Left over or opting out?

Squeeze, mismatch and surplus in Chinese marriage markets

Pauline Rossi and Yun Xiao*

Abstract

Marriage is declining in China. Among singles, the probability to marry in 2019 was twice as low as in 1999. We estimate a Choo and Siow (2006b) model using census data to quantify the relative roles of changes in population structure and changes in marital surplus, i.e., value of marriage. We find that the increase in the supply of educated people explains half of the decline, partly due to a mismatch between educated women and less-educated men. The deterioration of female-to-male ratio, known as the marriage squeeze, explains an additional 18% for men. The decrease in surplus explains the rest.

Keywords: Marriage markets, Sex ratio, Education, China

*Rossi (corresponding author): CREST, Ecole Polytechnique, France (email: pauline.rossi@ensae.fr); Xiao: University of Gothenburg, Sweden (email: yun.xiao@economics.gu.se). We are deeply grateful to Manasi Chhabra for outstanding research assistance and to Pascaline Dupas, Andreas Kotsadam, Roland Rathelot and Tom Vogl for detailed feedback on our first draft. Our research benefited from helpful comments from several seminar audiences, and from discussions with Li Chen and Lisa Eklund. Rossi gratefully acknowledges funding from the European Union (ERC, P3OPLE, ERC-2021-StG 101039252). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

1 Introduction

Marriage in China has been on a steady decline for a decade. Policy makers are concerned by high rates of celibacy among young women and young men for several reasons. On the women side, celibacy contributes to low birth rates in a country where out-of-wedlock births are rare. Fertility rates have remained below replacement levels for more than thirty years, leading to serious demographic challenges (Dollar, Huang, and Yao 2020). On the men side, celibacy is associated with unhappiness, adverse health outcomes including premature deaths, and crime in many contexts (Chang, Kan, and Zhang 2021; Edlund et al. 2013; Va et al. 2011; Zhou and Hesketh 2017). The fact that people marry later, or even not at all, therefore has broad implications (Shu and Chen 2023).

A large, multidisciplinary literature aiming to understand the reasons behind this decline has identified three key factors. The first factor is the *marriage squeeze*, which refers to the relative scarcity of men compared to women on the marriage market. The One-Child Policy combined with a strong preference for sons has generated a substantial sex imbalance: there were approximately 35 millions more men than women in the 2020 census, fueling concerns about “leftover" men (*shengnan*) (National Bureau of Statistics of China 2021). The second factor is the *marriage mismatch*, which occurs when singles struggle to find a suitable partner due to a shortage of their preferred type. Chinese traditions encourage female hypergamy; marriages in which the groom has less education than the bride are frowned upon. As education has risen faster for women than for men in recent decades, there is a shortage of suitable husbands for highly educated women, who end up “leftover" (*shengnu*). The third factor is the decrease in *marriage surplus*, namely the perceived benefits of being married versus being single. Changing social norms and preferences, greater gender equality in employment opportunities, as well as increasing costs of having children, lead more and more individuals to “opt out" of marriage, at least temporarily.

In this paper, we quantify the role of each factor. We use the Choo and Siow (2006b) model, which allows us to identify and estimate the marital surplus using data on the number of singles and the number of matches, by education category, in a given year. We need to observe how many males and females of each category are single at the beginning of year t, whether they match during the year and the category of their partner. This information can be constructed for the year 1999 using the 1% sample of the 2000 Chinese Census, and for the year 2019 combining aggregate statistics from the 2020 Chinese Census and survey data from the China Family Panel Studies. We proceed in three steps: (1) estimate the marital surplus in 1999 and 2019; (2) predict counterfactual marriages in 2019 under the assumption that the surplus remained constant at 1999 levels during the 20-year period; and (3) simulate different scenarios changing the population structure (the sex ratio and the level of education, by gender) of new cohorts joining the market between 1999 and 2019.

We find that the marital surplus indeed decreased between 1999 and 2019. The strongest variation (-28%) is observed for matches between less educated men and less educated women, the category that used to generate the highest surplus and be the most prevalent in 1999. Interestingly, matches between less educated men and more educated women, which used to generate the lowest surplus and be the least prevalent in 1999, are the only type of matches experiencing a substantial improvement (+12%). This suggests that the preference for female hypergamy has weakened over time. When we simulate the evolution of the market holding surplus constant, we find that there would be 131 million singles in 2019 compared to 150 million in the real data.

In the counterfactual analysis, we are mainly interested in the annual matching rate, that is the probability to find a match during the year for a single man and a single woman. This probability was divided by two between 1999 and 2019, decreasing from 11% to 6% for men, and from 19% to 9% for women. The difference between men and women reflects the fact that men typically marry older than women, and hence stay longer on the marriage market. What we want to explain is not the difference in levels but rather the trends in both male and female matching rates. When we hold marital surplus constant at 1999 levels and update the flow of singles to reflect the observed composition of new cohorts, we predict a decrease in matching rates by 3p.p. for men and by 4.4p.p. for women, representing 60% and 46% of the real declines, respectively. The remaining 40% (resp. 54%) of the decline for men (resp. women) can be attributed to changes in surplus. When we update only the education of new cohorts while maintaining the sex ratio balanced, we can explain 46% and 54% of the real declines for men and women, respectively. Changes in the supply of educated men and women may reduce matching rates because, first, educated people marry later, and second, educated women do not marry down. To disentangle both channels, we raise education levels in the same way for both genders, keeping the gender gap in education constant at the 1999 level. We predict a decrease in matching rates by 1.5p.p. for men and 3.5p.p. for women, representing 30% and

37% of the real declines, respectively. This implies that a substantial part (roughly one third) of the contribution of education can be explained by a mismatch. Finally, updating only the sex ratio of new cohorts while maintaining the education at 1999 levels has a negative impact on male matching rates, contributing to the decline by 18%. The impact on female matching rates is positive, implying that, in the absence of changes in education and marital surplus, single women would have been more likely to find a match in 2019 than in 1999.

Our article contributes to the literature on marriage markets in China by studying the main drivers of the decline in marriage rates in a unified, quantitative framework. In the literature, each factor tends to be studied separately, by researchers in different disciplines using different methods. On the one hand, quantitative models are used to study the marriage squeeze. The evolution of the marriage market, taking into account the cohort sizes on both sides, is simulated. The goal is to estimate whether and when the imbalanced sex ratio has become a first-order factor. Projections depend on assumptions on age at marriage and spousal age gap. There is a broad consensus that the marriage squeeze matters and will peak in the 2040s-50s. Estimates for the most recent period (pre-2020) are mixed, some papers arguing that sex ratios already make a difference (Guilmoto 2012) whereas others argue that they do not (Jiang et al. 2016). The methodology is often very sophisticated, but the underlying model fails to account for changes in the value of marriage. On the other hand, qualitative methods are used to document the changing strategies employed by individuals on the market (Qian and Qian 2014; You, Yi, and Chen 2016). The key research question is to determine to what extent younger generations, and in particular women, are willing and able to reject marriage. Some studies emphasize that, besides individual preferences and social norms, the pool of potential partners is an important factor influencing matches (Eklund and Attané 2017; Li, Jiang, and Feldman 2017; Song, Skaggs, and Frazier 2017; To 2015). However, as pointed out by Schwartz (2013), the methodology does not allow to quantify the relative importance of different factors.

We also relate to the literature on matching models in economics. The methodology proposed by Choo and Siow (2006b) to decompose changes in marriage patterns into changes in population composition and changes in surplus has been applied in Canada (Choo and Siow 2006a), in the US (Chiappori, Salanié, and Weiss 2017; Cornelson and Siow 2016) and in Denmark (Bruze, Svarer, and Weiss 2015). In the Chinese context, Choo and Siow (2006b) has been used to study the impact of the 1958-61 famine on the marital attractiveness of smaller cohorts (Brandt, Siow, and Vogel 2016), the marriage penalty for educated women (Brandt et al. 2018), and the value of transmitting family name (Yang and Spencer 2022). Other papers have studied the impact of the marriage squeeze on savings, marriage and intra-household allocations (Ong, Yang, and Zhang 2020; Porter 2016; Wei and Zhang 2011). Our paper is the first attempt to quantify how much of the decline in Chinese marriage can be attributed to changes in sex ratio and changes in education. In particular, we isolate the role of the mismatch between highly educated women and less educated men, contributing to the nascent literature in economics studying hypergamy (Almås et al. 2023).

Beyond the Chinese context, the decline in marriage has been observed throughout East and South East Asia (Jones 2017). Like China, some of these countries experienced deep changes in attitudes toward marriage as well as a deterioration of sex ratios at birth (e.g. Vietnam) and a rapid expansion of education (e.g. Japan, South Korea, Singapore, and more recently Thailand and Vietnam). Our quantification exercise can easily be replicated to analyse the evolution of marriage outside China.

2 Methodology

2.1 Set-up

Choo and Siow (2006b) propose an empirical model with transferable utility and no search friction. The main advantage is to identify the gains from matching using only data on matches and without requiring data on transfers between partners, which are typically unobserved.

Denote μ_{ij} the number of matches between a man of type i and a woman of type j observed in the data. μ_{i0} and μ_{0j} are the number of men of type i and women of type j who remain unmatched, respectively. Denote π_{ij} the *marriage surplus*, defined as the total gain to marriage per partner for a pair (i, j) relative to the total gain per partner from remaining unmarried. The key result is the following relationship:

$$\pi_{ij} = \ln(\frac{\mu_{ij}}{\sqrt{\mu_{i0}\mu_{0j}}})$$

The intuition behind this result is the following: if we observe many marriages between type i and type j, these unions must be valuable. However, it can also be that there are lots of men of type i and lots of women of type j. We need to “control for" the scale dimension in order to isolate the value dimension. We use the number of unmatched individuals of both types to do so. The value of a match is high if these matches are often observed *relative to* the frequency of types in the population.

The identification relies on assumptions regarding the distribution and the separability of unobserved heterogeneity in matching payoffs. These assumptions impose restrictions on the patterns of substitution between the different choices of partners. They also rule out the possibility that the unobserved characteristics of a specific man and the unobserved characteristics of a specific woman interact when generating the match surplus. In the model, the total surplus π_{ij} is determined by deep preference parameters. It does not depend on the population vector. What depends on the population vector is the transfer between spouses, which guarantees that the demand for a spouse of a given type equals the supply.

We provide more details on the modeling of the matching market, the identification, and the interpretation of π_{ij} in Appendix A.

2.2 From observed matches to estimation of surplus

Suppose that we have two types, H and L. We need to observe (i) the population vector, i.e. the number of singles of each type at the beginning of the period: m_H, m_L, f_H and f_L; and (ii) the number of matches in all four categories: μ_{HH}, μ_{LH}, μ_{HL} and μ_{LL}. We can then infer the number of unmatched individuals, by type. In practice, this means filling in the population matrix; then, we use the formula to estimate the surplus matrix, as illustrated below.

$$\begin{array}{cc} & \text{Women} \\ \text{Men} & \begin{pmatrix}\mu_{HH} & \mu_{HL}\\ \mu_{LH} & \mu_{LL}\end{pmatrix} \quad \begin{matrix}\mu_{H0}\\ \mu_{L0}\end{matrix} \quad \xrightarrow{\pi_{i,j}=ln(\frac{\mu_{i,j}}{\sqrt{\mu_{i0}\mu_{0j}}})} \begin{pmatrix}\pi_{HH} & \pi_{HL}\\ \pi_{LH} & \pi_{LL}\end{pmatrix} \\ & \begin{matrix}\mu_{0H} & \mu_{0L}\end{matrix} \end{array}$$

2.3 From estimated surplus to simulation of matches

Once we have estimated the surplus in a given market, we can perform simulations by keeping the surplus constant (in blue) and changing the population vector (in red). This exercise requires solving the following system of 8 equations and 8 unknowns, where i and j take values H and L.

$$\begin{array}{ll} (1) & \mu'_{iH}+\mu'_{iL}+\mu'_{i0}=m'_i \\ (2) & \mu'_{Hj}+\mu'_{Lj}+\mu'_{0j}=f'_j \\ (3) & {\mu'_{ij}}^2=\mu'_{i0}\mu'_{0j}exp(2\pi_{i,j}) \end{array} \xrightarrow{\text{CS algorithm}} \begin{array}{cc} \begin{pmatrix} \mu'_{HH} & \mu'_{HL} \\ \mu'_{LH} & \mu'_{LL} \end{pmatrix} & \begin{array}{c} \mu'_{H0} \\ \mu'_{L0} \end{array} \\ \begin{array}{cc} \mu'_{0H} & \mu'_{0L} \end{array} & \end{array}$$

Choo and Siow (2006a) propose a simple algorithm to solve this system (which also works when there are more than 2 types). The algorithm predicts the number of matches in all 4 categories and the number of unmatched individuals by type. We can then answer the question: what would have happened if surplus had stayed constant while population structure changed?

2.4 From simulated matches to construction of counterfactual markets

In order to study the evolution of the counterfactual marriage market over several years, we need to model the evolution of the stock of singles. At period 0, we estimate the surplus, π_0, using the observed number of matched and unmatched individuals. At the beginning of the next period, the number of singles, Singles_1, is the sum of previously unmatched individuals and the new cohort entering the market. We simulate matches in period 1 using Singles_1 and π_0, and we predict the number of unmatched individuals at the end of the period. We can iterate the process at every period using the following loop:

$$(\text{Unmatched}_{t-1}+\text{New Cohort}_t=\text{Singles}_t)\text{ combined with }\pi_0 \xrightarrow{\text{CS algorithm}} \text{Unmatched}_t$$

Singles_t has two components: (i) an endogenous component, the *stock* of individuals who were already on the market and did not match in the previous year; its structure (size, sex ratio, type, age) is an outcome determined by the surplus and population vectors in previous years; (ii) an exogenous component, the *flow* of individuals entering the market in that year as they become adults; the structure of this incoming cohort (size, sex ratio, type) changes over time for reasons unrelated to the marriage market.

With data on the marriage market in 1999 and 2019 as well as data on incoming cohorts during the 2000-2019 period, we can (i) take the observed matches in 1999 and estimate the surplus in 1999, (ii) simulate matches year-by-year holding surplus constant and updating the singles vector with simulated values for the stock and observed values for the flow, and (iii) compare the final simulated matches and the matches observed in 2019. This allows us to decompose the evolution into changes in surplus and changes in the structure of incoming cohorts.

3 Data and Descriptive Statistics

3.1 Data

We use three data sources. First, the 1% sample of the 2000 National Population Census (Minnesota Population Center 2020) provides information on *who married whom* in 1999. We combine microdata on current marital status and year of first marriage to identify who was single at the beginning of 1999, who married during the year and with whom. We observe the number of male and female singles, by education category, and the number of matches, by education of *both* spouses. We are therefore able to construct (i) the vector of singles, (ii) the matching rate, and (iii) the frequency of match types in 1999. We observe around 55,000 marriages in 1999, which allows us to have very reliable estimates of the characteristics of matches. Second, microdata from the 2020 National Population Census are not publicly available yet. Instead, we use aggregate statistics available in the China Population Census Yearbook 2020 (National Bureau of Statistics of China 2020) to infer *who married* in 2019, but we cannot know with whom.[1] We observe the number of male and female singles, by education category, and the number of matches. So we can construct (i) the vector of singles and (ii) the matching rate in 2019. Third, in order to construct the frequency of match types in 2019, we use the 2020 China Family Panel Studies (CFPS). This is a household survey representative of 95% of the Chinese population and providing the relevant information (marital status and year of marriage). However, the sample size is 1/1000 of the census 1% sample. That is why we prefer to use census data as much as possible. Still, we observe around 150 marriages in 2019, which is enough to get estimates of the frequency of each type.[2]

We define the incoming cohort in year t as the cohort of new adults, i.e. turning 18 during the previous year. Although the minimum legal marriage age is 20 for women and 22 for men, we observe some marriages at age 19 in the data. Using the legal age to define the pool of potential partners would therefore raise issues when we count the number of matches. Next, we define the type as *Low* (L) if the individual has less than high school (HS) education, and *High* (H) if the individual has ever attended HS. High school typically starts at age 16 so the decision to enroll or not is taken before entering the marriage market. Types are therefore predetermined. This

[1] The aggregates are official statistics based on a 10% sample of the total population, who were administered with a long questionnaire covering detailed questions about marital status and marriage age.

[2] We checked that the vectors of singles and the matching rates are nearly the same when we estimate them using census data and using CFPS data. Note that we need the frequency of match types to compute the surplus in 2019, but not to perform the counterfactual analysis.

is why we choose high school rather than college. The incoming cohorts during the 2000-2019 period are born between 1981 and 2000. We use the 2020 aggregate statistics to get information on the number of individuals, by gender and education, in each cohort. Apart from age, we make one restriction when defining the pool of singles. We exclude individuals who have been previously married (divorced or widowed) because the marriage markets are quite segmented in China and the policy concern is about first marriages.[3] Finally, we focus on 2019 because marriage rates after 2020 were exceptionally low due to covid-19 and our goal is to capture long-term trends. In principle, we could have used the 2010 National Population Census to add a point in 2009 but the 0.2% sample we could access was not representative. We provide more details on the datasets and variables in Appendix B.

Between 1999 and 2019, the annual matching rate among singles declined by 5 percentage point for men, from 10.8% to 5.8%, and by 9.5 percentage points for women, from 18.5% to 9%. The annual matching rate is closely related to the distribution of ages at marriage. For instance, if the rate is equal to λ and constant between age 19 and 30, the proportion who marries before turning 30 is equal to $1-(1-\lambda)^{11}$. The decrease in matching rates would therefore correspond to a decrease in the proportion married before turning 30 from 72% to 48% for men, and from 89% to 65% for women. In the next sections, we discuss three often-cited explanations for the decrease and provide some descriptive statistics supporting them.

3.2 Explanation 1: Imbalanced sex ratios

The right panel of Figure 1a plots the evolution of the sex ratio by cohort using the 2020 census aggregates. For cohorts active on the marriage market in 1999 (born in 1980 and before), the sex ratio is stable around 1.03 men per woman. For the incoming cohorts, born between 1981 and 2000, we can distinguish two periods: before 1990, the sex ratio remains relatively balanced, whereas after 1990, the sex ratio becomes more and more imbalanced over time, reaching an all-time high of 1.13 men per woman. We therefore expect the male matching rate to decrease between 1999 and 2019, and in particular after 2010, as imbalanced cohorts enter the market. In the counterfactual exercise, we modify the sex ratio of incoming cohorts to quantify the contribution of the marriage squeeze.

[3] Among marriages happening in 2019, 92% are first marriages for both spouses. Among single men (women) in 2020, 12% (9%) were divorced and 11% (30%) were widowed. Over 99.5% of singles aged 70 and older remain single.

3.3 Explanation 2: Education expansion

The left panel of Figure 1a plots the evolution of the share of individuals ever attending high school, by gender and cohort. Two features of the evolution are striking: first, education increased dramatically over the past decades, and second, the gender gap reversed. In the 1970 cohort, 21% of women and 26% of men have ever attended high school. In the 2000 cohort, these proportions are 81% and 75%, respectively. Women have overtaken men in terms of education since cohorts born in the late 1980s. This implies that the composition of cohorts active on the marriage market is very different in 1999 (the average education is low and lower for women) and in 2019 (the average education is high and higher for women).

This change in the supply of educated men and women may affect the matching rates in two ways. First, high- and low-educated people may have different values of marriage *in general*: they marry later and have different outside options. Second, they have different preferences in terms of partner type. In particular, in presence of female hypergamy, the female advantage in education generates a mismatch between educated women and less-educated men. A first hint that female hypergamy exists is that, when looking at the distribution of marriage types in 1999, we find that marriages between L-type men and H-type women are very rare: they account for only 5% of all marriages. The opposite combination, between H-type men and L-type women, is twice as frequent (9.5%). However, these proportions reflect both the population structure and the marital surplus. We use our methodology to test whether LH matches indeed generate a lower surplus than HL matches. Then we consider counterfactual scenarios in which the education of incoming cohorts grew less, or grew equally for men and women.

3.4 Explanation 3: Decrease in the value of marriage

Before turning to the estimation results, we illustrate changes in the value of marriage, relative to being single, using data from the China General Social Survey. In particular, one question asks whether the respondent agrees with the statement that “Even the worst marriage is better than singlehood." The survey was conducted in 2006 and respondents were between 20 and 60 years old. Figure 1b plots the share of men and women agreeing with the above statement, by cohort. For older cohorts, the levels are high, around 40%, indicating that the outside option of remaining single used to have a very low value for a substantial share of people. The trend is clearly downward: for younger cohorts, less than 30% agree with the statement.[4]

If we do the same analysis by education level, we find that highly educated individuals agree less than relatively low educated individuals, for both men and women, and the decline over time is particularly pronounced for highly educated women. Therefore, in the incoming cohorts, we expect the value of marriage to decline over time, because of composition effects – the share with high school education increases – and because of changes in the value of marriage within each group. By estimating the model, we get a better measure of the marital surplus in the sense that (i) the measure has a structural interpretation and (ii) the measure is match-specific.

4 Results

4.1 Estimation

In Figure 2a, we report the surpluses in 1999 and 2019 calculated using equation 1 (see matrices in Appendix C.1 for more details). Figure 2b reports changes over time in percentages points. Note that all surpluses are negative, meaning that the average value from matching is lower than the average value from remaining single for both partners. This is because, in a given year, most people remain single. Only the individuals who derive a large idiosyncratic payoff from a given match do choose to match; the majority do not.[5]

In 1999, the ranking of the different types of matches is: $LH < HL < HH < LL$. This indicates that (i) there is a strong preference for assortative matching; (ii) among assorted couples, low educated types derive more value from matching than high educated types; (iii) there is a strong distaste for female hypogamy.

In 2019, the surpluses generated by the most prevalent matches, LL and HH, have decreased by 28% and 13%, respectively. The strongest decline is observed among low-educated individuals, which challenges the common wisdom that opting out of marriage is a trend specific to highly educated groups. Moreover, the surplus generated by LH has not decreased, quite the opposite: it has increased by 12%, almost closing the gap with the surplus of HL, which remained stable over time. This evolution suggests that the strong distaste for women marrying down attenuates as the female advantage in education grows. Our finding is consistent with recent qualitative research arguing that norms around hypergamy are not experienced as strongly by younger

[4] Note that we cannot separate cohort and age in this dataset; we emphasize the cohort interpretation to be consistent with qualitative studies documenting generational gaps rather than life cycle trends. Unfortunately, we do not have information for our youngest incoming cohorts (born between 1986 and 2000).

[5] In most applications of Choo and Siow (2006b), estimated surpluses are negative.

cohorts (Eklund 2018). One explanation is that parents, who used to put pressure on their daughters to marry up, are now mainly worried that their daughters may not marry at all.

Overall, the analysis of surpluses shows that deep preferences for marriage have changed in ways that are sometimes unexpected and impossible to infer from descriptive statistics. Changes in surplus should explain part of the decline in marriage. Next, we turn to the counterfactual exercise to quantify precisely how much.

4.2 Counterfactual number of singles

We start by simulating the evolution of the stock of singles: we fix the surplus at 1999 levels and, in each year t, we allow individuals born in $t - 19$ to join the market. Figure 3 plots the evolution over 20 years and compares the final simulated value with the real value observed in 2019. If marital surplus had stayed constant during the 20-year period, there would be 18 million fewer singles in 2019, a drop by 12% compared to the real number. The structure would also be different: there would be 10 million fewer H-type men, 3 million *more* L-type men, 4 million fewer H-type women and 7 million fewer L-type women.

As expected, changes in surplus contributed to raising the total number of singles, as more men and women decided to *opt out*. However, a countervailing force was at play for L-type men and H-type women: the attenuation of the distaste for female hypogamy, which made LH matches more likely to happen than in 1999.

4.3 Counterfactual matching rate

The analysis of the stock of singles indicates that, even if marital surplus had stayed constant, the number of singles would have increased substantially, from roughly 100 million in 1999 to 130 million in 2019. This is because incoming cohorts are more numerous and have a different structure in terms of sex ratio and education. In order to focus on the role of the structure, we turn to a statistics that does not vary with the size of the population: the matching rate. Again, we fix the surplus at 1999 levels, and we consider 4 simulations. First, we update the population vector using both the sex ratio and the education rate of the incoming cohorts. This simulation gives the total contribution of changes in the population structure. Second, we update only the sex ratio and we attribute to the incoming cohorts the average level of education of cohorts active on the 1999 market. Third, we update only the education rates and we attribute to the incoming cohorts the sex ratio among cohorts active on the 1999 market. This allows us to decompose the contribution of the population structure into both components. Finally, we decompose the contribution of education into the absolute increase for both men and women and the relative increase for women with respect to men.

The results are shown in Figure 4a; we provide details on the different simulated markets in the figure note and we report the corresponding matrices in Appendix C.2. Figure 4a plots the counterfactual matching rates in each simulation, for men and women. We compare them with the real matching rates observed in 1999 (shown at the extreme left) and in 2019 (shown at the extreme right). We superimpose the 2019 rates on the other bars in order to highlight the magnitude of the change between 1999 and 2019. Figure 4b plots the fraction of the real change that can be explained by each simulation.

In simulation 1, we find that marriage rates would have decreased by 3p.p. for men and by 4.4p.p for women if surpluses had stayed constant. Given that the true decline is 5p.p. for men and 10p.p. for women, the change in the population vector explains 60% and 46% of the change in matching rate for men and women, respectively. The remaining part is due to the change in marital surplus documented in the previous section. Changes in population explain a larger share of the decline for men than for women because the change in sex ratio contributes in opposite ways: alone, it reduces matching rates by 18% for men and increases matching rates by 13% for women (simulation 2). The change in education alone is by far the main component for both men and women, explaining 46% and 54% of the total decline, respectively (simulation 3).[6]

Finally, to separate the contribution of the general increase in education from the marriage mismatch between high-educated women and low-educated men, we assume that education has increased equally for both women and men (simulation 4). We attribute to the incoming cohorts the observed education rate for men, and the same rate minus 4.4p.p. for women. This captures the overall increase in education while maintaining the gender gap in the share of H-type constant at the level of cohorts active on the 1999 market. When education increases while the gender gap does not, marriage rates drop by 1.5p.p. for men and by 3.5p.p for women, explaining 30% and 37% of the real change between 1999 and 2019. These numbers represent about two thirds of the contribution of changes in supply of educated men and women estimated in simulation 3. The remaining one third is attributed to the increase in the gender gap, which generates a

[6] The fact that simulations 2 and 3 do not perfectly add up to simulation 1 suggests that there is a (very small) interaction effect between education and sex ratio. When we update both components separately, we overestimate the decline by 0.1p.p. for men and we underestimate the decline by 0.4p.p. for women compared to the scenario in which we update them jointly.

mismatch between highly educated women and less educated men.

4.4 Extensions

Heterogeneity by urban-rural market. So far, we have considered the country as a unique marriage market. However, the literature has emphasized differences between urban and rural areas (Shu and Chen 2023). We repeat our analysis separately for cities and towns on the one hand and for villages on the other hand, using the classification of places of residence in the census. The assumption is that both markets are separate and determined at age 19. Individuals may move before turning 19 (typically leaving rural areas to start college education in urban areas) but afterwards, they live and marry in the same market. We find that the sex ratio of incoming cohorts deteriorated much more in the rural market, while both markets experienced a strong increase in education and a reversal of the gender gap. In terms of contribution to the decline in matching rates, our conclusions regarding education are unchanged: the absolute increase and the relatively higher increase for women matter everywhere. However, the marriage squeeze matters only in the rural market.

Defining type by *hukou* instead of education. The results of the decomposition exercise are specific to our definition of types as "ever attended high school". Another partition of the population would potentially lead to a different conclusion regarding the contribution of changes in the population structure. In that sense, the contribution of the surplus is a residual and an upper bound: we managed to explain 60% (resp. 46%) of the decline in marriage for men (resp. women) using only one dimension of the population vector. Another potentially interesting dimension would be *hukou* (agricultural or non-agricultural). Unfortunately, there are two data limitations. First, aggregate data from the 2020 census are not available by *hukou* status; therefore, we can only exploit the CFPS survey data which has fewer observations. Second, the *hukou* status is observed at the time of the survey, not at the time of marriage. If people change their status upon marriage, we mis-classify them. Keeping these caveats in mind, using CFPS data, we find that the share of individuals with an agricultural *hukou* increased between the 1981 cohort and the 2000 cohort, and is persistently higher for men. This is consistent with higher fertility and more imbalanced sex ratio in rural areas, where most agricultural *hukou*-holders live. Nonetheless, magnitudes are small and the counterfactual exercise assigns a limited role to changes in *hukou* composition in explaining the change in marriage. We conclude that "high school attendance" is by far the most relevant dimension to define types in this context.[7]

5 Conclusion

Our results help explain the decline in annual marriage rates over the past two decades in China. Quantitatively, roughly 40% of the decline is attributable to a decrease in marital surplus, while 60% is attributable to changes in population structure for men. The proportions are reversed for women and this is because the deterioration of sex ratio adversely affected men and not women. For both genders, the most important structural component is the increased supply of educated men and women. The contribution of education to the decline can further be decomposed into (i) an overall increase in education levels, accounting for two thirds, and (ii) a higher increase for women than for men, accounting for the remaining one third. In other words, the marriage decline is not only driven by an excess of men in general, but more specifically by an excess of uneducated men. Uneducated men are left over and educated women opt out partly because they do not find a suitable partner. Looking ahead, the role of marriage squeeze may become more prominent as the accumulation of imbalanced cohorts over time will increasingly impact the pool of singles. On the other hand, mismatch may become less influential as preferences evolve, and the option of marrying someone with a lower level of education may become more acceptable for women.

This application illustrates the potential of the methodology developed by Choo and Siow (2006b) to deepen our understanding of marriage markets. The methodology is computationally simple to implement and does not require detailed data. We can think of two avenues for future research: (i) applying it more systematically to quantify the drivers of marriage rates in different countries; (ii) explaining the evolution of surplus over time by considering, for instance, the role of changes in fertility preferences, divorce rights, housing markets or labor markets.

[7] We also considered defining type by ethnicity. However, changes along this dimension make little difference in the aggregate given that Han Chinese account for more than 90% of the population.

References

Almås, Ingvild, Andreas Kotsadam, Espen R. Moen, and Knut Røed. 2023. "The Economics of Hypergamy." *Journal of Human Resources* 58 (1): 260–281.

Brandt, Loren, Hongbin Li, Laura Turner, Jiaqi Zou, et al. 2018. "Are China's "leftover women" really leftover? An investigation of marriage market penalties in modern-day China," accessed November 7, 2023. https://cdn.dal.ca/content/dam/dalhousie/pdf/faculty/science/economics/seminars/2018/Seminar2018-11-02Turner.pdf.

Brandt, Loren, Aloysius Siow, and Carl Vogel. 2016. "Large demographic shocks and small changes in the marriage market." *Journal of the European Economic Association* 14 (6): 1437–1468.

Bruze, Gustaf, Michael Svarer, and Yoram Weiss. 2015. "The dynamics of marriage and divorce." *Journal of Labor Economics* 33 (1): 123–170.

Chang, Simon, Kamhon Kan, and Xiaobo Zhang. 2021. "Too many men, too-short lives: The effect of the male-biased sex ratio on mortality." *Journal of Human Resources,* ISSN: 0022-166X.

Chiappori, Pierre-André, Bernard Salanié, and Yoram Weiss. 2017. "Partner choice, investment in children, and the marital college premium." *American Economic Review* 107 (8): 2109–2167.

Choo, Eugene, and Aloysius Siow. 2006a. "Estimating a marriage matching model with spillover effects." *Demography* 43 (3): 463–490.

———. 2006b. "Who marries whom and why." *Journal of political Economy* 114 (1): 175–201.

Cornelson, Kirsten, and Aloysius Siow. 2016. "A quantitative review of Marriage Markets: How Inequality is Remaking the American Family by Carbone and Cahn." *Journal of Economic Literature* 54, no. 1 (March): 193–207.

Dollar, David, Yiping Huang, and Yang Yao. 2020. *China 2049: Economic Challenges of a Rising Global Power.* Brookings Institution Press.

Edlund, Lena, Hongbin Li, Junjian Yi, and Junsen Zhang. 2013. "Sex ratios and crime: Evidence from China." *Review of Economics and Statistics* 95 (5): 1520–1534.

Eklund, Lisa. 2018. "The sex ratio question and the unfolding of a moral panic? Notions of power, choice and self in mate selection among women and men in higher education in China." In *Scarce women and surplus men in China and India: Macro Demographics versus Local Dynamics,* edited by S. Srinivasan and S. Li, 105–125. Springer.

Eklund, Lisa, and Isabelle Attané. 2017. "Marriage squeeze and mate selection in China." Chap. 10 in *Handbook on the Family and Marriage in China,* edited by Xiaowei Zang and Lucy Xia Zhao. Handbooks of Research on Contemporary China Series. New York, NY: Edward Elgar Publishing.

Galichon, Alfred, and Bernard Salanié. 2022. "Cupid's Invisible Hand: Social Surplus and Identification in Matching Models." *The Review of Economic Studies* 89 (5): 2600–2629.

Guilmoto, Christophe Z. 2012. "Skewed sex ratios at birth and future marriage squeeze in China and India, 2005–2100." *Demography* 49 (1): 77–100.

Jiang, Quanbao, Xiaomin Li, Shuzhuo Li, and Marcus W Feldman. 2016. "China's marriage squeeze: A decomposition into age and sex structure." *Social indicators research* 127:793–807.

Jones, W, Gavin. 2017. "Changing marriage patterns in Asia." In *Routledge Handbook of Asian Demography.* Routledge.

Li, Shuzhuo, Quanbao Jiang, and Marcus Feldman. 2017. "Son preference and the marriage squeeze in China." Chap. 9 in *Handbook on the Family and Marriage in China,* edited by Xiaowei Zang and Lucy Xia Zhao. Handbooks of Research on Contemporary China Series. New York, NY: Edward Elgar Publishing.

McFadden, D. 1973. "Conditional logit analysis of qualitative choice behaviour." In *Frontiers in Econometrics,* edited by P. Zarembka, 105–142. New York, NY, USA: Academic Press.

Minnesota Population Center. 2020. *Integrated Public Use Microdata Series, International: Version 7.3 [Fifth National Population Census of China].* Accessed April 27, 2023. https://doi.org/10.18128/D020.V7.3.

National Bureau of Statistics of China. 2020. *China Population Census Yearbook 2020.* Beijing, China: National Bureau of Statistics of China. Accessed May 30, 2023. https://www.stats.gov.cn/sj/pcsj/rkpc/7rp/indexch.htm.

———. 2021. *Press Release: Main data of the Seventh National Population Census.* Accessed January 10, 2024. https://www.stats.gov.cn/english/PressRelease/202105/t20210510_1817185.html.

Ong, David, Yu Alan Yang, and Junsen Zhang. 2020. "Hard to get: The scarcity of women and the competition for high-income men in urban China." *Journal of Development Economics* 144:102434.

Porter, Maria. 2016. "How do sex ratios in China influence marriage decisions and intra-household resource allocation?" *Review of Economics of the Household* 14:337–371.

Qian, Yue, and Zhenchao Qian. 2014. "The gender divide in urban China: Singlehood and assortative mating by age and education." *Demographic Research* 31:1337–1364.

Schwartz, Christine R. 2013. "Trends and variation in assortative mating: Causes and consequences." *Annual Review of Sociology* 39:451–470.

Shu, Xiaoling, and Jingjing Chen. 2023. *Chinese Marriages in Transition: From Patriarchy to New Familism.* Rutgers University Press.

Song, Lijun, Rachel Skaggs, and Cleothia Frazier. 2017. "Educational homogamy." Chap. 7 in *Handbook on the Family and Marriage in China,* edited by Xiaowei Zang and Lucy Xia Zhao. Handbooks of Research on Contemporary China Series. New York, NY: Edward Elgar Publishing.

To, Sandy. 2015. *China's Leftover Women: Late Marriage among Professional Women and its Consequences.* Routledge.

Va, Puthiery, Wan-Shui Yang, Sarah Nechuta, Wong-Ho Chow, Hui Cai, Gong Yang, Shan Gao, Yu-Tang Gao, Wei Zheng, Xiao-Ou Shu, et al. 2011. "Marital status and mortality among middle age and elderly men and women in urban Shanghai." *PloS one* 6 (11): e26600.

Vallette, Adrien. 2023. *Male Childlessness and the Marriage Market.* Master thesis.

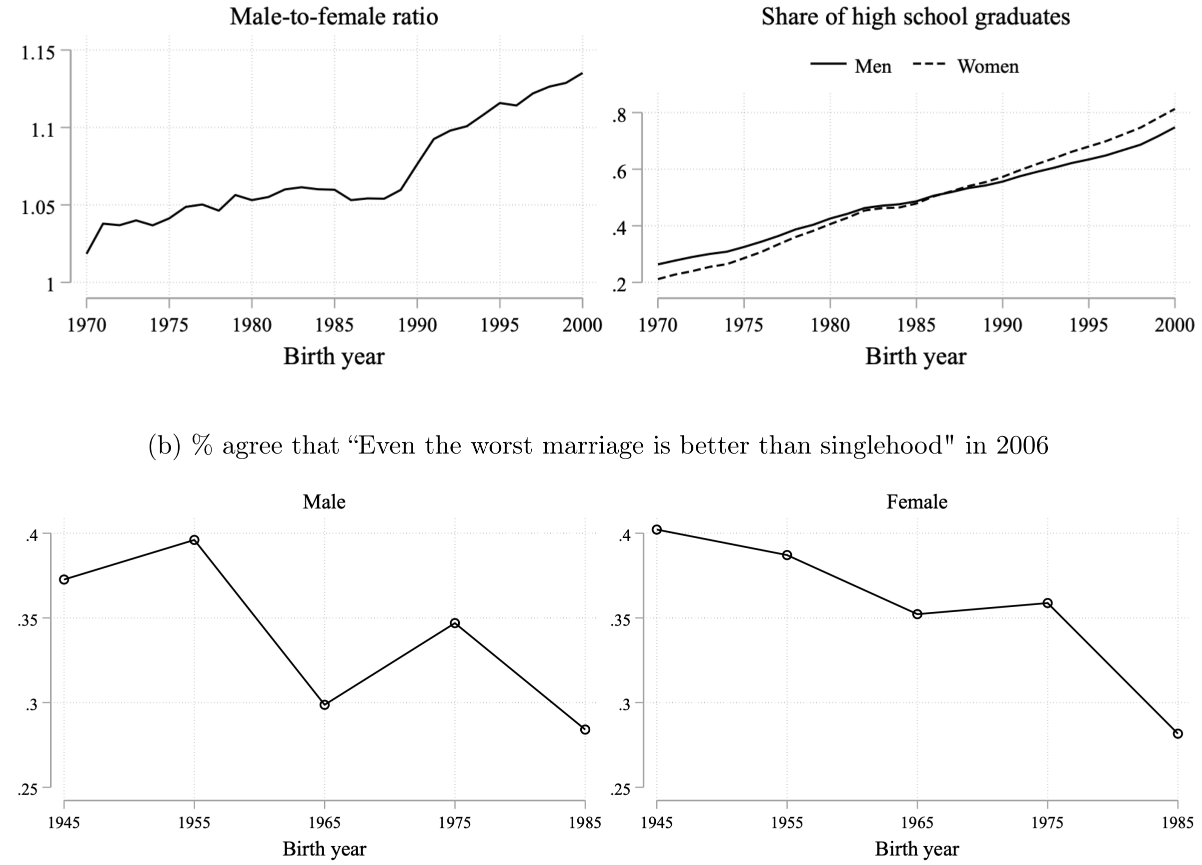
Wei, Shang-Jin, and Xiaobo Zhang. 2011. "The competitive saving motive: Evidence from rising sex ratios and savings rates in China." *Journal of political Economy* 119 (3): 511–564.

Yang, Wei, and Byron G Spencer. 2022. "In the name of the mother: Measuring the cultural value of family continuity in China." *Available at SSRN 4302537,* accessed October 10, 2023. https://ssrn.com/abstract=4554270.

You, Jing, Xuejie Yi, and Meng Chen. 2016. "Love, life, and "leftover ladies" in urban China."

Zhou, Xudong, and Therese Hesketh. 2017. "High sex ratios in rural China: declining well-being with age in never-married men." *Philosophical Transactions of the Royal Society B: Biological Sciences* 372 (1729): 20160324.

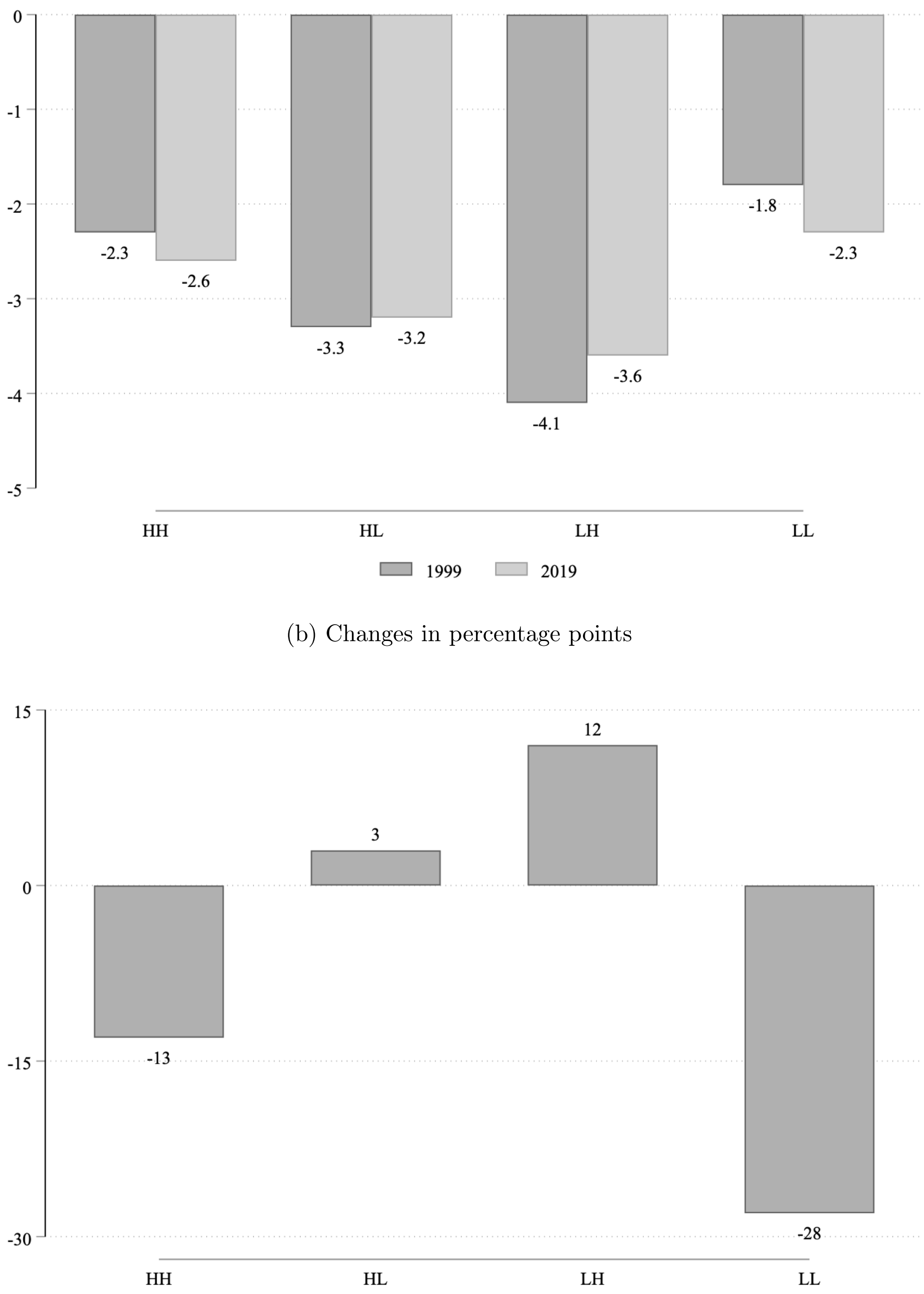
Figure 1: Population structure and attitudes toward marriage, by cohort


(a) Population structure

(b) % agree that “Even the worst marriage is better than singlehood" in 2006

Note: Figure (a) shows the male-to-female ratio (left panel) and share of high school educated by gender (right panel) by birth year, calculated with the aggregated statistics from the 2020 Census. Figure (b) shows the share of male respondents (right panel) and female respondents (left panel) agreeing with the statement "Even the worst marriage is better than singlehood" in CGSS 2006. The dot for birth year t represents the average over all individuals born between $t-9$ and t. For example, the number for birth year 1945 represents the average over individuals born between 1936 and 1945. Cohorts active on the marriage market in 1999 are born in 1980 and before. Cohorts entering the market between 1999 and 2019 are born between 1981 and 2000.

Figure 2: Estimated marriage surplus, by marriage type



(a) Levels in 1999 and 2019

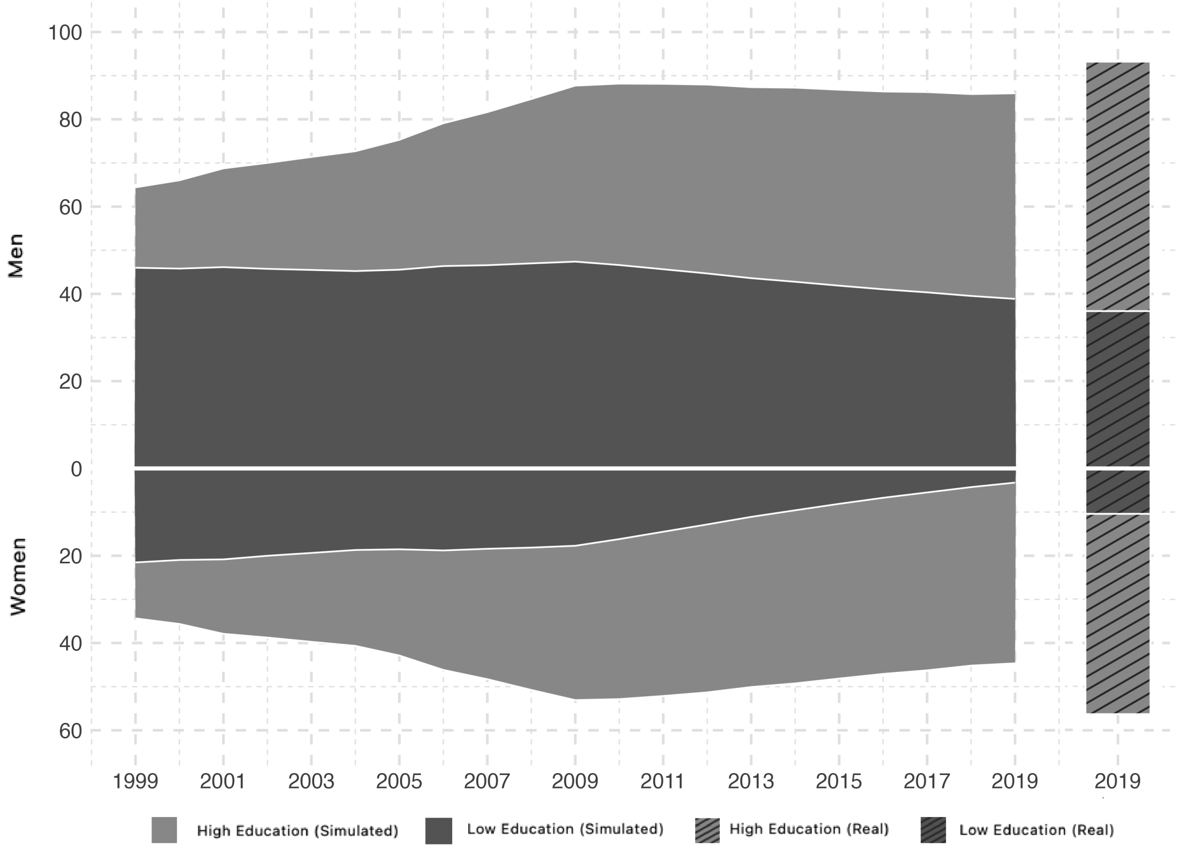
(b) Changes in percentage points

Note: Figure (a) shows the *marriage surplus* π_{ij} in 1999 and 2019, defined as the total systematic gain to marriage per partner for a pair (i, j) relative to the total systematic gain per partner from remaining unmarried. Figure (b) shows the change in marriage surplus between 1990 and 2019, in percentage points.

- HH: marriages between H-type men and H-type women (19% of marriages in 1999 and 57% in 2019)
- HL: marriages between H-type men and L-type women (10% of marriages in 1999 and 11% in 2019)
- LH: marriages between L-type men and H-type women (5% of marriages in 1999 and 14% in 2019)
- LL: marriages between L-type men and L-type women (66% of marriages in 1999 and 18% in 2019)

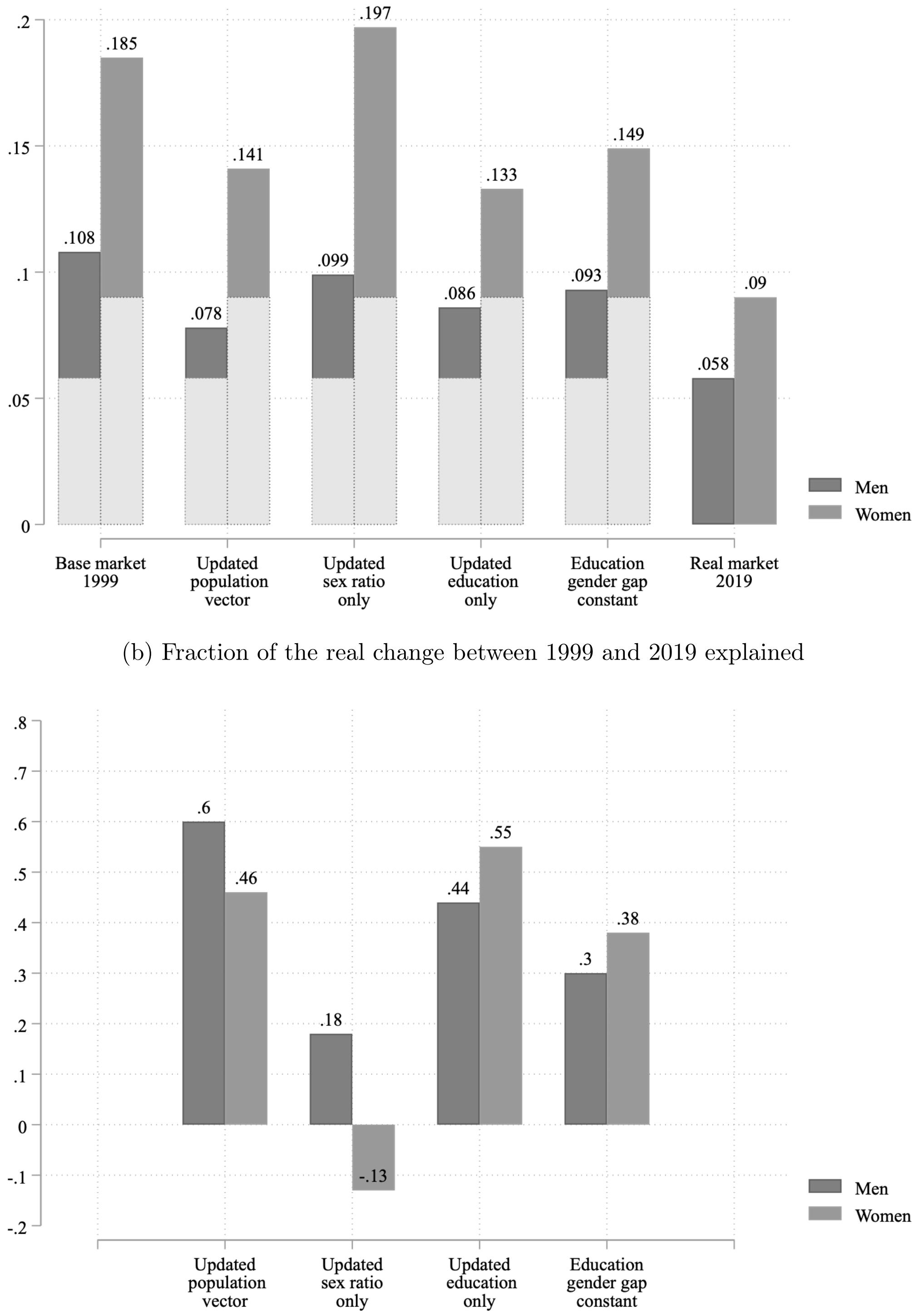
H-type means "ever attending high school" and L-type means "never attending high school". See Section 2 and Appendix C.1 for details on the methodology.

Figure 3: Number of singles in the counterfactual market keeping surplus constant at 1999 levels



Note: The figure plots the evolution of the number of singles in millions, by gender and education type, in the counterfactual market. We keep the surplus constant at 1999 levels and simulate the evolution of the market year-by-year. The bar at the extreme right represents the real number of singles observed in 2019. See Section 2 and Appendix C.2 for details on the methodology.

Figure 4: Counterfactual matching rates keeping surplus constant at 1999 levels



(a) Predicted matching rates

(b) Fraction of the real change between 1999 and 2019 explained

Note: See Section 2 and Appendix C.2 for details on the methodology. Each counterfactual market is characterized by the following vector of incoming cohorts and matching surplus:

	% H-type men	% H-type women	Sex ratio	Surplus
Base market 1999	0.299*	0.255*	1.03*	π_{1999}
1: Update population vector	Men_{t-19}	$Women_{t-19}$	$Cohort_{t-19}$	π_{1999}
2: Update sex ratio only	0.299*	0.255*	$Cohort_{t-19}$	π_{1999}
3: Update education only	Men_{t-19}	$Women_{t-19}$	1.03*	π_{1999}
4: Keep education gender gap constant	Men_{t-19}	Men_{t-19}-0.044*	1.03*	π_{1999}
Real market 2019	Men_{t-19}	$Women_{t-19}$	$Cohort_{t-19}$	π_{2019}

* mean of cohorts active in the 1999 market (born in 1966-1980)

Appendix A Choo and Siow model of matching markets

This appendix draws upon Vallette (2023) which summarizes the model developed by Choo and Siow (2006a, 2006b). This is a model of the marriage market with transferable utility and no friction: everyone can meet any potential partner, access all pertinent information, and transfer utility to their partners without any restriction.

Let us suppose there are I types of men and J types of women, where each type is defined by the intersection of observable characteristics. In the original framework, the types are constrained to be discrete; then, some extensions have allowed for continuous types. The number of types, denoted by I and J, is finite. Each individual belongs to a specific type (and only one).

A matching function $\mu(M, F : \Pi)$ relates the marriage distribution μ to population vectors M and F, and to marriage surplus Π. μ is represented as a matrix $(I+1) \times (J+1)$, where each element μ_{ij} indicates the number of matches between a man of type $i \in \{1, \ldots, I\}$ and a woman of type $j \in \{1, \ldots, J\}$ during a given time period. μ_{i0} and μ_{0j} represent the number of unmatched men of type i and unmatched women of type j, respectively. This matching function is deemed "feasible" if all individuals in the market are matched with either 0 or 1 partner, and "stable" if no matched individual would prefer to be single, and if no pair of individuals would prefer to be matched together instead of with their current partner.

Feasibility requires the matching function to satisfy the following conditions:

$$\mu_{i0} + \sum_{j=1}^{J} \mu_{ij} = m_i, \forall i, \tag{1}$$

$$\mu_{0j} + \sum_{i=1}^{I} \mu_{ij} = f_j, \forall j, \tag{2}$$

$$\mu_{0j}, \mu_{i0}, \mu_{ij} \geq 0, \forall i, j. \tag{3}$$

where m_i is the number of men of type i and f_j is the number of women of type j.

The utility of man s of type i marrying a woman of type j, V_{ijs} and the utility of a woman v of type j marrying a i-type man, V_{ijv}, are given by:

$$\begin{aligned} V_{ijs} &= \alpha_{ij} - \tau_{ij} + \epsilon_{ijs} \\ V_{ijv} &= \gamma_{ij} + \tau_{ij} + \epsilon_{ijv} \end{aligned} \tag{4}$$

The utility consists of three parts: (1) a gross systematic payoff α_{ij} or γ_{ij}, which is exogenous and depends only on the types of both spouses; (2) a (possibly negative) transfer τ_{ij} from i-type men to j-type women, which is determined endogenously at equilibrium; (3) an individual-specific random component ϵ_{ijs} or ϵ_{ijv}, which captures unobserved characteristics/preferences of individual s or v.

If they stay single, individuals receive a type-specific payoff as well as an individual-specific payoff. The singlehood payoffs for the i-type man s and the j-type woman v are given by:

$$\begin{aligned} V_{i0s} &= \alpha_{i0} + \epsilon_{i0s} \\ V_{0jv} &= \gamma_{0j} + \epsilon_{0jv} \end{aligned} \tag{5}$$

Thus, individuals will choose according to:

$$\begin{aligned} V_{is} &= \max_{y}\{V_{i0s}, ..., V_{ijs}, ..., V_{iJs}\} \\ V_{jv} &= \max_{x}\{V_{0jv}, ..., V_{ijv}, ..., V_{Ijv}\} \end{aligned} \tag{6}$$

There are two important assumptions on the unobserved heterogeneity components ϵ. First, conditional on (i, j), the joint surplus created by a match between a i-type man and a j-type woman does not depend on interactions between their unobserved characteristics. In other words, the match can happen because the unobservable characteristics of this specific woman make her attractive for all men of group i or because this specific man has strong unobserved preferences for all women of group j. What is ruled out by assumption in the model is that the match happens because this specific man has unobserved preferences for unobserved characteristics of this specific woman. This is known as the *separability* assumption. Second, the ϵ terms are independent and identically distributed according to the Type I Extreme Value distribution (McFadden 1973). This parametric assumption imposes restrictions on the patterns of substitution between the different choices of partners. Galichon and Salanié (2022) show that the second assumption can be relaxed.

Knowing the distribution of ϵ, we can deduce the quasi demand and supply for all types on the market. Choo and Siow (2006b), building on earlier work by McFadden (1973), show that the quasi demand equation by men of type i for women of type j is given by:

$$\ln \mu_{ij}^{d} = ln \mu_{i0}^{d} + \alpha_{ij} - \alpha_{i0} - \tau_{ij} \tag{7}$$

where $\ln \mu_{ij}^d$ is the demand for j-type women by i-type men and $ln\mu_{i0}^d$ is the number of unmarried i-type men. The quasi demand is decreasing with τ_{ij}. Similarly, they show that the quasi supply equation of women of type j for men of type i is given by:

$$\ln \mu_{ij}^s = ln\mu_{0j}^s + \gamma_{ij} - \gamma_{0j} + \tau_{ij} \tag{8}$$

where $\ln \mu_{ij}^s$ is the supply of j-type women for i-type men and $ln\mu_{0j}^s$ is the number of unmarried j-type women. The quasi supply is increasing with τ_{ij}.

The marriage market clears when the demand for women by men is equal to the supply of women for men for all types, given the equilibrium transfers τ_{ij}. That is, for all i, j, $\mu_{ij} = \mu_{ij}^d = \mu_{ij}^s$. Using this condition and summing the last two demand and supply equations, we get:

$$\ln \mu_{ij} - \frac{\ln \mu_{i0} + \ln \mu_{0j}}{2} = \frac{\alpha_{ij} - \alpha_{i0} + \gamma_{ij} - \gamma_{0j}}{2} \tag{9}$$

Let us denote $\pi_{ij} = \frac{\alpha_{ij} - \alpha_{i0} + \gamma_{ij} - \gamma_{0j}}{2}$, and call it the *marriage surplus.* π_{ij} has a clear interpretation: it captures the per-capita systematic net gains to marriage (relative to remaining single) for a couple consisting of an i-type man and j-type woman. We then have the following equality:

$$\pi_{ij} = \ln(\frac{\mu_{ij}}{\sqrt{\mu_{i0}\mu_{0j}}}) \tag{10}$$

At equilibrium, this simple equation links the marriage surplus to the number of newlywed couples divided by the geometric average of men and women who stay unmarried. Choo and Siow (2006b) conclude that this ratio of observable marriage market outcomes is a sufficient statistic for quantifying the quality of marriage matches.

$\pi_{i,j}$ is determined by exogenous preference parameters (α, γ) and is unaffected by changes in the quantities of men and women of different types. What does depend on the population vector is the transfer between spouses, which guarantees that the demand for a spouse of a given type equals the supply. Allowing for *spillover effects* is an important contribution of this model compared to previous quantitative models of the marriage market: matching outcomes between i-type men and j-type women are affected not only by the quantities of i-type men and j-type women but also by the quantities of men and women of all other types.

Appendix B Datasets and variables

Appendix B.1 Marriage market in 1999

We construct the marriage market in 1999 using the 1% sample of the 2000 Census (Minnesota Population Center 2020). We restrict the sample to individuals aged at least 19, hence adults (18 and older) at the beginning of 1999. The marriage market in 1999 consists of individuals who are single at the start of 1999; in 2000, these individuals are either still single or newlywed couples who married in 1999 and 2000. We identify the latter group by exploiting information on the year of the first marriage. We classify individuals as matched in 1999 if they married during that year, and as unmatched if they married in 2000 or remained single in 2000. With this definition, we calculate the annual matching rate for both men and women shown in Figure 4a (base market 1999).

To estimate the marital surplus in 1999, we need to know the number of matches by education of both spouses. We combine (i) information on demographic characteristics for everyone living in the same household and (ii) information on each member's relationship to the household head in order to identify the couples within a household. We proceed in three steps. First, we match the household head with the spouse of the household head and identify 34,257 couples who married in 1999. Second, we match the child of the household head with the child-in-law of the household head. We identify 20,851 additional couples who married in 1999. Third, we match the parents of the household head; however, and not surprisingly, none of these couples married in 1999. With this procedure, we have 55,108 matched couples, which account for 70% of the 78,513 men and 77,803 women who married in 1999 according to the census. The educational distributions in our sub-sample of matched couples closely resembles the distributions in the entire sample of men and women who married in 1999. Appendix C.1 shows the distribution of the four different types of matches in 1999 based on the sub-sample of matched couples.

The remaining unmatched couples correspond to the following situations. First, a couple does not live in the same household; this is relatively rare, e.g. 9% of married household heads do not have a spouse in the household. Second, the couple resides in a collective household (larger dwelling units with several individuals who are not related by family links); this is also rare, e.g. about 3% of individuals aged 19 and above live in collective households and only one-third among them are married. Third, household heads have multiple children or children-in-law; this is the main reason why we cannot identify the couples correctly. We tried to increase the number of matched couples by making assumptions on who is married with whom among the multiple children.[8] However, we ended up with a sub-sample of matched couples that was less comparable to the entire sample of couples in terms of education. We therefore believe that maximizing the fraction of couples matched can be counterproductive in terms of representativeness. Our main assumption in constructing the distribution by type of marriage is that couples who can be properly matched with our 3-step procedure are representative of all couples.

Appendix B.2 Marriage market in 2019

We construct the marriage market in 2019 using the aggregated statistics of the 2020 Census (National Bureau of Statistics of China 2021). The marriage market in 2019 consists of those who married in 2019 and 2020, as well as those reported as single in 2020. The census aggregates report the number of singles in 2020 and the number of individuals married in 2019 and 2020 by gender. Hence, we can calculate the annual matching rate in 2019 for both genders, as reported in Figure 4a (real market 2019).

To derive the real vector of singles reported in Figure 3, we require additional data on the number of individuals by educational levels. This information is only available for the singles in 2020, not for those who married in 2019 and 2020. Thus, we approximate our real vector of singles on the 2019 market using the observed vector of singles in 2020. Considering the smooth evolution of population structure across incoming cohorts (see Figure 1a), we expect both vectors to be similar.

The aggregated data of the 2020 Census does not allow us to identify the spouse. Instead, we turn to the 2020 wave of the China Family Panel Studies (CFPS) for this purpose. Once again, we restrict the sample to those aged 19 and above in 2020, who reported being single in 2020 or entered their first marriage in 2019 and 2020. Those married in 2019 are considered as being matched on the 2019 market. We do not have to identify the couples in CFPS: the survey directly lists the spouse's attributes of all individuals. We can therefore compute the distribution of the types of matches on the entire sample of couples (see Appendix C.1).

[8] This is the approach of Brandt et al. (2018). For instance, they match individuals within a household based on marriage duration or they draw "replacement" spouses from the age-education distribution observed among the matched couples.

Appendix B.3 Counterfactual stock of singles

When we model the evolution of the stock of singles in Figure 3, we need to take into account the possibility that some of the previously unmatched individuals die. We calculate the death rates among singles using the 2000 census, by gender: the annual death rate is equal to 0.3% for single men and 0.08% for single women.

Appendix C Matrices

Appendix C.1 Real markets

Surplus in 1999 calculated using 2000 census data

$$\begin{array}{cccc} & \text{Women} & & \\ \text{Men} & \begin{pmatrix}1.5 & 0.7\\ 0.4 & 5.2\end{pmatrix} & \begin{array}{c}18.4\\ 46.0\end{array} & \xrightarrow{\pi_{i,j}=ln(\frac{\mu_{i,j}}{\sqrt{\mu_{i0}\mu_{0j}}})} \begin{pmatrix}-2.3 & -3.3\\ -4.1 & -1.8\end{pmatrix} \\ & \begin{array}{cc}12.8 & 21.5\end{array} & & \end{array}$$

Numbers for the real market in 1999 are based on the 1% sample of the 2000 Census and expressed in millions of individuals. We multiply the number of observations in the sample by 100 to get an estimate of the total number of individuals in the population.

Surplus in 2019 calculated using 2020 CFPS data

$$\begin{array}{cccc} & \text{Women} & & \\ \text{Men} & \begin{pmatrix}4.1 & 0.8\\ 1.0 & 1.3\end{pmatrix} & \begin{array}{c}60.5\\ 26.8\end{array} & \xrightarrow{\pi_{i,j}=ln(\frac{\mu_{i,j}}{\sqrt{\mu_{i0}\mu_{0j}}})} \begin{pmatrix}-2.6 & -3.2\\ -3.6 & -2.3\end{pmatrix} \\ & \begin{array}{cc}50.0 & 5.6\end{array} & & \end{array}$$

Numbers for the real market in 2019 are based on the 2020 wave of CFPS and expressed in millions of individuals. We multiply the number of observations in the sample by 54000 to get an estimate of the total number of individuals in the population.

Note that, unlike the 2020 census which covers 100% of the population, CFPS is representative of 95% of the Chinese population: specific provinces like Tibet and Inner Mongolia are not included. This is why we use CFPS data only to estimate the surplus in 2019 and to discuss the evolution of surplus over time. We prefer to use the 2020 census to compute the characteristics of the incoming cohorts and to implement the counterfactual analysis.

Appendix C.2 Simulated markets

We start with the number of unmatched individuals at the end of 1999 as shown in Appendix C.1. We fix π at the 1999 level. We consider the following four simulations for the incoming cohort in year 2000 (cohort born in 1981):

	Nb men	Nb women	% H men	% H women	SR	Surplus
1: Update population vector	9.8	9.3	0.442	0.428	1.05	π_{1999}
2: Update sex ratio only	9.8	9.3	0.299*	0.255*	1.05	π_{1999}
3: Update education only	9.8	9.5	0.442	0.428	1.03*	π_{1999}
4: Keep education gender gap constant	9.8	9.5	0.442	0.398	1.03*	π_{1999}

After running the CS algorithm, we obtain the number of unmatched individuals at the end of 2000. We repeat the same exercise incorporating new cohorts one by one.

Simulation 1: update population vector

Numbers for the incoming cohort are observed in the 2020 census aggregates and expressed in millions of individuals.

$$\begin{aligned}
&(1) \quad \mu'_{HH} + \mu'_{HL} + \mu'_{H0} = 18.4 + 4.3(= 9.8 \times 0.442)\\
&(2) \quad \mu'_{LH} + \mu'_{LL} + \mu'_{L0} = 46.0 + 5.5(= 9.8 \times (1 - 0.442))\\
&(3) \quad \mu'_{HH} + \mu'_{LH} + \mu'_{0H} = 12.8 + 4.0(= 9.3 \times 0.428)\\
&(4) \quad \mu'_{HL} + \mu'_{LL} + \mu'_{0L} = 21.5 + 5.3(= 9.3 \times (1 - 0.428))\\
&(5) \quad {\mu'_{HH}}^2 = \mu'_{H0}\mu'_{0H}exp(2 \times -2.3)\\
&(6) \quad {\mu'_{HL}}^2 = \mu'_{H0}\mu'_{0L}exp(2 \times -3.3)\\
&(7) \quad {\mu'_{LH}}^2 = \mu'_{L0}\mu'_{0H}exp(2 \times -4.1)\\
&(8) \quad {\mu'_{LL}}^2 = \mu'_{L0}\mu'_{0L}exp(2 \times -1.8)
\end{aligned}$$

$$\xrightarrow{\text{CS algorithm}} \begin{array}{cc} \begin{pmatrix} 1.7 & 0.8 \\ 0.4 & 5.1 \end{pmatrix} & \begin{matrix} 20.2 \\ 45.8 \end{matrix} \\ \begin{matrix} 14.7 & 21.0 \end{matrix} & \end{array}$$

Simulation 2: update sex ratio only

We keep the total number of women and men constant as observed in the census and we change the fraction of H-type to reflect the mean of cohorts active in the 1999 market (born in 1966-80).

$$\begin{array}{ll}
(1) & \mu'_{HH}+\mu'_{HL}+\mu'_{H0} = 18.4 + 2.9(= 9.8 \times 0.299) \\
(2) & \mu'_{LH}+\mu'_{LL}+\mu'_{L0} = 46.0 + 6.9(= 9.8 \times (1-0.299)) \\
(3) & \mu'_{HH}+\mu'_{LH}+\mu'_{0H} = 12.8 + 2.4(= 9.3 \times 0.255) \\
(4) & \mu'_{HL}+\mu'_{LL}+\mu'_{0L} = 21.5 + 6.9(= 9.3 \times (1-0.255)) \\
(5) & {\mu'_{HH}}^2 = \mu'_{H0}\mu'_{0H}exp(2 \times -2.3) \\
(6) & {\mu'_{HL}}^2 = \mu'_{H0}\mu'_{0L}exp(2 \times -3.3) \\
(7) & {\mu'_{LH}}^2 = \mu'_{L0}\mu'_{0H}exp(2 \times -4.1) \\
(8) & {\mu'_{LL}}^2 = \mu'_{L0}\mu'_{0L}exp(2 \times -1.8)
\end{array}
\xrightarrow{\text{CS algorithm}}
\begin{array}{cc}
\begin{pmatrix} 1.6 & 0.8 \\ 0.4 & 5.3 \end{pmatrix} & \begin{array}{c} 19.0 \\ 47.0 \end{array} \\
\begin{array}{cc} 13.2 & 22.4 \end{array} &
\end{array}$$

Simulation 3: update education only

We keep the total number of men constant as observed in the census and we change the number of women to reflect the sex ratio among cohorts active in the 1999 market (born in 1966-80). We keep the fraction of H-type as observed in the census.

$$\begin{array}{ll}
(1) & \mu'_{HH}+\mu'_{HL}+\mu'_{H0} = 18.4 + 4.3(= 9.8 \times 0.442) \\
(2) & \mu'_{LH}+\mu'_{LL}+\mu'_{L0} = 46.0 + 5.5(= 9.8 \times (1-0.442)) \\
(3) & \mu'_{HH}+\mu'_{LH}+\mu'_{0H} = 12.8 + 4.1(= 9.5 \times 0.428) \\
(4) & \mu'_{HL}+\mu'_{LL}+\mu'_{0L} = 21.5 + 5.4(= 9.5 \times (1-0.428)) \\
(5) & {\mu'_{HH}}^2 = \mu'_{H0}\mu'_{0H}exp(2 \times -2.3) \\
(6) & {\mu'_{HL}}^2 = \mu'_{H0}\mu'_{0L}exp(2 \times -3.3) \\
(7) & {\mu'_{LH}}^2 = \mu'_{L0}\mu'_{0H}exp(2 \times -4.1) \\
(8) & {\mu'_{LL}}^2 = \mu'_{L0}\mu'_{0L}exp(2 \times -1.8)
\end{array}
\xrightarrow{\text{CS algorithm}}
\begin{array}{cc}
\begin{pmatrix} 1.7 & 0.8 \\ 0.4 & 5.1 \end{pmatrix} & \begin{array}{c} 20.2 \\ 45.8 \end{array} \\
\begin{array}{cc} 14.7 & 21.1 \end{array} &
\end{array}$$

Simulation 4: keep education gender gap constant

We keep the total number of men and the fraction of H-type men constant as observed in the census and we change: (i) the number of women to reflect the sex ratio among cohorts active in the 1999 market and (ii) the fraction of H-type among women to reflect the gender gap in education (4.4p.p.) among cohorts active in the 1999 market.

$$\begin{array}{ll}
(1) & \mu'_{HH} + \mu'_{HL} + \mu'_{H0} = 18.4 + 4.3(= 9.8 \times 0.442) \\
(2) & \mu'_{LH} + \mu'_{LL} + \mu'_{L0} = 46.0 + 5.5(= 9.8 \times (1 - 0.442)) \\
(3) & \mu'_{HH} + \mu'_{LH} + \mu'_{0H} = 12.8 + 3.8(= 9.5 \times 0.398) \\
(4) & \mu'_{HL} + \mu'_{LL} + \mu'_{0L} = 21.5 + 5.7(= 9.5 \times (1 - 0.398)) \\
(5) & {\mu'_{HH}}^2 = \mu'_{H0}\mu'_{0H} exp(2 \times -2.3) \\
(6) & {\mu'_{HL}}^2 = \mu'_{H0}\mu'_{0L} exp(2 \times -3.3) \\
(7) & {\mu'_{LH}}^2 = \mu'_{L0}\mu'_{0H} exp(2 \times -4.1) \\
(8) & {\mu'_{LL}}^2 = \mu'_{L0}\mu'_{0L} exp(2 \times -1.8)
\end{array}$$

$$\xrightarrow{\text{CS algorithm}} \begin{array}{cc} \begin{pmatrix} 1.7 & 0.8 \\ 0.4 & 5.1 \end{pmatrix} & \begin{array}{c} 20.2 \\ 45.8 \end{array} \\ \begin{array}{cc} 14.5 & 21.3 \end{array} & \end{array}$$