Son preference and sex ratios: How many 'missing women' are missing?

Jas Ellis*
DESTIN and EOPP
London School of Economics
j.j.ellis@lse.ac.uk
personal.lse.ac.uk/ellisj/

September 9, 2008
Original version: July 2008

Abstract

When parents prefer sons, heterogeneity in the probability of having sons can lead to excess girls. I argue that this may lead to under-counting the number of 'missing women'. First, I prove that relaxing assumptions on population homogeneity means that son preference can lead to skewed sex ratios. Second, I measure significant heterogeneity in the sex ratio at birth: ten percent of women have probabilities of having boys that are less than 42% or more than 61%. Third, existing work measures significant differences in parents' son preferences between countries. I exploit these differences in parental behaviour to simulate sex ratios in the presence of heterogeneity. I measure that parents' son preferences account for 1.5% of differences between sex ratios worldwide (significant at 10%). The presence of this effect may imply that sex ratios are more biased than previously estimated, since previous comparisons use benchmarks that already contain too few girls. Therefore there may be more women missing due to discrimination than we thought.

*Many thanks to Karl Ove Moene and seminar participants in DESTIN and EOPP. This work was supported by ESRC Studentship PTA-031-2004-00297.

1 Introduction

Since recognition of the 'missing women' problem by Sen [1989], several explanations have been made for the high proportion of boys in a number of countries. Recent work has highlighted biological factors as a possible cause of differences in sex ratios (the number of boys per girl), notably the Hepatitis-B virus [Oster, 2005]. Conversely, a majority of authors conclude that social norms are the proximate cause, as these lead to lower survival rates for girls.[1] To date, however, the effect of son preference in fertility decisions has been neglected, despite evidence that parents' sex preferences are mainly determined by cultural background [Ellis, 2008].

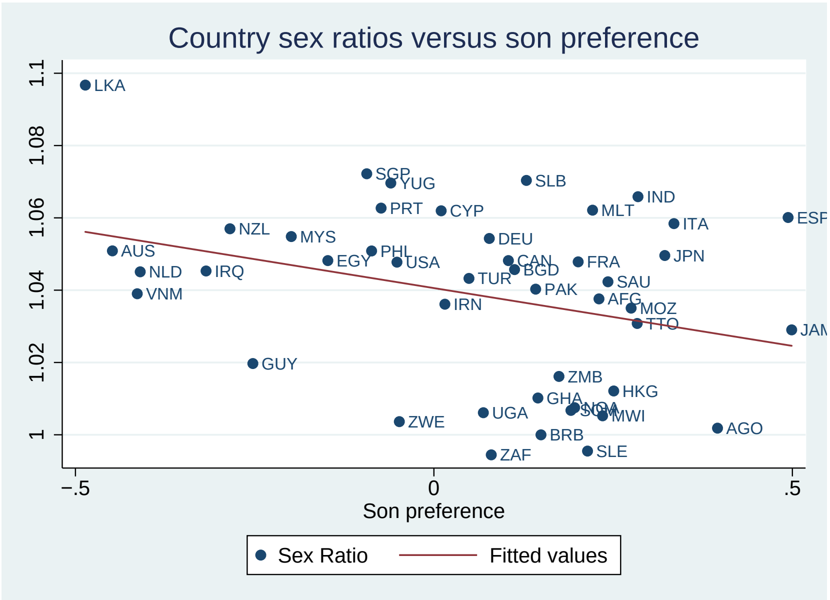
This paper estimates the effect of parental fertility decisions on sex ratios worldwide when women are heterogeneous in the probability of bearing sons. If women in a population have boys with differing probabilities, then son-preferring fertility behaviour will lead to excess girls. The extra girls borne by women that are more likely to have girls outnumber the reduction of girls borne to women likely to have boys. Previous theoretical work [Weiler, 1959; Goodman, 1961; Yamaguchi, 1989] has recognised this phenomenon in principle; I know of no attempt to quantify the effect in practice.

Like previous authors [reviewed in James, 2000], I find significant heterogeneity in the probability of having a son, suggesting that ten percent of women have boys with probabilities outside the interval [0.42, 0.61]. Accounting for parental behaviour, this heterogeneity leads to sex ratios in the range 1.043–1.051, explaining 1.5% of differences worldwide (significant at the 10% level). As the theory suggests, son preference is associated with excess girls. Thus, since previous estimates of missing women have used comparisons without accounting for these excess girls at birth, the number of women missing due to explicit discrimination may have been under-counted.

Ellis [2008] provides theoretical and empirical evidence that parents' prefer-

[1] Proponents of cultural explanations include Sen [1992]; Das Gupta et al. [2002]; Arokiasamy [2004]; Das Gupta [2005]; Qian [2006]; Chamarbagwala and Ranger [2006] and Lin et al. [2008].

Figure 1: Son preferences versus sex ratios. Son preference (after two children) is estimated from a birth-hazard regression for foreign-born women in the UK grouped by country of origin. Child sex ratio is derived from World Development Indicators (1997), under-15 male population divided by under-15 female population. Correlation is significant at 5% (robust to inclusion/exclusion of outliers).[a]



[a]Son preference is $-\delta_c$ from the Cox Proportional Hazards regression $\log \lambda_i = \beta' X_i + \sum_c \gamma_c d_{ic} + \sum_c \delta_c \text{SONS}_i * d_{ic}$, where woman i from country c has birth hazard $\theta_{ic}(t) = \lambda_{ic}\theta_0(t)$ relative to the (unspecified) baseline $\theta_0(t)$. See Ellis [2008], Section 4. Countries with son preference coefficients absolutely greater than 0.5 are omitted here for clarity.

ence for sons is mainly driven by cultural factors. He provides uniquely comparable estimates of son preference between countries, based on the behaviour of foreign-born women in the UK. Figure 1 plots son preference after two children versus sex ratios in those women's countries of origin. The correlation is significant and negative (and robust to omission of outliers). This suggests that culturally-driven son preference may lead to a reduction in the sex ratio.

Consider an extreme example: women continue to have children until they have a son. If boys and girls are equally likely for every women, the sex ratio in aggregate will be 1 [Weiler, 1959; Goodman, 1961; Sheps, 1963]. However, if half of women only ever have boys and half only girls, the former will obtain their son at the first birth. The latter will continue to have girls until some maximum family size is reached, and *girls will outnumber boys*. I prove a more general form of this result in Section 2. The phenomenon relies on the existence of heterogeneity in the probability of a son.

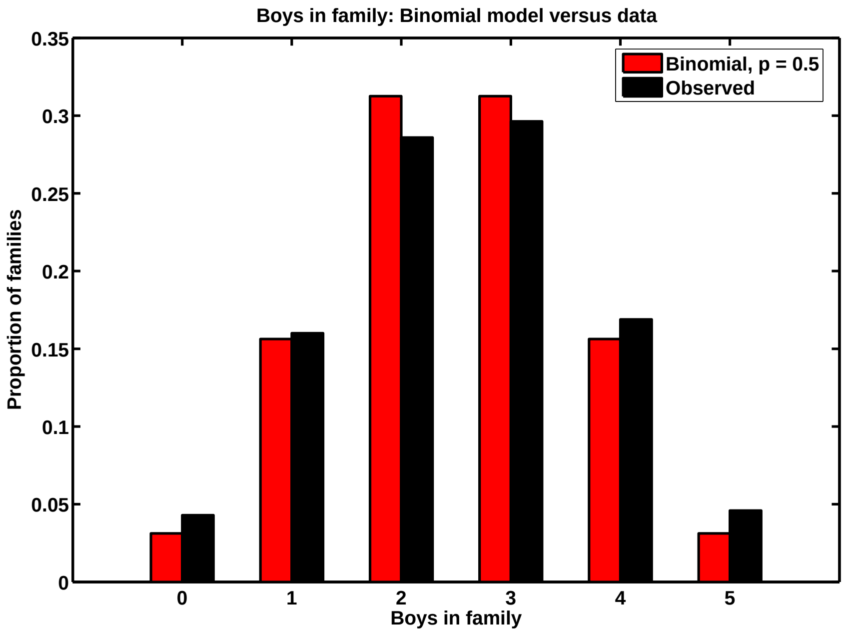
Figure 2 suggests that within-population heterogeneity of probability of having a son is indeed a possibility. In Section 3, I provide estimates of such heterogeneity, and find homogeneity is rejected at the 0.01% level. I derive an estimator based on modelling childbirth as a limited dependent variable problem with random effects. A woman i has an unobserved factor X_i, and child j is a boy if $X_i + \epsilon_{ij} > 0$, with ϵ_{ij} being drawn from an independent standard normal distribution. Thus, woman i gives birth to boys with probability $\Phi(X_i)$.[2] By assuming a distributional form for X, parameters may be obtained by maximum likelihood estimation. I find significant heterogeneity amongst a sample of 116,513 British-born women: 10% of women have boys with probabilities outside [0.42, 0.61]. My measurements are closely in line with previous work, despite the difference in estimation technique.

In Section 4, I simulate the effect of son preferences on aggregate sex ratios, using the estimates for heterogeneity derived in Section 3. To obtain preferences for sons, I estimate the son preferences of immigrant women in the UK. By grouping the women by country of origin, I measure the fertility behaviour in response to their existing family compositions. I can then calculate the sex ratio that would emerge in a population of women behaving this way. This simulated sex ratio is then compared with the sex ratio in the women's countries of origin.

The contribution of this paper is to establish that fertility behaviour does feasibly affect sex ratios in practice. Moreover, the bias is towards the less-favoured sex. Therefore, even though the effect I find is likely outweighed by discriminatory behaviour (such as selective abortion, infanticide or neglect), it is important because missing women cannot be measured correctly unless sex ratios at birth are properly accounted for. This reinforces the arguments made by Mayer [1999] and Griffiths et al. [2000], that sex ratios alone should be treated

[2] Φ is the standard normal cumulative distribution function (CDF).

Figure 2: Family composition (5 children) under a Binomial model ($B \sim$ Bernoulli(0.5)), versus actual data (see Table 1). The data show more dispersion than predicted by the model with homogeneous probabilities of having sons.



with caution as a measure of women's position in society: I maintain sex ratios must be treated with care when measuring the number of missing women.

Related Literature

There have been numerous attempts to quantify the number of missing women worldwide, notably Drèze and Sen [1989], who arrive at a figure of 100 million, and Coale [1991] finding a reduced figure of 60 million. Oster [2005] takes account of Hepatitis-B and its effect on the probabilities of having sons, coming to a still lower figure of 32 million. All three works take probabilities of sons to be homogeneous within countries.[3]

Meanwhile, there is a growing literature that implies within-population heterogeneity in the probability of having boys. Factors thought to affect the 'parental' sex ratio include status and personal traits such as dominance [Kemper, 1994; James, 1994; Grant, 1996, cited in Edlund, 1999], parental perception of the adult sex ratio [James, 1995], times of war [Graffelman and Hoekstra, 2000], Hepatitis B [Oster, 2005, 2006], maternal partnership status [Norberg, 2004] and maternal diet [Matthews et al., 2008]. Further, Lindsey and Altham [1998] find that a binomial model, as implied by homogeneity in son-probabilities, has a poor fit to family composition data. They find the number of sons in families is 'overdispersed', suggesting heterogeneity. The demographic literature on probability heterogeneity is comprehensively surveyed by James [2000]. The theoretical result that population heterogeneity creates a link between son preference and aggregate sex ratios is not new, going back to Weiler [1959] and Goodman [1961], but the finding is not recognised in the economic literature.

There is wide acknowledgement that cultural factors such as son preferences and the status of women drive mortality-rate differentials and incentivise female infanticide and selective abortion [Sen, 1992; Das Gupta et al., 2002, for exam-

[3]Nonetheless, Oster's work does in fact imply heterogeneity with at least two types: virus carriers and non-carriers.

ple]. Despite this recognition, no one has yet attempted to measure the effect of parents' fertility decisions on sex ratios. This omission surely results from an awareness of the fact that with *homogeneous* probabilities of having boys, son preference will have no effect on the sex ratio in large populations [Sheps, 1963; Leung, 1988].

The present paper sits at the juncture of two strands of literature. The first strand considers sex ratios and the biological factors affecting an individual woman's probability of having boys. However, instead of accounting for biological differences amongst women from different countries (as in Oster [2005]), I treat biological differences as unobserved but existing within every population. In this paper, differences between countries are *cultural*. Here, I follow the new literature on cultural differences between countries, such as [Fernández and Fogli, 2005, 2006; Ellis, 2007], which determine that cultural background is a significant predictor of fertility outcomes. Of particular relevance to the present work is Ellis [2008], which finds strong evidence that son preferences amongst immigrant women in the UK is driven by cultural factors, rather than economic mechanisms. This accords with Chamarbagwala and Ranger [2006], who argue that several cultural aspects such as religious composition and caste structure explain in part the high sex ratios found in some parts of India. Their suggested mechanism relies on selective abortion, infanticide and neglect of girls leading to an increase in the sex ratio (an excess of boys).

The results of this paper imply that son preferences lead to an excess of girls. This finding is of compatible with the results of Coale [1991]; Chamarbagwala and Ranger [2006] and others if discrimination outweighs the contribution of fertility decisions. In fact, the work here reinforces previous studies. I demonstrate the existence of a mechanism that opposes the effect of discrimination on the sex ratio. Thus, previous estimates may understate the importance of discrimination, and the number of missing women may be under-measured.

2 Theory

In this section I prove that parents' fertility decisions can affect population sex ratios, with a bias *against* the favoured sex. A necessary and sufficient condition is that the probability of a son is heterogeneous within the population.[4] This finding dates back to Weiler [1959] and Goodman [1961, both papers cited in Yamaguchi, 1989]; my proof is presented here for clarity and to support the intuition. In this paper, I focus on *lexis variation*, whereby women differ in their probabilities of having sons, but each woman's probability does not change over time.[5]

2.1 Background

Much existing research has focussed on homogeneous populations. In large samples, homogeneity entails that the population sex ratio matches the 'natural' ratio of boys to girls — the ratio which would prevail if parents only cared about the *size* of their families, not their sexes [Sheps, 1963; Leung, 1988]. With homogeneity, the probability that a further child is a boy is independent of the current family composition. *Why you choose to have a child* doesn't affect the probability you'll have a son. (This argument establishes that heterogeneity is necessary for son preference to affect sex ratios.)

Heterogeneity implies that the composition of a family *is not* independent of its size. When one sex is preferred, a woman's childbearing decisions are co-related to the probability she has boys.

Without loss of generality, let parents prefer sons. Then, those having boys are more likely to cease having children than those having girls. So those likely to have sons will, on average, have smaller families than those likely to have daughters. The difference in family sizes gives an excess of girls relative to the

[4] Throughout this paper, the 'probability of having a son' refers only to the biological chance of conceiving and bearing a boy. I assume that selective abortion never occurs. In the empirical sections, child mortality is also ignored (following Gangadharan and Maitra [2003]). Neither of these factors are likely to affect my results, since these phenomena are uncommon in the UK. (However, see Dubuc and Coleman [2007].)

[5] James [2000] discusses in detail the implications of different types of variation.

case without sex preferences.

2.1.1 Progression Rates

Of women with a certain number of children, the proportion going on to have further children is known as the *progression rate.* Rates can be compared between women who have different family compositions. Usually, women having boys are found to have lower progression rates, indicating a preference for sons[6] and this is rationalised by the microeconomic models of Leung [1991] and Ellis [2008]. For the model here, I choose the most extreme form of this behaviour: women have children until they bear a son (this may be called a *1-boy stopping rule*). This formulation simplifies the algebra considerably, but is not necessary for the intuition.[7]

2.2 Model

2.2.1 Homogeneity

First, consider the case of homogeneous probabilities. Let there be N women, each with natural probability p of having a boy at any birth. Since each woman i continues bearing children until a son is born, the number of boys in each family (b_i) will be one. Thus, the number of boys in the population will be

$$B \stackrel{def}{=} \sum_{i=1}^{N} b_i = N$$

The number of girls in each family takes a geometric distribution.[8] $g_i \sim \text{Geom}(p)$. The total number of girls, $G \stackrel{def}{=} \sum_{i=1}^{N} g_i$, tends to $N\frac{(1-p)}{p}$ as $N \to \infty$, because the expected number of girls in each family is $\frac{1-p}{p}$. Therefore the sex ratio $\frac{B}{G}$ converges to $\frac{p}{1-p}$, as when parents have no son preference. Equivalently,

[6] See, for example, Leung [1988]; Gangadharan and Maitra [2003]; Das Gupta [2005].

[7] Any n-boy stopping rule would yield exactly the same bias in the sex ratio (proof omitted). The effect of mixed-family stopping rule (eg, n boys, m girls) would be smaller but qualitatively similar (as long as $n > m$). The intuition is unchanged provided that after a son a woman is less likely to continue having children than after a daughter.

[8] I take the geometric defined as the number of *failures* before a success, not the number of *attempts*. $f_{\text{Geom}(p)}(g) = (1-p)^g p$.

$\frac{B}{B+G} \to p$.

2.2.2 Heterogeneity

Now, consider a mean preserving spread in the probabilities, with a fraction α having boys with probability p_1, and the remainder with probability $p_2 = \frac{p-\alpha p_1}{1-\alpha}$. With no son preference, childbearing ends independently of the composition of each family and (hence) independently of the woman's type. The population sex ratio will again be $\frac{p}{1-p}$.

With son preference, each woman still has one son, so $\tilde{B} = N$. However the distribution of girls is not the same for all women, since $g_i \sim \text{Geom}(p_i)$, and $\text{E}[g_i] = \frac{1-p_i}{p_i}$. Thus, as $N \to \infty$,

$$\tilde{G} \stackrel{def}{=} \sum_{i=1}^{N} g_i \to N\alpha \frac{1-p_1}{p_1} + N(1-\alpha)\frac{1-p_2}{p_2}$$

Hence the population sex ratio is:

$$\begin{aligned} \frac{\tilde{B}}{\tilde{G}} &\to \left(\alpha\frac{1-p_1}{p_1} + (1-\alpha)\frac{1-p_2}{p_2}\right)^{-1} \\ &= \frac{p_1 p_2}{\alpha(1-p_1)p_2 + (1-\alpha)p_1(1-p_2)} \\ &= \frac{p_1 p_2}{(1-\alpha)p_1 + \alpha p_2 - p_1 p_2} \end{aligned}$$

$\frac{\tilde{B}}{\tilde{G}}$ is necessarily *smaller* than the sex ratio with no preference, as the following theorem shows.

Theorem 1. *In a large population of women, let a proportion α have boys with probability p_1, and the remainder have boys with probability p_2. Then $\tilde{B}/\tilde{G} < B/G$ if and only if $p_1 \neq p_2$. Heterogeneity is a necessary and sufficient condition for son preference to bias in the population sex ratio.*

Proof. I begin with the following inequality, due to Cauchy (it is also implied by Jensen's inequality). $2p_1p - 2 \leq p_1^2 + p_2^2$ holds with equality only when $p_1 = p_2$, ie, the case of homogeneity. Otherwise, the inequality is strict:

$$\begin{aligned}
2p_1p_2 &< p_1^2 + p_2^2 \\
(2\alpha - 2\alpha^2)p_1p_2 &< \alpha(1-\alpha)(p_1^2 + p_2^2) \\
(1 - (1-\alpha)^2 - \alpha^2)p_1p_2 &< \alpha(1-\alpha)(p_1^2 + p_2^2) \\
p_1p_2 &< \alpha(1-\alpha)p_1^2 + (1-\alpha)^2p_1p_2 \\
&\quad + \alpha^2p_1p_2 + \alpha(1-\alpha)p_2^2 \\
p_1p_2 - \alpha p_1^2p_2 - (1-\alpha)p_1p_2^2 &< \alpha(1-\alpha)p_1^2 + (1-\alpha)^2p_1p_2 \\
&\quad + \alpha^2p_1p_2 + \alpha(1-\alpha)p_2^2 \\
&\quad - \alpha p_2^2p_2 - (1-\alpha)p_1p_2^2 \\
p_1p_2(1 - \alpha p_1 - (1-\alpha)p_2) &< (\alpha p_1 + (1-\alpha)p_2)((1-\alpha)p_1 + \alpha p_2 - p_1p_2) \\
\frac{p_1p_2}{(1-\alpha)p_1 + \alpha p_2 - p_1p_2} &< \frac{\alpha p_1 + (1-\alpha)p_2}{1 - \alpha p_1 - (1-\alpha)p_2} \\
\therefore \frac{\tilde{B}}{\tilde{G}} &< \frac{B}{G}
\end{aligned}$$

Therefore the sex ratio is biased in favour of girls. Note that, under homogeneity of probabilities, each of the inequalities becomes an equality. The parental decisions only affect the sex ratio under heterogeneity. □

We see that a mean preserving spread in the probability of having sons skews the sex ratio in favour of girls when parents prefer boys. In the next sections, I take this theoretical result and estimate the magnitude of the effect in reality.

One caveat must be noted. Here, I treat women as biologically heterogeneous but socially identical: they share the same preferences and behave the same. This is unlikely to be the case in practice. No work to date has considered heterogeneity in son preference within groups, despite evidence of variations between groups [Ellis, 2008]. However, heterogeneity in behaviour would likely result in a averaged outcome; the effects modelled here would apply for any subgroups behaving similarly. The overall effect would be an aggregate of the subgroup effects.

3 Estimating Heterogeneity

In this section, I attempt to estimate within-population heterogeneity in the probability of having a son. Previous work by economists has considered child-bearing as a Bernoulli trial with a probability p of bearing a son, and $1 - p$ of bearing a daughter; authors have allowed p to differ between populations but not within them. Here, this restriction is reversed: p varies within populations, but the distribution is the same across countries.

I assume that, throughout their life, each woman has the same probability of having sons. In other words, I assume only *lexis variation.* Therefore, since I observe multiple outcomes (children) for some women, I am able to estimate the underlying distribution of probabilities, given some functional assumptions.

3.1 Empirical model

Consider a population of women, with each woman i having some underlying factor X_i which affects the likelihood that she bears a son when she has a child. For simplicity, let X be distributed as a normal random variable with mean μ and variance σ^2:

$$X_i \sim \mathrm{N}(\mu, \sigma^2)$$

When woman i bears a child j, an independent, identically distributed draw is made. I take this ϵ_{ij} also to be Gaussian, and without loss of generality, normalise the mean and variance to 0 and 1 respectively:

$$\epsilon_{ij} \sim \mathrm{N}(0, 1)$$

A boy is born if the sum of X and ϵ is greater than zero. That is, if B_{ij} is a random variable indicating birth of a boy,

$$B_{ij} \stackrel{def}{=} \begin{cases} 0 \text{ if } X_i + \epsilon_{ij} \leq 0 \\ 1 \text{ if } X_i + \epsilon_{ij} > 0 \end{cases}$$

Let Φ and ϕ be the cumulative distribution and probability density functions of the standard normal distribution.[9] The probability of woman i having a boy at any birth is

$$\begin{aligned} p_i &= \mathrm{P}[X_i + \epsilon_{ij} > 0] \\ &= \mathrm{P}[\epsilon_{ij} < -X_i] \\ &= 1 - \Phi(-X_i) \\ &= \Phi(X_i) \end{aligned}$$

Thus, B_{ij} may be treated as a Bernoulli trial with probability of success $p_i = \Phi(X_i)$.[10]

3.2 Likelihood functions

If X_i were known, the likelihood of observing a certain family composition for woman i can be easily computed. Suppose observed sons are indicated by $b_{i1} \dots b_{ik_i}$. Then, since individual births are independent,

$$\begin{aligned} \mathcal{L}(b_{ij}|X_i) &\stackrel{def}{=} \mathrm{P}[B_{ij} = b_{ij} \text{ for } j = 1 \dots k_i] \\ &= \prod_{j=1}^{k_i} \mathrm{P}[B_{ij} = b_{ij}] \\ &= \prod_{j=1}^{k_i} \Phi(X_i)^{b_{ij}} (1 - \Phi(X_i))^{(1-b_{ij})} \\ &= \Phi(X_i)^{k_i^{\mathrm{B}}} (1 - \Phi(X_i))^{k_i^{\mathrm{G}}} \end{aligned}$$

[9] When necessary, the CDF and PDF of X will be denoted Φ_X and ϕ_X.

[10] This 'Probit' derivation is inherently isomorphic to the formulation $B_{ij} \sim \mathrm{Bernoulli}(p_i)$ with p having CDF $\Phi_X(\Phi^{-1}(p))$. Whilst somewhat arbitrary, the approach taken here is easily comprehensible, treating X as an unobserved variable in an LDV problem. An alternative method would be to estimate a distribution for p directly: possible distributions include the Beta, Kumaraswamy, Raised Cosine, Triangular, Truncated Normal, and Uniform (the support must be inside [0,1]). As well as being analytically simple, my chosen functional form gives p to be distributed on the *whole* of [0,1] with extreme values being unlikely, regardless of parametrisation. None of these other distributions display this property.

k_i^{B} and k_i^{G} are the numbers of boys and girls borne by women i. ($k_i^{\mathrm{B}} + k_i^{\mathrm{G}} = k_i$.) Since only the numbers of boys and girls matter, the likelihood may be rewritten $\mathcal{L}(k_i^{\mathrm{B}}, k_i^{\mathrm{G}} | X_i) = \binom{k_i^{\mathrm{B}} + k_i^{\mathrm{G}}}{k_i^{\mathrm{B}}} \Phi(X_i)^{k_i^{\mathrm{B}}} (1 - \Phi(X_i))^{k_i^{\mathrm{G}}}$, with the combination factor accounting for the number of different ways that composition can occur.[11]

The unconditional likelihood is obtained by integrating over X, given its distribution:

$$\mathcal{L}_i = \mathcal{L}(k_i^{\mathrm{B}}, k_i^{\mathrm{G}}; \mu, \sigma) = \binom{k_i^{\mathrm{B}} + k_i^{\mathrm{G}}}{k_i^{\mathrm{B}}} \int_{\mathbb{R}} \Phi(x)^{k_i^{\mathrm{B}}} (1 - \Phi(x))^{k_i^{\mathrm{G}}} \, d\Phi_X(x) \tag{1}$$

The sample likelihood is the product of the women's individual likelihoods, $\mathcal{L}(\mu, \sigma) = \prod_i \mathcal{L}_i$, because the X_i are independent. Estimates of the parameters of the distribution of X may be made by maximising this likelihood: $(\hat{\mu}, \hat{\sigma}) = \arg\max_{\mu} \mathcal{L}(\mu, \sigma)$.[12]

3.3 Data and Estimations

Using data from the UK Labour Force Survey 1996–2005, I construct fertility histories for 116,513 British-born women aged 16–55. When a household enters the survey, a matrix of household relationships is recorded, so I match women with their natural children under the age of 16.[13] Table 1 records the compositions of these families. As can be seen in Equation 1, my estimator requires only the number of boys and girls in each household.

My estimates of the parameters underlying X are derived by Maximum Likelihood Estimation, as outlined in the previous section, and standard errors by the bootstrap method (200 bootstrap samples). Baseline estimates are shown in Table 2, with Figure 3 giving estimated values for the bootstrap samples.

[11] Under the assumption of only lexis variation (constant probabilities for each woman), including uncompleted families does not affect the estimator.

[12] Recall that μ and σ parameterise the distribution of X. They appear in Equation 1 in the via the integrating density, $d\Phi_X(x)$.

[13] The survey is conducted as a rolling panel with households appearing in the survey for five quarters. I use only the households' first-quarter responses. Birth histories ignore mortality and the possibility some children are absent from the household. I expect these omissions to have only minor effects on my results. I take the oldest child under the age of 16 to be the woman's first, following the example of Gangadharan and Maitra [2003]. The same caveats apply to the birth histories used in Section 4.2.

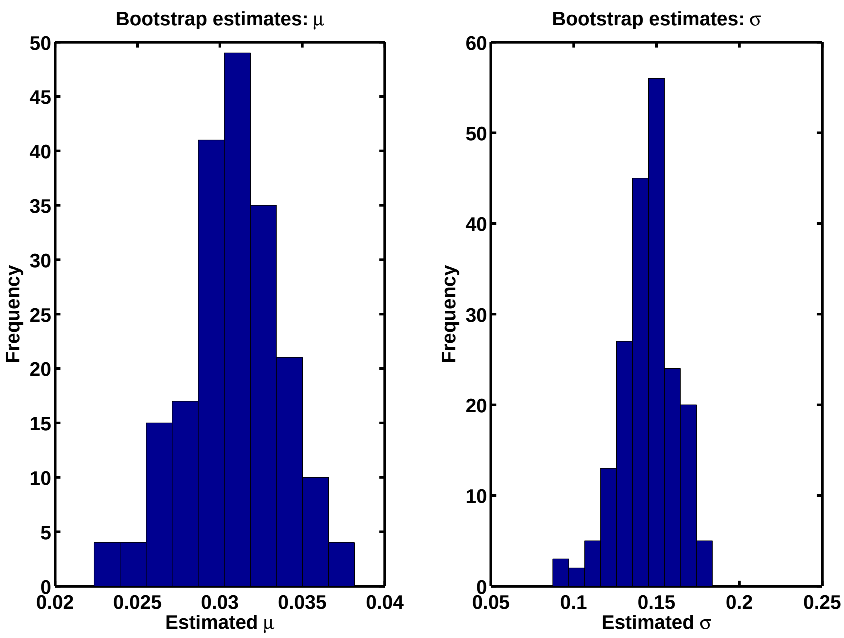
Children	Number of Boys 0	1	2	3	4	5	6	7	8	9	10
1	24526	25710	–	–	–	–	–	–	–	–	–
2	11205	24283	12014	–	–	–	–	–	–	–	–
3	1812	5100	5394	2182	–	–	–	–	–	–	–
4	233	744	1188	904	328	–	–	–	–	–	–
5	29	108	193	200	114	31	–	–	–	–	–
6	2	14	30	56	32	14	7	–	–	–	–
7	0	1	9	7	10	9	4	1	–	–	–
8	0	0	4	2	4	3	2	0	0	–	–
9	0	0	0	1	0	0	0	0	1	0	–
10	0	0	0	0	0	1	1	0	0	0	0

Table 1: Family composition data from the UK Labour Force Survey 1996–2005. British-born women aged 16–55.

Table 2: Maximum likelihood estimates of distribution of $X \sim \mathrm{N}(\mu, \sigma^2)$. The underlying probability that a woman has sons is $p = \Phi(X)$. Bootstrap standard errors in parenthesis (200 samples).

	Estimate	Std. Dev.	90% CI
$\hat{\mu}$	0.0305	(0.00285)	[0.0258, 0.0351]
$\hat{\sigma}$	0.145	(0.0165)	[0.118, 0.172]

Figure 3: Bootstrap estimates for parameters μ and σ of X.



Two implications are immediate: first, the median woman has natural probability of $\Phi(\mu) = 0.512$ of giving birth to a boy, yielding a 'natural' sex ratio of 1.050 boys per girl. This figure is lower than the usually cited ratio of 1.06 and the difference is statistically different at the 5% level.[14] However, my estimate lies within the range 1.03–1.06 given by Edlund [1999] as 'biologically normal'. The coherence of my estimate $\hat{\mu}$ with existing studies acts as a 'sanity check' on my methodology.

The second implication is more important: the variation in X, and hence p, is large. The baseline estimate gives a standard deviation for X of 0.145. This implies a 90% confidence interval (CI) of [0.42 , 0.61] for p. *Five percent of women are likely to have a boy with a greater than 61% chance, and five percent of women are likely to have a boy with a less than 42% chance.* Even taking a minimal value of σ (at the lower end of the 90% CI interval, $\sigma = 0.118$) yields a 90% confidence interval of [0.44, 0.59] for p.

I test for the significance of this heterogeneity with a likelihood ratio test. Homogeneity entails that X takes a degenerate distribution with value μ. Then all women would have boys with probability $\Phi(\mu)$. This case is equivalent to $\sigma = 0$, though now the likelihood function in Equation 1 is not well defined. However, the likelihood of a family comprising k_i^{B} boys and k_i^{G} girls nonetheless exists:

$$\mathcal{L}(k_i^{\mathrm{B}}, k_i^{\mathrm{G}}; \mu) = \binom{k_i^{\mathrm{B}} + k_i^{\mathrm{G}}}{k_i^{\mathrm{B}}} \Phi(\mu)^{k_i^{\mathrm{B}}} (1 - \Phi(\mu))^{k_i^{\mathrm{G}}}$$

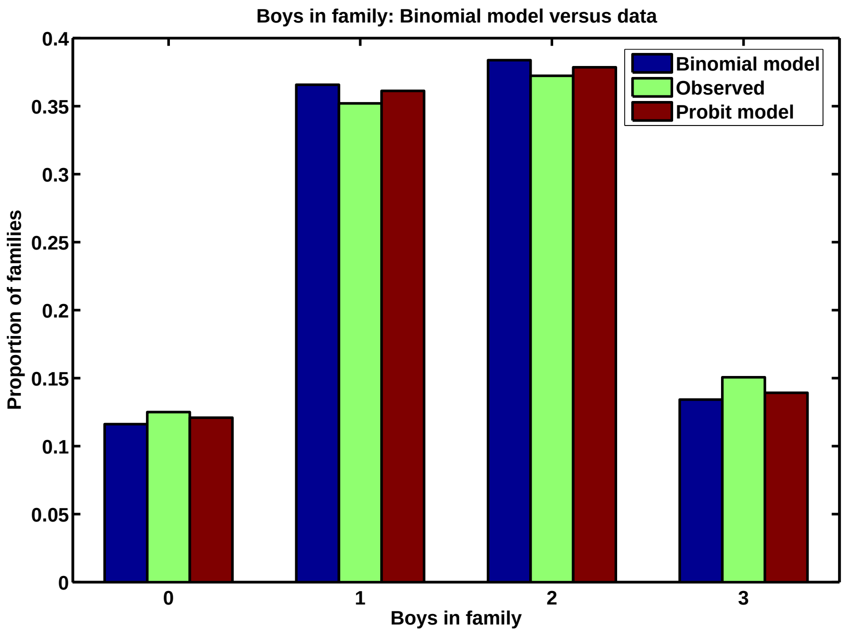
The restricted sample likelihood is analogous to the sample likelihood, being $\mathcal{L}(\mu) = \prod_i \mathcal{L}(k_i^{\mathrm{B}}, k_i^{\mathrm{G}}; \mu)$. MLE on this restricted model yields an estimate of $\check{\mu} = 0.0303$. Thus, the likelihood ratio statistic can be computed:

$$\Lambda = \frac{\max_{\mu} \mathcal{L}(\mu)}{\max_{\mu,\sigma} \mathcal{L}(\mu, \sigma)}$$

The restricted likelihood reduces the dimension of the problem by one, so the

[14] The median probability will be different from the mean, $\mathrm{E}_X\, \Phi(X)$. However, in this case they are identical to three significant figures.

Figure 4: Family composition (3 children) under: (a) Binomial model: $B \sim$ Bernoulli($\Phi(0.0303)$), (b) 'Probit' form heterogeneity $B \sim$ Bernoulli($\Phi(X)$), $X \sim \text{N}(0.0305, 0.145^2)$, and (c) Actual data (Table 1).



we have (asymptotically) $-2 \log \Lambda \sim \chi_1^2$. I am able to reject the null hypothesis that there is no heterogeneity at the 0.01% significance level ($-2 \log \Lambda = 16.8$). This result suggests that my model does indeed perform better than one in which all women have the same probability of having sons, as can be seen in Figure 4.

Lindsey and Altham [1998] find significant 'overdispersion' in family composition relative to a binomial model, as I do. Using their data, I am able to make a second estimate of heterogeneity in the probability of having sons.[15] Their data gives $\hat{\mu}' = 0.0367$ and $\hat{\sigma}' = 0.127$. The mean is significantly different from my original estimate at the 1% significance level, but the standard deviation estimate is not significantly different. Predicted son-probabilities are not practically dissimilar, with a 90% CI for p of [0.43, 0.60].

James [2000] surveys a variety of estimates of the standard deviation of p,

[15] Theirs is a sample of almost one million families from Saxony in the period 1876–1885, collected by Arthur Geissler.

centred on 0.05. My original estimates ($\hat{\mu} = 0.0305$,$\hat{\sigma} = 0.145$) give a standard deviation of 0.0572, and my estimates with the Lindsey and Altham data give 0.0502.

I can also compare Lindsey and Altham's model with my own. Under a fixed parametrisation with respect to family size, their Beta-Binomial model performs very similarly using Pearson's χ^2 test. However, as they note, they merely perform a data-fitting exercise and they highlight the lack of biological explanation behind their results. My model is perhaps more amenable to a biological explanation, since it is based on individual effects for each woman. Lindsey and Altham's model performs better than mine when parameters are allowed to vary with family size. They suggest that sex preferences are not the cause, since final children are omitted from their sample, though it is impossible to test this assertion with their data. In Appendix A, I propose a robustness check to deal with this concern.

Since Lindsey and Altham find family-size effects, women may not have constant probabilities of bearing sons over throughout their lives. This implies some Markov variation in the probabilities of birth [James, 2000], with p changing (monotonically) with birth order. My estimation technique does not account for such effects.

In sum, I conclude [after James, 2000] that lexis variation does exist in women's probability of having sons. In the next section I use my estimates to calculate the effect of parents' son preferences on the aggregate sex ratio in practice.

4 Sex ratio simulation

If parents have preferences for sons, heterogeneity in the chance that individual women have boys will lead to a skew in the sex ratio, as proved in Section 2. Here I attempt to calibrate the size of this effect using a simulation.

4.1 Procedure

I construct predictions of the sex ratio in countries worldwide based on two pieces of data: first, the underlying distribution of probabilities of having sons (estimated in the previous section), and the observed sex preferences of women from a variety countries (following Ellis [2008]).

Two assumptions underlie this calibration. First, I take as given that my heterogeneity estimates apply to women from all countries. My justification is that estimates from recent UK data (1996–2005) are similar to estimates based on data from Saxony in the late 19$^{\text{th}}$ century [Lindsey and Altham, 1998]. Both sets of estimates fall in line with previous estimations of lexis variation in the probability of having sons James [2000].

Second, and more problematically, it is necessary to assume that the fertility behaviour of immigrants to the UK is the same as that of women in their countries of origin. This assertion somewhat stretches the external validity of Ellis [2008]. Problems include: (1) emigrants being unrepresentative, (2) cost differences between childrearing in the UK and elsewhere, and (3) cultural assimilation within the UK. However, the first and third of these concerns are likely to bias measures of son preference down, relative to what they would be in the countries of origin. Women moving to the UK may be more likely to have preferences similar to British women, and absorption of local norms will reduce son preference (British women appear to show little son preference — see Appendix A).

The second concern — price differentials — may bias the measure up if girls cost more to raise than boys [Ellis, 2008, Section 2]. Women already having girls may reduce their fertility due to a wealth effect. However, in light of the large cultural effects found in that paper, it is likely that son preferences of immigrant women in the UK are less extreme than those of their compatriots. Therefore the effects predicted in this section may give a lower bound for the effect of parents' preferences on sex ratios worldwide.

Initially I generate a population of women ($i = 1 \dots N$) and assign them probabilities of having sons (p_i) using the distribution derived in Section 3. Then, for each woman, I construct a latent family composition ($\bar{B}_{ij}$ for $j = 1 \dots k$; $\bar{B}_{ij}$ indicates woman i's j^{th} child would a son). This gives the sexes of the women's (first) k children. This is the 'biological' population, which stays the same throughout the simulation.

$$\begin{aligned} p_i &= \Phi(X_i) \text{ with } X_i \sim \mathrm{N}(\hat{\mu}, \hat{\sigma}) \\ \bar{B}_{ij} &\sim \text{Bernoulli}(p_i) \end{aligned}$$

I then allow the fertility behaviour to vary by country. Let q_{bj} be the proportion of women from a given country who continue to have children, after having b boys amongst j children. (I measure these q_{bj} in Section 4.2, below.) For example, I take Indian women who have a boy and two girls and measure the proportion who go on to have a fourth child. This measurement is q_{13}^{INDIA}.

Finally, I simulate each woman's fertility decisions according to the observed patterns the data. I simulate the sex ratios for each country separately. I use the q probabilities to generate 'actual' birth histories for each woman in the simulation. If woman already has j children and b boys, child $j + 1$ is born with probability q_{bj}. This continues until the woman fails to have a child. I denote a birth of a j^{th} child to woman i by C_{ij}.

$$C_{ij+1} \sim \begin{cases} \text{Bernoulli}(q_{b_i j_i}) \text{ if } C_{ij'} = 1 \text{ for all } j' \leq j \\ 0 \text{ otherwise} \end{cases}$$

$$B_{ij+1} = \begin{cases} \bar{B}_{ij+1} \text{ if } C_{ij+1} = 1 \\ \text{[missing] otherwise} \end{cases}$$

This procedure provides me with a simulated sample of birth histories based on the sex preferences of women from each country. The ratio of boys to girls in this sample is easy to compute, and I compare this with the child sex ratios found in reality. Country data is taken from the World Development Indicators (1997), with the sex ratio being the under-15 male population divided by the under-15 female population.

4.2 Fertility Behaviour

4.2.1 Estimation method

Estimates of parental behaviour are derived from the birth histories of foreign-born women, grouping women by country of origin. For each possible family composition (up to three children), I calculate the asymptote of the Kaplan-Meier failure rate (KMFR). The KMFR is a non-parametric measure of the proportion of women who go on to have another child. The naïve progression rate, defined as the number of women *observed* to have a further child, does not account for the time women remain under observation (ie, censoring when the survey happens shortly after a birth). Thus, the KMFR is a more robust measure of true continuation rates.

Here, I consider childbirth as an absorbing transition from one state to another. Women either 'survive' with the number of children they have, or 'fail', and have another at some time. Alternatively, they may exit the dataset before failure (censoring). The asymptotic Kaplan-Meier failure rate is defined as follows [Jenkins, 2005, p. 55]. Let $t_1 < \ldots < t_m < \ldots < t_M$ be the observed transition times for women from a given country with a given family composition. For simplicity, assume transitions are never contemporary. Let n_m be the number of women at risk of making a transition immediately prior to t_m. This does not include women no longer under observation. The Kaplan-Meier estimate of the proportion surviving to time t is then:

$$\hat{S}(t) \stackrel{def}{=} \prod_{m|t_m<t} \left(1 - \frac{1}{n_m}\right)$$

The proportion of women surviving by t_1 is simply one minus proportion who have made a transition, which is estimated by the number of exits (one) divided by the number at risk, n_1. So $\hat{S}(t_1) = 1 - \frac{1}{n_1}$. Similarly, the proportion surviving from t_1 to t_2 is $1 - \frac{1}{n_2}$, so the overall proportion surviving to t_2 is thus the product of these: $\hat{S}(t_1) = 1 - (\frac{1}{n_1})(\frac{1}{n_1})$.

I am interested in the proportion of women, q, who have a child at any point in the future, ie, the asymptote of $1 - S(t)$ as $t \to \infty$. Therefore I have:

$$q \stackrel{def}{=} 1 - \hat{S}(\infty) = 1 - \prod_{m} \left(1 - \frac{1}{n_m}\right)$$

4.2.2 Data

Data is taken from the UK Labour Force Survey 1996–2005. Household composition records are available, so parents can be matched with their natural children and birth histories compiled. Table 3 gives estimated continuation rates for women from 55 countries for which more than 30 records are present at the second birth.[16]

As demonstrated in Ellis [2008], considerable variation in behaviour is seen amongst the women from different countries. For example, Australian women show some preference for daughters: 50% of those having two sons have a third child, compared with 37% of those with two daughters. Those with mixed

[16]Not all countries of origin in the LFS are identified uniquely. Five groups of countries appear in my regressions; I define the sex ratio for these to be total boys over total girls. The groups are are:

families are more likely to stop. On the other hand, Singapore shows a different pattern: 33% of those with three daughters have a fourth child compared with none of those with three sons.

4.3 Simulation results

Using the fertility behaviour for women from different countries, I compute expected sex ratios for those countries. I use three sets of estimates for the underlying likelihoods of having sons: my benchmark derived in Section 3 ($\mu = 0.0305, \sigma = 0.145$), my estimate using data from Lindsey and Altham [1998] ($\mu = 0.0367, \sigma = 0.127$), and a conservative estimate of the heterogeneity in my sample, at the bottom of the 90% CI ($\mu = 0.0305, \sigma = 0.118$). I simulate births for one million women (the same latent birth histories are used for each country estimate). Results are presented in Table 4, alongside the child sex ratio from the World Development Indicators (under-15 male population divided by under-15 female population).

Simulated sex ratios range from 1.043 (Colombia) to 1.051 (Australia) in my benchmark model. The ranges are similar for all three sets of estimates, though they are higher overall for the estimates with Lindsey and Altham's data. The larger estimate for μ shifts the whole distribution in favour of having sons.

Figure 5 plots my simulated sex ratios against the data and fitted values from the regression $SR_{\text{PREDICTED}} = \alpha + \beta SR_{\text{DATA}}$. R^2 is 0.053, and the estimate $\hat{\beta} = 0.015$ is significantly different to zero at the 10% level, suggesting that parents' preferences do indeed have an effect on country sex ratios due to heterogeneity in the probability of having a son. However, the intercept, $\hat{\alpha}$ is

Grp02 'Other Caribbean Commonwealth': Antigua and Barbuda, Bahamas, Dominica, Grenada, Solomon Islands, St. Kitts and Nevis, St. Lucia, St. Vincent and the Grenadines.
Grp04 'Other Africa': Benin, Burkina Faso, Burundi, Cameroon, Cape Verde, Central African Republic, Chad, Congo, Rep., Cote d'Ivoire, Djibouti, Equatorial Guinea, Eritrea, Gabon, Guinea, Guinea-Bissau, Liberia, Madagascar, Mali, Mauritania, Mozambique, Namibia, Niger, Rwanda, Sao Tome and Principe, Senegal, Togo.
Grp07 'Other South America': Bolivia, Ecuador, Paraguay, Peru, Suriname
Grp08 'Other Middle East': Bahrain, Jordan, Kuwait, Oman, Qatar, Saudi Arabia, Syrian Arab Republic, United Arab Emirates, West Bank and Gaza, Yemen.
Grp12 Afghanistan, Bhutan, Maldives, Nepal.

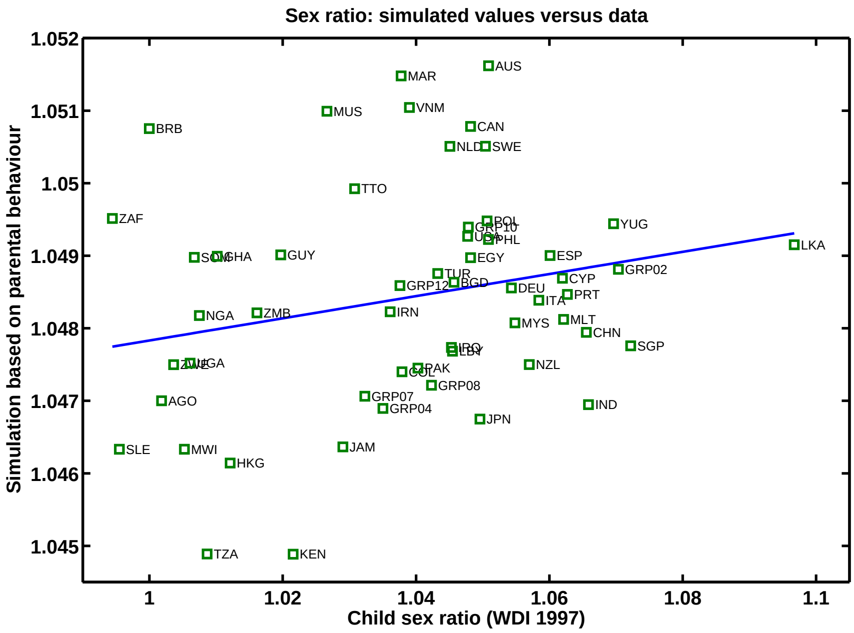
Table 3: Kaplan-Meier birth continuation rates by family composition. Foreign-born women in the UK LFS are grouped by country of origin. Statistics give the proportion of women having a further child given they already have b boys of j children.

	1 child		2 children			3 children			
Country	0 boys	1 boy	0 boys	1 boy	2 boys	0 boys	1 boy	2 boys	3 boys
AGO	0.93	0.88	0.67	0.33	0.33	–	–	–	–
AUS	0.76	0.78	0.37	0.44	0.50	0.40	0.13	0.34	1.00
BGD	0.91	0.93	0.84	0.82	0.81	0.74	0.73	0.71	0.61
BRB	0.62	0.70	0.43	0.22	0.55	–	–	–	–
CAN	0.76	0.75	0.37	0.34	0.34	0.00	0.13	0.17	0.60
CHN	0.58	0.60	0.40	0.34	0.07	–	–	–	–
COL	0.65	0.66	0.37	0.77	0.00	–	–	–	–
CYP	0.76	0.76	0.36	0.40	0.43	0.50	0.27	0.24	0.40
DEU	0.71	0.73	0.44	0.41	0.46	0.12	0.29	0.35	0.30
EGY	0.66	0.71	0.35	0.52	0.46	–	–	–	–
ESP	0.64	0.63	0.37	0.48	0.27	–	–	–	–
FRA	0.71	0.74	0.41	0.31	0.28	0.29	0.11	0.21	0.50
GHA	0.81	0.74	0.61	0.62	0.57	0.43	0.28	0.61	0.50
GRP02	0.68	0.59	0.15	0.38	0.28	–	–	–	–
GRP04	0.81	0.80	0.76	0.64	0.49	0.40	0.28	0.45	0.50
GRP07	0.78	0.82	0.43	0.37	0.27	–	–	–	–
GRP08	0.86	0.85	0.80	0.73	0.58	0.75	0.45	0.55	0.55
GRP12	0.90	0.93	1.00	0.69	0.63	–	–	–	–
GUY	0.67	0.64	0.25	0.38	0.54	–	–	–	–
HKG	0.76	0.67	0.47	0.32	0.35	0.53	0.20	0.19	0.30
IND	0.81	0.75	0.59	0.39	0.39	0.53	0.31	0.27	0.22
IRN	0.80	0.69	0.14	0.28	0.12	–	–	–	–
IRQ	1.00	0.81	0.73	0.56	0.75	–	–	–	–
ITA	0.56	0.72	0.45	0.33	0.20	–	–	–	–
JAM	0.57	0.65	0.71	0.43	0.30	0.70	0.41	0.28	0.53
JPN	0.67	0.64	0.44	0.22	0.13	–	–	–	–
KEN	0.83	0.75	0.64	0.29	0.31	0.53	0.19	0.20	0.21
LBY	0.85	0.71	0.55	0.69	0.72	–	–	–	–
LKA	0.73	0.70	0.16	0.29	0.41	–	–	–	–
MAR	0.91	0.81	0.34	0.53	0.75	–	–	–	–
MLT	0.76	0.62	0.52	0.38	0.41	–	–	–	–
MUS	0.68	0.65	0.30	0.33	0.61	–	–	–	–
MWI	0.91	0.76	0.64	0.59	0.42	–	–	–	–
MYS	0.86	0.75	0.42	0.32	0.46	0.00	0.42	0.27	0.38
NGA	0.89	0.86	0.79	0.60	0.68	0.73	0.47	0.53	0.55
NLD	0.76	0.77	0.40	0.39	0.74	–	–	–	–
NZL	0.71	0.62	0.52	0.37	0.46	–	–	–	–
PAK	0.92	0.87	0.84	0.78	0.74	0.80	0.68	0.65	0.76
PHL	0.54	0.71	0.21	0.25	0.21	–	–	–	–
POL	0.58	0.49	0.05	0.29	0.47	–	–	–	–
PRT	0.71	0.63	0.55	0.26	0.53	–	–	–	–
SGP	0.76	0.76	0.38	0.39	0.30	0.33	0.29	0.25	0.00
SLE	0.75	0.56	0.54	0.42	0.40	–	–	–	–
SOM	0.91	0.92	0.85	0.86	0.86	0.92	0.79	0.81	0.90
SWE	0.82	0.82	0.00	0.33	0.24	–	–	–	–
TTO	0.66	0.67	0.55	0.21	0.61	–	–	–	–
TUR	0.82	0.83	0.57	0.46	0.50	0.25	0.35	0.44	0.38
TZA	0.79	0.63	0.54	0.30	0.24	0.40	0.15	0.14	0.33
UGA	0.78	0.74	0.52	0.32	0.41	0.36	0.45	0.37	0.38
USA	0.78	0.71	0.43	0.39	0.49	0.25	0.41	0.44	0.20
VNM	0.79	0.88	0.41	0.40	0.56	–	–	–	–
YUG	0.84	0.77	0.49	0.43	0.58	–	–	–	–
ZAF	0.76	0.73	0.44	0.36	0.34	0.25	0.00	0.32	0.25
ZMB	0.77	0.77	0.71	0.48	0.53	0.25	0.52	0.18	1.00
ZWE	0.71	0.69	0.42	0.43	0.40	0.33	0.37	0.19	0.25

Table 4: Simulated sex ratios and actual data. Three different estimates of the underlying distribution of sexes are used: my benchmark, estimates from Lindsey and Altham's 1998 data, and a conservative estimate from my data. Parental behaviour is taken from UK LFS data (Table 3). Country sex ratios (under-15 male population divided by the under-15 female population) are taken from the World Development Indicators (1997).

		Predicted Values		
		$\mu = 0.0305$	$\mu = 0.0376$	$\mu = 0.0305$
Country	Data	$\sigma = 0.145$	$\sigma = 0.127$	$\sigma = 0.118$
AGO	1.002	1.047	1.060	1.048
AUS	1.051	1.052	1.063	1.051
BGD	1.046	1.049	1.061	1.049
BRB	1.000	1.051	1.062	1.050
CAN	1.048	1.051	1.063	1.051
CHN	1.066	1.048	1.060	1.049
COL	1.038	1.047	1.060	1.048
CYP	1.062	1.049	1.061	1.049
DEU	1.054	1.049	1.060	1.049
EGY	1.048	1.049	1.061	1.049
ESP	1.060	1.049	1.061	1.050
FRA	1.048	1.049	1.061	1.050
GHA	1.010	1.049	1.061	1.049
GRP02	1.070	1.049	1.061	1.049
GRP04	1.035	1.047	1.059	1.048
GRP07	1.032	1.047	1.059	1.048
GRP08	1.042	1.047	1.060	1.048
GRP12	1.038	1.049	1.061	1.049
GUY	1.020	1.049	1.061	1.049
HKG	1.012	1.046	1.059	1.047
IND	1.066	1.047	1.059	1.048
IRN	1.036	1.048	1.061	1.049
IRQ	1.045	1.048	1.060	1.049
ITA	1.058	1.048	1.060	1.049
JAM	1.029	1.046	1.059	1.048
JPN	1.050	1.047	1.059	1.048
KEN	1.022	1.045	1.058	1.047
LBY	1.045	1.048	1.060	1.048
LKA	1.097	1.049	1.061	1.049
MAR	1.038	1.051	1.063	1.051
MLT	1.062	1.048	1.061	1.049
MUS	1.027	1.051	1.063	1.051
MWI	1.005	1.046	1.059	1.047
MYS	1.055	1.048	1.060	1.049
NGA	1.007	1.048	1.060	1.049
NLD	1.045	1.051	1.062	1.050
NZL	1.057	1.047	1.060	1.048
PAK	1.040	1.047	1.060	1.048
PHL	1.051	1.049	1.061	1.049
POL	1.051	1.049	1.061	1.049
PRT	1.063	1.048	1.061	1.049
SGP	1.072	1.048	1.060	1.048
SLE	0.995	1.046	1.059	1.048
SOM	1.007	1.049	1.061	1.049
SWE	1.050	1.051	1.062	1.050
TTO	1.031	1.050	1.062	1.050
TUR	1.043	1.049	1.061	1.049
TZA	1.009	1.045	1.058	1.046
UGA	1.006	1.048	1.060	1.048
USA	1.048	1.049	1.062	1.050
VNM	1.039	1.051	1.063	1.051
YUG	1.070	1.049	1.062	1.050
ZAF	0.994	1.050	1.062	1.050
ZMB	1.016	1.048	1.060	1.048
ZWE	1.004	1.047	1.060	1.048

Figure 5: Simulated sex ratios and country data. Simulation is based on heterogeneity in the probability of a son as derived in Section 3 ($\mu = 0.0305, \sigma = 0.145$). Parental behaviour is given in Table 3.



1.033, and is significantly different from one at 0.1%, implying that my model does not capture the whole story. Regressions with the other distributional estimates produce similar results.

The estimated slope, 0.015, is shallow. It suggests that parental behaviour and underlying heterogeneity accounts for 1.5% of the difference in sex ratios between countries. There are several possible reasons for this small figure. First is data quality. Aside from the usual noise issues, I have used child sex ratios, not birth sex ratios. If son-preferring fertility behaviour is correlated with discriminatory behaviour (as is likely), countries that are expected to have excess girls at birth will also have high female mortality, reducing the number of girls as I measure them. Under this assumption, my estimate of 1.5% will be biased *downwards*, since mortality due to discrimination lowers the excess of girls I am looking for. Data quality would be one area in which this study could be improved.

My two identifying assumptions may be incorrect: behaviour of immigrants to the UK may not represent countries of origin in terms of preferences, or the underlying distribution of 'natural' probabilities of sons is not identical across countries for biological or social reasons [cf. Oster, 2005; Matthews et al., 2008]. However, as discussed in Section 4.1, the former of these caveats may work against finding a positive result, if immigrants in the UK prefer sons less than their compatriots. The second concern is allayed by James [2000] and others who find similar levels of heterogeneity.

Finally, in a model simulating at most four children, the theoretical minimum and maximum sex ratios are 1.030 and 1.068 respectively (under my benchmark son-probability estimates). These ratios occur when parents practice extremely selective behaviour: stopping only after one son or daughter. The range seen in the data is [0.994, 1.097], so child mortality plainly plays a larger role in affecting sex ratios than parental preferences do.

Nonetheless, 1.5% very probably represents a lower bound on the effect of parental preferences on aggregate sex ratios. Most importantly, mortality differences between girls and boys in the countries of origin will likely bias my estimate downwards. Moreover, parental behaviour of immigrants to the UK is likely to be less extreme than women in their originating countries.

5 Discussion

For almost fifty years, it has been recognised that parental preferences can influence aggregate sex ratios when women have heterogeneous 'natural' probabilities of having boys [Yamaguchi, 1989]. However, the literature on 'missing women' has, to date, failed to account for such effects. In the present paper I attempt to calibrate the size of this effect.

My simulation gives that 1.5% of differences in country sex ratios are explained by parental preferences. However, as I discuss in Section 4.3, this figure is likely to be a lower bound on the potential effect. The major causes are attenuated preferences amongst immigrants to the UK and the limitation of using child sex ratios rather than birth sex ratios. My estimates of heterogeneity in son-probability are broadly in line with previous work by Lindsey and Altham [1998] and authors suveyed by James [2000], suggesting that misspecifiction of that distribution is not a concern.

Due to limited data and the caveats given above, I do not compute any adjusted estimates of the number of missing women worldwide. Nonetheless, the implications of my findings are not heartening: there may be more missing women than previously thought. My model predicts that a country with strong son preferences will have a *low* sex ratio at birth, because heterogeneity in the probability of having a son leads to excess girls, relative to homogeneity. Previous comparisons have used a baseline that is *not* biased towards girls. Thus, countries such as China or India which are known to favour boys [Das Gupta et al., 2002] are missing *more* girls than we had believed. At the very least, the results presented here should caution the use of aggregate sex ratios as a measure of attitudes, particularly in light of the negative correlation with parental preferences (Figure 1).

As Das Gupta [2005] notes, social and cultural preferences for sons play a prime role in skewing sex ratios in several developing countries, whether due to selective abortion, infanticide, or higher mortality rates for girls. I present evidence that fertility decisions guided by the same parental preferences can also lead to biased sex ratios in aggregate. Since son preferences bias the ratio downward, previous estimates of the number of missing women may be too low. Future work should aim to better account for the effect of parents' preferences.

Bibliography

Perianayagam Arokiasamy, 2004. Regional Patterns of Sex Bias and Excess Female Child Mortality in India. *Population*, 59(6):833–863.

Rubiana Chamarbagwala and Martin Ranger, 2006. India's Missing Women: Disentangling Cultural, Political and Economic Variables. Working paper, Indiana University, Kelley School of Business, Department of Business Economics and Public Policy.

Ansley J. Coale, 1991. Excess Female Mortality and the Balance of the Sexes in the Population: An Estimate of the Number of "Missing Females". *Population and Development Review*, 17(3):517–523.

Monica Das Gupta, 2005. Explaining Asia's "Missing Women": A New Look at the Data. *Population and Development Review*, 31:529–535.

Monica Das Gupta, Jiang Zhenghua, Li Bohua, Xie Zhenming, Woo-jin Chung, and Bae Hwa-Ok, 2002. Why is son preference so persistent in East and South Asia? A cross-country study of China, India, and the Republic of Korea. Policy Research Working Paper 2942, The World Bank.

Jean Drèze and Amartya Sen, 1989. *Hunger and public action*. Oxford: Clarendon Press.

Sylvie Dubuc and David Coleman, 2007. An Increase in the Sex Ratio of Births to India-born Mothers in England and Wales: Evidence for Sex-Selective Abortion. *Population and Development Review*, 33(2):383–400.

Lena Edlund, 1999. Son Preference, Sex Ratios, and Marriage Patterns. *The Journal of Political Economy*, 107(6):1275–1304.

Jas Ellis, 2007. Prices, Norms and Preferences: The Influence of Cultural Values on Fertility. Working paper, London School of Economics. Avaliable at http://personal.lse.ac.uk/ellisj/.

Jas Ellis, 2008. Son Preference and Culture. Working paper, London School of Economics. Avaliable at http://personal.lse.ac.uk/ellisj/.

Raquel Fernández and Alessandra Fogli, 2005. Culture: An Empirical Investigation of Beliefs, Work, and Fertility. NBER Working Paper 11268, National Bureau of Economic Research.

Raquel Fernández and Alessandra Fogli, 2006. Fertility: The Role of Culture and Family Experience. *Journal of the European Economic Association*, 4:552–561.

Lata Gangadharan and Pushkar Maitra, 2003. Testing for Son Preference in South Africa. *Journal of African Economies*, 12(3):371–416.

L. A. Goodman, 1961. Some possible effects of birth control on the human sex ratio. *Annals of Human Genetics*, 25(1):75–81.

J. Graffelman and R. F. Hoekstra, 2000. A statistical analysis of the effect of warfare on the human secondary sex ratio. *Human Biology*, 73:433–445.

Valerie J. Grant, 1996. Sex determination and the maternal dominance hypothesis. *Human Reproduction*, 11:111–125.

Paula Griffiths, Zoe Matthews, and Andrew Hinde, 2000. Understanding the Sex Ratio in India: A Simulation Approach. *Demography*, 37(4):477–488.

William H. James, 1994. Parental dominance/social status, hormone levels, and sex ratio of offspring. In *Reproductive and interpersonal aspects of dominance and status*, edited by Lee Ellis, volume 2 of *Social stratification and socioeconomic inequality*. Westport, Conn.: Praeger.

William H. James, 1995. What stabilizes the sex ratio? *Annals of Human Genetics*, 59:243–249.

William H. James, 2000. The variation of the probability of a son within and across couples. *Human Reproduction*, 15(5):1184–1188.

Stephen P. Jenkins, 2005. *Survival Analysis*. Institute for Social and Economic Research, University of Essex.

Theodore D. Kemper, 1994. Social stratification, testosterone, and male sexuality. In *Reproductive and interpersonal aspects of dominance and status*, edited by Lee Ellis, volume 2 of *Social stratification and socioeconomic inequality*. Westport, Conn.: Praeger.

Siu Fai Leung, 1988. On tests for sex preferences. *Journal of Population Economics*, 1:95–114.

Siu Fai Leung, 1991. A Stochastic Dynamic Analysis of Parental Sex Preferences and Fertility. *Quarterly Journal of Economics*, 106(4):1063–88.

Ming-Jen Lin, Jin-Tan Liu, and Nancy Qian, 2008. More missing women, fewer dying girls: the impact of abortion on sex ratios at birth and excess female mortality in Taiwan. CEPR Discussion Paper 6667, Centre for Economic Policy Research.

J. K. Lindsey and P. M. E. Altham, 1998. Analysis of the Human Sex Ratio by Using Overdispersion Models. *Applied Statistics*, 47(1):149–157.

Fiona Matthews, Paul J. Johnson, and Andrew Neil, 2008. You are what your mother eats: evidence for maternal preconception diet influencing foetal sex in humans. *Proceedings of the Royal Society B: Biological Sciences*, 275(1643):1661–1668.

Peter Mayer, 1999. India's Falling Sex Ratios. *Population and Development Review*, 25(2):323–343.

Karen Norberg, 2004. Partnership status and the human sex ratio at birth. *Proceedings of the Royal Society B*, 271:2043–2410.

Emily Oster, 2005. Hepatitis B and the Case of the Missing Women. *Journal of Political Economy*, 113:1275–1304.

Emily Oster, 2006. On Explaining Asia's "Missing Women": Comment on Das Gupta. *Population and Development Review*, 32(2):323–327.

Nancy Qian, 2006. Missing Women and the Price of Tea in China: The Effect of Sex-Specific Earnings on Sex Imbalance. CEPR Discussion Paper 5986, Centre for Economic Policy Research.

Amartya Sen, 1989. Women's Survival as a Development Problem. *Bulletin of the American Academy of Arts and Sciences*, 43(2):14–29.

Amartya Sen, 1992. Missing women. *British Medical Journal*, 304:587–588.

M. Sheps, 1963. Effects on family size and sex ratio of preferences regarding the sex of children. *Population Studies*, 17:66–72.

H. Weiler, 1959. Sex Ratio and Birth Control. *American Journal of Sociology*, 65(3):298.

Kazuo Yamaguchi, 1989. A Formal Theory for Male-Preferring Stopping Rules of Childbearing: Sex Differences in Birth Order and in the Number of Siblings. *Demography*, 26(3):451–465.

A Robustness of heterogeneity estimates

If parents base their fertility decisions on their current family composition, there may be concerns that my estimates of heterogeneity in son-probability are biased. In particular, the measured heterogeneity might be a result of differences in behaviour due to preferences over family composition. Here, I test for the effect of observed parental behaviour in a simulation. The exercise is very similar to that of Section 4, however the outcome of interest if not the final sex ratio, but the estimates of underlying heterogeneity.

Following the model of Section 2, suppose that the true distribution of the underlying factor X were $\mathrm{N}(\tilde{\mu}, \tilde{\sigma}^2)$. I generate a population of such women and also a latent family composition for each woman. Next, I simulate each woman's fertility decisions according to the observed patterns amongst the women in my sample. If some woman has two boys and a girl, I set the probability of her having a fourth child to be that observed amongst actual women with those children, giving simulated birth histories for each of the women. See Section 4.1 for details.

The key step is to compute estimates $\hat{\mu}$ and $\hat{\sigma}$ based on these simulated birth histories by the MLE method of Section 3. By comparing $\hat{\mu}$ and $\hat{\sigma}$ to $\tilde{\mu}$ and $\tilde{\sigma}$, the robustness of the estimator to parental fertility decisions can be assessed. In the case where $\hat{\mu} \approx \tilde{\mu}$ and $\hat{\sigma} \approx \tilde{\sigma}$ for measured values, these estimates can be considered robust to son and daughter preferences.

Table 5 records birth progression rates for the UK-born women in my sample, grouped by existing family composition. These estimates are computed exactly as in Section 4.2. This data gives the proportion q_{bj} of women who go on to have another child if they already have b boys amongst j children.[17]

Figure 6 gives estimated $\hat{\mu}$ and $\hat{\sigma}$ based on various values of $\tilde{\mu}$ and $\tilde{\sigma}$ in a simulation with one million women. Tested values are encompass the parameter

[17] These British-born women appear to have preferences for mixed families: after two or three children, progression rates are lowest with children of each sex. Conversely, after four and five children some son preference is seen. It is feasible that these women having more children are not representative, and both the larger family and son preference is driven by some unobserved factor, such as being a second generation migrant.

Table 5: Kaplan-Meier birth continuation rates by family composition for UK-born women. Sample selection is that in Section . Statistics give the proportion of women having a further child given they already have b boys of j children.

Childen	0 boys	1 boy	2 boys	3 boys	4 boys	5 boys
1	0.70	0.69	–	–	–	–
2	0.40	0.33	0.42	–	–	–
3	0.35	0.30	0.30	0.34	–	–
4	0.38	0.35	0.30	0.28	0.31	–
5	0.58	0.46	0.44	0.40	0.34	0.34

confidence intervals derived by the bootstrap procedure of Section 3.3. For each parameter, estimates are very close to the posited values. Correlation in either case is 1.00 (three significant figures), and the slopes are also very close to one. I conclude that the preferences held by British-born women are very unlikely to bias my estimates of heterogeneity in the probability of having sons.

Figure 6: Estimated $\hat{\mu}$ and $\hat{\sigma}$ based on different possible 'true' values $\tilde{\mu}$ and $\tilde{\sigma}$. 1,000,000 women simulated, following the fertility behaviour given in Table 5.