University of Glasgow

Adam Smith
Business School

WORKING PAPER SERIES

Uncovering Son Preference Cultures from Migrants under Endogenous Destination Selection

Kai Liu, Xiangqing Liu, Anwen Zhang and Zhongliang Deng

Paper No. 2026-05
June 2026

Uncovering Son Preference Cultures from Migrants under Endogenous Destination Selection*

Kai Liu† Xiangqing Liu‡ Anwen Zhang§ Zhongliang Deng¶

June 2026

Abstract

This paper examines the cultural component of son preference through rural-to-urban migration in China, using a large representative dataset covering migrants from all origins to all destinations. Employing a generalized Roy model with a control function approach, we show that conventional epidemiological-approach estimates overstate cultural persistence. Correcting for endogenous destination selection reduces the estimated effect of origin-specific son preference on migrants' fertility behavior by 57–71%. Differences among migrants in the same destination therefore reflect both cultural influences and non-random sorting. We further find that cultural persistence weakens in higher-income destinations but strengthens in areas with greater migrant concentration.

JEL codes: J13, J61, Z13

Keywords: Son preference; cultural persistence; endogenous migration; destination selection; fertility behavior; sex ratios; rural-to-urban migration; China

*We thank Li Gan, Lingsheng Meng, Wenjie Wu, Yiming Xia, and the seminar and workshop participants at several institutions for helpful comments and suggestions. Zhang acknowledges financial support from the British Academy and Leverhulme Trust (Project No.: SRG1920\101705). All remaining errors are ours.

†Faculty of Economics, University of Cambridge. Email: kai.liu@econ.cam.ac.uk

‡International Business School Suzhou, Xi'an Jiaotong-Liverpool University. Email: xiangqing.liu@xjtlu.edu.cn

§Adam Smith Business School, University of Glasgow; GLO. Email: anwen.zhang@glasgow.ac.uk

¶Chinese Academy of Social Sciences. Email: dengzl@cass.org.cn

1 Introduction

Son preference has played an important role in shaping demographic patterns across the world. Practices such as sex-selective abortion and the neglect of girls have led to more than 100 million "missing women" globally (Sen, 2002). These gender imbalances have distorted marriage and labor markets (Angrist, 2002; Dahl and Moretti, 2008; Brainerd, 2017; Grosjean and Khattar, 2019; Huang et al., 2025), and have far-reaching social consequences, such as on crime (Edlund et al., 2013; Cameron et al., 2019) and savings (Wei and Zhang, 2011; Horioka and Terada-Hagiwara, 2017).

It remains unclear whether son preference should be viewed as a structural challenge or an isolated cultural norm. Culture—defined as the customary beliefs and values transmitted across generations within social groups (Guiso et al., 2006)—is highly persistent and shapes deep-seated preferences for male offspring. On the other hand, structural and contextual factors, such as economic conditions and institutional settings, can also influence son preference. For instance, in traditional agricultural societies, men were often more productive in physically intensive farm work and typically inherited land, reinforcing the economic value of sons relative to daughters (Alesina et al., 2013).

This paper provides evidence on the effect of culture on son preference, separating it from the influence of contemporaneous contextual factors and exploring its determinants. Our evidence is relevant for the broad policy debate about how to combat gender imbalance and its socio-economic consequences: if son preference is primarily driven by culture effect, interventions should focus on long-term shifts in social attitudes; on the other hand, if the culture effect is small, changes to the economic environment and institutional reforms may be more relevant and effective.

To identify the cultural component of son preference, we compare migrants from different rural origins who have moved to the same urban destination in the post-reform era of China. Our analysis draws on three key institutional features. First, China has experienced large-scale rural-to-urban migration, estimated at 145 million in 2009 (Meng, 2012). This mass movement occurs against a large and pre-existing rural-urban divide in son preference, offering a natural setting to examine cultural persistence (or assimilation) in the preference for sons.[1] Second, the One-Child Policy (OCP) affected sex ratios at birth to varying extent across different areas, which provides a credible proxy to measure son pref-

[1] Most rural migrants move to urban areas by their early 20s, exposing themselves to a new urban environment while retaining cultural norms and preferences shaped by their place of origin.

erence cultural norms across areas (Li et al., 2011; Das Gupta, 2010). Third, policy and technological factors made an individual's son preference more discernible from their fertility outcomes. The imposed birth quota and exception rules under the OCP, non-compliance under imperfect enforcement, and the availability of sex selection through ultrasound technology (Chen et al., 2013) together render both the decision to have a second child and that child's sex especially informative about an individual's son preference.

A well-established literature has used migrants to separate culture effects from the influence of contextual factors on a variety of outcomes following an epidemiological approach (Fernández, 2011). Despite the popularity of this approach, we tackle an identification problem that has been largely overlooked—endogenous selection of migration destinations. This issue of non-random selection, long recognized in the migration literature (Roy, 1951; Borjas, 1987), implies that observed differences between migrants of different origins may reflect unobserved individual characteristics rather than culture per se. How much of son preference in migrant fertility behavior is genuinely cultural transmission from their origin, versus an artifact of who migrates where?

To illustrate, consider a simple scenario where migrants to an urban destination originate from two rural regions $j = \{1, 2\}$. These rural origins differ in their cultural traits (e.g., sex ratio, s_j) and costs of migration, but are otherwise identical; importantly, unobserved individual skills (ϵ_i) are identically distributed across j prior to migration. Suppose that the destination context affects all migrants equally. Then, comparing outcomes (y_i) among migrants from different origins eliminates contextual differences, leaving the observed difference as a combination of the culture effect and a selection effect:

$$
\begin{aligned}
&E(y_i|\text{ migrants from j=2}) - E(y_i|\text{ migrants from j=1}) \\
&= \underbrace{\beta(s_2 - s_1)}_{\text{culture effect}} + \underbrace{E(\epsilon_i|\text{ migrants from j=2}) - E(\epsilon_i|\text{ migrants from j=1})}_{\text{selection effect}}. \qquad (1)
\end{aligned}
$$

If it is more costly to migrate from origin 2 than origin 1, a Roy model of migration will predict that migrants from 2 are likely to be positively selected in terms of ϵ_i relative to those from 1. This selection effect leads to an upward bias in the estimated culture effect (β). More generally, we show and provide empirical evidence that the prevailing culture effect may be confounded with selection into migration destinations.

To separate culture effect from selection, we combine the potential outcomes framework with a generalized Roy model of migration decisions. In this model, migration decisions are multidimensional: a migrant from a given origin chooses among *multiple* possible destinations. Consequently, the potential outcome is specific to each origin–destination pair, and migration selection manifests as a multi-market selection problem. Following Dahl (2002), we use a semi-parametric control function approach to correct for selection. The control functions are estimated nonparametrically using migration probabilities and do not rely on strong assumptions about the joint distribution of outcomes and choice shocks.

Our analysis leverages the China Migrants Dynamic Survey (CMDS), a large representative dataset covering migrants from all origins to *all* destinations within China. This dataset offers two advantages over previous studies that have compared migrants originating from different origins to a *single* destination. First, from a methodological perspective, observing migration flows from every origin to all potential destinations is essential for constructing the control functions and correcting for multi-market selection. Second, it allows us to identify the distribution and heterogeneity of culture effects across destinations. After all, cultural persistence is unlikely to be uniform; rather, it may vary depending on local institutional and economic contexts.

We have three sets of main findings. First, cultural persistence in son-favoring fertility behavior (e.g., gender outcomes at the second birth) tends to be overstated when endogenous selection of migration destination is ignored. Once selection is taken into account, the estimated effect of origin son preference is substantially smaller. Our results therefore suggest that observed fertility differences among migrants from different backgrounds living in the same destination cannot be interpreted as purely cultural; in some destinations, a significant share of the gap reflects endogenous sorting rather than genuine cultural differences.

Second, we show that most of the selection effects are driven by selection on unobservable characteristics rather than observables, indicating that controlling for observed characteristics is unlikely to reduce the bias. We also show that the selection pattern is inconsistent with Tiebout migration based on preferences, where people "vote with their feet" by choosing residential locations that best align with their preferences (Tiebout, 1956).

Third, we find that the spatial distribution of culture persistence can be explained by a horse race between migrants' peer environment and average income level in the destination. Our estimates suggest that, *ceteris paribus*, high initial local income reduces cultural persistence, and a large share of migrants in the destination increases cultural persistence. For migrants moving to an area with high initial income, the income effects tend to weaken the persistence of son-preference culture. At the same time, high-income areas also attract a large number of migrants, which reduces the exposure that migrants have to the local social environment. This, in turn, can strengthen the persistence of culture. Overall, culture effects are jointly shaped by the dynamics of social/peer environment and local income.

A large literature has exploited variations in migrants' backgrounds to estimate the impact of culture on a wide range of outcomes, including fertility (Fernández and Fogli, 2006), savings (Carroll et al., 1999), labor supply (Fernández, 2007; Fernández and Fogli, 2009; Blau et al., 2011; Boelmann et al., 2025), household living arrangements (Giuliano, 2007), and social and political preferences (Luttmer and Singhal, 2011; Alesina and Giuliano, 2011; Hauge et al., 2023). In her comprehensive review, Fernández (2011) refers to this broad methodology as the "epidemiological approach".

We contribute to this literature by formally assessing the bias introduced by general forms of migration selection in the estimation of cultural persistence. To the best of our knowledge, this is the first paper that has explicitly quantified the impact of multi-market migration selection on culture effect estimates. Evidence of migrant selection is scant in this literature and the discussion is often limited to specific types of selection (such as a binary mover/stayer migration decision).[2] Although we focus on son preference, our approach is general and applicable to a wide range of settings. That said, implementing the control-function strategy requires estimating migration probabilities based on observed flows from each origin to all potential destinations, which may not be feasible in small-scale surveys or datasets that follow migrants to only a single destination.

More broadly, this paper contributes to the understanding of son preference in fertility decisions. Our finding of a small culture effect indicates that migration and the associated exposure to the new environment are important in explaining fertility behavior. This coincides with a line of work showing that policy and economic environment can have large and lasting impacts on son preference, including

[2] For instance, Fernández (2011) pointed out that second-generation migrants "may not be randomly selected in the sense that their immigrant parents may be a selected sample from the distribution of beliefs in the country of ancestry." This discussion recognizes a mover/stayer decision, but abstracts from the endogenous choice of migration destinations. Luttmer and Singhal (2011) discussed how migrant selection based purely on preferences (Tiebout selection) would bias their estimated culture effects downward. We provide a formal test of Tiebout selection using our model (Section 6.1), finding little support for migration based on son preference.

property rights and land ownership (Bhalotra et al., 2019; Almond et al., 2019), relative income by gender (Rosenzweig and Schultz, 1982; Qian, 2008), marital institutions (Sun and Zhao, 2016; Alfano, 2017), and old age support (Guo et al., 2025).

Several papers have studied son-favoring fertility behavior among migrants. Dubuc and Coleman (2007) and Almond et al. (2013) show that migrants in the UK and Canada originating from Asia exhibit similar son-favoring fertility behavior to that of their home countries; Kim and Lee (2020) find that son preference formed at an early age (measured by gender gap in education achievement at parent's place of birth) is associated with son-favoring fertility behavior in adulthood in South Korea.[3] Similar to these papers, our estimates also point to a cultural component in son preference that does not go away when contexts change, although the effects are small after correcting for migration selection. The institutions in China also offer several advantages relative to other countries: conceptually, sex ratios at birth over this period are shown to be a credible proxy for son preference given the OCP and the availability of sex-selective technologies (Li et al., 2011; Chen et al., 2013); statistically, there are very large variations in sex ratios over time and across areas which provide useful sources of variation for identification; logistically, there is no linguistic barrier to migration and the urban labor market does not differentiate migrants based on their place of origin. In general, unlike international migration, domestic rural-to-urban migrants are subject to the same institutions in a destination.

The paper is organized as follows. Section 2 introduces the potential outcomes framework and clarifies identifying assumptions underlying the epidemiological approach. Section 3 provides institutional details relating to the OCP and rural-to-urban migration. Section 4 presents data and descriptive evidence for migrant selection. Empirical models and the estimates are presented in Section 5. Section 6 uses the estimated model to test Tiebout migration, quantify culture and selection effects, and understand the determinants of spatial heterogeneity in cultural persistence. Section 7 concludes.

[3] Arguably, observed gender gaps in educational achievement at parents' place of birth may reflect differences in local economic opportunities rather than son preference alone. Kim and Lee (2020) also test whether the effects of son preference may be overestimated when migration is purely based on preference (Tiebout migration). Our migration decisions are based on a generalized Roy model, which can accommodate but is not limited to Tiebout migration.

2 Potential Outcomes and Selection Bias

This section introduces the potential outcomes framework, laying the groundwork for our empirical analysis. We also show the assumptions under which comparing migrants from different origins to the same destination can be given a causal interpretation.

Consider an individual i who grew up in origin j. Let $y_{ij}(k)$ denote the individual's *potential* fertility outcome, which captures their potential demand for sons if they are externally assigned to work at a destination $k \in \{1, \ldots, n\}$. If the individual i chooses to migrate to destination k, then their observed outcome is given by $y_{ijk} = y_{ij}(k)$. For each destination k, we assume that their potential fertility is determined by:

$$y_{ij}(k) = \alpha_k + \beta_k s_j + \boldsymbol{x_i'}\boldsymbol{\gamma_k} + \epsilon_{ijk}, \tag{2}$$

where α_k is a destination-specific intercept, s_j is a proxy for son-favoring culture at the origin j, $\boldsymbol{x}_i$ is a vector of individual-specific covariates including gender, ethnic minority status, income, education, age and its squared term, and cohort dummies, and the error term ϵ_{ijk} captures *any* idiosyncratic and unobserved residual factors that affect the potential outcome (including, but not limited to, idiosyncratic preferences for sons above and beyond s_j).

The parameter α_k captures economic and institutional factors at the destination that affect the individual's demand for sons. These "contextual factors" may include the provision of local health services (e.g., access to sex-selection technologies) and local culture. The coefficients $\boldsymbol{\gamma_k}$ measure how these contextual factors vary by individual characteristics.

The parameter β_k captures the transmission of culture. We measure origin son preference using the sex ratio among new births in the area where the individual grew up and at the time when the individual was aged 16. Therefore, s_j is predetermined and is not affected by the fertility decisions of the same individual. Section 3 provides a background discussion to justify the use of sex ratios as a measure of culture over the period of our analysis.[4]

If $\beta_k = 0$, then the sex ratio at the origin does not have an impact on the potential demand

[4] Although s_j varies by cohort within an origin, we omit the cohort subscript in our notation for brevity. In our empirical specification (Section 5.1), we use a standardized sex ratio at age 16 as s_j that captures the relative position of sex ratios.

for sons; only contextual factors matter. On the other hand, a positive β_k means that the origin sex ratio continues to affect potential demand after the individual has been externally assigned to k. This indicates a persistent cultural component in the demand for sons. Note that β_k is allowed to vary by destination, which means that the rate of cultural transmission could be context-specific. For example, consider two identical individuals who are from the same origin j but assigned to two different destinations. Their potential demand for sons may differ due to their exposure to different contextual factors, as well as differential rates of transmission of the origin culture.

We assume that ϵ_{ijk} are i.i.d. across i and j, but, conditional on i (and j), they are allowed to correlate across k. This could capture, for instance, that an individual with a strong idiosyncratic preference for sons has a high demand for sons in all destinations. Although ϵ_{ijk} are assumed mean-independent of s_j and $\boldsymbol{x_i}$, in the self-selected sample of migrants in destination k, ϵ_{ijk} does not necessarily have zero means conditional on j and $\boldsymbol{x_i}$ (see Section 2.1). This is the key source of endogeneity in the model.

Equation (2) naturally introduces an exclusion restriction: conditional on $\boldsymbol{x}_i$, the only impact the origin has on the potential outcome is through the individual's exposure to the origin sex ratio, s_j. This assumption, although common in the literature, necessarily reduces otherwise complex origin influences to a scalar quantity.[5] In our empirical analysis, we show that our results are robust to the inclusion of other origin-level characteristics that might directly affect fertility, such as the strictness of the OCP at the origin.

2.1 Selection Bias from Comparing Migrants of Different Birth Places in the Same Destination

Potential outcomes are not observed in every destination k; we observe y_{ijk} if and only if the individual from origin j has chosen to migrate to destination k. Let d_{ijk} denote an indicator function of whether individual i has moved from j to k. By iterated expectations, equation (2) implies the mean realized

[5]For example, Fernández (2007) assumes that female immigrants' countries of origin influence their labor supply decisions solely through historical female labor force participation rates. Likewise, Alesina and Giuliano (2011) assume that immigrants' origin countries affect current political participation only through the strength of family ties. In Luttmer and Singhal (2011), the identification assumption is that there are no omitted factors correlated with average preferences for redistribution in the birth country that also influence immigrants' redistribution preferences in the host country. Although they focus on second-generation immigrants who are not directly exposed to institutional or economic conditions in their countries of origin, these assumptions may still be violated if non-cultural factors from the origin country have intergenerational effects.

outcome conditional on $d_{ijk} = 1$ and the remaining covariates can be written:

$$E(y_{ijk}|d_{ijk} = 1, \boldsymbol{x_i}) = \alpha_k + \beta_k s_j + \boldsymbol{x_i'\gamma_k} + E(\epsilon_{ijk}|d_{ijk} = 1, \boldsymbol{x_i}), \text{ for } k = 1, \ldots, n. \tag{3}$$

To identify the culture effect β_k, the literature compares the average outcomes between individuals who have migrated to the same destination but come from different places of origin ($j = \{1, 2\}$, for example). In the current framework, this amounts to taking the following difference:

$$\begin{aligned} &E(y_{i2k}|d_{i2k} = 1, \boldsymbol{x_i}) - E(y_{i1k}|d_{i1k} = 1, \boldsymbol{x_i}) \\ &= \underbrace{\beta_k(s_2 - s_1)}_{\text{culture effect}} + \underbrace{E(\epsilon_{i2k}|d_{i2k} = 1, \boldsymbol{x_i}) - E(\epsilon_{i1k}|d_{i1k} = 1, \boldsymbol{x_i})}_{\text{selection effect}}. \end{aligned} \tag{4}$$

Equation (4) makes it clear that the difference in the conditional means of the observed outcomes consists of the true culture effect and a selection effect. To consistently estimate the causal effect of culture, this approach requires that there is no selection effect:

$$E(\epsilon_{i2k}|d_{i2k} = 1, \boldsymbol{x_i}) = E(\epsilon_{i1k}|d_{i1k} = 1, \boldsymbol{x_i}) \tag{5}$$

This equality holds if selection into migrating to destination k is identical between the two places of origin, meaning that any selection bias from endogenous migration will be eliminated by the within-destination and cross-origin comparison. In other words, since two people choose to move to the same destination, their selection biases can be characterized by the same distribution, regardless of their place of origin.[6]

To further assess this assumption, consider the generalized Roy model of migration in Dahl (2002). For individual i from origin j, the decision to move to k is determined by:

$$d_{ijk} = \begin{cases} 1 \text{ if and only if } V_{jk} + e_{ijk} \geq V_{jm} + e_{ijm}, \forall m \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

[6] To make this assumption less stringent, some papers in the literature use second generation migrants, where migration decision was made by the previous generation (e.g., Fernández and Fogli (2009)). A sizable literature has documented large inter-generational correlations in skills and other traits (Black and Devereux, 2011; Heckman and Mosso, 2014). As long as son preferences are correlated between generations, the same selection problem applies.

where V_{jk} denotes conditional mean utility of moving from origin j to destination k (the subutility function), and e_{ijk} captures individual-specific deviations from the conditional mean. V_{jk} is predicted by $\boldsymbol{x_i}$ and additional covariates (see Section 5.1), and is a function of both j and k. The individual chooses to migrate from j to k if the utility of working in k exceeds the utilities of working in any other places. This implies that the conditional mean of idiosyncratic factors in the realized outcomes can be written:

$$E(\epsilon_{ijk}|d_{ijk}=1,\boldsymbol{x_i}) = E(\epsilon_{ijk}|e_{ijm}-e_{ijk} < V_{jk}-V_{jm},\forall m) = \Phi_{jk}(V_{j1}-V_{jk},\cdots,V_{jn}-V_{jk}) \tag{7}$$

The selection function Φ_{jk} is a generalization of the standard inverse Mills ratio correction term used in the two-step Heckman (1979) selection correction. In the current context of multi-market selection, Φ_{jk} is an unknown function of multiple indices $V_{j1}-V_{jk},\cdots,V_{jn}-V_{jk}$. If migration decisions are independent of potential outcomes, i.e., $e_{ijm} \perp \epsilon_{ijk}, \forall m$, then $\Phi_{jk}=0$ and there is no selection bias. In all other cases, $\Phi_{jk}\neq 0$, implying that the mean imputations of the unobservables drive endogeneity in observed outcomes conditional on the chosen destination. This could happen if migration decisions are driven by the idiosyncratic son preference which determines potential fertility at the same time (Tiebout migration). More broadly, any idiosyncratic factors that affect potential outcomes and migration decisions at the same time (such as human capital) will drive endogeneity in the model.

Importantly, within-destination and cross-origin comparison does not necessarily eliminate selection bias under either of the following two scenarios: (1) when the joint distribution of ϵ_{ijk} and $(e_{ij1}-e_{ijk},\cdots,e_{ijn}-e_{ijk})$ depends on j, and (2) when the multiple indices that determine Φ_{jk} differ by origin j. Either case implies that $\Phi_{jk}-\Phi_{j'k}\neq 0$.

2.2 The Direction of Selection Bias

The analysis above establishes that the within-destination, cross-origin estimator is biased whenever $\Phi_{jk}(\cdot)-\Phi_{j'k}(\cdot)\neq 0$. We now characterize the *direction* of that bias. Consider two origins j and j' with $s_j > s_{j'}$, both sending migrants to destination k. For simplicity, write the raw cross-origin difference in mean outcomes as:

$$\bar{y}_{jk} - \bar{y}_{j'k} = \underbrace{\beta_k(s_j - s_{j'})}_{\text{culture effect}} + \underbrace{\left[\Phi_{jk}(\cdot) - \Phi_{j'k}(\cdot)\right]}_{\text{selection bias}}. \tag{8}$$

Conditional on the same destination, the sign of the selection bias depends on how the idiosyncratic component of son preference correlates with the propensity to migrate from each origin, and in particular, on whether that correlation differs across origins. If the son-favoring cultural trait and the migration incentive share common economic and institutional determinants, e.g., origin j has a higher agricultural labor supply surplus than j', migrants from high-s origin j over-represent individuals with intense son preference. In essence, destination k selects more migrants from son-favoring individuals. Formally, $\Phi_{jk}(\cdot) > \Phi_{j'k}(\cdot)$, the bias is positive, and the raw estimate *overstates* the true culture effect. Selection could also go in the opposite direction. For instance, suppose the labor market at destination k offers high returns to skill, then it could attract more skilled migrants from j' proportionally than from j. To the extent that unobserved skill is negatively correlated with the idiosyncratic preference for sons, migrants from the high-s origin under-represent individuals with intense son preference relative to those from j'. In essence, destination k draws its least son-favoring migrants disproportionately from the high-s origin which makes $\Phi_{jk}(\cdot) < \Phi_{j'k}(\cdot)$. In this case, the bias is negative, and the raw estimate *understates* the true culture effect.

The discussion above concerns an origin-composition (Roy) channel of selection. The selection could also operate through destination-composition (Tiebout) channel of selection. Unlike the origin-composition channel which produces selection that covaries with the origin trait s_j, this destination-composition channel produces selection that covaries with the destination trait s_k. Conditional on the same origin j, individuals with a strong idiosyncratic preference for sons (a high realization of ϵ_{ijk}) can be drawn to destination k where prevailing norms are more son-favoring than those in k', so that this idiosyncratic component is positively correlated with the propensity to choose a high-son-favoring destination k. In previous literature, this has been discussed as a downward attenuation bias of the culture effects (Luttmer and Singhal, 2011).

The sign of the selection bias is therefore an empirical question, determined by relative strengths of various factors that predict different directions of selection.

3 Background: Son Preference, Sex Ratio, and the OCP

3.1 Son Preference and the OCP

China has a long history of son preference. Historically, sons have been valued for continuing the family lineage, performing ancestral rites, and providing economic security to parents in old age. Daughters, by contrast, were often regarded as "temporary members" of their natal households, expected to marry out and contribute to their husband's family (Das Gupta, 2005; Ebenstein, 2010).

Although these cultural norms are shared nationwide, the intensity and expression of son preference differ between rural and urban areas. In rural regions, son preference has historically been stronger due to the economic value of male labor in agriculture, patrilocal living arrangements, and the limited availability of formal old-age support systems. Parents often relied on sons for both farm work and financial assistance in later life. In contrast, urban families with higher levels of female education and labor force participation, more limited housing space, and access to state-provided social benefits tended to show a weaker preference for sons. This urban-rural divide was reinforced by the hukou (household registration) system, which classified individuals as rural or urban residents and restricted rural citizens from freely moving to cities prior to the 1980s.

The introduction of the OCP in 1979 interacted with these preexisting norms to reshape reproductive behavior. The OCP generally restricted couples to a single child, with limited exceptions. Enforcement was strictest in urban areas, where state control over employment, housing, and welfare benefits enabled effective monitoring and sanctions (Ebenstein, 2010; Zhang, 2017).[7] In rural areas, where agricultural labor needs remained high, implementation was more flexible. Beginning in 1984, many provinces adopted a "1.5-child policy", allowing rural couples to have a second child if the first was a girl (Zhang, 2017).[8] This accommodation implicitly acknowledged the existence of son preference

[7]Urban residents, particularly those working in state-owned enterprises, faced significant penalties for exceeding birth quotas, including loss of employment and denial of access to housing and education benefits.

[8]China's population policies have gone through major reforms recently. Following the revision of the "Population and Family Planning Law of the People's Republic of China" in December 2015, the nationwide implementation of the universal two-child policy took effect on January 1, 2016, marking the official retirement of the differentiated fertility policy of 1.5-child rule in rural areas. Since then, the law has undergone its second amendment based on the "Decision on Amending the 'Population and Family Planning Law of the People's Republic of China'" made at the 30th Session of the Standing Committee of the 13th National People's Congress on August 20, 2021. Article 18 stipulates that "The state advocates for marriage and childbirth at an appropriate age and for healthy pregnancy and childbirth. A couple is allowed to have three children." Meanwhile, Article 27 states that "The state adopts supportive measures in finance, taxation, insurance, education, housing, employment, and other aspects to alleviate the burden of family fertility, child-rearing, and education." Source: https://www.nhc.gov.cn/fzs/c100048/202303/fbbf58febae1445f820648195eab70be.shtml.

and the economic rationale for larger families in rural areas. Note that this “1.5-child policy” applied to all families with a rural hukou, including rural-to-urban migrants.[9]

The emergence of ultrasound technology in the early 1980s transformed these preferences into observable demographic outcomes. The widespread adoption of ultrasound technology, which covered over 60% of counties by 1985, enabled prenatal sex determination and, in turn, sex-selective abortion (Chen et al., 2013). Prior to its diffusion, China’s sex ratio at birth remained close to the natural biological range of 103–107 males per 100 females, despite son preference being a longstanding cultural norm (Coale and Banister, 1994). Following the introduction of ultrasound and the fertility restrictions imposed by the OCP, the sex ratio began to rise sharply, reaching 111 males per 100 females by 1990 and exceeding 120 in some provinces by 2000 (Hesketh and Xing, 2006).[10]

In summary, sex ratio is a credible proxy for, but not a perfect measure of son preference. Rural–urban differences in economic institutions, reinforced by limited mobility, contributed to divergence in sex ratios. But the interaction of binding fertility restrictions and access to sex-selective technology means that families with stronger son preference were more likely to act on them and translate those preferences into fertility behavior, which are aggregated into regional sex ratios. In this context, the sex ratio serves as a meaningful proxy for the intensity of son preference, even if it also captures institutional variation.

These institutional features also speak to the direction of selection bias in equation (8). Zhang (2017) finds that stricter OCP enforcement, while changing rural sex ratios, also raises male labor supply and rural out-migration rates. These suggest that the same economic and institutional features that generate son preference also create a household demand for sons to migrate as economic providers, the cultural trait and the migration incentive share common determinants. While it is not the only explanation for self-selection, positive selection on son preference is a plausible empirical consideration in this setting.

[9]Despite economic reforms, the hukou system continued to impose barriers for rural migrants to access social services such as education, healthcare, and housing in urban areas. Policy reforms in the 2010s started to address these challenges by easing hukou restrictions in small and medium-sized cities while maintaining tighter controls in megacities like Beijing and Shanghai.

[10]The mechanisms of selection also varied by region. In urban areas, where couples were generally limited to a single child, sex selection occurred primarily at the first birth, as families sought to ensure their only child was male. In rural areas, where the first child could be female without penalty, selection often occurred at the second birth, following the “1.5-child” allowance (Almond et al., 2019).

3.2 Construction of Sex Ratios

We combine a series of Chinese census data to construct sex ratios for each year at the origin province level, defined as the *share of males among newborns* in that year. As the census takes place every ten years, data are drawn from children born in 1976–1990 from the 1990 census 1% microdata, children born in 1991–2000 from the 2000 census 1% microdata, and aggregate birth count data for children born in 2001–2010 from the 2010 census as microdata for this census are not publicly available.[11] For each province, we separately define rural and urban sex ratios based on the child's agricultural or non-agricultural hukou status.[12]

3.3 The Geography of Son Preference

Rural vs. urban sex ratios. Figure 1a presents newborn sex ratios in rural (solid line) and urban (dashed line) areas. Up to around 1985, both lines hover close to the natural biological range of about 0.51–0.52 (i.e. 51–52 males per 100 total births), and the gap between them is narrow. From the late 1980s onward, both lines begin rising, but the rural line climbs considerably faster and higher than the urban one. By the late 1990s the rural sex ratio peaks at around 0.55–0.56, while the urban line reaches a more modest 0.53–0.54. In the early 2000s, the rural line shows some decline and convergence toward the urban line, though both remain above the natural baseline through 2010.

The figure reveals a pronounced rural–urban divergence in sex ratios. While cultural differences in son preference, reinforced by longstanding rural–urban segregation, likely contribute to this pattern, heterogeneity in the enforcement of the OCP and differential access to prenatal sex-detection technology also play an important role. Ultrasound technology started to diffuse across the country since the 1980s, which significantly reduced the cost of prenatal sex selection. Urban areas faced stricter fertility enforcement and gained earlier access to ultrasound technology, both of which are associated with more male-biased sex ratios. If that is the case, then the rural–urban gap shown in Figure 1a likely understates the true difference in son-preference norms between rural and urban areas, and should

[11] The 1990 and 2000 census microdata are available to approved applicants at https://international.ipums.org/international-action/sample_details/country/cn. The 2010 aggregate count census data are publicly available at https://www.stats.gov.cn/sj/pcsj/rkpc/6rp/indexch.htm. For the latter, we have data on the aggregate count of children aged 0, 1–4, and 5–9, but not by specific birth year. Therefore, for newborn sex ratios between 2001 and 2009, we substitute year-specific sex ratios with age-band sex ratios.

[12] For the 2010 aggregate count census data, individual information on hukou status is not available and we calculate these based on rural and urban status of the locality.

be treated as a conservative proxy for the underlying divergence in son-preference environments. In addition, Section 5.3 shows that our findings are robust to controlling for measures of OCP enforcement.

Focusing on rural areas, Figure 1b shows that the biased sex ratios for rural residents are mainly driven by the second birth. First-birth sex ratios (circles) stay essentially flat and close to 0.51 throughout the entire period, suggesting that sex-selective behavior at the first birth was rare, likely because many rural families wait to see the sex of the first child before engaging in sex-selective behavior under the 1.5-child rule (Das Gupta, 2005; Almond et al., 2019). In contrast, second births (diamonds) show increasingly higher sex ratios, from 0.52 in the late 1970s to over 0.60 in 2000, which point to households' desire to secure a son within the OCP's fertility constraints. The pattern is even stronger when the first child is female (triangles): sex ratios of the secondborn rise steeply from around 0.51 in the late 1970s to over 0.65 by 2000. In contrast, if the firstborn is a boy (squares), sex ratios of the secondborn remain largely flat around or occasionally below 0.50. Taken together, these patterns suggest a strong desire for having a son within or sometimes outside policy constraints, and parents who had a daughter first were highly motivated to ensure the second child was a son.

Geographic distribution of sex ratios in rural areas. Table A.4 shows the province-level rural sex ratios in selected years between 1986 and 2010, and Figure A.2 plots the distributions graphically. The proportion of rural male newborns is unevenly distributed across China's four regions, and the imbalance sharpens over time. East China and Central and South China consistently stand out as the high-ratio regions: provinces such as Anhui, Jiangxi, Henan, Hubei, Hunan, and Guangdong repeatedly post the largest proportions, frequently climbing above 0.55 from 1995 onward and reaching the dataset's peak of 0.59 in Anhui in 2000. North China occupies an intermediate position, with its provinces clustering near 0.52–0.53 and showing comparatively modest movement over the period. West China is generally the lowest region, anchored by provinces like Tibet, Xinjiang, and Qinghai that persistently sit at or below 0.52. Taken together, the regional pattern traces a broad gradient that rises from the sparsely populated western frontier toward the densely populated agricultural provinces of eastern and central-southern China, suggesting that son-preference pressures on the sex ratio at birth were most acute in the latter regions and weakest in the west.

Figure A.2 shows clearly the pronounced time trend in the spatial distribution of sex ratios between 1986 and 2010. Starting from a relatively compressed distribution with few outliers in 1986, the geographic spread of rural sex ratios widened progressively through the 1990s, and maintained high dispersion around 2000. After 2000, the distribution began to stabilize and partially compress, as the upper tail moderated and the lower tail crept upward, narrowing the cross-sectional spread to around 0.05 by 2010.

The geographic differences in son preference can be attributed to various factors, such as the labor intensity of agricultural activities (Ebenstein, 2021) or the influence of clan culture and patrilineal family structures (Das Gupta, 2010; Murphy et al., 2011), but exploring the origins of these differences is beyond the scope of this study. This substantial cultural diversity within the same country provides an opportunity to exploit internal migration flows to investigate the effects of culture.[13]

4 Migrants Data and Descriptive Evidence

4.1 The China Migrants Dynamic Survey

Beyond the census data which generates our sex ratio measures as described in Section 3.2, this study uses annual data from the China Migrants Dynamic Survey (CMDS) 2013–2017 as the primary data source.[14] The CMDS is a nationwide repeated cross-sectional survey of internal migrants in China conducted by the National Health Commission, covering 31 provinces (including autonomous regions and municipalities) across the country. The survey defines migrants as individuals who have resided in a location for more than one month but whose hukou (household registration) is registered in a different county. The CMDS adopts a stratified, multi-stage sampling design using Probability Proportional to Size (PPS) methods, ensuring representativeness of internal migrants aged 15–59 (Wang et al., 2021). The sampling unit is the individual. If two sampled individuals are from the same household, only one is interviewed. The questionnaire collects detailed information on the respondent's province of origin, family composition, migration history and fertility behavior. Approximately 200,000 migrants are surveyed annually, making the CMDS the largest and most comprehensive dataset on internal migration in China. Over the five-year period, the survey covered a total of 931,847 respondents.

[13] One potential concern is that the census records children's current place of residence rather than their province of birth. For example, a child born in province a in 1987 may have migrated to province b by the time of the 1990 census and therefore would not be included in the calculation of the newborn sex ratio for province a in 1987. However, Appendix Figure A.1 shows that fewer than 2% of children aged 0–9 have experienced cross-province migration, and these movements are not gender-specific. This suggests that cross-province migration is unlikely to bias our sex ratio estimates.

[14] Data were accessed upon approval at http://www.chinaldrk.org.cn/wjw/#/data/classify/population.

A key advantage of our data is that sampling is conducted at migrants' destinations, rather than at their hukou place of registration. In standard household surveys in China (such as the China Health and Nutrition Survey), respondents are sampled based on their hukou location, meaning that migrants who have moved to cities are either missed entirely or are represented only through responses provided by family members who remain in rural areas. Conversely, typical urban household surveys include only individuals with urban hukou, thereby excluding the large population of rural-to-urban migrants residing in cities. By sampling directly where migrants actually live and work, our data capture this otherwise overlooked group, enabling a more accurate characterization of migrant experiences and outcomes.

Several features of the CMDS make it well suited to our research. First, it tracks migrants moving from multiple origins to multiple destinations. For any given origin, we observe all destination choices, in contrast to most studies in the cultural-effects literature, which typically examine migrants from different origins moving to a single destination country. Observing the full migration flows is essential to correct for selection arising from endogenous migration, and to investigate how culture effects vary with destination characteristics. Second, the large sample size provides sufficient observations for each origin-destination flow. Third, the CMDS records the gender and age of all children of each migrant, enabling us to recover complete gender compositions, which is an advantage over census microdata, which only report detailed information on co-resident children.

Exposure to Origin Son-Preference Culture. Migrants' exposure to origin son-preference culture is defined as the sex ratios (share of males among newborns) of their origin province at age 16. We anchor exposure at age 16 for a number of reasons. First, by mid-adolescence, individuals have accumulated substantial exposure to the son-preference culture that the sex ratio proxies, while remaining young enough that the measure precedes the outcomes of interest. Second, age 16 corresponds to a developmental window in which attitudes and preferences stabilize, making it the appropriate point of measurement for later life outcomes. And lastly, in our setting, rural-to-urban migration starts rising substantially after age 16 (see Figure A.1). As a robustness check, we consider the average sex ratio during ages 12–16 as an alternative proxy for exposure to son-preference norms.

To account for common time trends and disentangle underlying son preference from cohort effects, we standardize sex ratios within each year and use the standardized sex ratio at age 16 as our measure of exposure to son preference. This procedure purges the common annual trend and retains only the cross-sectional variation in how skewed a given province's sex ratio is relative to its contemporaries. The resulting measure can be interpreted as a relative-intensity index of son preference, comparable across cohorts.

As shown in Appendix Figure A.3(a), older cohorts of rural-to-urban migrants in the CMDS sample were exposed to lower non-standardized sex ratios at age 16 than younger cohorts. In contrast, Appendix Figure A.3(b) shows that the distributions of standardized sex ratios at age 16 are highly comparable across cohorts. In addition, we include cohort fixed effects in the analysis to further account for potential cohort trends.

Sample selection. We focus on rural-to-urban migrants, defined as individuals residing in urban areas while holding an agricultural hukou (56.49% of the full sample). We start with 307,831 migrants born between 1970 and 1994 who have at least one child, for whom exposure to son-preference culture can be proxied by newborn sex ratios in their regions of origin between 1986 and 2010 (measured at age 16). For these cohorts, sex-selective technologies were already available in most counties, making sex ratios an informative measure of son preference.

As shown in Figure A.4, fertility timing is closely aligned with this life-cycle pattern: first births occur on average around age 25, while second births occur around age 29. In addition, the distribution of age at second birth conditional on the first child being a girl is largely the same as the unconditional age distribution, suggesting that the gender composition of earlier births is not strongly correlated with the timing of the following births.

Table A.3 reports mean individual characteristics by the gender of the firstborn child. Individuals with a firstborn daughter (N = 143,087) are very similar to those with a firstborn son (N = 164,744) in terms of education, gender, age, ethnicity, and income, indicating minimal selection on observables with respect to first-birth gender. However, households with a firstborn daughter are notably more likely to have a second child (50.4% vs 32.0%) and, conditional on doing so, are more likely to have a second son (65.7% vs 48.8%). These large differences are consistent with the presence of gender selection at the second birth.

We therefore restrict the sample to households whose first child is a daughter. The final sample contains 143,087 individuals, of whom 72,101 have a second child. Table 1 reports summary statistics for these two samples.

4.2 Descriptive Evidence of Migrant Selection

Rural-to-urban migration in China typically begins at a young age. Evidence from another nationally representative migrant survey in 2009 indicates that most rural migrants first move to urban areas in their late teens (Meng, 2012).[15] As a result, migrants are exposed to the urban environment relatively early in life. Meng notes that 55% of the male rural labor force have migrated by age 25, and 50% of their female counterpart by age 20; rural migrants stay in cities for about seven years.[16]

As discussed in Section 2, migration is an endogenous choice, and migrants from different origins likely sort into different destinations. The remainder of this section presents descriptive evidence consistent with Roy-type migrant sorting: migrants from a given origin differ in their educational composition across destinations, and origins of differing son-preference sort into destinations of differing economic conditions.

Figure 2 summarizes rural-to-urban migration flows across five eastern provinces by education level of the migrant, with rows denoting origins and columns denoting destinations.[17] Educational attainment is grouped into primary school or below (PS), middle school (MS), and high school or above (HS). The height of each bar represents the probability that migrants in a given origin–education group choose a particular destination, based on our CMDS sample.

Three main patterns emerge. First, within-province migration is generally more common than cross-province migration, with exceptions such as Anhui.[18] Second, migration choices vary systematically by education: more educated migrants disproportionately locate in major cities such as Shanghai, while Zhejiang attracts a relatively larger share of lower-educated migrants from nearby provinces (e.g., Jiangsu, Anhui, Fujian), despite being economically developed. Third, migration patterns differ markedly across origins even within a geographically compact region, implying that geography alone does not account for destination choices. For instance, Fujian migrants with high school or higher education are more likely to move to Shanghai than to the geographically closer Zhejiang. Overall, the education profiles of migrants from a given origin vary across destinations, providing evidence of

[15] This finding is based on data from the 2009 Rural Urban Migration in China and Indonesia (RUMiCI) project.

[16] Consistent with this pattern, the average duration of the current migration spell in our sample is 5.4 years.

[17] Shanghai is excluded as an origin because it contributes a negligible number of rural out-migrants.

[18] In the overall sample, slightly more than half of rural migrants move within the province (53.78%).

endogenous selection into migration destinations.

Figure 3 plots the relationship between migrants' average years of education by origin province and the proportion of male newborns in that province in 1990, for the four most popular destinations. Circle size reflects the number of migrants from each origin. In three of the four destinations, education is positively correlated with the origin sex ratio (weighted by migrant population). This suggests that, if education is positively associated with the outcome of interest, failing to control for it would lead the estimated effect of the origin sex ratio to overstate the true culture effect.

Figure A.5 provides further descriptive evidence on migrant sorting. The Sankey diagram groups origin provinces into quartiles by average rural sex ratio on the left and destination provinces into quartiles by average GDP per capita on the right, then traces the migration flows between them. The flows are far from evenly spread: a handful of bands are much thicker than the rest, indicating that a small number of origin–destination pairings absorb a disproportionate share of all migrants, while many other pairings carry only thin trickles. The richest destination quartile in particular draws very heavy flows, mainly from origins with above-median sex ratios. If migration destination is random, each origin quartile would distribute across destination quartiles in proportion to destination size. The flow patterns depict that migrants reaching a given destination quartile originate from systematically different points of the sex-ratio distribution. This non-random matching of origins to destinations is consistent with the origin-composition channel in Section 2.2, whereby selection covaries with the origin son-preference trait.

5 Empirical Model and Results

5.1 Correcting for Selection Bias

Our approach for correcting for selection bias is to control for the mean imputations of the unobservables (Φ_{jk} in equation (7)) that drive endogeneity in the model. When migrants can choose from n alternative destinations, the selection function Φ_{jk} depends on the correlation of the error terms from n-selection equations with the error term in the outcome equation, a high-dimensional construct. This makes it difficult to apply a conventional control function approach that derives the analytical expression of the selection function given distributional assumptions of the error terms.

Following Dahl (2002), we reduce the dimensionality of the error terms using order statistics and a semiparametric approach. Let $f_{jk}(\epsilon_{ijk},\ e_{ij1}-e_{ijk},\ldots,\ e_{ijn}-e_{ijk})$ denote the joint density function of the error terms in the potential outcome equation and the selection equations. Conditional on differences in the subutility functions $V_{j1}-V_{jk},\ldots,V_{jn}-V_{jk}$, f_{jk} can be transformed into a bivariate distribution of ϵ_{ijk} and a maximum order statistic:

$$\begin{aligned}
&f_{jk}\big(\epsilon_{ijk},\ e_{ij1}-e_{ijk},\ldots,\ e_{ijn}-e_{ijk}|V_{j1}-V_{jk},\ldots,V_{jn}-V_{jk}\big)\\
&=g_{jk}\big(\epsilon_{ijk},\max_{m=1,\ldots,n}(V_{jm}-V_{jk}+e_{ijm}-e_{ijk})|V_{j1}-V_{jk},\ldots,V_{jn}-V_{jk}\big)\\
&=g_{jk}\big(\epsilon_{ijk},\max_{m=1,\ldots,n}(V_{jm}-V_{jk}+e_{ijm}-e_{ijk})|p_{ij1},\ldots,p_{ijn}\big). \qquad (9)
\end{aligned}$$

where p_{ijk} is the probability that individual i born in origin j moves to the destination k conditional on $V_{j1}-V_{jk},\ldots,V_{jn}-V_{jk}$. The first equality reduces the dimensionality of the error terms by transforming them into an event in which the maximum of the collection of random variables $V_{j1}-V_{jk}+e_{ij1}-e_{ijk},\ldots,V_{jn}-V_{jk}+e_{ijn}-e_{ijk}$ is less than or equal to 0 (Lee, 1983). The second equality states that migration probabilities $p_{ij1},\ldots,p_{ijn}$ include all the information contained in the vector $V_{j1}-V_{jk},\ldots,V_{jn}-V_{jk}$, which determines an individual's migration choice.[19]

Equation (9) implies that the observed outcome for migrant i living in k can be written as a multiple-index model:

$$y_{ijk}=\alpha_k+\beta_k s_j+\boldsymbol{x_i'\gamma_k}+\sum_{j=1}^{n}[d_{ijk}\times\lambda_{jk}(p_{ij1},\ldots,p_{ijn})]+\omega_{ijk},\text{ for } k=1,\ldots,n. \qquad (10)$$

The selection correction functions for destination k, $\{\lambda_{jk}\}_{j=1}^{n}$, are unknown functions of the full set of migration probabilities. By construction, the error term ω_{ijk} has a zero conditional mean:

$$E(\omega_{ijk}|\boldsymbol{x_i},s_j,p_{ij1},\ldots,p_{ijn},d_{ijk}=1)=0,\text{ for } k=1,\ldots,n.$$

However, it is impractical to estimate a high-dimensional correction function using semiparametric methods.[20] Similar to Dahl (2002), we make index sufficiency assumptions that the joint distribution

[19]This equality holds subject to standard assumptions of the implicit function theorem; see Dahl (2002).

[20]For instance, there are 31 possible destinations ($n=31$) in our setting. If we use a second-order polynomial expansion to approximate the correction terms, then there will be 527 terms in each correction function λ_{jk}.

of ϵ_{ijk} and error terms in the selection equations depends only on a subset of $p_{ij1}, \ldots, p_{ijn}$.

We start by assuming that the probability of the first-best choice is a sufficient statistic for the joint distribution g_{jk}:

$$E(\tilde{\omega}_{ijk}|\boldsymbol{x_i}, s_j, p_{ijk}, d_{ijk} = 1) = 0, \text{ for } k = 1, \ldots, n.$$

where p_{ijk} is the probability that individual i born in province j chooses province k (their actual residence) as first-best destination. This assumption implies that the probability of the utility-maximising choice exhausts all the information about how $V_{j1} - V_{jk}, \ldots, V_{jn} - V_{jk}$ influences the conditional density function g_{jk}. Under this assumption, the multiple index model (10) can be reduced to a *single index* model:

$$y_{ijk} = \alpha_k + \beta_k s_j + \boldsymbol{x_i'\gamma_k} + \sum_{j=1}^{n}[d_{ijk} \times \tilde{\lambda}_{jk}(p_{ijk})] + \tilde{\omega}_{ijk}, \text{ for } k = 1, \ldots, n, \tag{11}$$

We can also allow other migration probabilities to influence the joint distribution g_{jk}. Our preferred model is a *double index* model where both the probability of the first-best choice p_{ijk} and the probability of intra-province migration p_{ijj} affect the selection function:

$$y_{ijk} = \alpha_k + \beta_k s_j + \boldsymbol{x_i'\gamma_k} + \sum_{j=1}^{n}[d_{ijk} \times \hat{\lambda}_{jk}(p_{ijk},\ p_{ijj})] + \hat{\omega}_{ijk}, \text{ for } k = 1, \ldots, n, \tag{12}$$

We choose to include the probability of intra-province migration p_{ijj} because intra-province migration is quite common in the data, and its inclusion in the correction function leads to a significant change in the estimated origin-exposure effect and improves the fit of the outcome regressions.[21]

Estimating the Migration Probabilities. The migration probabilities are not observable in the data and must be estimated. To avoid making strong functional form assumptions, we follow Dahl (2002) and estimate these probabilities nonparametrically.[22] In particular, we assume that migrants from one origin with the same characteristics are identically affected by the costs and benefits of

[21]Notice that for intra-province rural-to-urban migrants, p_{ijk} and p_{ijj} are equivalent and only one probability enters into the correction functions.

[22]One alternative is to specify and estimate a multinomial logit model to predict the migration probabilities. This approach requires the "independence of irrelevant alternatives" assumption to hold, which can be a strong assumption.

migration and have the same probability of choosing a specific destination. We calculate the migration probability for individuals from origin j by

$$p_{ijk} = Pr\big(d_{ijk} = 1|\text{cell}, j\big). \tag{13}$$

A cell is defined from individual characteristics that include educational attainment (primary school or below, middle school, high school or above), gender, and cohort (whether being born before 1980). As a result, there are 12 mutually exclusive cells. Conditional on origin, the probability that an individual in a cell moves to a specific destination is simply equal to the proportion of individuals in that cell who choose the same migration path. Note that our sample restriction criteria have already guaranteed that individuals in the sample have the same fertility history of the first child and hukou type. Therefore, we believe that it is reasonable to assume that migrants within each cell from the same origin are similar enough and have the same probability of choosing a specific migration path.[23]

The cells create variations in the estimated correction functions. Importantly, considering that the nonparametric migration probabilities are fully saturated, the marginal effects of individual characteristics $\boldsymbol{x_i}$ on p_{ijk} vary flexibly with *both* j and k. In other words, migration decisions to k are driven by interactions of $\boldsymbol{x_i}$ and j. Although the same characteristics enter the potential outcome equations via the term $\boldsymbol{x_i'\gamma_k}$, their marginal impacts on the potential outcomes do not vary by j.[24] Therefore, the interactions of $\boldsymbol{x_i}$ with the origin j are effective exclusion restrictions in our model.

Practical Choices. We use a second-order polynomial expansion to approximate the correction functions. In the single index model,

$$\tilde{\lambda}_{jk}(p_{ijk}) = \tilde{\lambda}_{jk0} + \tilde{\lambda}_{jk1}p_{ijk} + \tilde{\lambda}_{jk2}p_{ijk}^2 \tag{14}$$

[23] Practically we could use additional covariates to define finer level cells. However, with too many cells the migration probabilities within each cell become imprecisely estimated given the sample size. We have experimented with changing the number of cells (e.g., reducing the number of education categories to 2), and the results do not change much.

[24] This restriction follows from the exclusion restriction made in Section 2: conditional on observable individual- and location-specific characteristics, the origin can only affect the potential outcome at a destination through the culture effect.

and in the double index model,

$$\hat{\lambda}_{jk}(p_{ijk},\ p_{ijj}) = \hat{\lambda}_{jk0} + \hat{\lambda}_{jk1}p_{ijk} + \hat{\lambda}_{jk2}p_{ijj} + \hat{\lambda}_{jk3}p_{ijk}^2 + \hat{\lambda}_{jk4}p_{ijj}^2 + \hat{\lambda}_{jk5}p_{ijk}p_{ijj} \tag{15}$$

The parameters in each correction function are specific to j and k. Therefore, for each destination k, there is a total of n different correction functions to estimate, each corresponding to an origin.

In practice, estimating n different functions is challenging, as the number of individuals choosing a certain migration path can be small. To increase precision, we collapse 31 origin provinces into larger groups, and assume that within each group the parameters in the correction function are common. In our baseline specification, we estimate two sets of correction function parameters for any given destination k: one for inter-province migrants (originating from provinces other than k) and one for intra-province migrants (moving from rural to urban areas within k).[25] As a robustness check, we allow the correction function to differ by geographic region of the origin province, resulting in five different correction functions for each destination k.[26]

The constant terms in the correction functions are not separately identified from the intercept α_k of the potential outcome equation. This can be easily seen by plugging the correction functions into the estimating equations.[27] Therefore, what we can identify is the shape of the correction function terms across covariates, i.e. how selection varies with the migration probabilities.

5.2 Estimation Results

Tables 2 and 3 report the selection-corrected double index model and uncorrected estimates of the parameter β_k, on the gender of the second birth (=1 if a boy) and whether having a second child, respectively. We report the estimated β_k for each destination province, as well as the average β_k by geographic region. We find that the uncorrected estimates (column 1) are mostly positive, suggesting

[25] This is similar to Dahl (2002) who estimated two correction functions (movers, stayers) out of 51 states in the US.

[26] We divide the 31 origin provinces into 4 geographic regions: North China (Beijing, Tianjin, Hebei, Shanxi, Inner Mongolia, Liaoning, Jilin, Heilongjiang); East China (Shanghai, Jiangsu, Zhejiang, Anhui, Fujian, Jiangxi, Shandong); Central and South (Henan, Hubei, Hunan, Guangdong, Guangxi, Hainan); West China (Chongqing, Sichuan, Guizhou, Yunnan, Tibet, Shaanxi, Gansu, Qinghai, Ningxia, Xinjiang). This division is based on the Greater Administrative Regions, which were top-level administrative divisions (above province) during 1949 and 1954. Intra-province migrants still have a different correction function.

[27] For instance, there are two constant terms in the correction functions in the baseline, $\hat{\lambda}_{k0}^M$ (the constant term in the correction function for inter-province migrants) and $\hat{\lambda}_{k0}^S$ (the constant term in the correction function for intra-province migrants). The intercept in equation (12) becomes a composite intercept equal to $\alpha_k + \hat{\lambda}_{k0}^M$. The difference $\hat{\lambda}_{k0}^S - \hat{\lambda}_{k0}^M$ is absorbed as the coefficient on an indicator for intra-province migration.

that growing up in places with a high sex ratio is positively correlated with son-favoring fertility outcomes. The magnitudes of the uncorrected estimates tend to be smaller when the outcome is the gender of the second birth and when destination provinces are located in the southern region.

The average of all province-specific culture effects shows that correcting for endogenous migration selection substantially reduces, although not eliminates, the estimated cultural transmission of son preference. For having a secondborn son conditional on having two children and a firstborn daughter, the average uncorrected estimate across all destination provinces falls from 0.024 to 0.007 under the double index model (Table 2, columns 1 and 2, under panel "all regions"), representing a reduction of 71%. For the probability of having a second child conditional on a firstborn daughter, the mean uncorrected estimate of 0.047 falls to 0.020 after correction (Table 3, columns 1 and 2, under panel "all regions"), a reduction of 57%.

The uncorrected estimates combine the true culture effect with selection effect. As a direct test for the presence of selection bias, we perform a Hausman-type test for the difference between the uncorrected estimates (column 1) and selection-corrected estimates from the double index model (column 2) for each destination. Column (3) reports the χ^2-statistics and p-values. The difference is significant at the 10 percent significance level for the majority of provinces where the uncorrected estimates are statistically significant at the same level. The difference is insignificant for only 4 (out of 26) provinces in terms of whether having a second child, and 7 (out of 17) provinces in terms of gender of the second child. Column (4) reports the results of a Wald test for the null hypothesis that the correction functions are jointly zero in the outcome equation. The correction functions are jointly significant at 10 percent level for all but 1 province when the outcome is whether having a second child, and for 17 out of 31 provinces when the outcome is the gender of the second child. All these results point to the importance of disentangling the selection effect from the uncorrected estimates, which have been regarded as the culture effect by the existing literature.

To graphically compare the selection-corrected vs. uncorrected estimates more clearly at the destination province level, Figure 4 plots the selection-corrected and uncorrected effects. The top panels (a) and (b) show the results for the gender of the second child as an outcome, and the bottom panels (c) and (d) report the results for the outcome of whether having a second child. Within each set, we report selection-corrected estimates from the single index model (equation (11)) in panels (a) and (c), and double index model (equation (12)) results in panels (b) and (d). For each panel, we conduct an equality of coefficients Wald test between corrected and uncorrected estimates. The 45-degree line indicates equality of coefficients between uncorrected and corrected estimates.

All panels show that, for most destinations, the selection-corrected estimates are smaller than the uncorrected ones. This suggests a positive selection effect, that the true culture effect is smaller and less persistent in affecting the fertility outcome of migrants. This is consistent with a pattern that migrants from high-son-preference origins are, on average, more son-preferring than their son preference culture alone would predict.

The results also show that the choice of selection modeling matters. Under the single index model of first-best destination choice, equality of coefficients is rejected for having a second child (panel c, F=5.61, p=0.00), but not for the gender of the second child (panel a, F=0.77, p=0.82). Under the double index model which further controls for intra-province migration probability, equality is rejected for both outcomes (panels b and d). Given that intra-province migration is a prominent feature in our setting, the results suggest that the single index model is insufficient to fully control for selection into migration, and the double index model performs better in terms of capturing the heterogeneity in unobservables associated with the qualitatively distinct choice between intra- and inter-province migration.

The correlation between the selection-corrected and uncorrected estimates is modest for the outcome of gender of the second child. In our preferred model (panel b), the correlation coefficient is 0.624 (p-val=0.000). By comparison, the correlation is quite small when the outcome is whether having a second child ($\rho = 0.211$, p-val=0.083). This small correlation indicates that correcting for selection significantly reduces the extent of cultural persistence among provinces where uncorrected estimates are large. In this case, there is less geographic variation in the culture effects after removing the selection effects.

5.3 Sensitivity Analysis

To gauge the sensitivity of our results to the index sufficiency assumption and functional form assumptions, Figures 5-6 contrast our preferred estimates from the double index model with alternative models. Focusing on gender of the second child as the outcome variable, panel (a) of Figure 5 shows that adding an additional probability term (probability of migrating to the most popular destination) to the selection correction function does not affect our estimates much; the two sets of estimates line up closely on the 45-degree line. This seems to suggest that the double index model is sufficient to capture most of selection effects. Employing a more flexible correction function, in which inter-provincial migrants originating from different regions have separate correction functions, produces estimates that are broadly similar to our preferred estimates (panel b). However, using this more flexible correction function leads to less precise estimates, and, if anything, even smaller origin effects in magnitude.

As discussed in Section 2, our specification of the potential outcome imposes that the origin sex ratio is the only origin-specific factor that can affect potential fertility. In panel (c), we relax this exclusion restriction by allowing a set of other origin-specific characteristics (facing the same individuals at age 16) to affect potential outcomes. They include population size, GDP per capita, rural income, fiscal revenue, population growth rate, and OCP fines in rural areas. For instance, one concern is that the strictness of the OCP may correlate with sex ratios (Li et al., 2011; Ebenstein, 2010) in origin and continue to affect future fertility outcomes due to the continuity of the OCP policy. Including these origin characteristics does not change our estimates materially.[28]

Finally, panel (d) shows that the estimated culture effects are similar if we control for destination prefectural fixed effects (the next administrative level under province) instead of provincial fixed effects in our model. Whilst comparing migrants working in the same prefecture improves the measurement of contextual factors relative to comparing those in the same province, doing so does not appear to affect the estimated culture effect.

Figure 7 reports selection-corrected estimates where the origin son preferences are defined using sex ratio exposed to an individual at different ages. The estimated distributions of son preference across all destinations are stable and robust to the age of exposure. Appendix Figure A.6 further confirms that using age 12–16 average sex ratio as the exposure to culture produces similar results relative to the baseline.

Further robustness checks. The CMDS data only cover migrants but not those choosing to stay in the rural origin. Our estimated migration probability p_{ijk} could be biased (beyond a simple scaling effect) if the probability of staying differs systematically by individual characteristics within an origin.

[28]However, to the extent that origin characteristics may affect culture preferences, there is a risk of over-controlling if a large number of origin characteristics are included in the potential outcome equation (Luttmer and Singhal, 2011).

In addition, the probability of staying could contain additional information about selection, in which case the index sufficiency assumption may not hold when the stayer probability is omitted from the selection function. To address these issues, we use the 2000 population census data to estimate the stayer probability, for a sample of individuals that are similar to the CMDS data in terms of observables. We then use the stayer probability from the census to adjust the estimated migration probability p_{ijk} from the CMDS data. Figure A.7 shows that our results are not affected if we use the adjusted migration probabilities or include the stayer probability in the selection function.

Placebo test. Section 3.3 and Figure 1b show that the national first-birth sex ratios remain close to the natural biological range, indicating little sex-selective behavior on average. As a placebo test, we estimate the single- and double index model using the gender of the first child (=1 if a boy) as the outcome, with the estimation sample now consisting of all migrants with children. Appendix Figure A.8 shows that the selection-corrected and uncorrected culture effects are much smaller and concentrated around zero. This is as expected, given that there is no sex selection on the first birth on average. The selection corrected estimates are also not significantly different from the uncorrected estimate (p-val = 0.81). Overall, the placebo estimates provide reassuring evidence that the empirical relationship between origin culture and fertility outcome is not picking up spurious effects.

6 Empirical Insights from the Estimated Model

6.1 Testing Tiebout Migration

Tiebout migration posits that individuals "vote with their feet" by choosing residential locations that best align with their preferences for local public goods, fiscal policies, or community characteristics (Tiebout, 1956). In this framework, households are viewed as mobile agents who sort across jurisdictions based on the bundle of local amenities and governance attributes that maximize their utility. This mechanism of preference-based sorting has been shown to generate spatial clustering of individuals with similar tastes or values, thereby reinforcing local social norms and cultural homogeneity (Epple et al., 1999; Nechyba, 1997).

In our setting, the selection correction functions, λ_{jk}, represent the mean imputation of the idiosyncratic demand for sons among households who choose to move from location j to k (up to a scale factor). Under Tiebout-style migration, such mobility decisions should reflect systematic preference sorting: conditional on origin, households with a stronger idiosyncratic preference for sons are more likely to relocate to destinations where the prevailing community norms exhibit higher son preference. Consequently, we expect λ_{jk} to be positively correlated with son preference in the destination, controlling for origin fixed effects and individual characteristics.

To test this hypothesis, we estimate the following regression

$$\hat{\lambda}_{jk} = \pi_j + \pi_s s_k + \boldsymbol{x_i'}\boldsymbol{\pi_x} + e_{ijk} \tag{16}$$

where the dependent variable is the (normalized) estimated selection term from our preferred double index model, s_k measures local son preference at the destination, π_j are a full set of origin fixed effects, and $\boldsymbol{x_i}$ is the same set of individual characteristics that define cells in estimating the migration probabilities.[29] We cluster the standard errors at the origin-by-destination level.

Under Tiebout migration, we should expect $\pi_s > 0$: conditional on origin, those who move to areas with high son preference should have differentially high demand for sons. Table 4 reports the results. Column (1) reports the correlation without controlling for π_j. Column (2) controls for the origin fixed effects, π_j, and thereby compares migrants from the same origin. The estimated π_s reveals, if anything, that those who move to urban areas with high son preference have differentially low demand for sons. Column (3) compares migrants of the same characteristics and from the same origin, by controlling for individual characteristics fully interacted with the origin FE. The estimated π_s are negative for the gender of the second child, and insignificantly different from zero for whether having the second child.

Overall, we reject the null hypothesis that the estimated π_s is positive. Our evidence is thus at odds with the prediction of Tiebout migration selection by son preference, suggesting that migration does not generate spatial clustering of individuals based on their son preferences.

[29]Because the constant term in the correction function is subsumed into the intercept and not separately identified from α_k (Section 5.1), we exclude the constant term in the correction function $\hat{\lambda}_{jk}$.

6.2 Quantifying the Culture Effects and Selection Effects

Consider the mean outcome gap at a given destination k

$$E(y_{ijk} \mid s_j = \text{high}) - E(y_{ijk} \mid s_j = \text{low}),$$

where s_j denotes the cultural attribute of migrants from origin j. In the existing culture literature, holding all other factors fixed, this difference is typically interpreted as the "culture effect."

Using our estimates, we decompose this mean gap into components arising from both cultural transmission and endogenous sorting. To illustrate, consider toggling s_j from the 25th percentile (s_j^{low}) to the 75th percentile (s_j^{high}). Such a shift affects predicted outcomes through three channels:

(i) (True) culture effect. Holding the distribution of individuals and covariates fixed, the causal contribution of culture to the mean outcome gap is captured by

$$\beta_k\big(s_j^{\text{high}} - s_j^{\text{low}}\big),$$

which traces out the change in y that results purely from culture and their transmissions.

(ii) Covariate-composition effect. A shift in s_j may also change the distribution of other observed characteristics x_i among migrants to k. The part of the outcome difference attributable to changes in observed covariates is given by

$$\gamma_k\Big(E[x_i \mid s_j^{\text{high}}] - E[x_i \mid s_j^{\text{low}}]\Big).$$

(iii) Selection effect. Finally, when culture changes, different unobserved types of individuals may choose to migrate to k. These composition changes are summarized by differences in the selection term λ, so that

$$\lambda_k(s_j^{\text{high}}) - \lambda_k(s_j^{\text{low}})$$

captures the extent to which migration or sorting responds to changes in culture and, in turn, affects average outcomes.

Table 5 reports the decomposition of the mean outcome gap in the gender of the second child (Panel A) and whether having a second child (Panel B) among migrants in five destination provinces: Beijing, Shanghai, Guangdong, Jiangsu and Shandong. Focusing on gender of the second child (Panel A), the first row presents the total gap in outcomes, which varies substantially across regions, ranging from a sizable positive gap in Beijing (0.143) to even negative gaps in Guangdong. The second row reports the total gap predicted by the model, which closely tracks the raw gap for most destinations, indicating a good overall fit of the model.

The total gap is then decomposed into three main components: the culture effect, the covariate-composition effect, and the selection effect. The culture effect captures the change in outcomes attributable purely to differences in son-preference, holding population composition fixed. This component is positive in most regions but generally smaller than the total gap. For example, in Beijing, only about half of the total gap (0.058 out of 0.113) is explained by the culture effect, while in Shanghai the estimated culture effect is close to zero.

The covariate-composition effect reflects changes in observable household characteristics induced by shifts in the son-preference environment. This component is further broken down by income and education categories. Across all provinces, the individual covariate contributions are relatively small and often offsetting, resulting in a modest total covariate effect. In most cases, this channel explains little of the overall outcome gap, suggesting that changes in observable socioeconomic composition play a limited role.

The selection effect accounts for changes in outcomes driven by endogenous sorting of households across regions as son-preference varies. This component is quantitatively important in several provinces and, in some cases, dominates the decomposition. For instance, in Beijing, Shanghai, Jiangsu, and Shandong, the selection effect explains a substantial share of the total gap, indicating that migrants originating from strong son-preference areas also have high propensities for son-biased fertility in terms of unobservables.

Overall, the table highlights that differences in observed fertility outcomes across regions cannot be interpreted as purely cultural. While cultural preferences matter, a significant portion of the gap arises from endogenous sorting. This decomposition clarifies how conventional "culture effects" estimated from mean differences may conflate true preference transmission with selection mechanisms.

6.3 Heterogeneity and Determinants of Cultural Persistence

What factors determine the extent of cultural persistence or assimilation? As a first step toward answering this question, we explore heterogeneity in cultural persistence by individual characteristics and correlate the spatial variations of cultural persistence with local area characteristics. Although these correlations are descriptive and not causal mechanisms, they are informative about the potential factors that could affect the culture effects.

Individual characteristics. Figure 8 plots the culture effects in all urban destinations by education (panel a) and cohort (panel b). In each panel we report the inter-quartile and 90-10 percentile ranges of the estimated culture effects, in each subgroup relative to the base group. Panel (a) shows that culture effects are stronger for those with high school degree (HS) relative to those who only completed primary school (PS). The majority of the differences in the point estimates between these two subgroups are above zero. Differences between middle-school (MS) and PS are qualitatively similar albeit smaller in magnitudes. Panel (b) compares the estimates by cohort, showing that the culture effects are larger for the young cohort relative to the old. Therefore, we conclude that the culture of son preference is more persistent (and hence less cultural assimilation) among migrants who are younger and have more years of schooling.

Destination characteristics. Another interesting question is why some urban areas exhibit stronger culture effects than others. There are at least two factors underlying this difference. The first factor is the social or peer environment that migrants face in the urban destination. If there is a large share of migrants in the urban destination, this may reduce the likelihood of migrants interacting with local residents and their exposure to urban environments overall. The second factor is the urban income level, which has been regarded as a good indicator for the strength of son-favoring norms in urban areas.

We explore the role of these two factors using a regression analysis. As a proxy for migrants' peer environment, we use the proportion of inter-province migrants living in the urban destination in 2010. We measure the mean income levels in urban areas from the 1990 census. Because migration affects local wages, we believe that using the mean income level measured prior to the mass wave of migration is arguably better (more exogenous) than using wages from subsequent census.

Table 6 reports the results. When the two covariates enter into the regression individually (first two columns), the estimated coefficients are small and statistically insignificant. In column (3), we allow income and migrant share to affect culture effects jointly. Both marginal effects are significant: conditional on the proportion of inter-province migrants, a high income level in the urban destination reduces culture effects; conditional on income level, a high proportion of inter-province migrants strengthens cultural persistence.[30] The contrast between column (3) and the first two columns can be explained by a horse race between migrant share and income on culture: high local income reduces cultural persistence, but areas with high local income also attract a large share of migrants, which in turn reduces the exposure migrants have to the local environment.

7 Conclusion

This paper studies the cultural component of son preference by exploiting large-scale rural-to-urban migration in China and explicitly accounting for endogenous selection into migration destinations. Using a generalized Roy model and a control function approach, we show that the conventional epidemiological approach tends to overstate cultural persistence when migration selection is ignored. Once selection is corrected for, the estimated impact of origin-specific son preference on migrants' fertility behavior is substantially smaller, though not zero.

Our results suggest that differences in son-favoring fertility behavior among migrants living in the same urban context cannot be interpreted as purely cultural. A significant portion of these differences reflects non-random sorting across destinations based on unobservable characteristics rather than persistent preferences alone. These findings highlight a general identification challenge in the epidemiological approach to culture that applies beyond the context of son preference. Our framework can be applied to a wide range of settings where cultural traits are studied through migration.

At the same time, we find meaningful heterogeneity in cultural persistence across destinations. The degree to which origin-specific preferences persist depends on the interaction between local income levels and migrants' peer environments. Higher local income weakens cultural persistence, consistent with economic conditions reshaping fertility incentives, while a higher concentration of migrants strengthens

[30]The results are less precise for whether having the second child (column (6)) but are qualitatively similar. In Appendix Table A.2 we replace the share of inter-province migrants by the share of total migrants or share of inter-city migrants in the urban destination, and find similar results.

persistence by limiting exposure to local social norms. Culture, therefore, is neither fully portable nor fully malleable; its influence is shaped by the environments in which migrants are embedded.

Taken together, our findings have important implications for both research and policy. Methodologically, they underscore the importance of accounting for multidimensional migration selection when using migrants to identify culture effects. Substantively, they suggest that son preference is not an immutable cultural trait. Instead, migration, economic development, and institutional context play a central role in shaping fertility behavior, even for preferences that are often viewed as deeply rooted. Policies aimed at reducing gender imbalance may therefore be most effective when they focus on improving economic opportunities, social integration, and institutional support, rather than relying solely on long-run cultural change.

References

Alesina, A. and Giuliano, P. (2011). Family ties and political participation. *Journal of the European Economic Association*, 9(5):817–839.

Alesina, A., Giuliano, P., and Nunn, N. (2013). On the origins of gender roles: Women and the plough. *The Quarterly Journal of Economics*, 128(2):469–530.

Alfano, M. (2017). Daughters, dowries, deliveries: The effect of marital payments on fertility choices in India. *Journal of Development Economics*, 125:89–104.

Almond, D., Edlund, L., and Milligan, K. (2013). Son preference and the persistence of culture: evidence from south and east asian immigrants to canada. *Population and Development Review*, 39(1):75–95.

Almond, D., Li, H., and Zhang, S. (2019). Land reform and sex selection in China. *Journal of Political Economy*, 127(2):560–585.

Angrist, J. (2002). How do sex ratios affect marriage and labor markets? evidence from America's second generation. *The Quarterly Journal of Economics*, 117(3):997–1038.

Bhalotra, S., Chakravarty, A., Mookherjee, D., and Pino, F. J. (2019). Property rights and gender bias: Evidence from land reform in west bengal. *American Economic Journal: Applied Economics*, 11(2):205–237.

Black, S. E. and Devereux, P. J. (2011). Recent developments in intergenerational mobility. *Handbook of labor economics*, 4:1487–1541.

Blau, F. D., Kahn, L. M., and Papps, K. L. (2011). Gender, source country characteristics, and labor market assimilation among immigrants. *Review of Economics and Statistics*, 93(1):43–58.

Boelmann, B., Raute, A., and Schönberg, U. (2025). Wind of change? Cultural determinants of maternal labor supply. *American Economic Journal: Applied Economics*, 17(2):41–74.

Borjas, G. J. (1987). Self-Selection and the Earnings of Immigrants. *American Economic Review*, 77(4):531–553.

Brainerd, E. (2017). The lasting effect of sex ratio imbalance on marriage and family: Evidence from World War II in Russia. *Review of Economics and Statistics*, 99(2):229–242.

Cameron, L., Meng, X., and Zhang, D. (2019). China's sex ratio and crime: Behavioural change or financial necessity? *The Economic Journal*, 129(618):790–820.

Carroll, C. D., Rhee, B.-K., and Rhee, C. (1999). Does cultural origin affect saving behavior? Evidence from immigrants. *Economic Development and Cultural Change*, 48(1):33–50.

Chen, Y., Li, H., and Meng, L. (2013). Prenatal sex selection and missing girls in China: Evidence from the diffusion of diagnostic ultrasound. *Journal of Human Resources*, 48(1):36–70.

Coale, A. J. and Banister, J. (1994). Five decades of missing females in China. *Demography*, 31(3):459–479.

Dahl, G. B. (2002). Mobility and the return to education: Testing a Roy model with multiple markets. *Econometrica*, 70(6):2367–2420.

Dahl, G. B. and Moretti, E. (2008). The demand for sons. *The Review of Economic Studies*, 75(4):1085–1120.

Das Gupta, M. (2005). Explaining Asia's "missing women"': A new look at the data. *Population and development review*, 31(3):529–535.

Das Gupta, M. (2010). Family systems, political systems and Asia's 'missing girls' the construction of son preference and its unravelling. *Asian Population Studies*, 6(2):123–152.

Dubuc, S. and Coleman, D. (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.

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. (2021). The Historical Origins of Son Preference: Patrilocality and Missing Women. SSRN Working Paper 3829406, Social Science Research Network.

Edlund, L., Li, H., Yi, J., and Zhang, J. (2013). Sex ratios and crime: Evidence from China. *Review of Economics and Statistics*, 95(5):1520–1534.

Epple, D., Romer, T., and Sieg, H. (1999). An equilibrium analysis of fiscal competition and local public goods in a system of cities. *International Economic Review*, 40(1):177–198.

Fernández, R. (2007). Women, Work, and Culture. *Journal of the European Economic Association*, 5(2-3):305–332.

Fernández, R. (2011). Does culture matter? *Handbook of social economics*, 1:481–510.

Fernández, R. and Fogli, A. (2006). Fertility: The role of culture and family experience. *Journal of the European economic association*, 4(2-3):552–561.

Fernández, R. and Fogli, A. (2009). Culture: An empirical investigation of beliefs, work, and fertility. *American economic journal: Macroeconomics*, 1(1):146–77.

Giuliano, P. (2007). Living arrangements in Western Europe: Does cultural origin matter? *Journal of the European Economic Association*, 5(5):927–952.

Grosjean, P. and Khattar, R. (2019). It's raining men! Hallelujah? The long-run consequences of male-biased sex ratios. *The Review of Economic Studies*, 86(2):723–754.

Guiso, L., Sapienza, P., and Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives*, 20(2):23–48.

Guo, N., Huang, W., and Wang, R. (2025). Public pensions and family dynamics: Eldercare, child investment, and son preference in rural China. *Journal of Development Economics*, 172:103390.

Hauge, K. E., Kotsadam, A., and Riege, A. (2023). Culture and gender differences in willingness to compete. *The Economic Journal*, 133(654):2403–2426.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161.

Heckman, J. J. and Mosso, S. (2014). The economics of human development and social mobility. *Annu. Rev. Econ.*, 6(1):689–733.

Hesketh, T. and Xing, Z. W. (2006). Abnormal sex ratios in human populations: Causes and consequences. *Proceedings of the National Academy of Sciences*, 103(36):13271–13275.

Horioka, C. Y. and Terada-Hagiwara, A. (2017). The impact of sex ratios before marriage on household saving in two asian countries: The competitive saving motive revisited. *Review of Economics of the Household*, 15:739–757.

Huang, W., Pan, Y., and Zhou, Y. (2025). One-Child Policy, Marriage Distortion, and Welfare Loss. *The Review of Economics and Statistics*, 107(5):1341–1355.

Kim, S. and Lee, S.-H. (2020). Son preference and fertility decisions: Evidence from spatiotemporal variation in korea. *Demography*, 57(3):927–951.

Lee, L.-F. (1983). Generalized econometric models with selectivity. *Econometrica: Journal of the Econometric Society*, pages 507–512.

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(4):1535–1557.

Luttmer, E. F. P. and Singhal, M. (2011). Culture, context, and the taste for redistribution. *American Economic Journal: Economic Policy*, 3(1):157–179.

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

Murphy, R., Tao, R., and Lu, X. (2011). Son preference in rural China: Patrilineal families and socioeconomic change. *Population and development review*, 37(4):665–690.

Nechyba, T. J. (1997). Existence of equilibrium and stratification in local and hierarchical Tiebout economies with property taxes and voting. *Economic Theory*, 10:277–304.

Qian, N. (2008). Missing women and the price of tea in China: The effect of sex-specific earnings on sex imbalance. *The Quarterly Journal of Economics*, 123(3):1251–1285.

Rosenzweig, M. R. and Schultz, T. P. (1982). Market opportunities, genetic endowments, and intrafamily resource distribution: Child survival in rural India. *The American Economic Review*, 72(4):803–815.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford economic papers*, 3(2):135–146.

Sen, A. (2002). More than 100 million women are missing. In *Gender and Justice*, pages 219–222. Routledge.

Sun, A. and Zhao, Y. (2016). Divorce, abortion, and the child sex ratio: The impact of divorce reform in China. *Journal of Development Economics*, 120:53–69.

Tiebout, C. M. (1956). A pure theory of local expenditures. *Journal of Political Economy*, 64(5):416–424.

Wang, Z., Wang, Q., and Qi, J. (2021). Ten years of the China migrants dynamic survey: Retrospective and prospects. *China Population and Development Studies*, 5(3):296–305.

Wei, S.-J. and Zhang, X. (2011). The competitive saving motive: Evidence from rising sex ratios and savings rates in China. *Journal of Political Economy*, 119(3):511–564.

Zhang, J. (2017). The evolution of China's one-child policy and its effects on family outcomes. *Journal of Economic Perspectives*, 31(1):141–160.

Table 1: Summary Statistics

Sample	(1) Have a 1st-born daughter		(2) Have a 1st-born daughter and a 2nd child	
	Mean	SD	Mean	SD
Fertility outcomes				
Having the 2nd child (1 = yes)	0.504	(0.500)	1.000	—
Gender of the 2nd child (1 = boy)	—	—	0.657	(0.475)
Son preference culture exposure				
Origin sex ratio at age 16 (std)	0.046	(0.971)	0.152	(0.958)
Migrant demographics				
Gender (1 = female)	0.491	(0.500)	0.482	(0.500)
Edu: Primary	0.120	(0.326)	0.168	(0.374)
Edu: Junior secondary	0.570	(0.495)	0.614	(0.487)
Edu: Senior secondary	0.216	(0.412)	0.167	(0.373)
Edu: Tertiary	0.081	(0.273)	0.034	(0.182)
Ethnicity (1 = ethnic minorities)	0.070	(0.256)	0.074	(0.262)
Log monthly household income	8.636	(0.583)	8.641	(0.593)
Age at survey year	34.336	(6.126)	36.356	(5.547)
Observations	143087		72101	

Note: This table reports the means and standard deviations of variables used in the analysis.

Table 2: Baseline selection correction results (Gender of the second child)

Province	Uncorrected Coef.	Corrected Coef.	Test for Difference	Wald Test for Correction Terms	Province	Uncorrected Coef.	Corrected Coef.	Test for Difference	Wald Test for Correction Terms
ALL REGIONS					**CENTRAL AND SOUTH**				
Mean	0.024	0.007			Henan	0.015	0.005	0.321	3.049
Std. Dev.	0.022	0.019				(0.013)	(0.020)	(0.571)	(0.004)
					Hubei	0.004	-0.003	0.613	4.023
NORTH						(0.012)	(0.014)	(0.434)	(0.000)
Beijing	0.053***	0.024	7.515	2.752	Hunan	0.014	-0.008	2.849	1.646
	(0.014)	(0.017)	(0.006)	(0.019)		(0.024)	(0.028)	(0.091)	(0.124)
Tianjin	0.009	-0.001	1.575	0.840	Guangdong	-0.006	-0.008	0.208	0.604
	(0.014)	(0.016)	(0.210)	(0.522)		(0.007)	(0.009)	(0.648)	(0.753)
Hebei	0.039***	-0.006	9.151	1.793	Guangxi	-0.018	-0.009	0.822	1.287
	(0.014)	(0.020)	(0.002)	(0.088)		(0.012)	(0.016)	(0.365)	(0.257)
Shanxi	0.037***	-0.005	8.131	2.520	Hainan	0.028**	0.024*	0.390	1.507
	(0.011)	(0.018)	(0.004)	(0.016)		(0.011)	(0.014)	(0.532)	(0.164)
Inner Mongolia	0.038**	0.021	1.399	2.013	Mean	0.006	0.000		
	(0.018)	(0.022)	(0.237)	(0.054)	Std. Dev.	0.016	0.013		
Liaoning	0.018	-0.003	1.517	2.234					
	(0.018)	(0.024)	(0.218)	(0.032)	**WEST**				
Jilin	0.058***	0.029	3.190	1.652	Chongqing	-0.022	-0.026	0.045	1.076
	(0.014)	(0.021)	(0.074)	(0.124)		(0.027)	(0.033)	(0.831)	(0.380)
Heilongjiang	0.051***	-0.015	6.931	10.550	Sichuan	0.062***	0.043*	1.055	2.118
	(0.018)	(0.031)	(0.008)	(0.000)		(0.021)	(0.026)	(0.304)	(0.043)
Mean	0.038	0.006			Guizhou	0.013	0.001	1.316	2.113
Std. Dev.	0.017	0.017				(0.017)	(0.022)	(0.251)	(0.043)
					Yunnan	0.038***	0.003	13.068	4.110
EAST						(0.013)	(0.017)	(0.000)	(0.000)
Shanghai	0.014	0.000	4.243	1.658	Tibet	0.053**	0.038	0.715	1.868
	(0.015)	(0.015)	(0.039)	(0.145)		(0.026)	(0.031)	(0.398)	(0.079)
Jiangsu	0.038***	0.018*	11.242	3.139	Shaanxi	0.031**	0.031*	0.004	1.150
	(0.011)	(0.011)	(0.001)	(0.003)		(0.015)	(0.016)	(0.948)	(0.332)
Zhejiang	0.017	0.002	8.807	2.435	Gansu	0.021*	0.004	5.105	1.535
	(0.010)	(0.010)	(0.003)	(0.019)		(0.012)	(0.014)	(0.024)	(0.155)
Anhui	0.008	-0.009	4.070	5.435	Qinghai	0.039***	0.034**	0.186	1.652
	(0.019)	(0.020)	(0.044)	(0.000)		(0.009)	(0.015)	(0.666)	(0.120)
Fujian	0.026**	0.015	1.829	1.028	Ningxia	0.052***	0.032**	4.299	1.423
	(0.011)	(0.012)	(0.176)	(0.411)		(0.009)	(0.014)	(0.038)	(0.196)
Jiangxi	-0.013	-0.035**	5.198	2.515	Xinjiang	0.026***	0.020	0.438	1.598
	(0.014)	(0.016)	(0.023)	(0.016)		(0.009)	(0.013)	(0.508)	(0.134)
Shandong	-0.005	-0.002	0.087	2.518	Mean	0.031	0.018		
	(0.020)	(0.023)	(0.769)	(0.017)	Std. Dev.	0.024	0.022		
Mean	0.012	-0.002							
Std. Dev.	0.017	0.018							

Notes: This table compares the coefficients and standard errors of uncorrected origin effects and selection-corrected origin effects (estimated using the double index model) for each destination province in columns (1) and (2). The outcome is an indicator of whether the second child is male. Column (3) reports the χ^2 statistics and p-values from the test of coefficient equality between the two models. Column (4) reports F-statistics and p-values from a Wald test for the null hypothesis that all correction terms are jointly zero. Mean within a region is calculated as the average of the province-specific culture effects in that region. Mean across all regions represents the average of all province-specific culture effects.

Table 3: Baseline selection correction results (Whether having the second child)

Province	Uncorrected Coef.	Corrected Coef.	Test for Difference	Wald Test for Correction Terms	Province	Uncorrected Coef.	Corrected Coef.	Test for Difference	Wald Test for Correction Terms
		ALL REGIONS					**CENTRAL AND SOUTH**		
Mean	0.047	0.020			Henan	0.033***	0.013	1.903	13.041
Std. Dev.	0.029	0.019				(0.009)	(0.016)	(0.168)	(0.000)
					Hubei	0.013	0.009	0.318	21.973
		North				(0.012)	(0.012)	(0.573)	(0.000)
Beijing	0.040***	0.026**	3.783	1.241	Hunan	0.040**	0.020	4.789	1.955
	(0.009)	(0.012)	(0.052)	(0.289)		(0.018)	(0.019)	(0.029)	(0.061)
Tianjin	0.067***	0.055***	4.678	5.227	Guangdong	0.017***	0.001	18.313	8.903
	(0.011)	(0.013)	(0.031)	(0.000)		(0.007)	(0.008)	(0.000)	(0.000)
Hebei	0.062***	0.062***	0.002	10.637	Guangxi	-0.005	0.024**	18.131	10.047
	(0.012)	(0.015)	(0.964)	(0.000)		(0.011)	(0.012)	(0.000)	(0.000)
Shanxi	0.053***	0.021	9.017	64.419	Hainan	0.033***	0.015	8.711	18.335
	(0.010)	(0.014)	(0.003)	(0.000)		(0.010)	(0.011)	(0.003)	(0.000)
Inner Mongolia	0.107***	0.055***	29.480	8.000	Mean	0.022	0.014		
	(0.012)	(0.015)	(0.000)	(0.000)	Std. Dev.	0.017	0.008		
Liaoning	0.109***	-0.001	121.001	25.930					
	(0.012)	(0.013)	(0.000)	(0.000)			**WEST**		
Jilin	0.091***	0.024	45.670	13.287	Chongqing	0.044***	-0.007	15.469	5.558
	(0.016)	(0.015)	(0.000)	(0.000)		(0.016)	(0.021)	(0.000)	(0.000)
Heilongjiang	0.077***	-0.016	36.703	40.720	Sichuan	0.071***	0.005	18.490	38.473
	(0.011)	(0.016)	(0.000)	(0.000)		(0.015)	(0.021)	(0.000)	(0.000)
Mean	0.076	0.028			Guizhou	0.020	0.026*	1.048	5.664
Std. Dev.	0.025	0.028				(0.012)	(0.013)	(0.306)	(0.000)
					Yunnan	0.060***	0.006	49.363	9.819
		EAST				(0.011)	(0.013)	(0.000)	(0.000)
Shanghai	0.041***	0.027**	7.856	11.263	Tibet	0.022**	-0.001	11.933	4.355
	(0.010)	(0.011)	(0.005)	(0.000)		(0.010)	(0.010)	(0.001)	(0.000)
Jiangsu	0.030***	0.025***	1.456	15.892	Shaanxi	0.027**	0.010	5.112	8.886
	(0.010)	(0.010)	(0.228)	(0.000)		(0.012)	(0.013)	(0.024)	(0.000)
Zhejiang	0.047***	0.008	92.511	30.042	Gansu	0.015	-0.003	9.541	36.400
	(0.009)	(0.006)	(0.000)	(0.000)		(0.010)	(0.012)	(0.002)	(0.000)
Anhui	0.004	0.015	1.666	11.909	Qinghai	0.037***	0.013	7.713	3.812
	(0.015)	(0.015)	(0.197)	(0.000)		(0.006)	(0.010)	(0.005)	(0.001)
Fujian	0.051***	0.025**	18.986	8.884	Ningxia	0.029***	0.033***	0.318	4.115
	(0.011)	(0.010)	(0.000)	(0.000)		(0.008)	(0.010)	(0.573)	(0.000)
Jiangxi	0.047***	0.015	13.848	3.916	Xinjiang	0.086***	0.049***	29.245	5.608
	(0.012)	(0.014)	(0.000)	(0.000)		(0.007)	(0.009)	(0.000)	(0.000)
Shandong	0.086***	0.052***	14.011	13.000	Mean	0.041	0.013		
	(0.015)	(0.014)	(0.000)	(0.000)	Std. Dev.	0.024	0.018		
Mean	0.044	0.024							
Std. Dev.	0.025	0.014							

Notes: This table compares the coefficients and standard errors of uncorrected origin effects and selection-corrected origin effects (estimated using the double index model) for each destination province in columns (1) and (2). The outcome is an indicator of whether a migrant has a second child. Column (3) reports the χ^2 statistics and p-values from the test of coefficient equality between the two models. Column (4) reports F-statistics and p-values from a Wald test for the null hypothesis that all correction terms are jointly zero. Mean within a region is calculated as the average of the province-specific culture effects in that region. Mean across all regions represents the average of all province-specific culture effects.

Table 4: Selection correction function and sex ratio

	(1)	(2)	(3)
Panel A:	Dep. Var.: λ's for Gender of 2nd child		
Destination sex ratio at age 16 (std)	-0.005^{*}	-0.004^{*}	-0.004^{**}
	(0.003)	(0.002)	(0.002)
N	72,101	72,101	72,101
Panel B:	Dep. Var.: λ's for whether having the 2nd child		
Destination sex ratio at age 16 (std)	0.013^{**}	-0.003	-0.003
	(0.006)	(0.003)	(0.004)
N	143,087	143,087	143,087
Individual characteristics	Yes	Yes	Yes
Origin FEs		Yes	Yes
Individual characteristics * Origin FE			Yes

Notes This table shows the correlation between sex ratio at age 16 in the destination province and predicted λ's obtained from the multiple-index model. In all columns, we include individual characteristics that are used to define cells, such as gender, educational attainment, and cohort group. Column (2) also controls for origin fixed effects. Column (3) further controls for all interactions between individual characteristics and origin fixed effects. Standard errors are clustered at the origin-by-destination level. ***, **, and * denote statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

Table 5: Decomposition of gaps in outcome variables

	Beijing	Shanghai	Guangdong	Jiangsu	Shandong
Panel A: Gender of the second child					
Gap in outcome	0.143	0.013	-0.025	0.077	0.056
Total gap	0.113	0.026	-0.019	0.064	0.042
Culture effect	0.058	0.001	-0.018	0.038	0.005
Covariate-composition effect					
Household income	-0.003	0.002	0.000	-0.001	-0.001
Primary education	0.003	0.004	-0.000	0.001	0.007
Junior secondary	0.009	-0.008	0.000	0.001	-0.001
Senior secondary	-0.012	-0.004	-0.000	-0.000	0.000
Tertiary	-0.004	0.009	-0.001	-0.000	-0.000
...					
Total covariate effect	-0.020	-0.005	-0.004	-0.014	-0.003
Selection effect	0.075	0.030	0.002	0.040	0.040
Panel B: Whether having the second child					
Gap in outcome	0.108	0.064	0.067	0.058	0.086
Total gap	0.108	0.063	0.090	0.053	0.084
Culture effect	0.064	0.059	0.003	0.054	0.017
Covariate-composition effect					
Household income	-0.004	-0.002	0.000	-0.001	-0.002
Primary education	-0.000	-0.002	-0.000	0.002	0.001
Junior secondary	-0.008	-0.002	0.000	-0.014	-0.017
Senior secondary	0.009	-0.007	0.002	0.013	0.007
Tertiary	0.030	0.017	-0.009	0.006	0.023
...					
Total covariate effect	0.012	-0.021	0.038	-0.001	-0.024
Selection effect	0.033	0.025	0.049	0.000	0.090

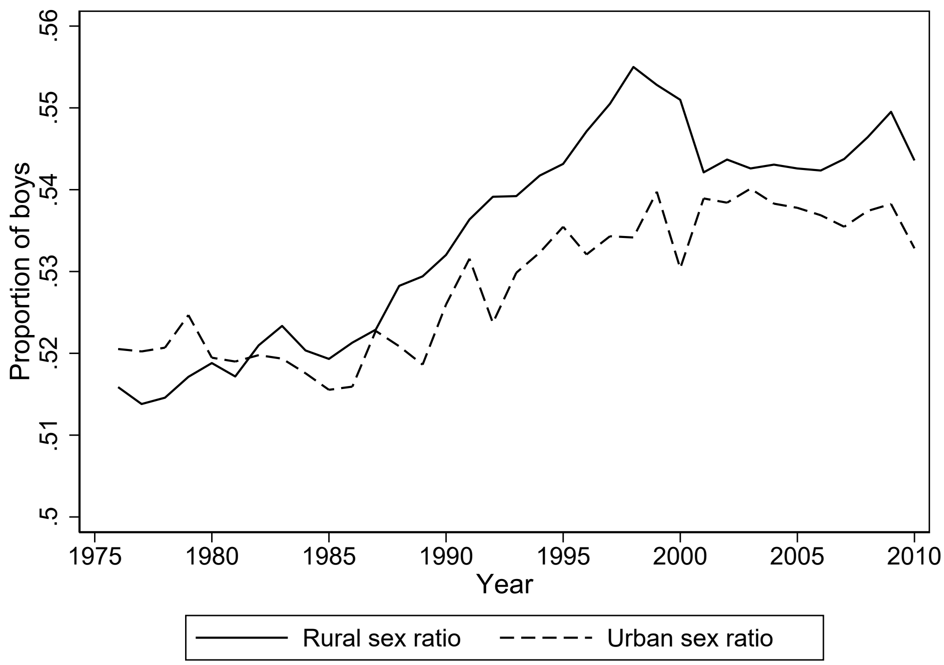
Notes This table presents the contribution of culture, covariates, and selection effects to the mean gap in outcome variables between the top and bottom quartiles of the sex-ratio distribution. Total covariate effects also include contributions from the following additional covariates: gender, ethnicity, age, age square, and cohort dummies. Total gap is the sum of culture effects, total covariate effects, and selection effects.

Table 6: Culture effect and destination characteristics

Dep. Var.	β_k for gender of 2nd child			β_k for whether having the 2nd child		
	(1)	(2)	(3)	(4)	(5)	(6)
Urban income	−0.030		−0.399**	−0.052		−0.305
(10k)	(0.123)		(0.192)	(0.125)		(0.205)
% of cross-province		0.037	0.122**		0.019	0.084
migrants		(0.032)	(0.051)		(0.033)	(0.055)
N	31	31	31	31	31	31

Notes This table shows the correlation between β_k and economic conditions and the proportion of migrants in the destination province. Outcome is the province-specific selection-corrected culture effects β_k obtained from the multiple index model. Urban income is measured using the urban income (in ten thousand RMB) in 1990. The proportion of cross-province migrants is calculated using data from the 2010 Census. ***, **, and * denote statistical significance at 0.01, 0.05, and 0.1 levels, respectively.

(a) Sex ratios in rural and urban areas: 1976-2010



(b) Rural sex ratio by birth order: 1976-2000

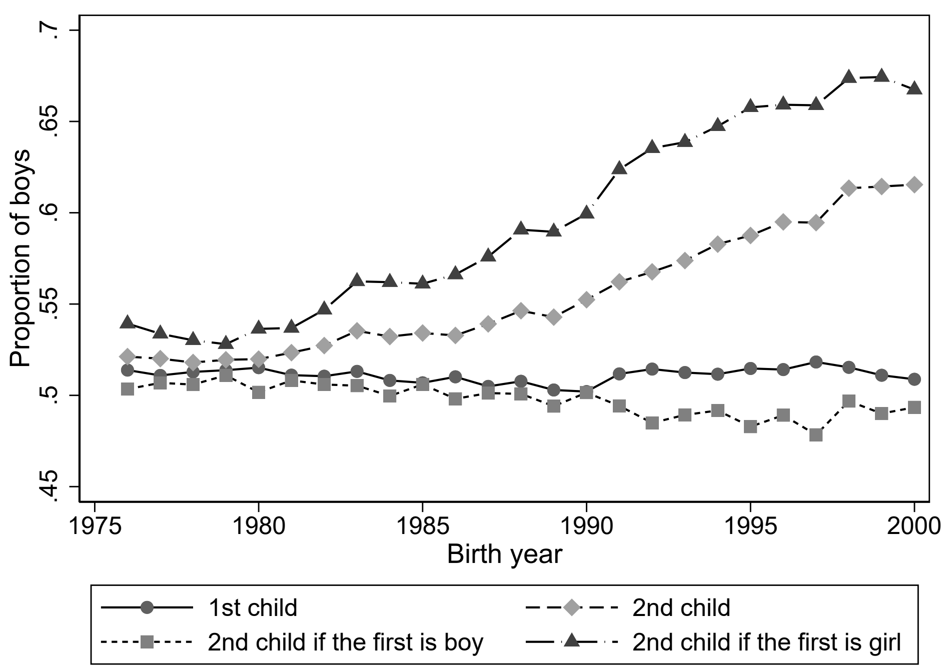


Figure 1: Sex ratios in China

Notes : Figure (a) shows the evolution of sex ratios in rural (solid line) and urban areas (dashed line) between 1976 and 2010. Data source: 1990 and 2000 Census 1% microdata, 2010 Census province-level data. Figure (b) shows the proportions of male newborns in rural areas by birth order in each year between 1976 and 2000. Data source: 1990 and 2000 Census 1% microdata.

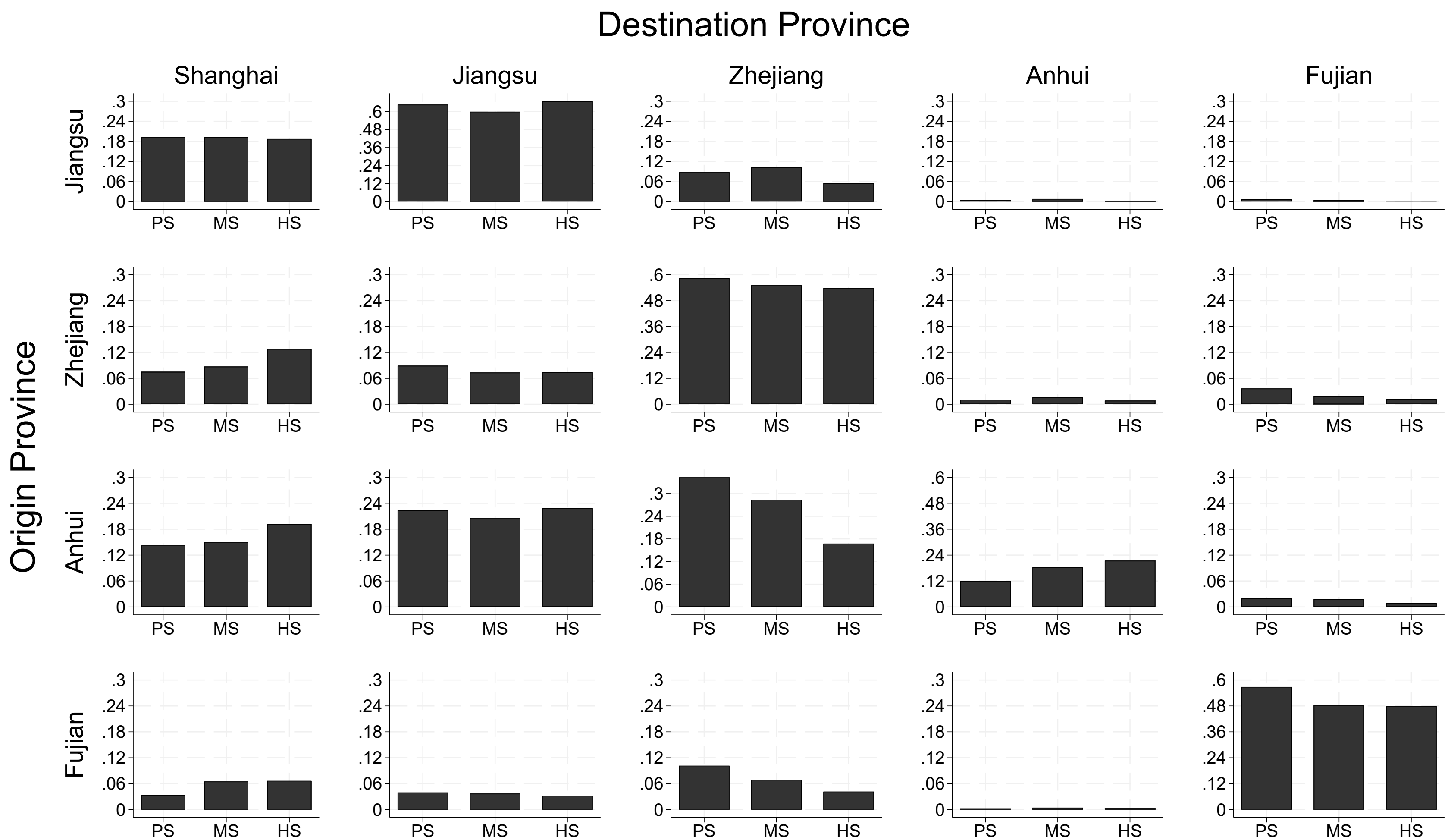


Figure 2: Migration patterns for 5 provinces

Notes This figure shows the probability that migrants from a given origin and educational level choose a specific destination. Educational groups are categorized as PS (primary school or below), MS (middle school), and HS (high school or above). The probabilities are calculated using the CMDS 2013-2017 sample of rural-to-urban migrants whose first child is a girl.

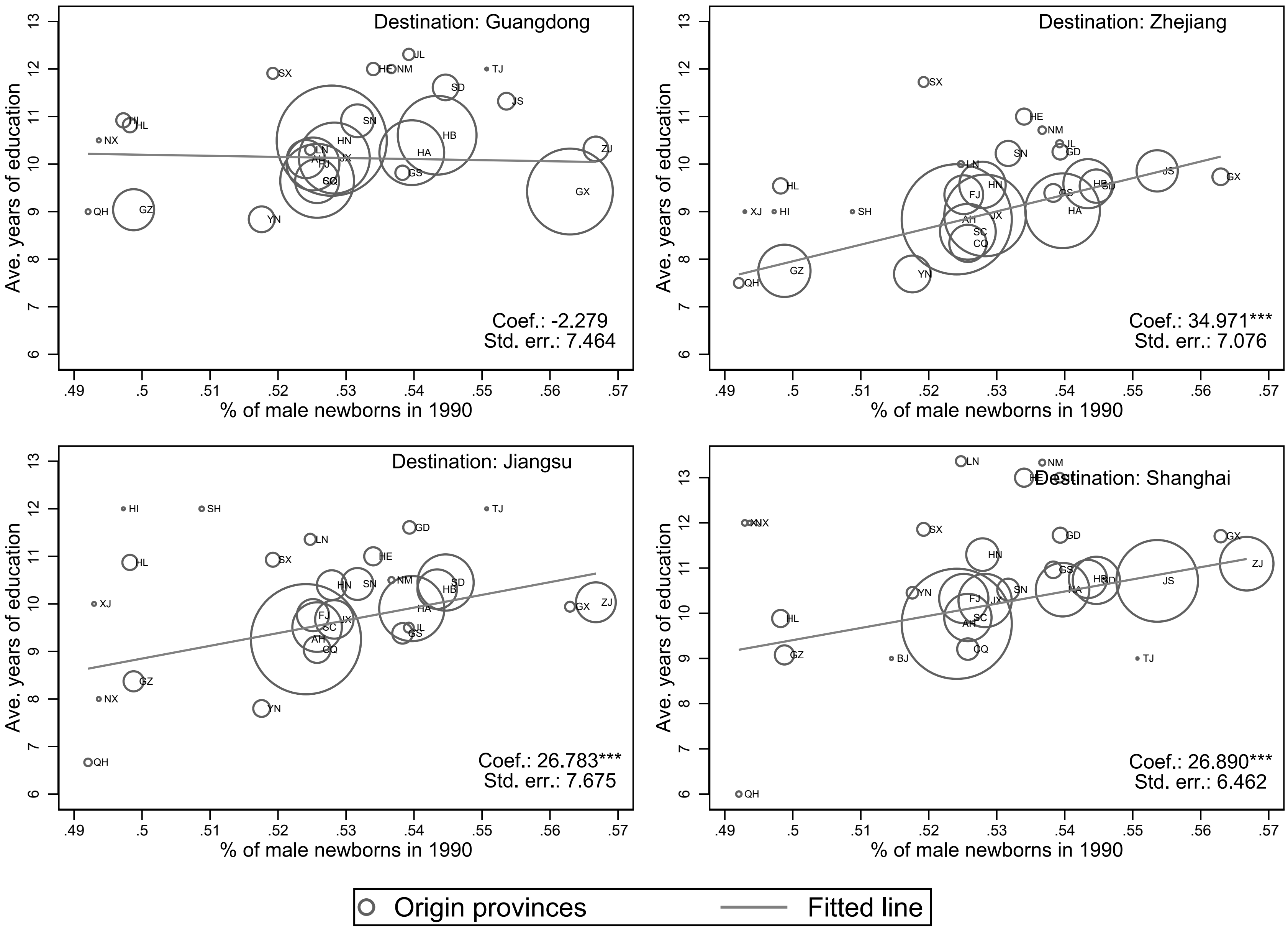
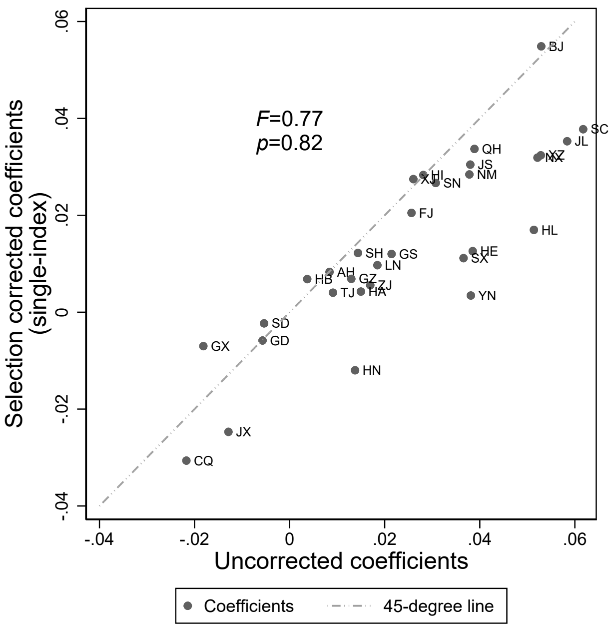
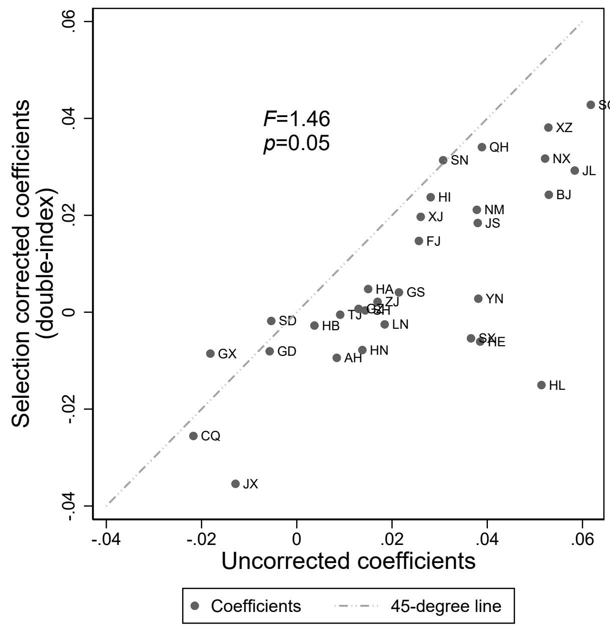


Figure 3: Origin sex ratio and average years of education of immigrants

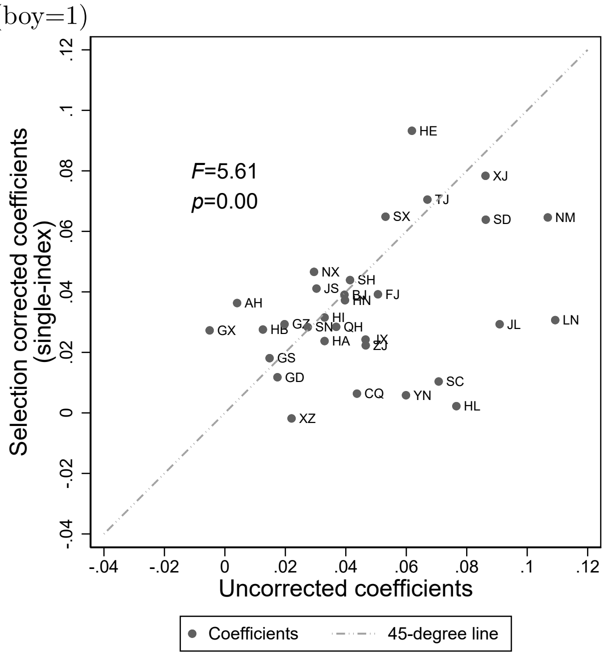
Notes This figure shows the relationship between the average years of education among immigrants from each origin province and the proportion of male newborns in that province in 1990. Average years of education is calculated from the CMDS 2013-2017 sample of rural-to-urban migrants whose first child is a girl. The size of the circle represents the population size of immigrants from each origin.



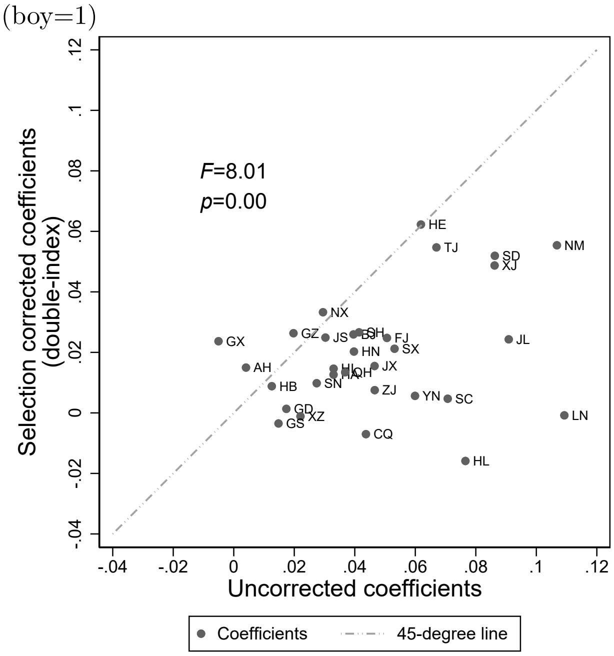
(a) Single index model, gender of the second child (boy=1)



(b) Double index model, gender of the second child (boy=1)



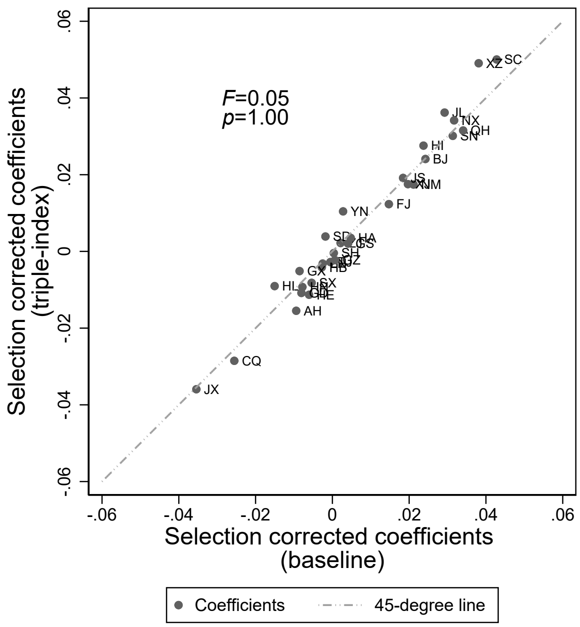
(c) Single index model, whether having a second child (yes=1)



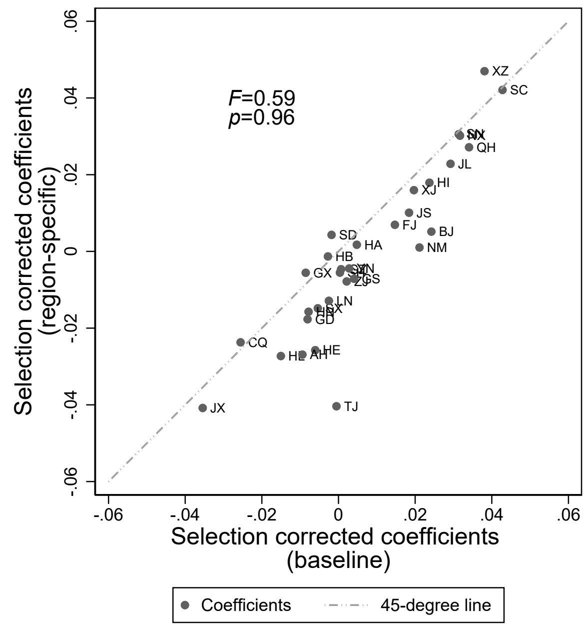
(d) Double index model, whether having a second child (yes=1)

Figure 4: Corrected vs. uncorrected cultural mobility: son preference

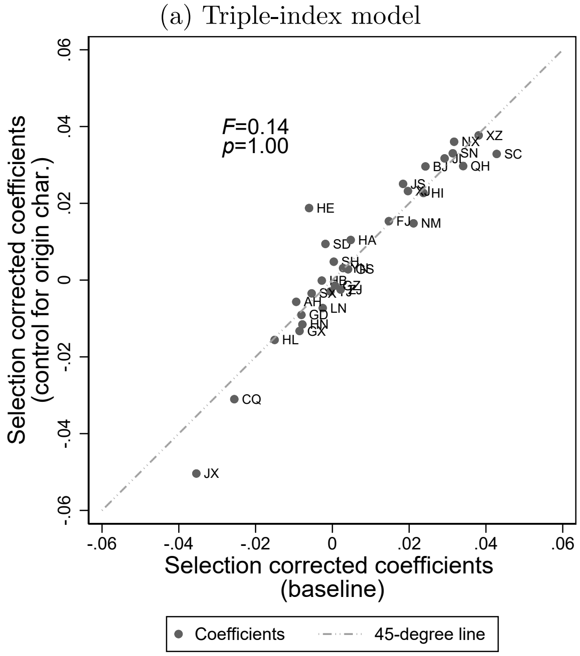
Notes : This figure plots the single index selection corrected results (Figure 4a and 4c) and double index selection corrected results (Figure 4b and 4d) against reduced form origin effects. Outcomes are gender of the second child (boy=1) and whether having the second child (yes=1), conditional on the first birth being a girl. The dashed line is the 45-degree line. F-statistic and p-value are from a Wald test of the hypothesis that all selection corrected coefficients are equal to the uncorrected ones. For the corresponding full names of each province, refer to Table A.1 in the Appendix.



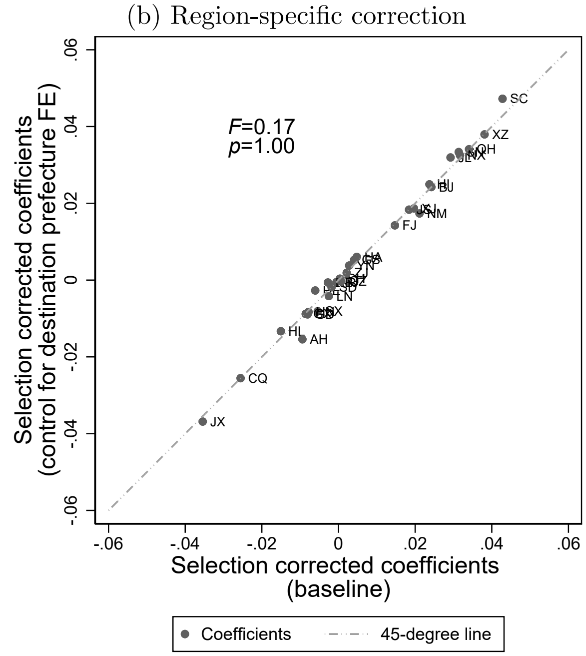
(a) Triple-index model



(b) Region-specific correction



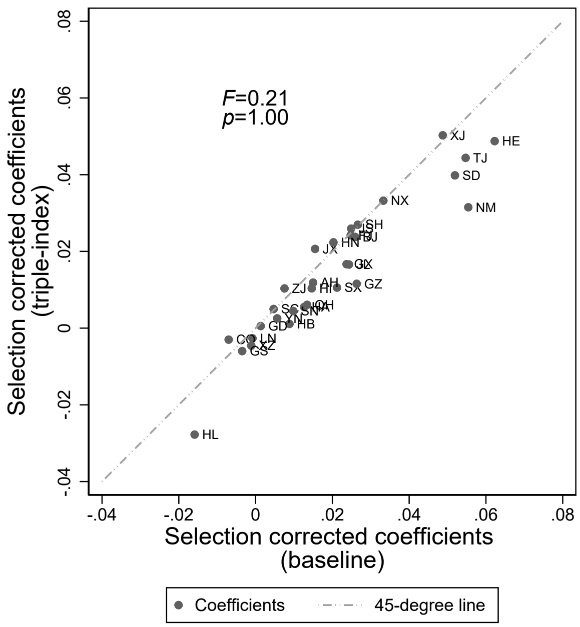
(c) Control for origin characteristics at 16



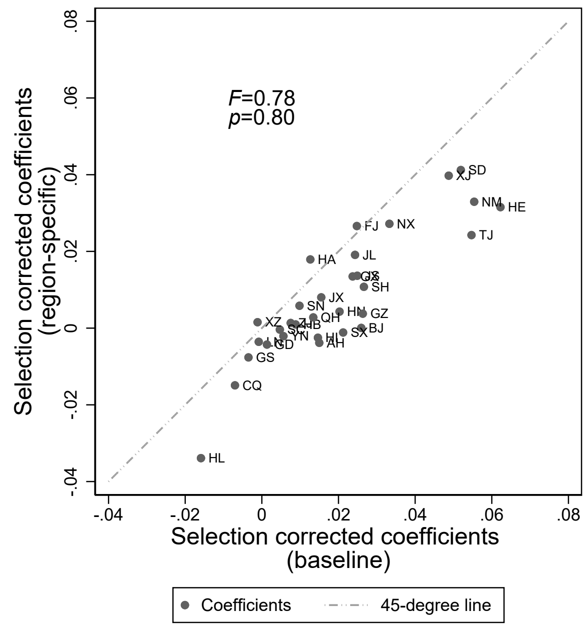
(d) Control for destination prefecture FEs

Figure 5: Robustness checks: Gender of the second child (boy=1)

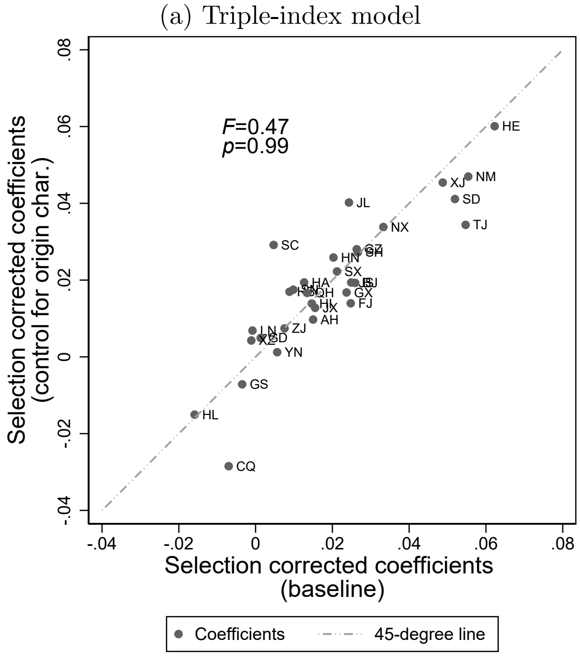
Notes : This figure plots selection corrected coefficients obtained from four robustness checks against baseline selection correction coefficients. The outcome is the gender of the second child (boy=1). For the corresponding full names of each province, refer to Table A.1 in the Appendix.



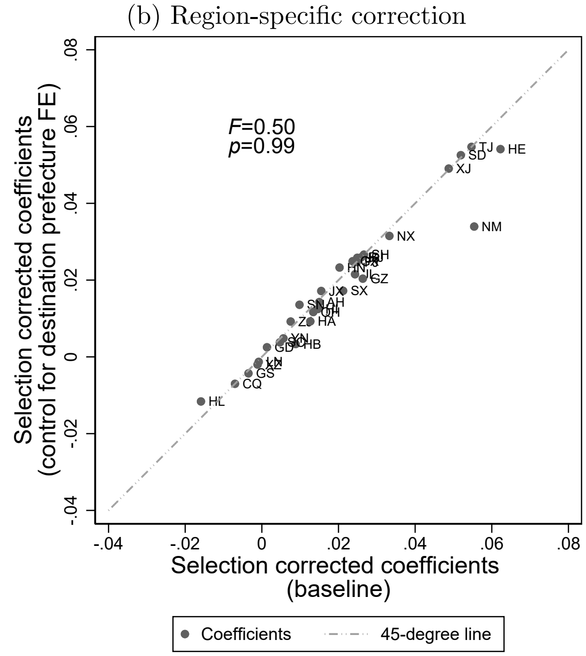
(a) Triple-index model



(b) Region-specific correction



(c) Control for origin characteristics at 16



(d) Control for destination prefecture FEs

Figure 6: Robustness checks: Whether having the second child (yes=1)

Notes : This figure plots selection corrected coefficients obtained from four robustness checks against baseline selection correction coefficients. The outcome is whether having the second child (yes=1). For the corresponding full names of each province, refer to Table A.1 in the Appendix.

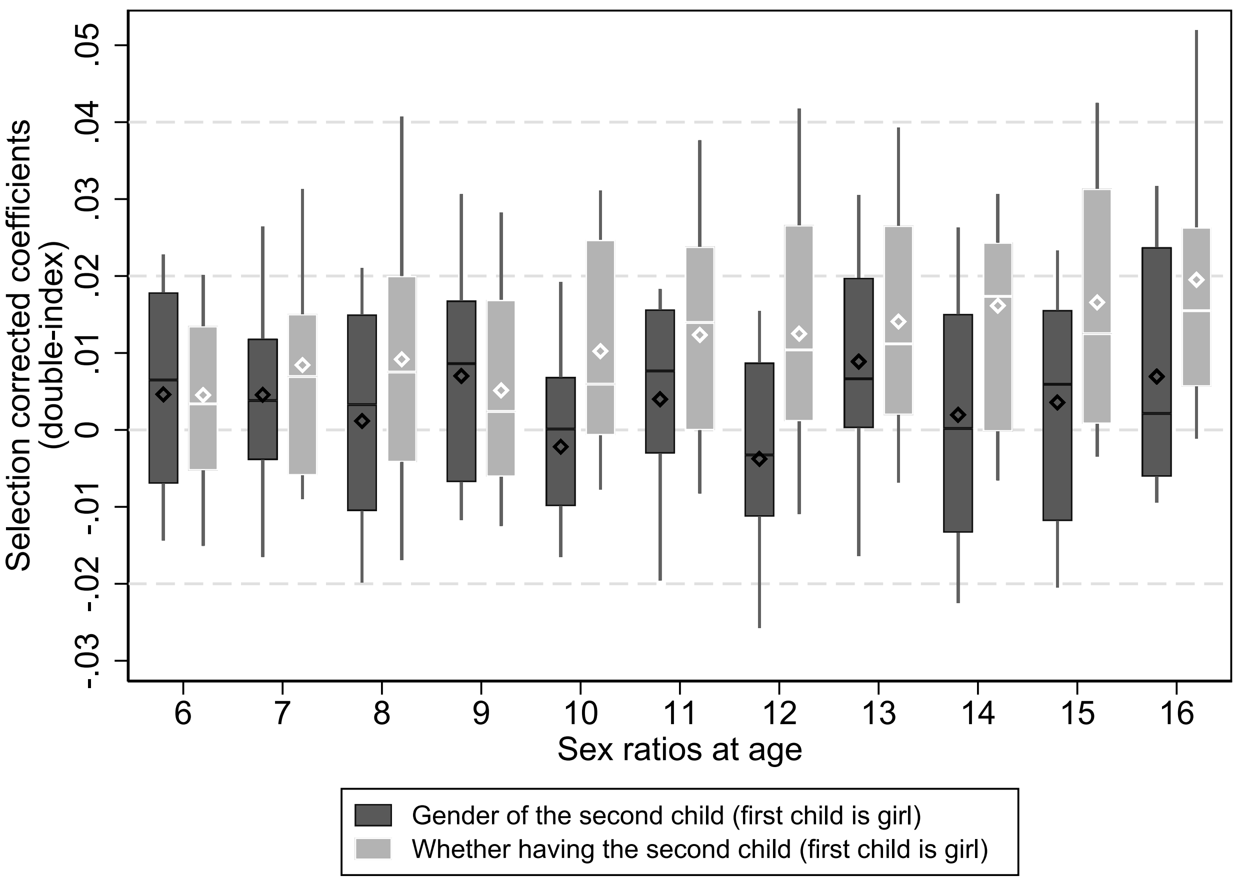
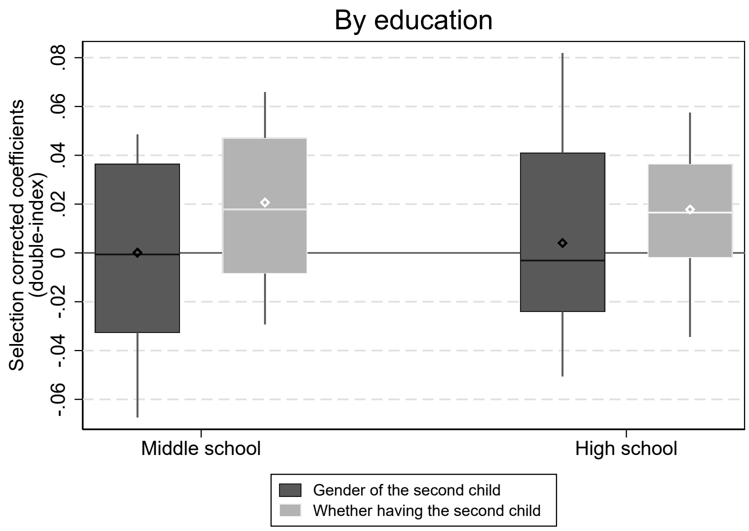
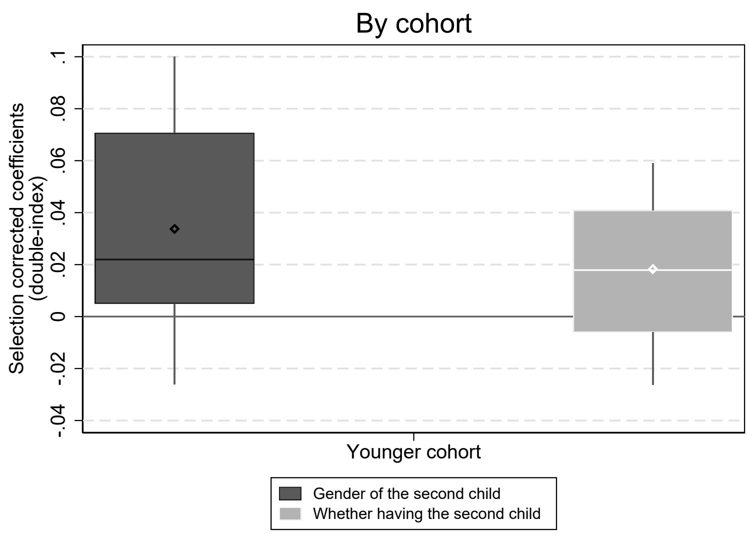


Figure 7: Measuring origin culture using sex ratios at different ages

Notes : This figure shows the means (diamond), medians (central line), 25th and 75th percentile ranges (box), and 10th and 90th percentile ranges (vertical line) of the distributions of selection corrected coefficients on origin sex ratios at different ages. Outcomes are gender of the second child (boy=1, dark grey boxes) and whether having the second child (yes=1, light grey boxes), conditional on the first child being a girl.



(a) By level of education



(b) By cohort

Figure 8: Heterogeneity of culture effects by individual characteristics

Notes : This figure shows the distribution of estimated culture effects for each subgroup relative to the base group, with diamonds representing means, central lines representing medians, boxes indicating the 25th–75th percentiles, and vertical lines indicating the 10th–90th percentiles. In panel (a), education groups are middle school and high school, with primary school as the base group. In panel (b), cohort groups are defined by birth year, with those born before 1980 as the base group.

A Online Appendix

Table A.1: Chinese Province Abbreviations

Abbreviations	Full name
AH	Anhui Province
BJ	Beijing Municipality
CQ	Chongqing Municipality
FJ	Fujian Province
GD	Guangdong Province
GS	Gansu Province
GX	Guangxi Zhuang Autonomous Region
GZ	Guizhou Province
HA	Henan Province
HB	Hubei Province
HE	Hebei Province
HI	Hainan Province
HL	Heilongjiang Province
HN	Hunan Province
JL	Jilin Province
JS	Jiangsu Province
JX	Jiangxi Province
LN	Liaoning Province
NM	Inner Mongolia Autonomous Region
NX	Ningxia Hui Autonomous Region
QH	Qinghai Province
SC	Sichuan Province
SD	Shandong Province
SH	Shanghai Municipality
SN	Shaanxi Province
SX	Shanxi Province
TJ	Tianjin Municipality
XJ	Xinjiang Uyghur Autonomous Region
XZ	Tibet Autonomous Region
YN	Yunnan Province
ZJ	Zhejiang Province

Table A.2: Culture effect and destination characteristics

Dep. Var.	β_k for gender of 2nd child				β_k for whether having the 2nd child			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Urban income			-0.297^{*}	-0.233			-0.157	-0.250
			(0.159)	(0.163)			(0.175)	(0.167)
% of cross-city migrants	0.056		0.121^{**}		0.013		0.048	
	(0.038)		(0.050)		(0.040)		(0.055)	
% of migrants		0.045		0.097^{*}		0.038		0.095^{*}
		(0.040)		(0.054)		(0.041)		(0.055)
N	31	31	31	31	31	31	31	31

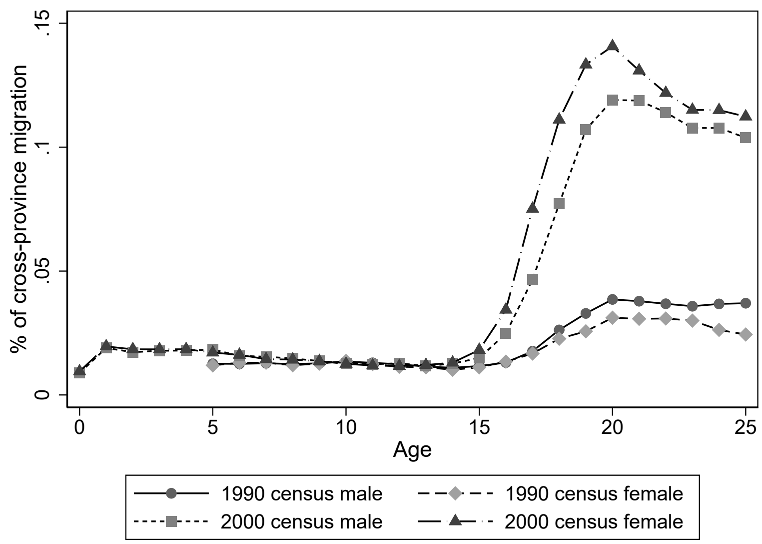
Notes This table shows the correlation between β_k and economic conditions and the proportion of migrants or cross-city migrants in the destination province. Outcome is the province-specific selection-corrected culture effects β_k obtained from the multiple index model. Urban income is measured using the urban income (in 10 thousand RMB) in 1990. The proportion of migrants or cross-city migrants is calculated using data from the 2010 Census. ***, **, and * denote statistical significance at 0.01, 0.05, and 0.1 levels, respectively.

Table A.3: Summary statistics by gender of the first child

	First child is girl	First child is boy	Difference
Edu: Primary	0.120	0.125	-0.005^{***}
Edu: Junior secondary	0.570	0.576	-0.006^{**}
Edu: Senior secondary	0.216	0.211	0.005^{***}
Edu: Tertiary	0.081	0.074	0.007^{***}
Gender (1 = female)	0.491	0.493	-0.002
Ethnicity (1 = ethnic minorities)	0.070	0.070	0.000
Log of monthly household income	8.636	8.633	0.003
Age at survey year	34.336	34.601	-0.265^{***}
Having the 2nd child (1 = yes)	0.504	0.320	0.184^{***}
Gender of the 2nd child (1 = boy)	0.657	0.488	0.169^{***}
Observations	143,087	164,744	

Notes This table reports mean values of migrant parents' characteristics for two subsamples: rural-to-urban migrants whose first child is a boy and those whose first child is a girl. The last column presents the difference in means between the two groups. Two-sample t-tests are conducted to assess equality of means. ***, **, and * denote statistical significance at 0.01, 0.05, and 0.1 levels.

Figure A.1: Cross-province migration rate and age



Notes Individuals above 5 years old were asked where they lived five years ago in the 1990 Census, and in the 2000 Census people were asked whether their birthplace and current residence were in the same province. Based on this information, this figure depicts the distribution of cross-province migration rates against age for females and males separately. Data source: 1990 and 2000 Census 1% microdata.

Table A.4: Proportions of rural male newborns by province: 1986, 1990, 1995, 2000, 2005, and 2010

Province	1986	1990	1995	2000	2005	2010
Beijing	0.526	0.515	0.521	0.528	0.524	0.526
Tianjin	0.511	0.551	0.532	0.526	0.541	0.536
Hebei	0.510	0.534	0.527	0.543	0.537	0.535
Shanxi	0.514	0.519	0.522	0.550	0.523	0.526
Inner Mongolia	0.504	0.537	0.537	0.502	0.525	0.536
Liaoning	0.511	0.525	0.521	0.542	0.528	0.527
Jilin	0.526	0.539	0.514	0.517	0.527	0.529
Heilongjiang	0.504	0.498	0.525	0.516	0.521	0.532
Shanghai	0.468	0.509	0.526	0.545	0.544	0.532
Jiangsu	0.518	0.554	0.559	0.552	0.548	0.541
Zhejiang	0.525	0.567	0.532	0.545	0.525	0.545
Anhui	0.519	0.524	0.552	0.579	0.556	0.567
Fujian	0.530	0.525	0.552	0.531	0.538	0.562
Jiangxi	0.525	0.528	0.554	0.571	0.568	0.558
Shandong	0.521	0.545	0.541	0.536	0.539	0.547
Henan	0.532	0.540	0.562	0.573	0.566	0.541
Hubei	0.524	0.543	0.564	0.574	0.557	0.558
Hunan	0.514	0.528	0.532	0.560	0.552	0.554
Guangdong	0.520	0.539	0.542	0.575	0.551	0.547
Guangxi	0.542	0.563	0.556	0.557	0.537	0.552
Hainan	0.528	0.497	0.557	0.557	0.551	0.552
Chongqing	0.525	0.526	0.526	0.535	0.539	0.534
Sichuan	0.525	0.526	0.537	0.535	0.531	0.529
Guizhou	0.513	0.499	0.528	0.516	0.539	0.553
Yunnan	0.522	0.518	0.528	0.532	0.532	0.530
Tibet	0.508	0.476	0.518	0.517	0.509	0.514
Shaanxi	0.528	0.532	0.571	0.549	0.545	0.537
Gansu	0.523	0.538	0.539	0.521	0.533	0.540
Qinghai	0.494	0.492	0.501	0.528	0.517	0.531
Ningxia	0.549	0.494	0.515	0.512	0.522	0.535
Xinjiang	0.504	0.493	0.503	0.516	0.514	0.514

Notes This table shows the rural proportion of male newborns in each province in 1986, 1990, 1995, 2000, 2005, and 2010. Data source: 1990 and 2000 Census %1 microdata, 2010 Census aggregated data.

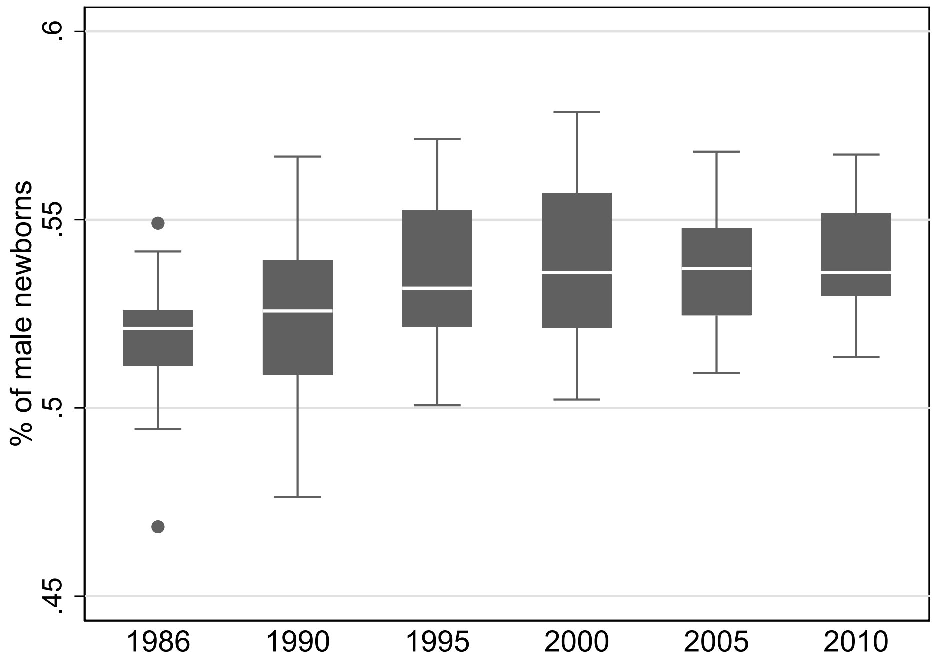
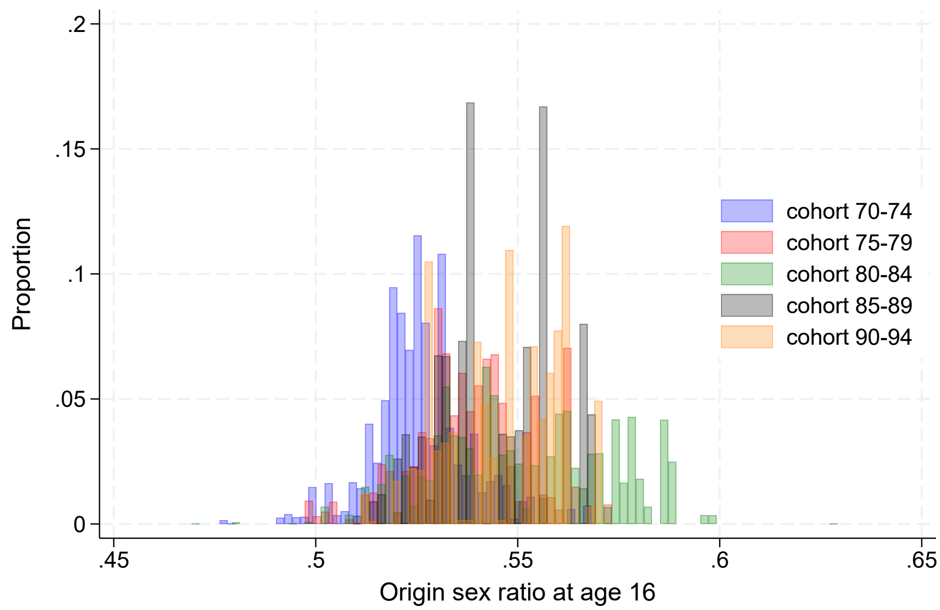
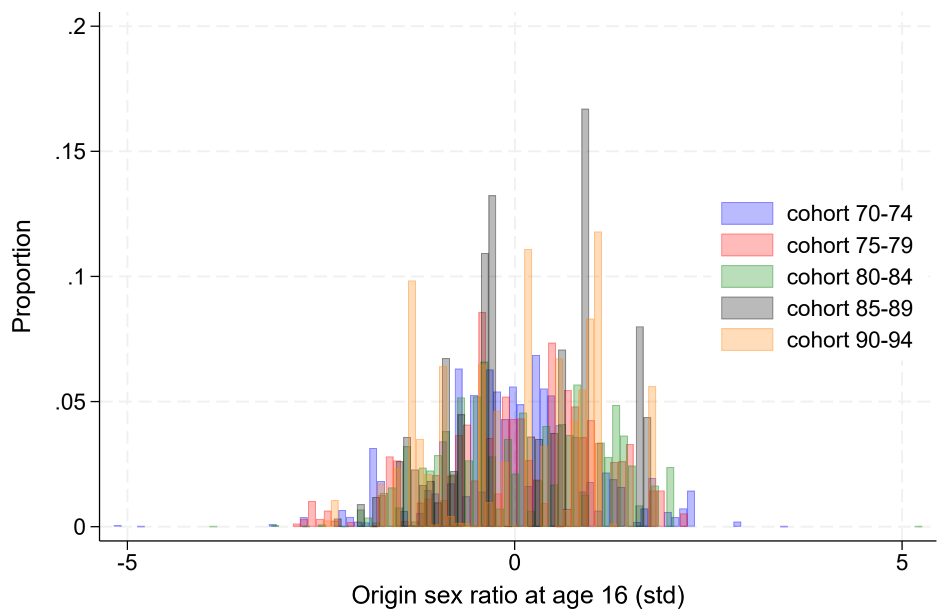


Figure A.2: Distribution of sex ratio across provinces, 1986-2010
Noyes This figure shows box plots of sex ratios across 31 provinces for 6 years. The box represents the interquartile range (25th-75th percentiles), the horizontal line indicates the median, and the whiskers extend to 1.5 times the interquartile range. Outliers are plotted as individual points. The sex ratio is defined as the proportion of males among newborns.

Figure A.3: Histograms of rural percentage of male newborns at 16 years old in origin, by cohort



(a) Distribution of rural percentage of male newborns at 16 years old in origin across 5 cohort groups, distinguished by different colours. Data source: CMDS 2013-2017 sample of rural-to-urban migrants born between 1970 and 1994, restricted to those whose first child is a girl.



(b) Distribution of standardized rural percentage of male newborns at 16 years old in origin across 5 cohort groups, distinguished by different colours. Data source: CMDS 2013-2017 sample of rural-to-urban migrants born between 1970 and 1994, restricted to those whose first child is a girl.

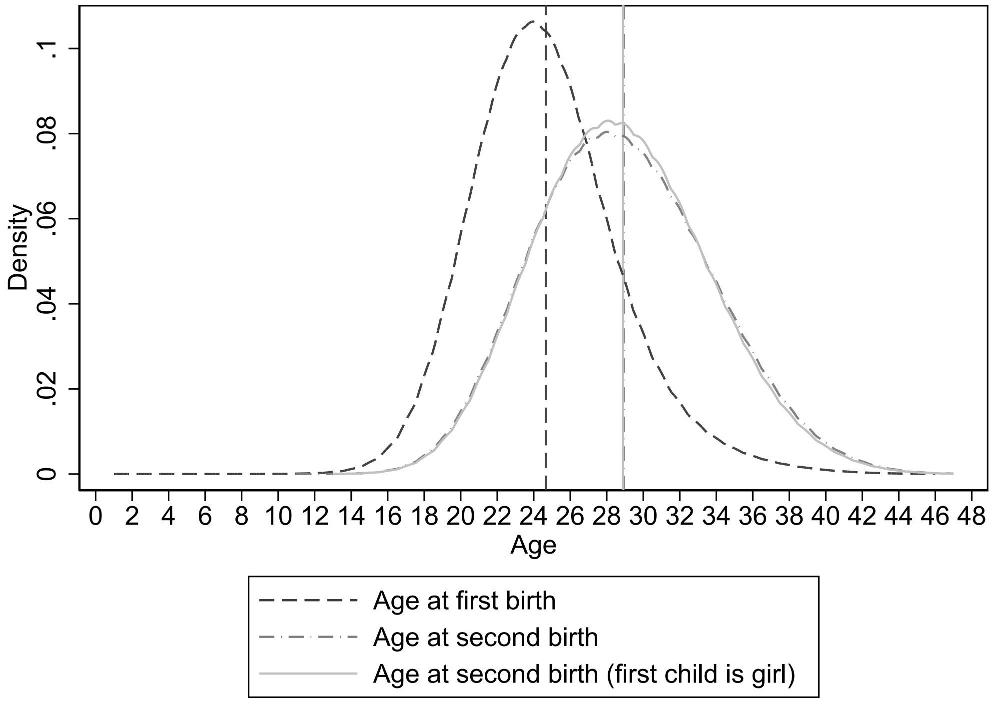
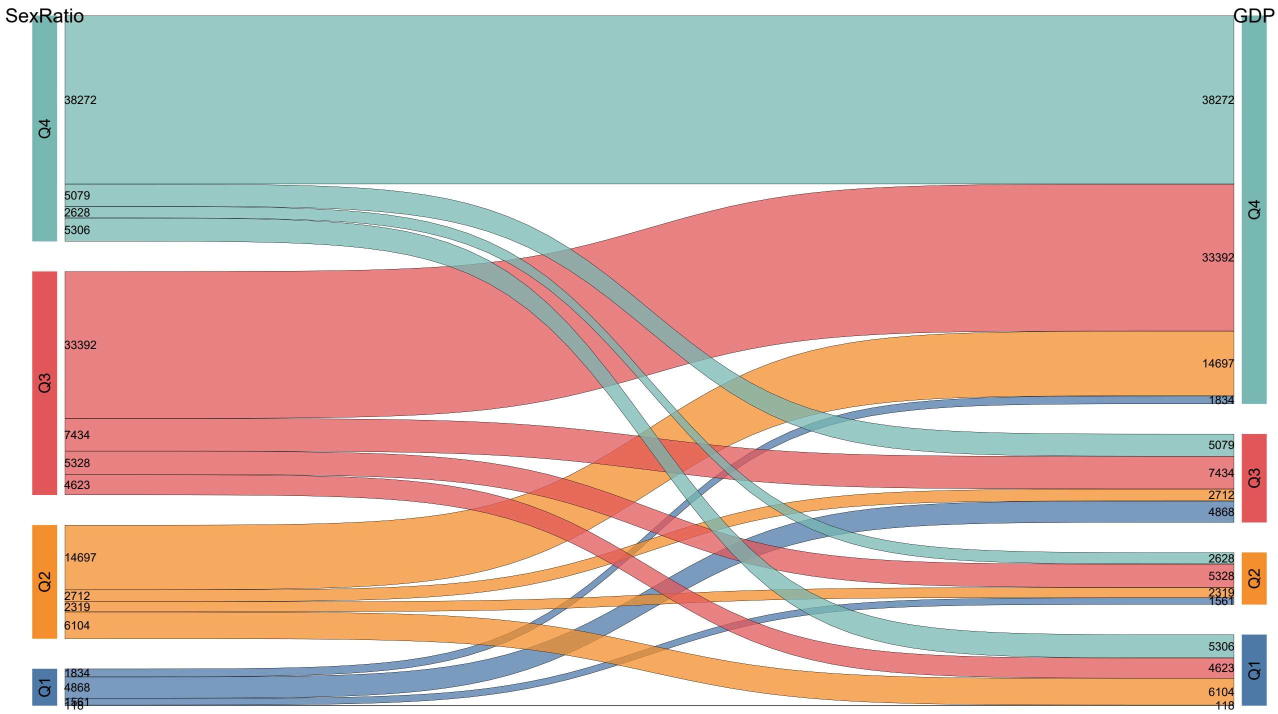


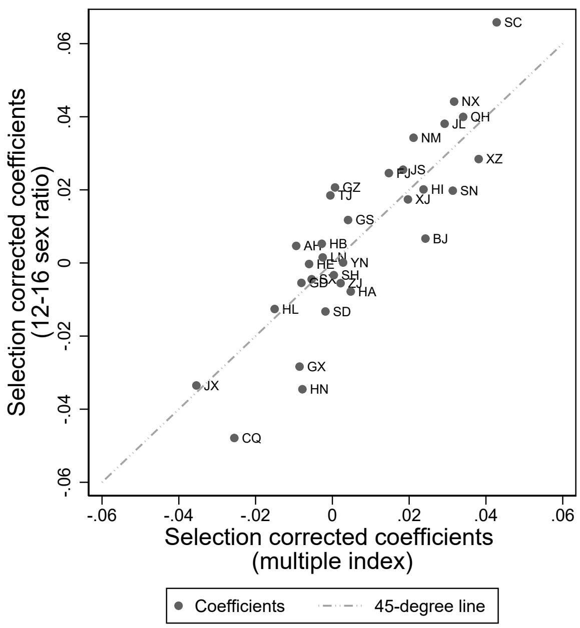
Figure A.4: Distribution of timing of births

Notes This figure plots the age distribution of age at birth of the first child, age at birth of the second child, and age at birth of the second child when the first child is a girl. Vertical lines are means. The sample includes rural-to-urban migrants born between 1970 and 1994 who had at least one child. Data source: CMDS 2013-2017.

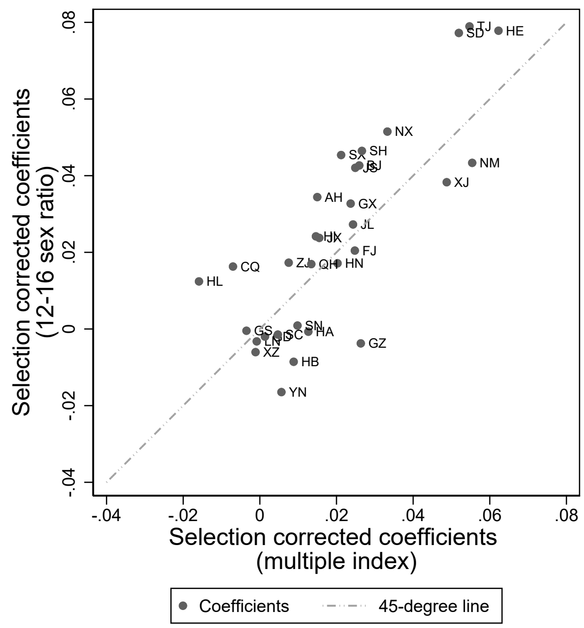
Figure A.5: Sankey Diagram for migration patterns



Notes On the left side of the Sankey diagram, all origin provinces are divided into four quartile groups based on average rural sex ratios during 1986-2010, while on the right side, all destination provinces are divided into four quartile groups based on average GDP per capita during 1986-2010. This diagram illustrates the migration flows among these origin and destination groups. Numbers represent the volume of migrants choosing each origin-destination combination. Data source: CMDS 2013-2017 sample of rural-to-urban migrants born between 1970 and 1994, restricted to those whose first child is a girl.

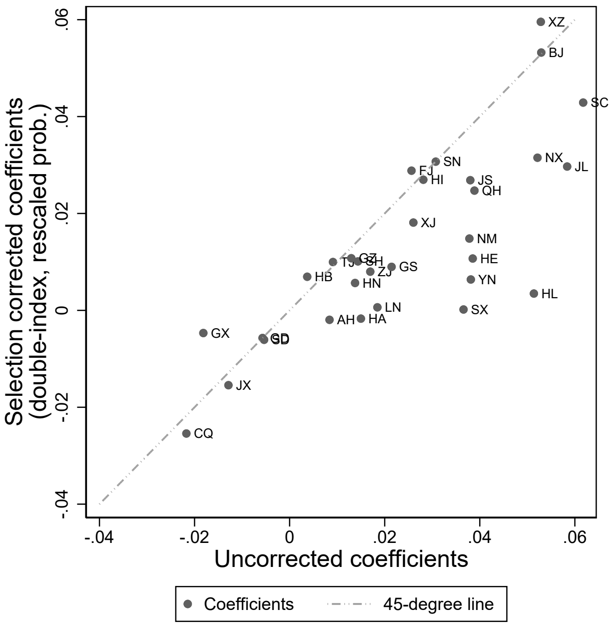


(a) Gender of the second child (boy=1)

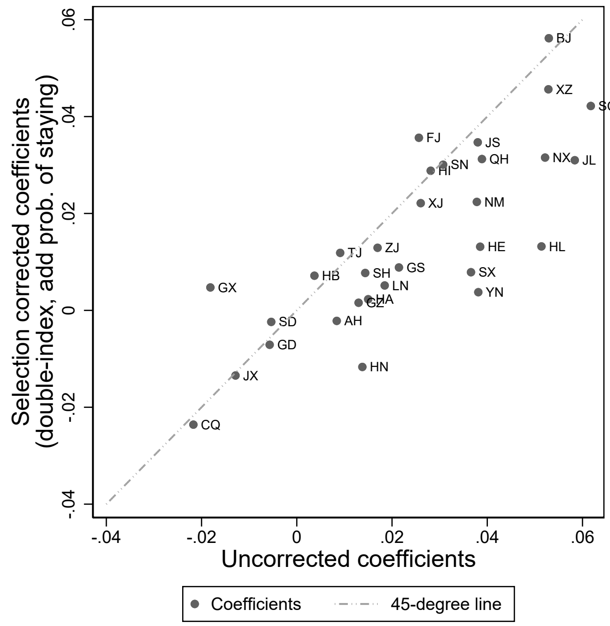


(b) Whether having a second child (yes=1)

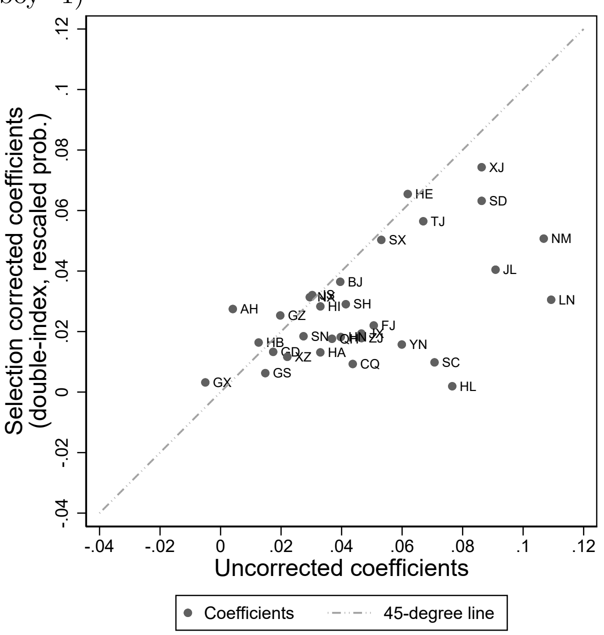
Figure A.6: Robustness check: Using sex ratios at origin between 12–16 years old as son preference
Notes : This figure plots selection corrected coefficients using origin sex ratio during 12–16 years old as the measure of son preference culture against baseline selection correction coefficients. Outcomes are the gender of the second child (boy=1) and whether having the second child (yes=1). For the corresponding full names of each province, refer to Table A.1 in the Appendix.



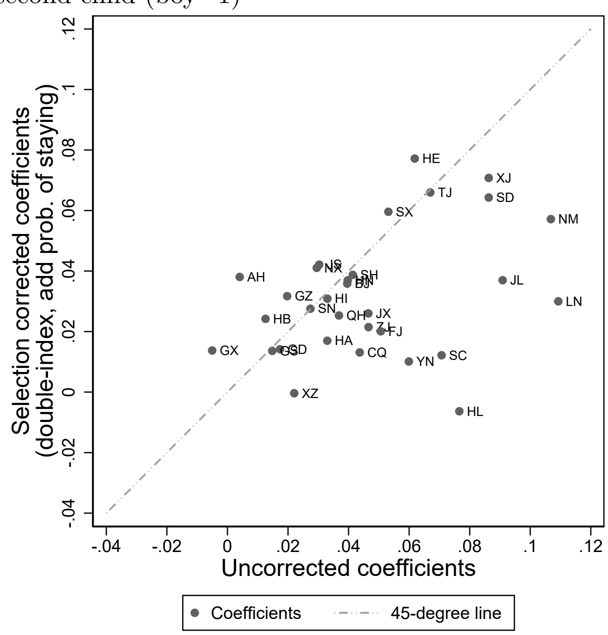
(a) Rescaling probabilities, gender of the second child (boy=1)



(b) Adding probability of non-moving, gender of the second child (boy=1)



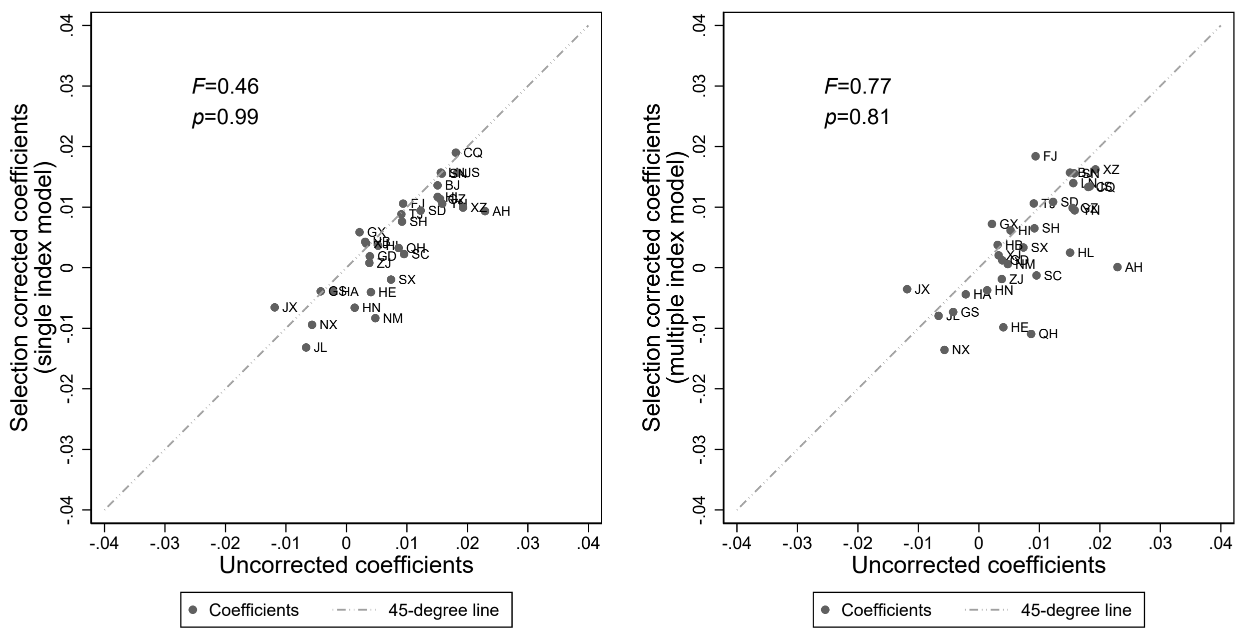
(c) Rescaling probabilities, whether having the second child (yes=1)



(d) Adding probability of non-moving, whether having the second child (yes=1)

Figure A.7: Robustness check: Considering the probability of non-moving

Notes : This figure shows the relationship between selection correction results considering probability of non-moving and reduced-form uncorrected results. In Figure A.7a and A.7c, probabilities of inter-province and intra-province migration are rescaled. In Figure A.7b and A.7d, probabilities of inter-province migration are rescaled and probabilities of intra-province migration are replaced using the probability of staying in rural areas of the origin. For the corresponding full names of each province, refer to Table A.1 in the Appendix.



(a) Single index model, gender of the first child (boy=1)

(b) Multiple index model, gender of the first child (boy=1)

Figure A.8: Placebo test: Corrected vs. uncorrected culture effects on gender of the first child

Notes : This figure plots the single index selection corrected results (Figure A.8a) and multiple index selection corrected results (Figure A.8b) against reduced form origin effects. Outcomes are gender of the first child (boy=1). F-statistic and p-value are from a Wald test of the hypothesis that all selection corrected coefficients are equal to the uncorrected ones. For the corresponding full names of each province, refer to Table A.1 in the Appendix.