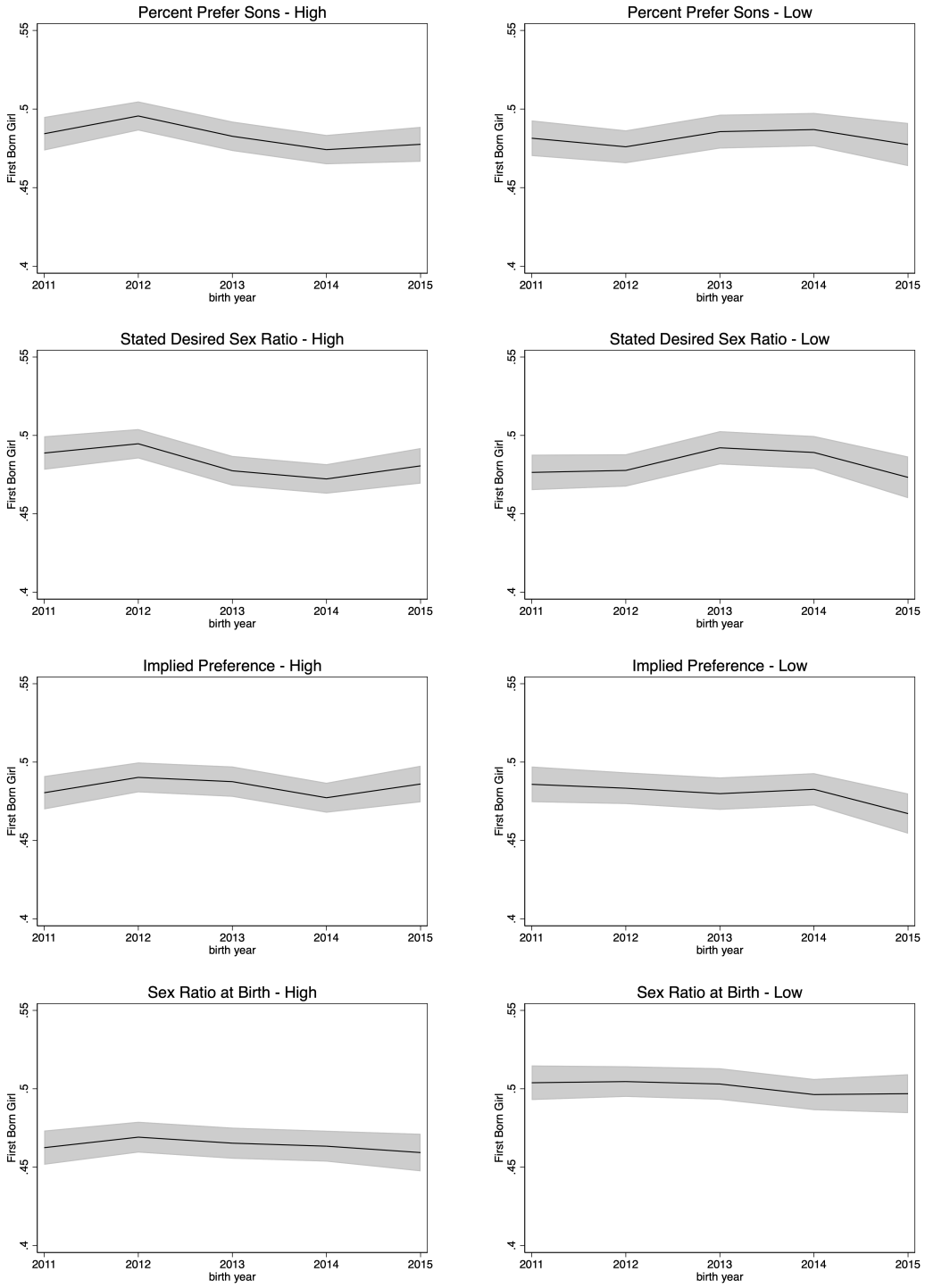

Figure A2: **District level** share of girls born at first birth order with 95% confidence interval, by district's son preference measure and intensity. Data from the 2015/16 Indian National Family Health Survey. The sample is restricted to children born within five years before the survey year.

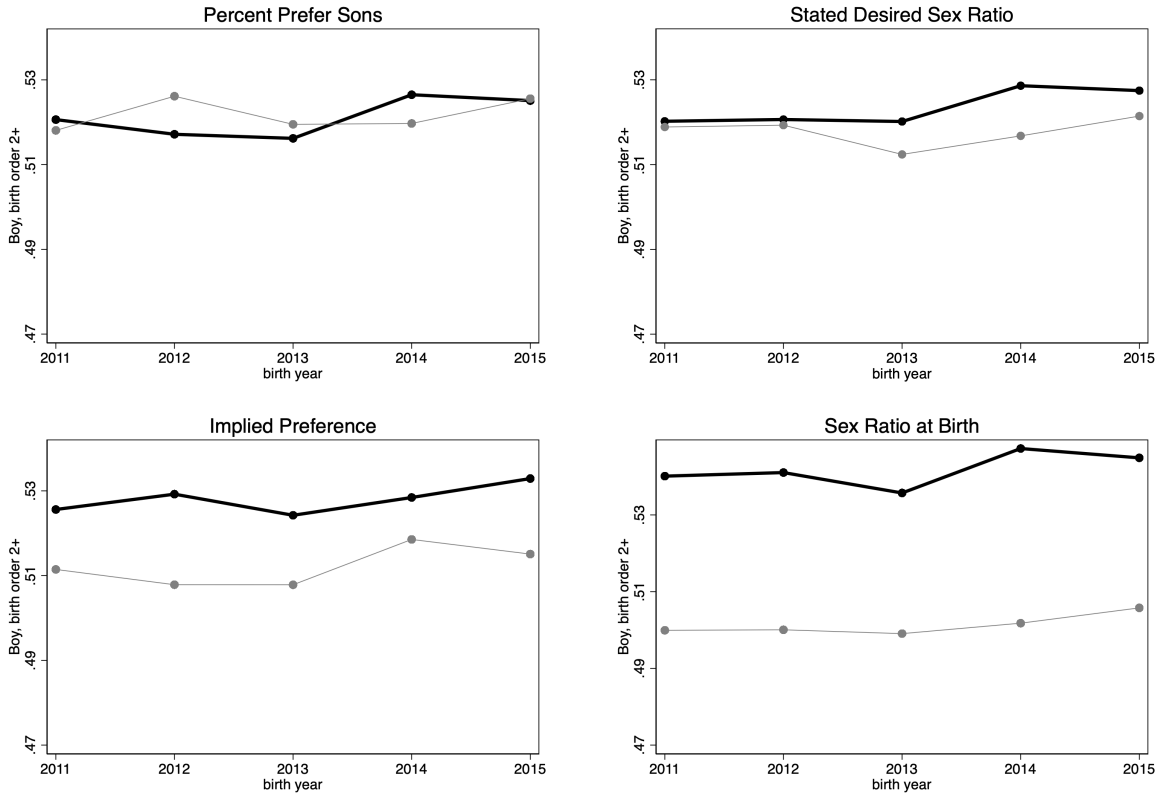

Figure A3: **District level** share of boys born at birth orders two and greater, by district's son preference measure and intensity. Bold lines represent districts with high son preference; gray (non-bold) lines represent districts with low son preference. Data from the 2015/16 Indian National Family Health Survey. The sample is restricted to children born within five years before the survey year.

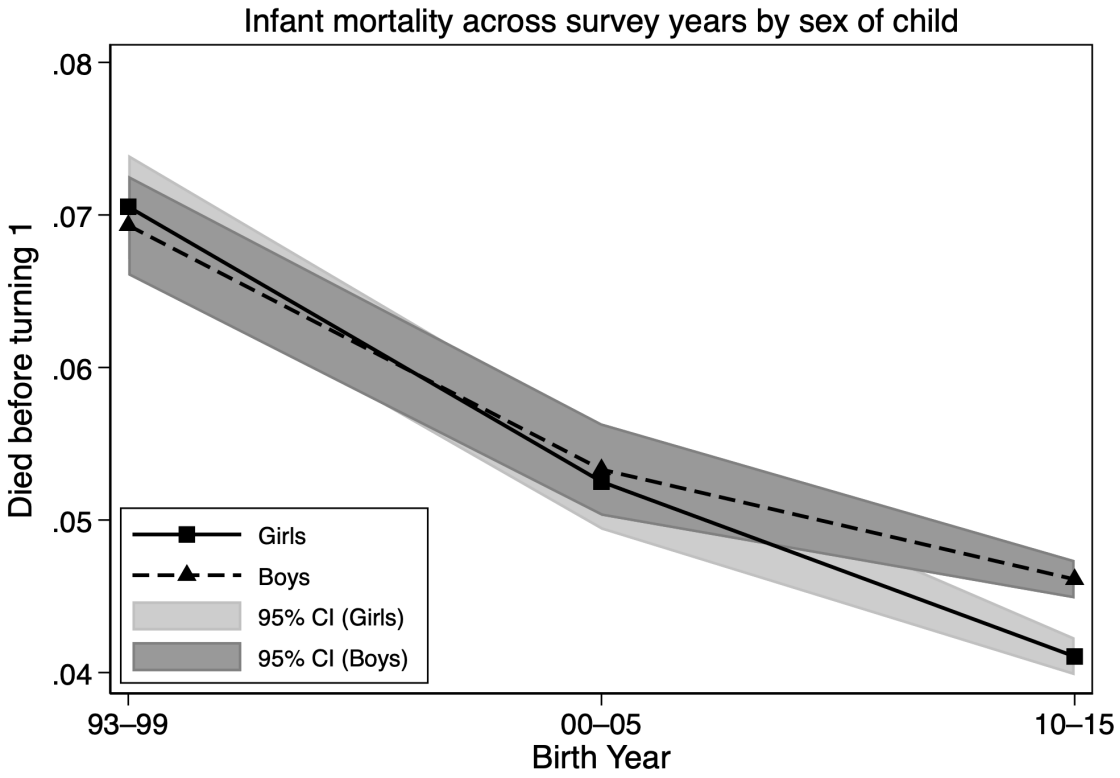

Figure A4: Trends in infant mortality rates by sex. Data from Indian NFHS from 1998, 2005, and 2015. The sample is restricted to children born within 1-5 years before the time of the survey.

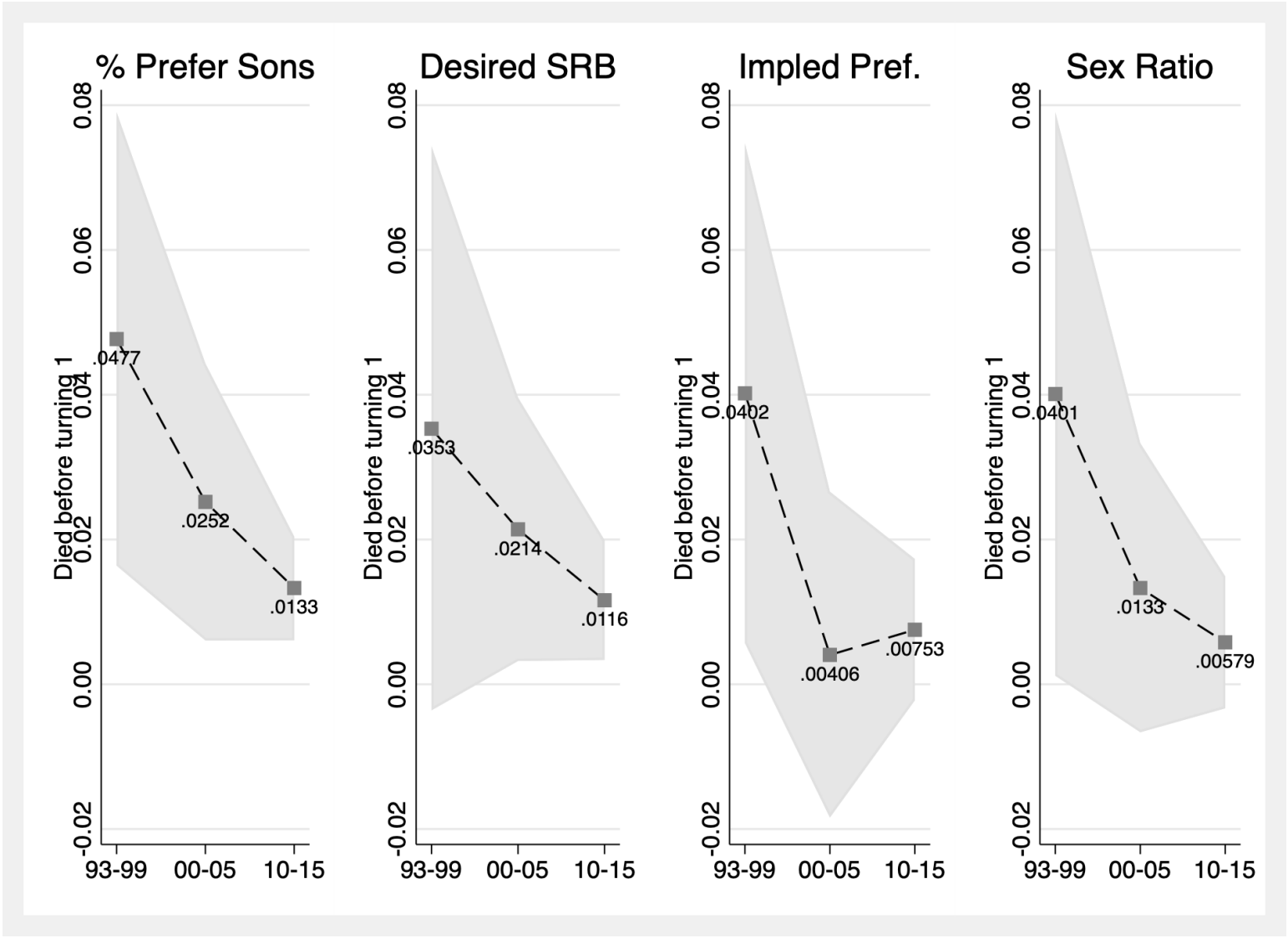

Figure A5: Evolution of gender gap in infant mortality by state's son preference measure. Figure graphs coefficient on $FBG_i \times PrefHigh_s \times girl_i$ in Equation 2 by child's birth year for each of the four son preference measures (% prefer sons, desired sex ratio, implied son preference, and sex ratio at birth). Measures of son preference are calculated using the 2015/16 Indian NFHS data and a description of how each measure is calculated is provided in Section 3. Data from Indian NFHS from 1998, 2005, and 2015. The sample is restricted to children born within 1-5 years before the time of the survey.

B Appendix: Implied Preference Explored

The paper introduces a simple metric to capture a family's implied preference for sons by estimating Equation 1, which compares the desire for more children among families with no sons to those with no daughters. This coefficient β captures the increase in childbearing desire attributable to not yet having a son and the implied preference for at least one son. This estimate is identified using variation across families with same-sex children; and decisions made by families with mixed-sex children are not incorporated.

The framework can be extended to explore whether families prefer to have more than one son. Specifically, one could ask whether families with mixed-sex children but only one son express a greater desire for additional children compared to mixed-sex families with more than one son. To test this, I augment the implied preference specification as follows:

$$\begin{aligned} want_i = \ & \beta\, no\, sons_i + \gamma(mixed_i \times only\, one\, son_i) + mixed_i \\ & + only\, one\, son_i + i.Nchildren_i + \varepsilon_i \quad (A1) \end{aligned}$$

Equation A1 extends the original implied preference specification by including an indicator variable for mixed-sex families, and an indicator variable for families with only one son, and the interaction of the two terms, $mixed_i \times only\, one\, son_i$, which captures families with mixed-sex children but only one son. The coefficient γ captures the additional desire for more children among mixed-sex families with only one son, relative to mixed-sex families with more than one son. A positive γ suggests a preference for at least two sons. It is important to note that this estimate is only identified for larger families — for families with exactly two children, *mixed* and $mixed \times only\, one\, son$ indicators are perfectly collinear. This raises some concerns regarding the generalizability of this measure and also highlights the importance of controlling for overall fertility. In my sample, 64% of mothers have 1 or 2 children, and only 36% of them have 3 or more children. Thus, preference for at least 2 sons is likely to correlate with high fertility states. To remove any influence of fertility differences

across states from the implied preference measures, I also include child fixed effects.

This extended analysis yields two coefficients of interest, β and γ, capturing preference for at least one and at least two sons, respectively. It is not obvious how, or if, one should combine the two coefficients into one measure. Thus, I study the gender health gap associations with each measure (high β and high γ) separately. Table A1 presents the results from various measures of implied preference. Column 1 reproduces the results from the main paper where implied preference is estimated using Equation 1. Column 2 investigates if the simple calculation of implied preference is influenced by different fertility levels and adds number of children fixed effects in Equation 1. I do not find this to be the case as the coefficient of interest on excess female mortality in firstborn girl families remains similar. Column 3 estimates the coefficient capturing a preference for at least one son (β) from Equation A1 – the measure's intensity is perfectly correlated with that in column 2 and hence yields an identical result. Column 4 investigates if girls in firstborn girl families from states with an implied preference for at least two sons (high γ) suffer a higher infant mortality rate. I do not find this to be the case.

Table A2 investigates if each of the discussed measures of implied preference is associated with sex selection at higher birth orders. While the implied preference measures for at least one son (β) from both Equation 1 and A1 are associated with more boys in firstborn girl families at the second birth order, implied preference for at least two sons, γ, shown in column 4 is not associated with more sex selection. States with high implied preference for two sons do not seem to sex select or discriminate against daughters, which is inconsistent with the substitution hypothesis. One potential explanation is that these high fertility states are more likely to have to have at least one son and the presence of one son is enough to help protect vulnerable daughters. Alternatively, if these states sex select, they may do so at higher birth orders, but the nature of the data is unable to pick up later birth order sex selection. Note that the data focus on families with children born within last 5 years (to focus on current son preference measures and gender gaps), since these are relatively newer

families, they are less likely to have completed their fertility.

Table A1: Infant mortality gender gap by Implied Preference Measures

	(1)	(2)	(3)	(4)
Measure	β Eq. 1	β Eq. 1 i.Nchild	β Eq. A1	γ Eq. A1
FBG×Pref High×girl	0.00753	0.00713	0.00713	0.00429
	(0.00471)	(0.00525)	(0.00525)	(0.00505)
FBG×girl	-0.00443	-0.00435	-0.00435	-0.00202
	(0.00399)	(0.00447)	(0.00447)	(0.00432)
FBG×Pref High	-0.00671**	-0.00607	-0.00607	-0.00133
	(0.00320)	(0.00357)	(0.00357)	(0.00311)
FBG	-0.00497*	-0.00526	-0.00526	-0.00900***
	(0.00285)	(0.00319)	(0.00319)	(0.00253)
Wild-boot p-value	0.0720	0.132	0.132	0.699
Observations	203,187	203,187	203,187	203,187

Dependent variable is an indicator for whether the child died before turning 1 years old. Data source: 2015/16 Indian NFHS. Sample restricted to children who were born between 1-5 years before the survey. State clustered standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1.

Table A2: Is high birth order child a boy by Implied Preference Measures

	(1)	(2)	(3)	(4)
Measure	β Eq. 1	β Eq. 1 i.Nchild	β Eq. A1	γ Eq. A1
FBG×Pref High×birth order = 2	0.0175*	0.0195*	0.0195*	-0.00599
	(0.00975)	(0.0104)	(0.0104)	(0.0132)
FBG×Pref High×birth order $\geq$ 3	-0.00153	-0.00120	-0.00120	-0.00824
	(0.00846)	(0.00930)	(0.00930)	(0.00879)
Wild-boot p-val birthorder =2	0.0910	0.0830	0.0830	0.639
Wild-boot p-val birthoder $\geq$ 3	0.845	0.903	0.903	0.398
Observations	256,847	256,847	256,847	256,847

Dependent variable is an indicator for whether the child is a boy. Data source: 2015/16 Indian NFHS. Sample restricted to children who were born within 5 years before the survey. State clustered standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1.