Young Women in Cities

Yumi Koh* Jing Li† Yifan Wu‡ Junjian Yi§ Hanzhe Zhang¶

July 29, 2023

Abstract

Young women outnumber young men in cities in many countries during periods of economic growth and urbanization. This gender imbalance among young urbanites is more pronounced in larger cities. We use the gradual rollout of special economic zones across China as a quasi-experiment to establish the causes of this gender imbalance. Our analysis suggests that a key contributor is gender-differential incentives to migrate due to rural women's higher likelihood of marrying and marrying up in cities when urbanization creates more economic opportunities and an abundance of high-income marriage-age men.

Keywords: Urbanization, migration, gender imbalance, labor market, marriage market

JEL classifications: O15, J12

*School of Economics, University of Seoul
†School of Economics, Singapore Management University
‡Strategic Development Department, Shanghai Trust
§China Center of Economic Research, National School of Development, Peking University
¶Department of Economics, Michigan State University

1 Introduction

Young women outnumber young men in cities in many countries, such as all Central and South American countries (Tacoli 2012); Germany and Russia (Wiest et al. 2013); India (PTI 2017); Scandinavian countries (Pettay et al. 2021); Vietnam (Nguyen 2022); and China. This gender imbalance among young urbanites is largely driven by the fact that young women in rural areas are more likely than their male counterparts to migrate to urban areas. For example, Figure 1 depicts the net female share (female minus male share) by age for migrants and locals in Chinese cities in 2000: It is positive for migrants between ages 16 and 25, negative for older migrants, and close to zero for locals.

Furthermore, the gender imbalance is more pronounced among young migrants to larger cities (Figures 2a and 2b). Consequently, the gender imbalance among youths is bigger in larger cities (Figure 2c). The increasing gender imbalance in larger cities remains after taking into account heterogeneities in industrial composition across cities by controlling for industry fixed effects (Figure 2d).

Why are there gender-differential migration incentives that result in the observed gender imbalance in cities? Though rural-to-urban migration is extensively studied, discussion of gender differences in migration incentives is limited.[1] Exploiting the staggered rollout of special economic zones (hereafter SEZs) in China from 1996 to 2000 as a quasi-experiment, we examine a set of potential gender-specific factors. We find that the gender difference in marital incentives but not other plausible explanations—such as education, amenities, or changes in industrial composition in the labor market—plays the most important role.

Conceptually, massive rural-to-urban migration during urbanization creates a large pool of unmarried individuals. In particular, when high-skilled single males are drawn to urban areas

[1] The conventional rationale for rural-to-urban migration includes (i) structural changes in industrial composition—i.e., the technological advancements that release labor from agriculture to manufacturing (Henderson 2003) and (ii) the agglomeration economies of a large labor market—i.e., the concentration of workers causes positive spillovers that improve labor productivity and result in a large urban-rural wage gap (Duranton and Puga 2004; Combes et al. 2012; Combes and Gobillon 2015). However, these classic considerations do not focus on gender differences.

due to skill-biased technological progress, this creates a ripple effect of an increased likelihood of young women who seek to marry someone of higher socioeconomic status in urban areas.[2] Consequently, the interaction between labor and marriage markets amplifies the perceived advantages of living in urban areas and offers stronger incentives for young women to migrate to urban areas. To our knowledge, this is the first paper to highlight how the interplay between labor and marriage markets differentially impacts migration incentives for young men and women during the urbanization process.

It is challenging to empirically identify the causal link between urbanization and gender imbalance among young individuals. The birth and growth of a city depend on a comprehensive set of location-specific factors, such as culture, geography, transportation, natural resources, climate, and history (Mumford 1961), which are often difficult to measure. It is possible that these factors correlate with gender-differential preferences for living in large urban areas, which could give rise to an endogeneity problem.

In this paper, the empirical challenge is addressed by our exploitation of the gradual rollout of SEZs across China in the early development stage as a quasi-experiment. The SEZs create location-time variations in the extent of urbanization, which affect relative economic attractiveness across locations and individuals' migration incentives. We track the migration timing of individuals in a 1% sample of the Chinese Population Census 2000 to estimate the impact of SEZ-induced economic shocks on the size of various subpopulations between 1995 and 2000 in both a two-way fixed effects (hereafter TWFE) and a staggered difference-in-differences (hereafter DiD) setting. The identification assumption is that the timing of SEZs established during this time period is quasi-random. Both approaches reveal an economically and statistically significantly positive treatment effect of SEZs on the size of the population and the inflow of young migrants: The opening of an SEZ in a county increases the inflow of young female migrants by 38% to 43% and that of young male migrants by 31% to 35%.

To further account for cross-individual migration incentives, we construct a Bartik-like com-

[2]Hypergamy, the tendency for women to seek or marry men of higher socioeconomic status, has been observed across various cultures for centuries (Becker 1991; Weiss et al. 2018).

posite explanatory variable by converting destination-specific SEZ shocks into an origin-specific push factor based on the initial migration network.[3] We then examine its impact on the individual-specific migration responses and gender differences of these effects in a first-difference (hereafter FD) setup. In this setup, another empirical challenge arises from the possibly endogenous Bartik weight (Adao et al. 2019; Goldsmith-Pinkham et al. 2020; Borusyak et al. 2022). The concern is that even if the time variation in SEZs can be considered a quasi-random assignment, the exposure shares constructed from the migration network may still be endogenous and consequently result in biased estimates.

We overcome this empirical challenge by implementing the approach proposed by Borusyak and Hull (2023). Specifically, we separately control for an *expected* push factor by averaging 5,000 simulated push factors constructed by randomly drawing a subset of SEZ shocks out of SEZs established between 1996 and 2006. Our identification assumption rests on the quasi-random timing of SEZs established between 1996 and 2006 and conditional independence between the composite push factor and the remaining unobserved residual. We find that a one standard deviation increase in the constructed push factor leads to a 0.26 percentage-point increase in the probability of emigrating for males and a 0.46 percentage-point increase for females; considering that about 10% of the population migrates, these are significant effects and create a large gender difference.

We examine various explanations that could account for the gender imbalance among young migrants. Specifically, excluding migrants who move for education purposes does not affect the baseline finding, which rules out a gender-varying education-motivated explanation. In addition, the baseline result is not driven by educated females who are found to enjoy amenities more (Diamond 2016), which suggests that the baseline finding is not channelled through amenities-related considerations.[4] Moreover, controlling for changes in industrial composition slightly

[3]The term "push factor" is defined in Section 3.2. It captures an external urbanization push force that incentivizes individuals to emigrate out of their county of residence and is termed to be in line with the literature, such as Burchardi et al. (2019).

[4]This finding is also in line with Figure C.1 in the Online Appendix, which shows that the gender disparity in the tendency to emigrate among the young cohort is mainly driven by the less educated.

reduces the baseline finding, but does not completely eliminate it. In contrast, the pattern disappears when the analysis focuses on the middle-aged cohort, who are more likely to already be married. We also find that females are more likely to marry and more likely to marry up than males in response to a stronger push factor. This is consistent with a marriage-market explanation in which young females' migration decisions are largely driven by higher expected returns from the marriage market in larger urban areas.

Our study carries weight for both scholars and policymakers since urbanization and feminization are closely linked to numerous social issues. The disparities that accompany rapid urbanization are notable between urban and rural areas and between males and females. From the perspective of the marriage market, the presence of relatively more young females in urban areas benefits males and disadvantages females in cities; the reverse is true in rural areas. This widening spatial inequality and gender divide may have far-reaching social implications on marriage and birth outcomes. Spatial mismatch in gender may further cause a decline in overall social and family stability.[5] Second, we highlight the interplay of the labor market and the marriage market and find that a higher proportion of young females in urban areas and consequent marriage-market outcomes can have a reciprocal effect on the labor market. More specifically, the presence of more women in urban areas can potentially affect the types of jobs available, competition for job search, and the gender wage gap in cities.

This study contributes to three strands of the literature. First, it adds to the literature on gender differences in urbanization. Starting from Marshall (1890), researchers have documented cities' advantages in higher productivity, higher wages, and better amenities, which provide incentives to migrate and settle (Rosenthal and Strange 2004; Combes and Gobillon 2015; Diamond 2016; Couture and Handbury 2017). A few studies highlight gender differences in responding to such urban advantages, which result in differences between urban and rural areas in terms of gender gaps in labor participation, wages, and entrepreneurship (Phimister 2005; Rosenthal and Strange 2012; Bacolod 2017). In this paper, we conclude that the gender difference in the

[5]This implication is also consistent with the findings of Meng and Zhao (2019) on bride drain in the rural marriage market of China.

migration incentives of rural youths is also likely driven by different returns from the marriage market in cities—a perspective that, to our knowledge, has not been examined in the literature.

Second, this paper also relates to studies on the role of the marriage market in urbanization. Fan and Zou (2021) develop a spatial equilibrium model that captures the interactions between the labor market and the marriage market and show that the declining preference for marriage and narrowing gender wage gap contribute to the spatial dispersion of population. Costa and Kahn (2000) document that, conditional on being married, "power couples"—couples in which both husband and wife have college degrees—are more likely to reside in larger cities, which suggests the presence of an urban skill premium channelled through positive assortative matching in the marriage market. Different from previous studies, we treat marriage as an endogenous choice in which the returns to marriage differ across gender and location.

Third, this paper contributes to the literature on gender differences in premarital investment and consequent marriage and labor market outcomes. Previous studies mainly focus on premarital investment in the form of wealth, education, or their interactions (Peters and Siow 2002; Iyigun and Walsh 2007; Chiappori et al. 2009; Zhang 2021; Bhaskar et al. 2023; Zhang and Zou 2023). Similar to Dupuy (2021), our paper considers migration decisions as premarital investments. In contrast to studies that mostly suggest the theoretical importance of premarital investments, we employ empirical methods to causally identify their importance.

The rest of the paper is organized as follows. Section 2 introduces the institutional background. Section 3 describes the data. Section 4 lays out the empirical design. Section 5 reports the empirical results, and Section 6 concludes.

2 Institutional Background

2.1 SEZs in China

China's SEZs were first established in the late 1970s as part of China's economic reform and opening-up policy (Shirk et al. 1993). The first SEZ was established in Shenzhen in 1979, and

was followed by SEZs in Zhuhai, Shantou, and Xiamen in 1980. These four cities were chosen because of their proximity to Hong Kong and Taiwan, and were intended to serve as pilot projects for China's economic reforms (Xu 2011).

SEZs were designed to attract foreign investment and promote exports. These zones were equipped with special economic policies and incentives that aimed to facilitate economic growth (Wang 2013; Alder et al. 2016; Lu et al. 2019).[6] The economic performance of the four initial SEZs was remarkable. For example, between 1980 and 1990, Shenzhen's gross domestic product grew at an average rate of approximately 28% per year (National Bureau of Statistics of China 2021).

The success of the four initial SEZs led to their proliferation in other regions. In the 1990s, the central government embraced SEZ development as a national strategy, with the intention of achieving geographic diversity. Figure 3 depicts the number of SEZs established per year from 1984 to 2000 in the top panel and the cumulative area of SEZs established across years in the bottom panel. The number and area of SEZs increased significantly in the 1990s. Figure 4 illustrates the temporal geographic expansion of SEZs with four panels representing different years (1990, 1995, 2000, and 2005). Between 1995 and 2005, SEZs were established across China in a dispersed manner. Our identification strategy relies on the quasi-random variation in the timing of SEZ establishment during this period, conditional on their selection, as described in Section 4.[7]

SEZs played a significant role in promoting urbanization in China by attracting a large influx of foreign investment and creating job opportunities in both the manufacturing and service

[6]Such incentives included preferential tax policies that lower or exempt corporate taxes; simplified customs procedures to facilitate trade; reduced bureaucratic procedures that ease business operations; preferential land use, such as lower land-use fees and priority access to land; access to credit and other financial resources; and openness to foreign investment by allowing foreign investors to own their entire enterprises without the need for a Chinese partner.

[7]SEZs were primarily located in China's eastern coastal regions during the initial stages of economic reform. The establishment and success of these SEZs have contributed to the prosperity of the coastal regions but also exacerbated economic disparities between different regions. Consequently, efforts were made in the 1990s to establish SEZs in a more balanced manner across the country, aiming to reduce regional disparities. During this stage, the timing of their establishment varied primarily based on the bureaucratic processes involved in approving the SEZs (Crane et al. 2018).

sectors; this, in turn, attracted a large number of migrants from other parts of the country. As the SEZs became economically more competent, they began to implement policies to attract and retain workers, such as offering higher wages, better benefits, and improved working conditions (Xu 2011; Wang 2013; Alder et al. 2016; Lu et al. 2019).

2.2 The Hukou System and Migration in China

Before the economic reform in 1979, migration within China was rare under the *hukou* system, which has been in place since the 1950s (see Young (2013) for a review of China's hukou system). The system is based on household registration, which assigns every Chinese citizen a place of origin that is recorded on their hukou or household registration document. The hukou system divides the population into two categories: rural and urban. A rural (resp., an urban) hukou is assigned to individuals who were born and raised in the countryside or smaller towns (resp., in cities). The hukou system serves various purposes, such as determining access to public services, such as education and healthcare, and tracking population movement. The system restricts access to public services and job opportunities based on an individual's place of origin, which makes it difficult for people to move from rural to urban areas and receive the same level of services as they would in their place of origin.

China's economic reform in the late 1970s brought significant changes to the hukou system. With the establishment of SEZs and the transition to a market-oriented economy, there was growing demand for labor in urban areas and many rural residents began to migrate to cities in search of work. However, the hukou system remained a significant barrier to migration and mobility, since rural residents were not allowed to obtain an urban hukou; this restricted their access to social welfare benefits and public services. To address this issue, the government began to implement reforms to the hukou system in the 1980s. One of the key changes was the introduction of a temporary residency system, which allowed rural residents to obtain temporary urban residency permits. These permits granted them access to certain social welfare benefits and public services in urban areas, although they were not equivalent to an urban hukou.

In the 1990s, the government introduced further reforms to the hukou system to promote greater social inclusion and mobility. One of the significant changes was the introduction of the "floating population" concept, which recognized the existence of migrant workers and granted them certain legal rights and protections. In recent years, there have been further reforms to the hukou system. These include the expansion of social welfare benefits and public services to nonlocal migrants and the promotion of greater labor mobility.

With economic growth and relaxation of the hukou system, China witnessed an unprecedented wave of internal migration in the 1990s and 2000s. The urban population share increased from 19.39% in 1980 to 26.22% in 2000 (National Bureau of Statistics of China 2021). In particular, based on our calculations using the Chinese Population Census 2000, the total number of young cross-county migrants grew about 13-fold from 1995 to 2000 (Figure 5). According to the population census conducted by the National Bureau of Statistics in 2020, the number of migrants was estimated to be 376 million (Liu 2021).

3 Data and Variables

3.1 Data Overview

Our empirical analysis relies on two primary datasets. The first is a 1% sample of the Chinese Population Census 2000. From the sample, we use information on various demographic and migration-related variables from each surveyed individual: gender, year of birth, education level, marital status, migration status, migration year, county of residence, as well as the county of origin and destination for those who migrated.

We focus on the 2000 census sample for three reasons. First, it contains one of the largest samples—1% of the population—compared with census samples available for analysis in later years. Second, the 2000 census sample provides information on both the origin and the destination of a migrant, which is an advantage over the census data collected in earlier years. Third,

the 2000 census allows us to accurately identify the migration year up to 1995.[8] For this reason, our focus is on the changes that took place between 1995 and 2000 or between 1996 and 2000. The specific time frame depends on whether we need to exclude the initial year, 1995, in order to construct the composite push factor variable based on the initial migration networks. Fourth, during the sample period, hukou restrictions on migration were significantly relaxed compared with earlier years, which allows us to observe a large sample of migrants. Last, the country has not yet been affected by its WTO accession in 2001.

The second dataset is the list of SEZs published by the National Development and Reform Commission of China.[9] The list provides information on the SEZ ID, name, approval date, approval authority, and major industries recommended for each SEZ. To identify counties that were treated with SEZs, we track the geographic boundaries of each SEZ established during our sample period and match them with the corresponding counties.

We construct our data at two levels. First, at county-year level, we track the changes in population size and SEZ treatment status for each county across each year from 1995 to 2000. Throughout this paper, we define individuals as migrants if they moved across counties. We focus on cross-county migrants because we track SEZ shocks at county level. We then track the year in which the migration took place and summarize the year-specific aggregate population size after accounting for migration inflows and outflows. In panel (a) of Table 1, we present the summary statistics on demographics in 2000 at county level. While there were more male migrants overall, there were more female migrants in the age group of 16 to 25. Specifically, we show that on average, 54% of young migrants in a county were females. Among all counties in our sample, 16% have at least one SEZ as of 2000.

Second, we form a panel dataset of individuals who were between the ages of 16 and 25 at any point during the period 1996 to 2000 to track changes in their migration status and marriage

[8]The 2000 census asked respondents if they had always lived in their birth town. If not, they were asked to provide the destination county and the year of their move. Due to the questionnaire design, all years before 1995 were grouped as "moved before 1995," so we can only determine the exact moving year up to 1995.

[9]The list was first published in 2006 (NDRC 2006) and later updated in 2018 (NDRC 2018). To ensure that we use the full information, we combine both lists and track SEZ establishments during our sample period.

outcomes over time at individual level. In panel (b) of Table 1, we present individual-level summary statistics. On average, 49% of the sample were females. The average age is 22.73, and 72% had an education level of middle school or below. 10% migrated across counties between 1996 and 2000, and 25% of them got married during the same time period. $\Delta(\text{Marry up})_{i,c}$ is an indicator of marrying up and is set to one if someone married to a partner who has a higher level of education during the years 1996 to 2000 and zero otherwise. Among the married, 16% married someone of strictly higher education level.

3.2 External Urbanization Push Factor: Δ(Push Factor)

The designation of other counties as SEZs can lead to increased urbanization and increased economic opportunities outside an individual's current county of residence. Consequently, the opportunities that arise from urbanization in other areas can act as an "external urbanization push factor" ("push factor" for short) that encourages individuals to consider emigrating from their present county of residence. To quantify this push factor, we construct a composite variable, Δ(Push Factor), which captures the extent to which individuals are induced to emigrate from their county of residence.

Conceptually, this variable combines an $N_c \times N_c$ matrix $\boldsymbol{M}$ and an $N_c \times 1$ vector $\boldsymbol{V}$ in a multiplicative way as $\boldsymbol{MV}$, where N_c is the number of counties. $\boldsymbol{M}$ captures the normalized cross-county migration flows in 1995 as predetermined weights. Specifically, element $m_{ij} \in \boldsymbol{M}$ represents the proportion of people who moved from county i to county j in 1995, normalized by the total number of people who moved out of county i. All diagonal elements of matrix $\boldsymbol{M}$ are zero. Element j of vector $\boldsymbol{V}$ equals 1 if county j received its SEZ shock for the first time between 1996 and 2000, and 0 otherwise. Hence, element i of Δ(Push Factor) captures the degree of urbanization caused by SEZ shocks that occurred outside county i between 1996 and 2000 and propelled individuals in county i to emigrate according to the predetermined migration network.

However, the migration flow matrix constructed at county level suffers from sparsity due

to relatively few observed migration flows between specific pairs of counties.[10] To resolve this issue, we aggregate migration flow and SEZ shock data at province level, which allows us to capture broader patterns and trends in migration while reducing the impact of sparse data on our analysis. By taking this approach, we can obtain a more robust and reliable estimation of the effects of SEZs on migration patterns.

Specifically, we capture migration flows at province level using an $N_p \times N_p$ matrix, where N_p is the number of provinces. The $N_p \times 1$ vector of SEZ shocks consists of elements that capture the count of counties in each province treated with SEZs for the first time between 1996 and 2000. Then we construct Δ(Push Factor) at province level, which can be mapped to county-level data based on the province a county is located in.

Note that such values are applicable to counties that did not receive SEZ shocks between 1996 and 2000. Because we want Δ(Push Factor) to only reflect the *external* conditions outside a county, for counties in province p that were treated with SEZs between 1996 and 2000, we subtract 1 from the p^{th} element in the SEZ shock vector $\boldsymbol{V}$ to net out its own internal shock, while keeping other values. We recompute the values of Δ(Push Factor) accordingly for all counties that were affected by their own SEZ shocks, such that the values only reflect external conditions outside the county, which are either within or outside its own province.

In summary, higher Δ(Push Factor) indicates the presence of greater economic opportunities and stronger urbanization in other places. We thus expect that such increased attractiveness of other locations will act as a push factor that encourages residents in the current location to emigrate.

[10] In the data, 65.35% of counties either have zero migration outflows or receive zero inflows, which leaves a large share of the migration weight undetermined.

4 Empirical Design

4.1 Specification for County-level Analysis

Exploiting the heterogeneous timing of SEZ establishments, we first analyze aggregate-level data to see how the population size of various subgroups changed in response to SEZ shocks. Our specification for the TWFE model is

$$y_{c,t} = \alpha + \beta \cdot \mathbb{1}(\text{SEZ})_{c,t} + FE_c + FE_{r,t} + \epsilon_{c,t}, \tag{1}$$

where subscripts c, r, and t capture county, prefecture city, and year, respectively. For the dependent variable $y_{c,t}$, we use the log population size of various groups of interest (total population, total migrants, young female migrants, and young male migrants) for a county in a year. $\mathbb{1}(\text{SEZ})_{c,t}$ is a binary indicator that equals one if county c has ever been treated with an SEZ and zero otherwise.[11] Lastly, FE_c, $FE_{r,t}$, and $\epsilon_{c,t}$ represent county fixed effects, prefecture city-by-year fixed effects, and the idiosyncratic error, respectively. Standard errors are corrected for heteroskedasticity and are clustered at county level.

The TWFE model is commonly used in the literature to identify the average treatment effect on the treated, conditional on satisfaction of the parallel-trends assumption. However, de Chaisemartin and D'Haultfoeuille (2020) and Goodman-Bacon (2021) find that a TWFE model does not yield an interpretable causal parameter when there are cross-sectional variations in treatment timing and heterogeneous treatment effects. To complement our TWFE model findings, we also estimate a staggered DiD model, following the approach of Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021). Specifically, we group counties into various cohorts, depending on the first year an SEZ was established in a county, and use the not-yet-treated cohort as the control group. We then compute the average treatment effect, which aggregates the average treatment effect on the treated across different cohorts and years. The

[11] In the data, some counties were affected by the establishment of multiple SEZs. To capture the extensive margin of SEZ establishment, we thus focus on the first year a county was treated with an SEZ.

identification relies on the parallel trends between the treatment cohort and the not-yet-treated cohort, which is implied if the timing of SEZs established during the sample period is quasi-random.

4.2 Specification for Individual-level Analysis

Although the TWFE and staggered DiD models allow us to quantify the impact of SEZs on population size and migration flows, one limitation is that they only capture aggregate changes. Next, we estimate an FD model using individual-level data, specified as follows:

$$\Delta(Migrate)_{i,c} = \beta\Delta(\text{Push Factor})_c + \gamma\Delta(\text{Push Factor})_c \times \mathbb{1}(\text{Female})_i + \delta_i + \epsilon_{i,c}. \tag{2}$$

For individual i living in county c in 1996, the binary indicator $\Delta(Migrate)_{i,c}$ equals one if they migrated from county c to another county between 1996 and 2000. $\Delta(\text{Push Factor})_c$ captures the changes in the external urbanization push factor between 1996 and 2000, which motivates residents of county c to emigrate. The coefficient of interest is γ, which captures whether females respond differently to the push factor than males. The variable δ_i represents age fixed effects, and $\epsilon_{i,c}$ is an idiosyncratic shock. Standard errors are corrected for heteroskedasticity and clustered at county level.

The FD model estimator helps to analyze how SEZ-induced changes in urbanization intensity affect the migration decisions of individuals while accounting for cross-individual age heterogeneity. In an alternative specification, we also control for $\Delta(SEZ)_c$, which is a binary indicator that equals one if county c received its own SEZ shock for the first time between 1996 and 2000 and zero otherwise. The additional control variable reflects the internal urbanization shock generated by the county's own SEZ establishment, which acts as a "staying" force.

A bias may arise from $\Delta(\text{Push Factor})_c$ which combines the changes in SEZ shocks and exposure to such shocks using predetermined migration flows in a composite form. Even if the SEZ shocks may be as good as random, omitted variable bias may still arise if *exposure* to

exogenous shocks is not random (Adao et al. 2019; Goldsmith-Pinkham et al. 2020; Borusyak et al. 2022).

To address this concern, we follow the estimation strategy proposed by Borusyak and Hull (2023). The idea is to randomly draw shocks that may plausibly have occurred, recompute the composite index, and repeat this across many simulations to compute their averaged value, denoted the "expected $\Delta(\text{Push Factor})_c$." Then we can avoid omitted variable bias by including the expected $\Delta(\text{Push Factor})_c$ as a control variable in the regression. Given that we have additional SEZs established beyond our sample period, we include the actual SEZ establishment shocks in later years between 2001 and 2006 to form a larger pool from which the counterfactual SEZ shocks are drawn in our simulations. We elaborate on the details of the simulation algorithm in Appendix A and present a comparison of the values between $\Delta(\text{Push Factor})_c$ and expected $\Delta(\text{Push Factor})_c$ in Figure C.2 in the Appendix.

Therefore, identification of the FD model relies on the assumption that out of all SEZs established between 1996 and 2006, whether they are established before or after 2000 is as good as random. Moreover, conditional on the expected push factor, the composite Bartik-like push factor based on realized SEZs between 1996 and 2000 is not correlated with the remaining unobserved residual.

5 Results

5.1 Gender Imbalance in Migration

We first analyze the impact of SEZ shocks on population size and whether the extent of migration differed between young males and young females. In Table 2, we use the sample of counties from 1995 to 2000. In panel (a) of Table 2, we present the TWFE model estimates. Column (1) shows that the establishment of an SEZ has a positive and statistically significant impact on population size at county level. Column (2) shows that the increase in population is largely driven by the inflow of migrants. In particular, columns (3) and (4) show that young females aged 16 to

25 migrated more to the counties treated with SEZs compared with their male counterparts. Specifically, the opening of an SEZ in a county increases the inflow of young female migrants by 43.28% and that of young male migrants by 35.17%.

In panel (b), we report estimates from the staggered DiD model, in which we use the not-yet-treated cohort as the control group. Although the estimates are slightly smaller than those from the TWFE model, both estimations produce consistent findings whereby the opening of an SEZ increased both the population and inflow of migrants. Moreover, the staggered DiD model also shows that there was a larger inflow of young females to the treated area than young males: The opening of an SEZ in a county increases the inflow of young female migrants by 37.80% and that of young male migrants by 31.20%.

Figure 6 presents event study graphs, which show dynamic coefficients from the staggered DiD model. The dependent variable in panel (a) (resp., panel (b)) is the log of young female (resp., male) migrants. Almost all coefficients during the pre-treatment period are insignificantly different from zero in both panels, in line with the parallel-trends assumption. Estimated coefficients increase sharply with the establishment of an SEZ.

In Table 3, we present the baseline estimates from the FD model using a sample of all individuals who were between the ages of 16 and 25 at any point between 1996 and 2000. Column (1) shows that when the push factor becomes larger, young individuals are more likely to migrate to other counties. Moreover, such a tendency is more pronounced among young females. In column (2), we address endogeneity concerns by including the expected terms following the method of Borusyak and Hull (2023). Even after addressing such endogeneity concerns, we find that the main effects are still consistent and statistically significant. Specifically, a one standard deviation increase in the push factor leads to a $(0.0022 \times 1.17 \times 100 =)$ 0.26 percentage point increase in the tendency to emigrate for young males. The effect is $(0.0017 \times 1.17 \times 100 =)$ 0.20 percentage points larger for young females.

In columns (3) and (4), we extend our analysis to further include a control variable, $\Delta(\text{SEZ})_c$, which captures the internal SEZ shock that serves as a staying force. Indeed, we find that when

the county of residence becomes treated with an SEZ, this reduces the likelihood of migrating out. Even with this additional control variable, we document robust evidence that young females are more responsive to the external urbanization push factor to emigrate out of their original place of residence than young males: The gender difference in the effects is still 0.20 percentage points. These estimates are in line with our expectations.[12]

5.2 Potential Explanations

In this section, we explore possible explanations for the observed gender imbalance pattern among young migrants. We examine each potential explanation by conducting heterogeneity analyses on different subgroups to gain a deeper understanding of the phenomenon.

5.2.1 Education

One possible explanation for the gender imbalance in migration is related to education. That is, females may place a higher value on education, and education quality may be better in larger cities. Another possibility is that the gender gap in returns to education is greater in larger cities, with women experiencing a higher return than men. To investigate such a possibility, we re-estimate our FD model using the sample that excludes individuals who responded that they migrated for education-related purposes. Panel (a) of Table 4 shows that the estimates remain mostly unchanged compared with the baseline estimates in Table 3. Therefore, education-driven motives do not appear to be the primary cause of gender imbalance among young migrants in our study.

5.2.2 Amenities

Another possible explanation is that larger cities may offer amenities that are more attractive to young females than to young males. For example, certain types of amenities that support family life, such as access to good schools or green parks, may be valued more by young females

[12]In Table D.1, we restrict the sample to include only rural counties and obtain similar results.

and larger cities may offer more of such amenities. Based on U.S. data, Reynolds and Weinstein (2021) find that males and females largely share preferences for natural amenities (e.g. coastal location, climate), but there are important gender differences among young singles regarding their importance relative to nonnatural amenities (e.g. public transportation, safety, progressive gender-role attitudes).

To analyze the role of amenities, we estimate the FD model for the high-skilled and low-skilled samples separately and check whether the migration tendency is stronger among the high-skilled group, and particularly so for females. This is because high-skilled individuals are known to value amenities more than low-skilled individuals (Diamond 2016). We use education level to proxy for skills. Individuals with an education level of middle school or below are classified as having low skills, and those with an education level above middle school are classified as having high skills. In panels (b) and (c) of Table 4, we find that less educated females show a much stronger tendency than their male counterparts to emigrate in response to the push factor, but the opposite is true for the highly educated sample. Therefore, we do not find supporting evidence for the explanation based on amenities in our empirical context.

5.2.3 Changes in Industrial Composition

SEZs often involve the implementation of policies and incentives that are designed to attract investment and promote economic development in specific industries. Therefore, SEZ establishment can potentially change the industrial composition of the affected area and result in non-gender-neutral growth in labor demand. That is, gender imbalance among migrants may arise due to increased demand for young female workers, given the SEZ-induced industrial composition changes.[13]

In order to examine this possibility, we construct two measures, $\Delta(\text{Labor Mkt})^F$ and $\Delta(\text{Labor Mkt})^M$, to control for possible changes in gender-specific labor market conditions

[13]For instance, "the feminization of SEZ production is attributed to three broad factors in the literature: women's relative 'cheapness' owning to the gender wage gap, rising international competition, and gendered norms and stereotypes that segment work by sex and assign women to low-skill and low-paying work" (Farole and Akinci 2011, p.#251).

driven by SEZ establishment. We detail the variable construction in Appendix A2. In essence, $\Delta(\text{Labor Mkt})^F$ and $\Delta(\text{Labor Mkt})^M$ reflect the changes in labor market conditions that attract female and male workers to migrate, respectively. The idea follows and extends the standard structure of a Bartik instrument (Bartik 1991; Blanchard et al. 1992). We first capture the SEZ-induced changes in industrial composition reflected in the national aggregate data. We then project these aggregate changes onto each county based on its exposure, measured as the industry-specific employment share within each county. We further consider each industry's dependence on either female or male labor, captured by the industry-specific share of workers of each gender. These three factors contribute to a vector that summarizes the changes in labor market conditions for female or male workers in each destination. We then convert the force induced by these changes to a labor market-specific push factor for each county of origin based on the origin-destination migration flow matrix.

To account for such gender-specific demand changes in the labor market induced by SEZs, we include $\Delta(\text{Labor Mkt})^F$ and $\Delta(\text{Labor Mkt})^M$ as controls and report the estimates in panel (d) of Table 4. After controlling for gender-specific labor market conditions, we find that the estimated impact of the push factor on emigration incentives is smaller in magnitude for both males and females, which is consistent with the change in industrial composition playing a role. However, the evidence suggests that even when this factor is considered, young females are still more responsive to the push factor than their male counterparts. Therefore, changes in industrial composition cannot fully explain the gender imbalance observed in our baseline results.

5.2.4 The Marriage Market

When SEZs create jobs and stimulate labor market growth, population in that location increases because workers are attracted to the area in search of employment opportunities. Such an influx of people creates a larger pool of unmarried individuals, which can affect the dynamics of the local marriage market. In particular, when high-skilled males are drawn to large urban areas due to skill-biased technological progress and a male-dominated skill pool, this leads to a ripple effect

on young females' migration choices. Essentially, the prospect of marrying up also increases in large urban areas. Hypergamy—the tendency whereby women tend to court or marry men of a higher socioeconomic status than their own—has existed across the world for centuries (Becker 1991; Weiss et al. 2018). Therefore, this interplay between labor and marriage markets increases the expected returns from living in urban areas and strengthens the incentives for young females, especially the low-skilled, to migrate to large urban areas. In Appendix E, we present an equilibrium model that is consistent with the observed pattern whereby young females are more likely to migrate and marry up.

To explore this channel, we re-estimate our FD model using individuals aged between 30 and 39 in 2000 and present the results in panel (e) of Table 4. The idea is that if individuals are driven by marriage incentives to emigrate to cities, the gender imbalance in the tendency to migrate, as documented in our baseline findings, should not apply to the subgroup of individuals who are likely to be married. We find that for individuals belonging to this age group, females indeed show a weaker migration tendency than their male counterparts; this clearly contrasts with our baseline results for young individuals.

Furthermore, we examine the influence of external urbanization as a push factor on marriage outcomes and present our findings in Table 5. In panel (a), we observe a decrease in the probability of marriage for men but an increase for women due to the push factor, which is also in line with model predictions. In panel (b), we present the estimated impact of the push factor on the probability of marrying up. We find that the push factor reduces the likelihood of males marrying up but increases that of females.[14] These findings support the notion that young females are more likely driven by marriage market incentives to migrate to large cities in search of improved marriage prospects.

[14]In Table D.2 in the Appendix, we check the robustness of this finding using a sample of older individuals who were between the ages of 20 and 35 at any time from 1996 to 2000 and obtain consistent findings.

6 Conclusion

In this paper, we analyze the extent of gender imbalance in migration among young individuals and explore underlying explanations that can rationalize this phenomenon. Using the gradual rollout of SEZs across China between 1996 and 2000, we find that more young women than young men migrated from rural areas to urban areas. Such gender-differential migration patterns are shown to be largely driven by marital incentives. Our study suggests that such a gender imbalance in migration patterns can further trigger growing disparities in the marriage market between urban and rural areas and between males and females. Such widening inequality and gender divide in the marriage market may have far-reaching implications for social and family stability.

References

Adao, R., Kolesár, M., and Morales, E. (2019). Shift-share designs: Theory and inference. *The Quarterly Journal of Economics*, 134(4):1949–2010.

Alder, S., Shao, L., and Zilibotti, F. (2016). Economic reforms and industrial policy in a panel of Chinese cities. *Journal of Economic Growth*, 21(4):305–349.

Bacolod, M. (2017). Skills, the gender wage gap, and cities. *Journal of Regional Science*, 57(2):290–318.

Bartik, T. J. (1991). *Who benefits from state and local economic development policies?* WE Upjohn Institute for Employment Research.

Becker, G. S. (1973). A theory of marriage: Part I. *Journal of Political Economy*, 81(4):813–846.

Becker, G. S. (1991). *A Treatise on the Family: Enlarged Edition*. Harvard university press.

Bhaskar, V., Li, W., and Yi, J. (2023). Multidimensional pre-marital investments with imperfect commitment. *Journal of Political Economy*.

Blanchard, O. J., Katz, L. F., Hall, R. E., and Eichengreen, B. (1992). Regional evolutions. *Brookings Papers on Economic Activity*, 1992(1):1–75.

Borusyak, K. and Hull, P. (2023). Non-random exposure to exogenous shocks: Theory and applications. *Econometrica*, Forthcoming.

Borusyak, K., Hull, P., and Jaravel, X. (2022). Quasi-experimental shift-share research designs. *The Review of Economic Studies*, 89(1):181–213.

Burchardi, K. B., Chaney, T., and Hassan, T. A. (2019). Migrants, ancestors, and foreign investments. *The Review of Economic Studies*, 86(4):1448–1486.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

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

Combes, P.-P., Duranton, G., Gobillon, L., Puga, D., and Roux, S. (2012). The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica*, 80(6):2543–2594.

Combes, P.-P. and Gobillon, L. (2015). The empirics of agglomeration economies. *Handbook of Regional and Urban Economics*, pages 247–348.

Costa, D. L. and Kahn, M. E. (2000). Power couples: changes in the locational choice of the college educated, 1940–1990. *The Quarterly Journal of Economics*, 115(4):1287–1315.

Couture, V. and Handbury, J. (2017). Urban revival in America, 2000 to 2010. Technical report, National Bureau of Economic Research.

Crane, B., Albrecht, C., Duffin, K. M., and Albrecht, C. (2018). China's special economic zones: an analysis of policy to reduce regional disparities. *Regional Studies, Regional Science*, 5(1):98–107.

de Chaisemartin, C. and D'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96.

Diamond, R. (2016). The determinants and welfare implications of US workers' diverging location choices by skill: 1980-2000. *American Economic Review*, 106(3):479–524.

Dupuy, A. (2021). Migration in China: To work or to wed? *Journal of Applied Econometrics*, 36(4):393–415.

Duranton, G. and Puga, D. (2004). Micro-foundations of urban agglomeration economies. In *Handbook of Regional and Urban Economics*, volume 4, pages 2063–2117. Elsevier.

Fan, J. and Zou, B. (2021). The dual local markets: Family, jobs, and the spatial distribution of skills. Mimeo.

Farole, T. and Akinci, G. (2011). *Special Economic Zones: Progress, Emerging Challenges, and Future Directions*. World Bank Group, Washington DC.

Goldsmith-Pinkham, P., Sorkin, I., and Swift, H. (2020). Bartik instruments: What, when, why, and how. *American Economic Review*, 110(8):2586–2624.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Henderson, J. (2003). Urbanization and economic development. *Annals of Economics and Finance*, 4:275–341.

Iyigun, M. and Walsh, R. P. (2007). Endogenous gender power, household labor supply and the demographic transition. *Journal of Development Economics*, 82(1):138–155.

Liu, M. (2021). The seventh national population census reveals that China's floating population has reached 376 million.

Lu, Y., Wang, J., and Zhu, L. (2019). Place-based policies, creation, and agglomeration economies: Evidence from China's economic zone program. *American Economic Journal: Economic Policy*, 11(3):325–60.

Marshall, A. (1890). *Principles of Economics*. Macmillan, London.

Meng, L. and Zhao, M. Q. (2019). Bride drain: An unintended consequence of China's urban-rural divide. *Labour Economics*, 58:69–80.

Mumford, L. (1961). *The city in history: Its origins, its transformations, and its prospects*, volume 67. Houghton Mifflin Harcourt.

National Bureau of Statistics of China (2021). *China Statistical Yearbook 2021*. China Statistics Press.

NDRC (2006). Catalog of announcement reviews for Chinese development zones (2006).

NDRC (2018). Catalog of announcement reviews for Chinese development zones (2018).

Nguyen, M.-N. (2022). Sex ratio among urban and rural population in Vietnam 2010-2021. Accessed: 2023-05-23.

Peters, M. and Siow, A. (2002). Competing premarital investments. *Journal of Political Economy*, 110(3):592–608.

Pettay, J. E., Lummaa, V., Lynch, R., and Loehr, J. (2021). Female-biased sex ratios in urban centers create a "fertility trap" in post-war finland. *Behavioral Ecology*, 32(4):590–598.

Phimister, E. (2005). Urban effects on participation and wages: Are there gender differences? *Journal of Urban Economics*, 58(3):513–536.

PTI (2017). Sex-ratio in urban areas worse than rural areas: Survey. Accessed: 2023-05-23.

Reynolds, C. L. and Weinstein, A. L. (2021). Gender differences in quality of life and preferences for location-specific amenities across cities. *Journal of Regional Science*, 61(5):916–943.

Rosenthal, S. S. and Strange, W. C. (2004). Evidence on the nature and sources of agglomeration economies. In *Handbook of Regional and Urban Economics*, pages 2119–2177. Elsevier.

Rosenthal, S. S. and Strange, W. C. (2012). Female entrepreneurship, agglomeration, and a new spatial mismatch. *Review of Economics and Statistics*, 94(3):764–788.

Sant'Anna, P. H. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122.

Shirk, S. L. et al. (1993). *The Political Logic of Economic Reform in China*, volume 24. Univ. of California Press.

Tacoli, C. (2012). *Urbanization, gender and urban poverty: paid work and unpaid carework in the city*. Human Settlements Group, International Institute for Environment and Development.

Wang, J. (2013). The economic impact of special economic zones: Evidence from Chinese municipalities. *Journal of Development Economics*, 101:133–147.

Weiss, Y., Yi, J., and Zhang, J. (2018). Cross-border marriage costs and marriage behavior: Theory and evidence. *International Economic Review*, 59(2):757–784.

Wiest, K., Leibert, T., Johansson, M., Rauhut, D., Ponnikas, J., Timar, J., Velkey, G., and Győrffy, I. (2013). Selective migration and unbalanced sex ratio in rural regions. *SEMIGRA. ESPON & Leibniz Institute for Regional Geography, Leibniz*.

Xu, C. (2011). The fundamental institutions of China's reforms and development. *Journal of Economic Literature*, 49(4):1076–1151.

Young, J. (2013). *China's Hukou System: Markets, Migrants and Institutional Change*. Palgrave Macmillan London.

Zhang, H. (2021). An investment-and-marriage model with differential fecundity: On the college gender gap. *Journal of Political Economy*, 129(5):1464–1486.

Zhang, H. and Zou, B. (2023). A marriage-market perspective on risk-taking and career choices. *European Economic Review*, 152:104379.

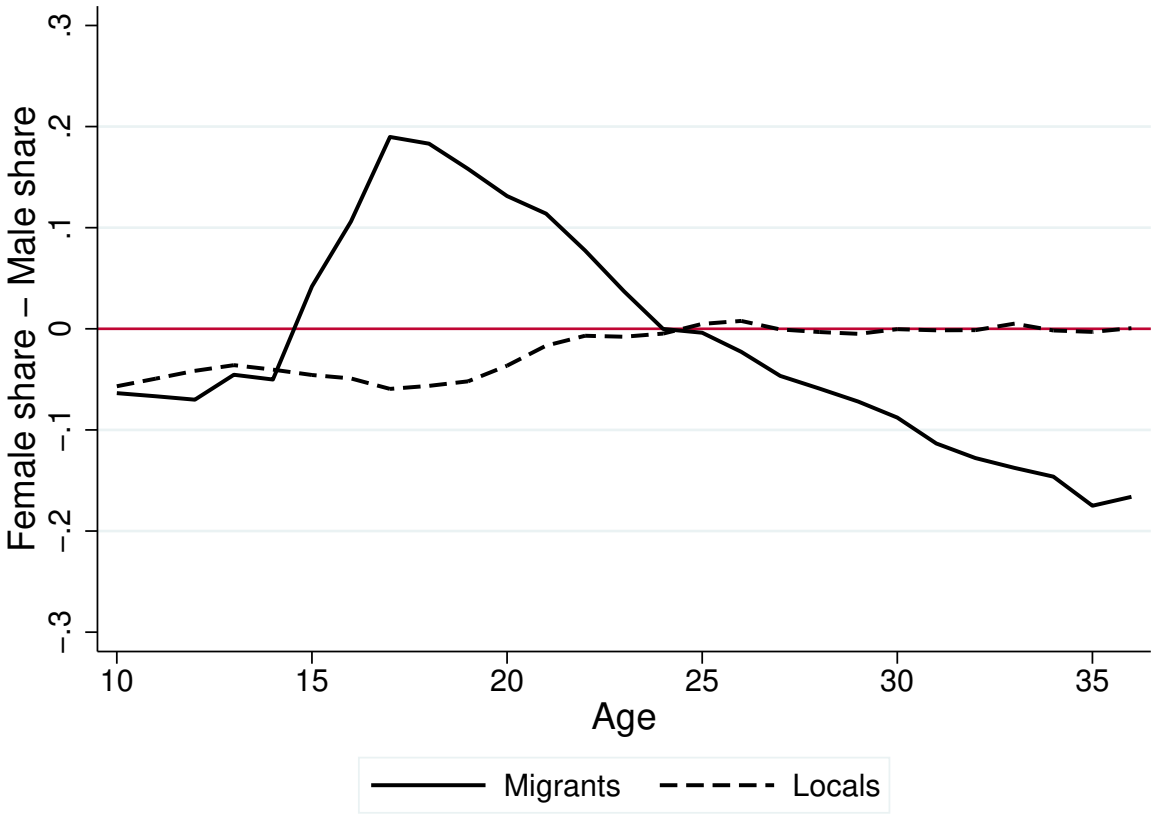
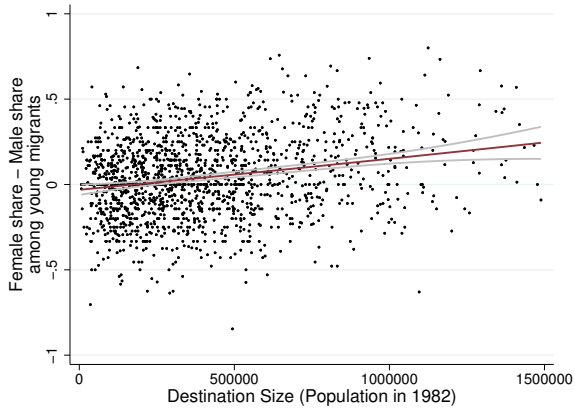


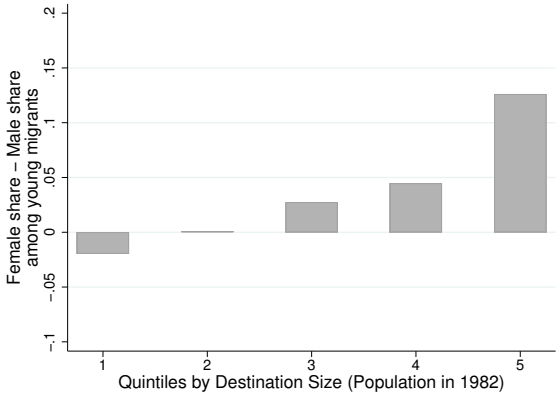
Figure 1: Gender Disparity by Age for Migrants and Locals in China in 2000

Data source: The Chinese Population Census 2000.

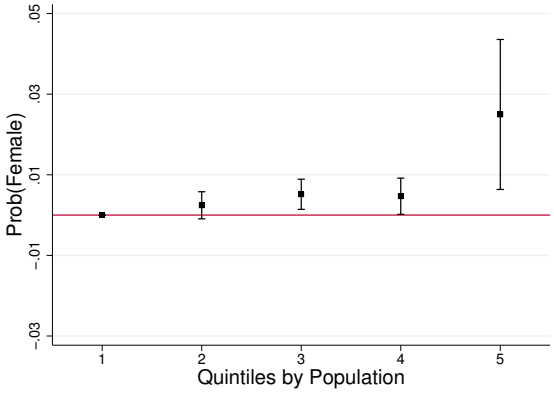
Notes: A person is defined as a migrant if they moved across counties and as a local otherwise. We calculate the difference between the female share and the male share for a given age. This value is zero when the gender distribution is perfectly balanced, positive when there is an excess of females, and negative when there is an excess of males.



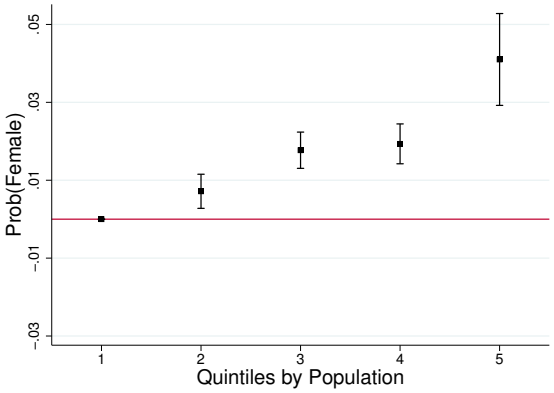
(a) By Destination Size



(b) By Quintile



(c) Coefficient Plot

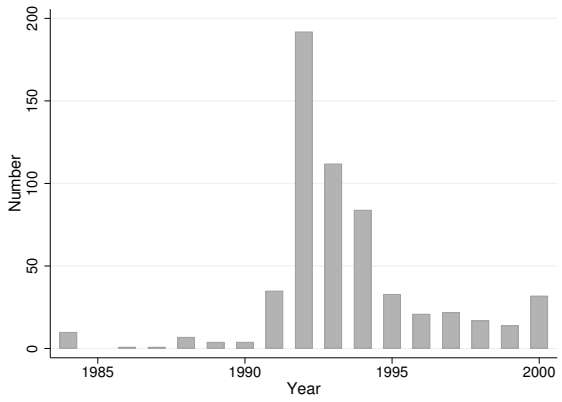


(d) Coefficient Plot, with Industry Fixed Effects

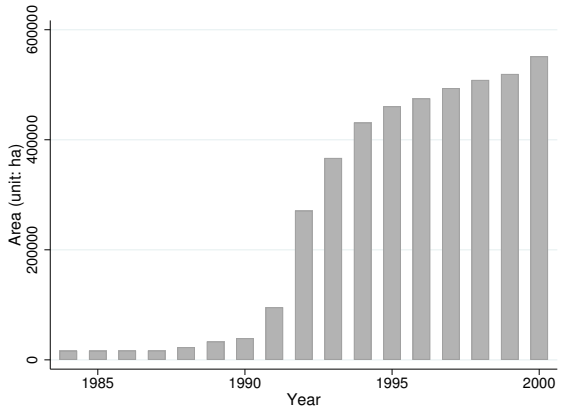
Figure 2: Gender Imbalance among Young Individuals in Larger Cities

Data source: The Chinese Population Census 2000.

Notes: In panels (a) and (b), we use a sample of young migrants aged 16 to 25 in the year 2000. In panel (a), the scatter plot displays counties represented by markers, along with a quadratic fitted line and corresponding confidence intervals. In panel (b), each bar represents the average net female share among young migrants across counties within a specific quintile. Quintile 1 (quintile 5) consists of the smallest (largest) counties in terms of population in 1982. In panel (c), we use a sample of all individuals aged 16 to 25 in the year 2000 and estimate the following: $\mathbb{1}(\text{Female})_i = \alpha_1 + \alpha_2\mathbb{1}(Q2) + \alpha_3\mathbb{1}(Q3) + \alpha_4\mathbb{1}(Q4) + \alpha_5\mathbb{1}(Q5) + \epsilon_i$. Quintiles are created using the county population in 2000, where quintile 1 (quintile 5) consists of the smallest (largest) counties. Standard errors are clustered at county level. Coefficients and vertical confidence intervals are displayed for each quintile, with quintile 1 serving as the reference category. In panel (d) we conduct a similar analysis, but further include fixed effects for individuals' labor-market industry.



(a) Number of Counties with SEZs

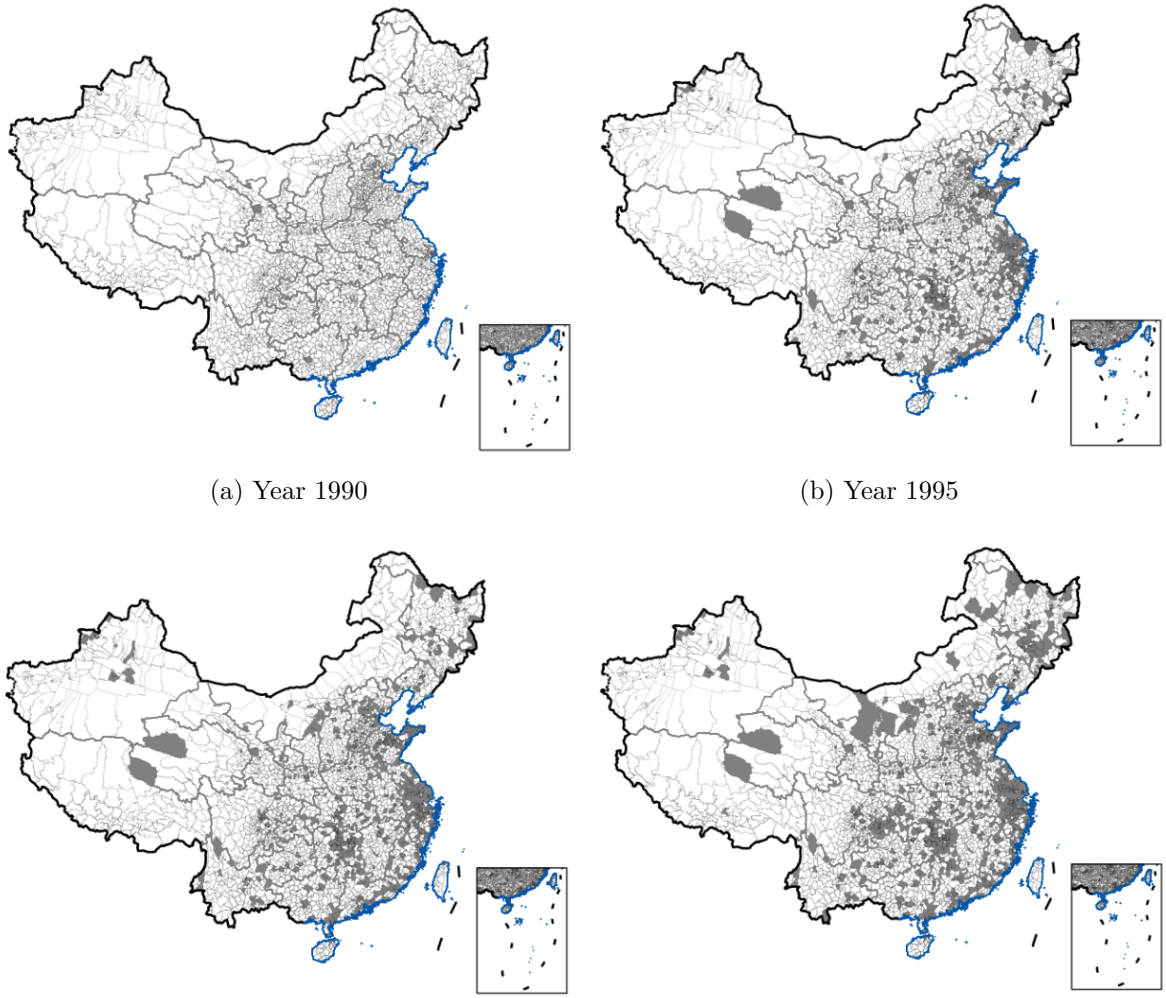


(b) Total Area of SEZs

Figure 3: Establishment of SEZs Across Years

Data source: The National Development and Reform Commission of China. (NDRC 2006)

Notes: Panel (a) displays the total number of SEZs that were newly established for the first time in a county during the specified year. Panel (b) shows the cumulative area of SEZs in a given year (unit: ha).



(a) Year 1990 (b) Year 1995

(c) Year 2000 (d) Year 2005

Figure 4: Geographic Spread of SEZs over Time in China

Data source: The National Development and Reform Commission of China.

Notes: The map displays China and its county borders. For each specified year (1990, 1995, 2000, and 2005), we denote SEZ-designated counties during that year using a darker shade of gray.

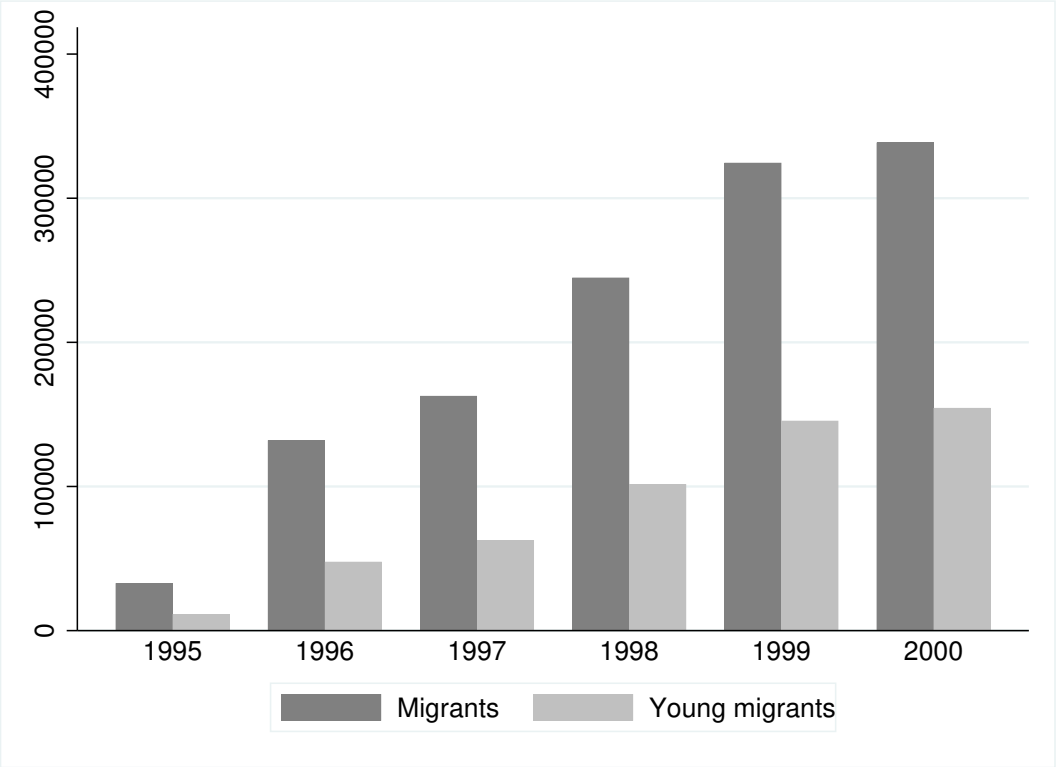
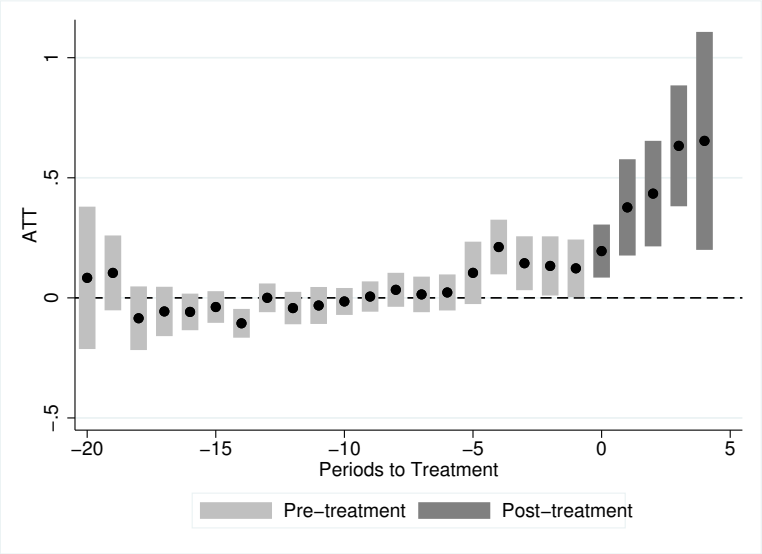


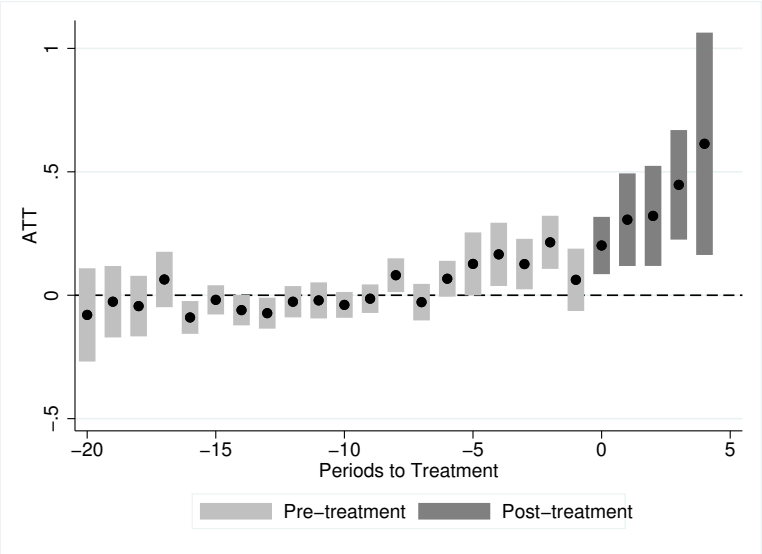
Figure 5: Total Number of Migrants

Data source: The Chinese Population Census 2000.

Notes: Darker gray bars indicate the count of all migrants who moved across counties in the specified year. The lighter gray bars indicate the count of young migrants aged between 16 and 25 for the specified year.



(a) Young Female Migrants



(b) Young Male Migrants

Figure 6: Time-varying Effects of SEZ Establishment on Young Migrants

Data source: The Chinese Population Census 2000.

Notes: In panel (a) we present dynamic ATTs (i.e., average treatment on the treated) obtained from estimating the specification in column (3) of panel (b) in Table 2. The dependent variable is log young female migrants and we follow Callaway and Sant'Anna (2021) in our estimation, using the not-yet-treated as the control group. In panel (b) we similarly present dynamic ATTs when the dependent variable is log young male migrants. The specification corresponds to column (4) of panel (b) in Table 2.

Table 1: Summary Statistics

Panel (a) County-level	Mean	SD	N
log(Population)	8.01	0.89	2,870
log(Migrants)	4.29	1.42	2,870
log(Female migrants)	3.56	1.41	2,870
log(Male migrants)	3.67	1.38	2,870
log(Female migrants aged 16-25)	2.31	1.39	2,870
log(Male migrants aged 16-25)	2.17	1.42	2,870
Female share	0.49	0.01	2,870
Female share among migrants	0.47	0.11	2,846
Female share among migrants aged 16-25	0.54	0.20	2,765
$\mathbb{1}$(Received SEZ shock by 2000)	0.16	0.37	2,870

Panel (b) Individual-level	Mean	SD	N
$\mathbb{1}$(Female)	0.49	0.50	2,562,835
Age	22.73	4.18	2,562,835
$\mathbb{1}$(Middle school or below)	0.72	0.45	2,562,835
Δ(Push factor)	3.14	1.17	2,562,835
Δ(Migrate)	0.10	0.30	2,562,835
Δ(Marry)	0.25	0.44	2,562,835
Δ(Marry up)	0.04	0.20	2,141,339

Data source: The Chinese Population Census 2000.
Notes: In panel (a), we report the county-level values in year 2000. In panel (b), we use the same sample as that for the regression analysis in Table 3: observations of young individuals who were between the ages of 16 and 25 at any point during the period 1996 to 2000.

Table 2: Effects of SEZ on Population

Panel (a)	Population	Migrants		
	(1)	(2)	(3)	(4)
Dep. variable	$\log(\text{Total})_{c,t}$	$\log(\text{Total})_{c,t}$	$\log(\text{Young F.})_{c,t}$	$\log(\text{Young M.})_{c,t}$
$\mathbb{1}(\text{SEZ})_{c,t}$	0.0546***	0.4982***	0.4328***	0.3517***
	(0.0089)	(0.0716)	(0.0682)	(0.0688)
Observations	25,695	25,695	25,695	25,695
Root MSE	0.125	0.560	0.449	0.437
Method	TWFE	TWFE	TWFE	TWFE

Panel (b)	Population	Migrants		
	(1)	(2)	(3)	(4)
Dep. variable	$\log(\text{Total})_{c,t}$	$\log(\text{Total})_{c,t}$	$\log(\text{Young F.})_{c,t}$	$\log(\text{Young M.})_{c,t}$
$\mathbb{1}(\text{SEZ})_{c,t}$	0.0084**	0.3919***	0.3780***	0.3120***
	(0.0037)	(0.0976)	(0.0698)	(0.0651)
Observations	6,987	6,987	6,987	6,987
Method	DiD	DiD	DiD	DiD
Control Group	Not-yet	Not-yet	Not-yet	Not-yet

Notes: In both panels, we use the yearly sample of counties in China from 1995 to 2000. In panel (a), we include county fixed effects and prefecture-by-year fixed effects across all specifications. In columns (3) and (4) of both panels, the term "young females" ("young males") refers to individuals between the ages of 16 and 25. In panel (b) we estimate a staggered DiD model following Callaway and Sant'Anna (2021) and use the not-yet-treated as the control group. Standard errors in parentheses are corrected for heteroskedasticity and clustered at county level. Asterisks ***, **, * denote p <0.01, p <0.05, p <0.1, respectively.

Table 3: Migration Outcomes

$\Delta(\text{Migrate})_{i,c}$	(1) Unadjusted OLS	(2) Controlled OLS	(3) Unadjusted OLS	(4) Controlled OLS
$\Delta(\text{Push Factor})_c$	0.0020*	0.0022**	0.0022**	0.0024**
	(0.0011)	(0.0011)	(0.0011)	(0.0011)
$\Delta(\text{Push Factor})_c \times \mathbb{1}(\text{Female})_i$	0.0034***	0.0017***	0.0035***	0.0017***
	(0.0002)	(0.0004)	(0.0002)	(0.0004)
Expected $\Delta(\text{Push Factor})_c$		0.0037***		0.0038***
		(0.0013)		(0.0013)
Expected $\Delta(\text{Push Factor})_c \times \mathbb{1}(\text{Female})_i$		0.0021***		0.0021***
		(0.0004)		(0.0004)
$\Delta(\text{SEZ})_c$			-0.0239***	-0.0245***
			(0.0081)	(0.0080)
$\Delta(\text{SEZ})_c \times \mathbb{1}(\text{Female})_i$			-0.0037	-0.0038
			(0.0047)	(0.0046)
Observations	2,562,835	2,562,835	2,562,835	2,562,835
Root MSE	0.299	0.299	0.299	0.299
Mean(y)	0.100	0.100	0.100	0.100

Notes: The dependent variable $\Delta(\text{Migrate})_{i,c}$ equals one for individual i if they emigrated from county c to another county between 1996 and 2000. The sample consists of all individuals who were between the ages of 16 and 25 at any point during the period 1996 to 2000. Across all specifications, we include fixed effects for age in year 2000. Standard errors in parentheses are corrected for heteroskedasticity and clustered at county level. Asterisks ***, **, * denote $p < 0.01$, $p < 0.05$, $p < 0.1$, respectively.

Table 4: Potential Explanations

Dep. var: $\Delta(\text{Migrate})_{i,c}$	(1) Unadjusted OLS	(2) Controlled OLS	(3) Unadjusted OLS	(4) Controlled OLS
Panel (a) Education – Excluding Student Migrants				
$\Delta(\text{Push Factor})_c$	0.0022**	0.0022**	0.0024**	0.0024**
	(0.0011)	(0.0011)	(0.0011)	(0.0011)
$\Delta(\text{Push Factor})_c \times \mathbb{1}(\text{Female})_i$	0.0035***	0.0018***	0.0035***	0.0019***
	(0.0002)	(0.0004)	(0.0002)	(0.0004)
Panel (b) Amenities – Sample of Less-educated Individuals				
$\Delta(\text{Push Factor})_c$	-0.0003	0.0000	-0.0001	0.0003
	(0.0012)	(0.0012)	(0.0012)	(0.0012)
$\Delta(\text{Push Factor})_c \times \mathbb{1}(\text{Female})_i$	0.0051***	0.0029***	0.0052***	0.0029***
	(0.0003)	(0.0005)	(0.0003)	(0.0005)
Panel (c) Amenities – Sample of Highly-educated Individuals				
$\Delta(\text{Push Factor})_c$	0.0068***	0.0069***	0.0070***	0.0071***
	(0.0014)	(0.0015)	(0.0014)	(0.0015)
$\Delta(\text{Push Factor})_c \times \mathbb{1}(\text{Female})_i$	-0.0016***	-0.0015***	-0.0016***	-0.0015***
	(0.0002)	(0.0004)	(0.0003)	(0.0004)
Panel (d) Industrial Composition Change – Including Labor Market Conditions				
$\Delta(\text{Push Factor})_c$	0.0026**	0.0018	0.0029**	0.0021*
	(0.0012)	(0.0012)	(0.0012)	(0.0012)
$\Delta(\text{Push Factor})_c \times \mathbb{1}(\text{Female})_i$	0.0016***	0.0012***	0.0016***	0.0012***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Panel (e) Marriage Market – Sample of Older Individuals				
$\Delta(\text{Push Factor})_c$	0.0045***	0.0037***	0.0047***	0.0039***
	(0.0006)	(0.0007)	(0.0006)	(0.0007)
$\Delta(\text{Push Factor})_c \times \mathbb{1}(\text{Female})_i$	-0.0050***	-0.0034***	-0.0050***	-0.0035***
	(0.0001)	(0.0002)	(0.0001)	(0.0002)
Expected terms	No	Yes	No	Yes
Own SEZ shocks	No	No	Yes	Yes

Notes: The dependent variable $\Delta(\text{Migrate})_{i,c}$ equals one for individual i if they emigrated from county c to another county between 1996 and 2000. There are 2,429,582; 1,845,444; 717,391; 2,562,835; and 2,186,997 observations in panels (a), (b), (c), (d), and (e), respectively. Across specifications, we include fixed effects for age in year 2000. Standard errors in parentheses are corrected for heteroskedasticity and clustered at county level. Asterisks ***, **, * denote $p < 0.01$, $p < 0.05$, $p < 0.1$.

Table 5: Marital Outcomes

	(1) Unadjusted OLS	(2) Controlled OLS	(3) Unadjusted OLS	(4) Controlled OLS
Panel (a) Dep. var: $\Delta(\text{Marry})_{i,c}$				
$\Delta(\text{Push Factor})_c$	-0.0076***	-0.0069***	-0.0076***	-0.0068***
	(0.0006)	(0.0007)	(0.0006)	(0.0006)
$\Delta(\text{Push Factor})_c \times \mathbb{1}(\text{Female})_i$	0.0132***	0.0104***	0.0133***	0.0104***
	(0.0003)	(0.0005)	(0.0003)	(0.0005)
Panel (b) Dep. var: $\Delta(\text{Marry up})_{i,c}$				
$\Delta(\text{Push Factor})_c$	-0.0049***	-0.0025***	-0.0049***	-0.0025***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
$\Delta(\text{Push Factor})_c \times \mathbb{1}(\text{Female})_i$	0.0105***	0.0051***	0.0106***	0.0052***
	(0.0002)	(0.0004)	(0.0002)	(0.0004)
Expected Terms	No	Yes	No	Yes
Own SEZ shocks	No	No	Yes	Yes

Notes: In panel (a), the dependent variable $\Delta(\text{Marry})_{i,c}$ equals one for individual i who was living in county c in 1996 and married sometime between 1996 and 2000. The sample consists of 2,562,835 individuals who were between the ages of 16 and 25 at any point during the period of 1996 to 2000. In panel (b), the dependent variable $\Delta(\text{Marry up})_{i,c}$ equals one for individual i if they married sometime between 1996 and 2000 and married a partner who has a higher education level. The sample comprises 2,141,339 individuals who meet the following criteria: (1) They were between the ages of 16 and 25 at any time from 1996 to 2000. (2) They were either consistently single from 1996 to 2000 or they married between 1996 and 2000, with their spouse's education type identified. The identification was possible because both the surveyed individual and their spouse were included in the Census data and surveyed together. Across all specifications in both panels, we include fixed effects for age in year 2000. Standard errors in parentheses are corrected for heteroskedasticity and clustered at county level. Asterisks ***, **, * denote $p < 0.01$, $p < 0.05$, $p < 0.1$.

Appendices

A Simulation Algorithm for Expected Δ(Push Factor)

A total of 87 counties received SEZ shocks between 1996 and 2000 and another 518 counties received SEZ shocks between 2001 and 2006. To compute the expected $\Delta(\text{Push Factor})_c$, we run simulations following Borusyak and Hull (2023). For each simulation s, we go through the following steps:

1. Given a set of 605 counties in which SEZs were established for the first time between 1996 and 2006, we randomly draw a set of 87 counties X^s that were as likely to have received the SEZ shocks before 2001.
2. Compute corresponding $\boldsymbol{V}^s$ according to the set of shocks X^s drawn.
3. Compute $\Delta(\text{Push Factor})^s = \boldsymbol{M}\boldsymbol{V}^s$
4. Repeat steps 1 to 3 5,000 times.
5. Compute the expected instruments:

$$\text{Expected } \Delta(\text{Push Factor}) = \frac{1}{N^s}\sum_{s=1}^{N^s} \Delta(\text{Push Factor})^s$$

Note that for each simulation, we separately compute the values for counties that received or did not receive their own internal SEZ shock.

We contrast the realized Δ(Push Factor) and the expected Δ(Push Factor) in Figure C.2. As expected, individuals from Shandong, Hubei, Jiangxi, and Fujian are more affected by the push of urbanization taking place in nearby fast-growing provinces or metropolitan cities, such as Beijing, Tianjin, Shanghai, Zhejiang, and Guangdong. The realized Δ(Push Factor) is positively correlated with the expected Δ(Push Factor), but also presents obvious discrepancies; e.g., Yunnan province.

B Gender-specific Labor Market Opportunities

Similar to Δ(Push Factor), we construct $\Delta(\text{Labor Mkt})^F$ and $\Delta(\text{Labor Mkt})^M$ at province level and project province-specific values to county level based on the province a county belongs to. $\Delta(\text{Labor Mkt})^F$ and $\Delta(\text{Labor Mkt})^M$ are analogously constructed, except for using either industry-specific female worker share or male worker share, respectively. We explain the construction of $\Delta(\text{Labor Mkt})^F$ below, and $\Delta(\text{Labor Mkt})^M$ follows analogously.

$$\Delta(\text{Labor Mkt})^F = \boldsymbol{M}\boldsymbol{B}^F, \tag{B.1}$$

where $\boldsymbol{M}$ is the normalized matrix of migration flows defined in Section 3.2. $\boldsymbol{B}^F$ is an $N_p \times 1$ vector, which is constructed in a manner that is similar to a Bartik instrument. Specifically, element $b_p^F \in \boldsymbol{B}^F$ is computed as $b_p^F = \sum_{k=1}^{K} b_{pk}^F$, where k indexes an industry, K is the total number of industries, and

$$b_{pk}^F = z_{pk} \cdot \{g_k^{treat} w_p^{treat} + g_k^{not} w_p^{not}\} \cdot f_k. \tag{B.2}$$

In equation (B.2), z_{pk} is industry k's share in province p in 1990, measured as the number of workers in industry k in province p divided by the number of all workers in province p in 1990. The term in braces shows the employment growth between 1990 and 2000 in industry k as a weighted average between treated and not-yet-treated groups. Specifically,

$$g_k^{treat} = \log(\text{employment in 2000})_k^{treat} - \log(\text{employment in 1990})_k^{treat} \tag{B.3}$$

is the employment growth in industry k for the treated at national level and g_k^{not} is defined analogously for the not-yet-treated. Weight w_p^{treat} is the share of the population in province p living in counties that were treated as of year 1990, and $w_p^{not} = 1 - w_p^{treat}$. Lastly, the f_k term is the percentage of female workers in industry k in 1990, which is shown in Figure C.3.

C Additional Figures

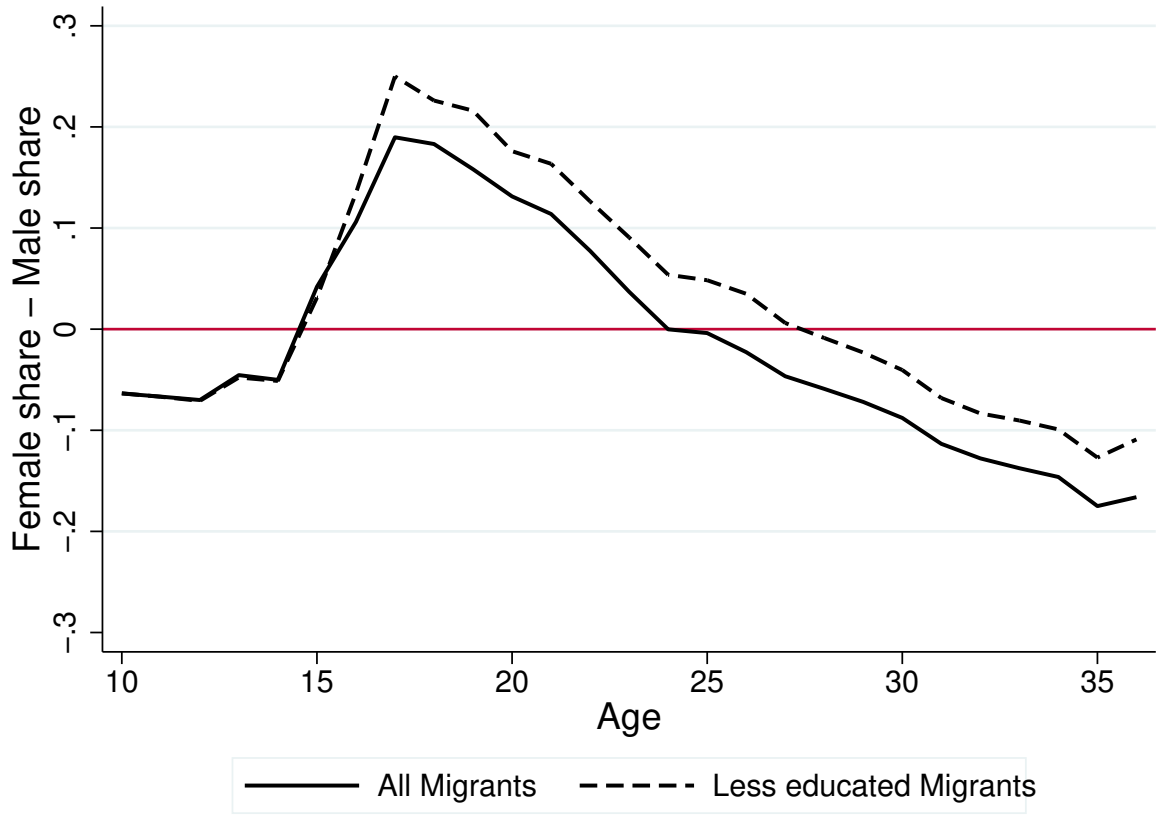


Figure C.1: Gender Disparity by Age in China in 2000 by Education

Data source: The Chinese Population Census 2000.
Notes: We focus on migrants who moved across counties. The solid line shows the difference between the female share and the male share for a given age among all migrants. The dashed line shows the difference between the female share and the male share for a given age among less-educated migrants. Individuals with an education level of middle school graduate or below are considered to be less educated.

(a) Realized Δ(Push Factor)

(b) Expected Δ(Push Factor)

Figure C.2: Δ(Push Factor) – Realized and Expected

Notes: Panel (a) displays the values of Δ(Push Factor) calculated following the method outlined in Section 3.2 using actual data. Darker gray indicates higher values of Δ(Push Factor), which indicates stronger motivation for residents of the respective province to emigrate to other locations. Panel (b) shows the expected values of Δ(Push Factor) derived following Borusyak and Hull (2023) by implementing the simulation algorithm outlined in Appendix Section A.

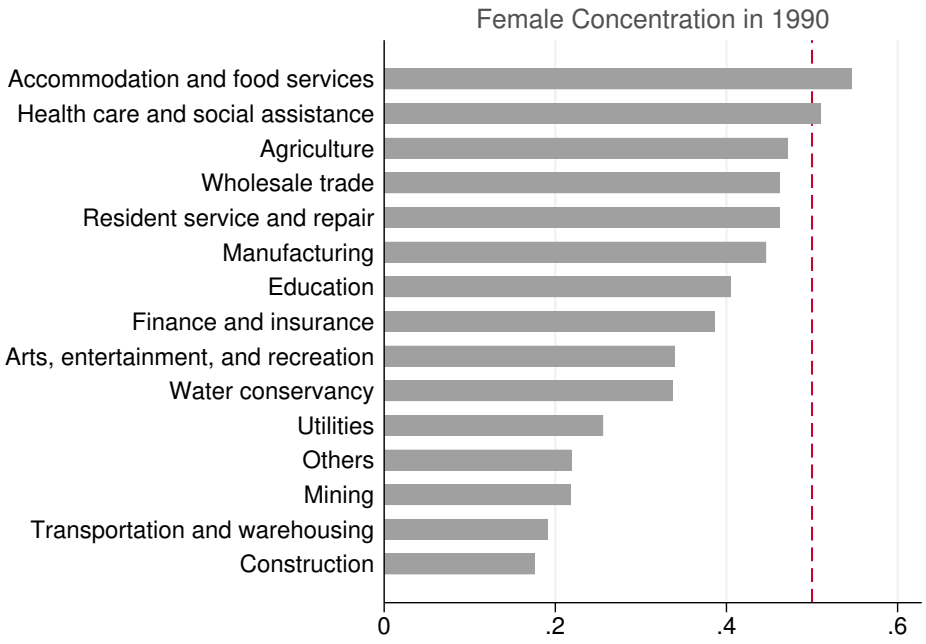


Figure C.3: Industry-specific Female Share in 1990

Data source: The Chinese Population Census 1990.

Notes: For each industry, we compute the share of female workers using Chinese Population Census 1990 data. The vertical dashed line indicates the value 0.5, which suggests a perfectly balanced ratio of male to female workers.

D Additional Tables

Table D.1: Changes in Migration Status Using Only Rural Counties

$\Delta(\text{Migrate})_{i,c}$	(1) Unadjusted OLS	(2) Controlled OLS	(3) Unadjusted OLS	(4) Controlled OLS
$\Delta(\text{Push Factor})_c$	0.0027**	0.0018*	0.0029***	0.0021*
	(0.0011)	(0.0011)	(0.0011)	(0.0011)
$\Delta(\text{Push Factor})_c \times \mathbb{1}(\text{Female})_i$	0.0035***	0.0016***	0.0036***	0.0016***
	(0.0002)	(0.0003)	(0.0002)	(0.0004)
Expected $\Delta(\text{Push Factor})_c$		0.0091***		0.0092***
		(0.0015)		(0.0015)
Expected $\Delta(\text{Push Factor})_c \times \mathbb{1}(\text{Female})_i$		0.0025***		0.0025***
		(0.0004)		(0.0004)
$\Delta(\text{SEZ})_c$			-0.0235***	-0.0255***
			(0.0074)	(0.0077)
$\Delta(\text{SEZ})_c \times \mathbb{1}(\text{Female})_i$			-0.0051*	-0.0055**
			(0.0027)	(0.0026)
Observations	1,571,397	1,571,397	1,571,397	1,571,397
Root MSE	0.283	0.283	0.283	0.283

Notes: The dependent variable $\Delta(\text{Migrate})_{i,c}$ equals one for individual i if they emigrated from county c to another county between 1996 and 2000. The sample consists of all individuals who were between the ages of 16 and 25 at any point during the period 1996 to 2000 and living in counties that had a population of less than 700,000 as of 1982. Across all specifications, we include fixed effects for age in year 2000. Standard errors in parentheses are corrected for heteroskedasticity and clustered at county level. Asterisks ***, **, * denote p <0.01, p <0.05, p <0.1, respectively.

Table D.2: Tendency to Marry Up – Individuals Aged 20-35 Years

$\Delta(\text{Marry up})_{i,c}$	(1) Unadjusted OLS	(2) Controlled OLS	(3) Unadjusted OLS	(4) Controlled OLS
$\Delta(\text{Push Factor})_c$	-0.0033***	-0.0017***	-0.0033***	-0.0017***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
$\Delta(\text{Push Factor})_c \times \mathbb{1}(\text{Female})_i$	0.0063***	0.0028***	0.0063***	0.0028***
	(0.0001)	(0.0003)	(0.0001)	(0.0003)
Observations	3,327,375	3,327,375	3,327,375	3,327,375
Root MSE	0.177	0.177	0.177	0.177
Mean(y)	0.034	0.034	0.034	0.034

Notes: The dependent variable $\Delta(\text{Marry up})_{i,c}$ equals one for individual i if they married sometime between 1996 and 2000 and married a partner who has a higher education level. The sample comprises 3,327,375 individuals who meet the following criteria: (1) they were between the ages of 20 and 35 at any time from 1996 to 2000 and (2) they were either consistently single throughout the period from 1996 to 2000 or they married between 1996 and 2000, with their spouse's education type identified. The identification was possible because both the surveyed individual and their spouse were included in the Census data and surveyed together. Across all specifications in both panels, we include fixed effects for age in year 2000. Standard errors in parentheses are corrected for heteroskedasticity and clustered at county level. Asterisks ***, **, * denote $p < 0.01$, $p < 0.05$, $p < 0.1$

E An Equilibrium Marriage-market Model

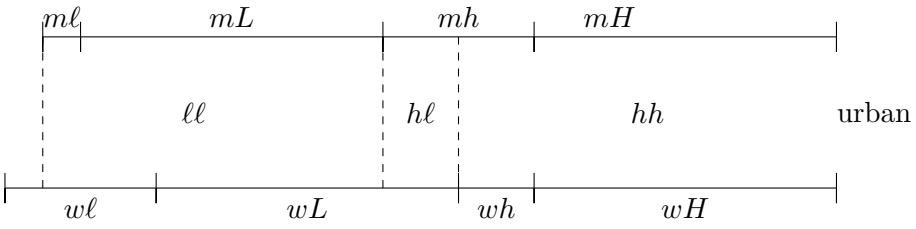
In this section, we present an equilibrium model that is consistent with the pattern whereby young females are more likely to migrate and marry up than young males.

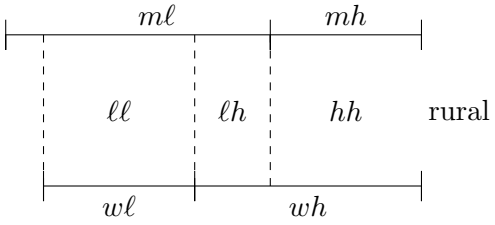
Each person is endowed with one of two skill/education types, high and low. Denote an original rural person's skill by h and ℓ and an urban person's skill by H and L. Each rural person is endowed with a heterogeneous differential y that indicates the nonmarital (education, amenity, and/or labor-market) gain from moving to the city. Mass distributions of nonmarital gain y are denoted as $F_{g\theta}$, which are gender- and skill-specific on support $[-Y, Y]$. Suppose there are equal masses of men and women in both rural and urban areas. The total marriage surplus $s_{\theta_m\theta_w}$ is determined by husband's and wife's skill types θ_m and θ_w, but the division of this surplus between the couple is determined in equilibrium. Without loss of generality, assume that they are not location-specific. In addition, assume strict surplus supermodularity to guarantee positive assortative matching: $s_{hh} + s_{\ell\ell} > s_{h\ell} + s_{\ell h}$.

A person's gain from moving is additive in the differential y and equilibrium marital gain that is determined according to Becker (1973): A *stable outcome* of the marriage market (G_m, G_w), where G_m and G_w denote the mass of males and females, respectively. The stable outcome consists of *stable matching* and *stable marriage payoffs*. Stable matching G satisfies *feasibility*: $\sum_{\theta_w\in\Theta_w} G_{\theta_m\theta_w} \leq G_{m\theta_m}$ for any $\theta_m \in \Theta_m$ and $\sum_{\theta_m\in\Theta_m} G_{\theta_m\theta_w} \leq G_{w\theta_w}$ for any $\theta_w \in \Theta_w$. Stable marriage payoffs v_m and v_w satisfy (i) *individual rationality*: $v_{m\theta_m} \geq 0$ for any $\theta_m \in \Theta_m$ and $v_{w\theta_w} \geq 0$ for any $\theta_w \in \Theta_w$ (every person receives at least as much as they would have if they had remained single); (ii) *pairwise efficiency*: $v_{m\theta_m} + v_{w\theta_w} = s_{\theta_m\theta_w}$ (every married couple divides the entire marriage surplus); and (iii) *Pareto efficiency*: $v_{m\theta_m} + v_{w\theta_w} \geq s_{\theta_m\theta_w}$ for all $\theta_m \in \Theta_m$ and $\theta_w \in \Theta_w$ (no man-woman pair not married to each other can simultaneously improve their marriage payoffs by marrying each other).

Claim. When $F_{w\ell}(s_{\ell\ell}) < F_{mL}(-s_{\ell\ell})$ and $F_{mh}(s_{\ell\ell}+\kappa) < F_{wh}(-\kappa)$, where $\kappa = (s_{hh}-s_{h\ell})-(s_{\ell h}-s_{\ell\ell})$, there is a unique equilibrium in which more skilled men than women and more unskilled women than men migrate.

Proof. The equilibrium matching is depicted as follows.





In the scenario depicted, the rural marriage payoffs are as follows: $v_{m\ell} = 0$, $v_{w\ell} = s_{\ell\ell}$, $v_{wh} = s_{\ell h}$, $v_{mh} = s_{hh} - s_{\ell h}$. The urban marriage payoffs are $V_{w\ell} = 0$, $V_{m\ell} = s_{\ell\ell}$, $V_{mh} = s_{h\ell} - s_{\ell\ell}$, $V_{wh} = s_{hh} - (s_{h\ell} - s_{\ell\ell})$. The marital benefits of moving to the city are

$$V_{w\ell} - v_{w\ell} = -s_{\ell\ell},$$

$$V_{m\ell} - v_{m\ell} = s_{\ell\ell},$$

$$V_{mh} - v_{mh} = (s_{h\ell} - s_{\ell\ell}) - (s_{hh} - (s_{\ell h} - s_{\ell\ell})) = -s_{\ell\ell} - [(s_{hh} - s_{h\ell}) - (s_{\ell h} - s_{\ell\ell})],$$

$$V_{wh} - v_{wh} = (s_{hh} - (s_{h\ell} - s_{\ell\ell})) - s_{\ell h} = (s_{hh} - s_{h\ell}) - (s_{\ell h} - s_{\ell\ell}),$$

where $V_{wh} - v_{wh} > V_{w\ell} - v_{w\ell}$ if s is supermodular. If $y > s_{\ell\ell}$, a rural low-skilled woman would move, so mass $F_{w\ell}(Y) - F_{w\ell}(s_{\ell\ell})$ of rural low-skilled women move. To sustain the equilibrium such that more skilled men than women and more unskilled women than men move, we must have the two conditions as specified. □