Preferences and Choice Constraints in Marital Sorting: Evidence From Korea

Soohyung Lee[1]
Department of Economics
University of Maryland, College Park
LeeS@econ.umd.edu

This Version: July, 2008

Abstract

Marital sorting along education, income and other salient dimensions is well-documented for many countries. Understanding the mechanisms behind such sorting is important because the degree of marital sorting may influence income inequality, intergenerational mobility, and household labor supply, as well as other economic outcomes. Marital sorting is often thought to arise from some combination of people's preferences and constraints on their choice sets. However, separating these two causes of marital sorting is difficult because typical data sets provide information on either a person's spouse or a person's dating partners, but not both. This paper circumvents this difficulty by using a novel data set from a major Korean matchmaking company which contains both types of information. The paper analyzes gender-specific marital preferences by estimating a marriage model. Using the estimated model, I find that constraints on people's choice sets may account for a substantial fraction of observed sorting along education and industry in the general population. The recent development of new search technologies, such as online dating services, alleviates these constraints and thus may reduce marital sorting along these dimensions. I also find evidence that changing individual-level income inequality has a very limited impact on marital sorting, implying that such changes are unlikely to be amplified at the household-level by endogenous marital sorting.

[1]I thank Pete Klenow, Luigi Pistaferri, John Pencavel and Michèle Tertilt for their advice and support throughout this project. I have benefited from discussions with Ran Abramitzky, Manuel Amador, Takeshi Amemiya, Orazio Attanasio, Doug Bernheim, Nick Bloom, Tim Bresnahan, David Card, Giacomo De Giorgi, Raquel Fernández, Doireann Fitzgerald, Chris Flinn, Bob Hall, John Hatfield, Han Hong, Joseph Hotz, Caroline Hoxby, Erik Hurst, Nir Jaimovich, Seema Jayachandran, Jakub Kastl, John Knowles, Yuan Chuan Lien, Tom MaCurdy, Aprajit Mahajan, Ben Malin, Ted Miguel, Sri Nagavarapu, Muriel Niederle, Minjung Park, Alex Ponce-Rodriguez, Felix Reichling, Azeem Shaikh, Frank Wolak, Joanne Yoong and participants of the labor and development reading group and macro bag lunch at Stanford University. I thank Ken Judd, Hyunok Lee and Zsolt Sándor for sharing their computational expertise; and the B.F. Haley and E.S. Shaw Fellowship and Stanford Graduate Research Opportunity Fellowship for financial support. I am indebted to Woong Jin Lee, Heui Gil Lee, Kang Yong Ahn, and Hye-Rim Kim for sharing the data.

1 Introduction

Sorting in marriages along age, education, income and other salient dimensions is well documented for many countries.[2] Understanding the mechanisms behind such sorting is important because the degree of marital sorting may influence income inequality, intergenerational mobility, and household labor supply, as well as other economic outcomes. Using a novel data set from a major Korean matchmaking company, this paper addresses the following questions: (1) How do people value various attributes such as education, income or even beauty when seeking a spouse? (2) How do changes in people's choice sets affect marital sorting?

The key findings of the paper are as follows:

- People consider a large number of attributes when choosing a spouse. Men and women value given attributes differently, but in general people prefer partners who are similar to themselves. Somewhat surprisingly, preferences inferred from first-date outcomes are highly predictive of final marriage decisions.
- Conditioning on these preferences, simulation results show that expanding people's choice sets can significantly reduce sorting observed in the general population along education, industry and geographic location. This suggests that constraints on choice sets may account for a substantial fraction of observed sorting along these dimensions, although the same is not true along other dimensions, such as age and marital history. In addition, changing individual-level income inequality has a very limited impact on marital sorting, implying that such changes are unlikely to be amplified at the household-level by endogenous marital sorting.

To answer the questions posed above, one needs to disentangle the different mechanisms underlying sorting. Separately identifying these mechanisms is often difficult. Consider, for example, sorting along education. Such sorting can arise because of people's preferences for education. Alternatively, people may only have the opportunity to meet potential spouses with similar educational backgrounds (choice-set constraints).[3] Different underlying explanations may imply different responses to changes in economic conditions that affect people's choice sets.

Consider the introduction of new search technologies such as online dating or matchmaking services that have been rapidly increasing in popularity, in the United States and elsewhere.

[2] See Blossfeld and Timm (2003) or Kalmijn (1998) for a detailed survey of these findings

[3] Sorting along education may also arise from people having preferences for attributes that happen to be correlated with education (e.g., income), even if people do not care for education per se. This possibility is examined in Section 6.

These services generally allow users to view a variety of potential spouses using a large-scale online database, leading to an expansion of the users' choice sets. If sorting is entirely the result of preferences, then the adoption of new search technologies will have little effect on sorting along education. However, if sorting is largely due to constraints on individuals' choice sets, then such technologies may significantly change the degree of sorting.

In principle, it is possible to distinguish between marital preferences and choice-set constraints only by examining both people's choice sets (who they date) and their marriage decisions (who they marry). Suppose male college graduates date women regardless of their educational attainment in order to explore whether they would be a good match, but tend to eventually marry college graduates. If only final choices are observed, it is not possible to determine whether sorting by education is due to preferences or due to choice-set constraints. On the other hand, observing only dates can identify preferences for dating partners but not necessarily preferences for marriage. In this example, analyzing only the dating behavior of male college graduates would lead to the conclusion that they do not value spousal education. The proper conclusion, that male college graduates prefer to marry similarly-educated women, can be reached only by observing both dates and marriages.

In the previous empirical literature, researchers have generally been able to observe either people's choice sets or final marital outcomes, but not both. For instance, typical population-based data sets, such as the Census or household surveys, do not provide information about people's choice sets. On the other hand, context-specific data from speed-dating experiments and online dating services may provide information about people's choice sets, but not about their ultimate choice of spouse.

This paper overcomes this difficulty by exploiting an unusually rich data set from a major matchmaking company in Korea. The data set provides detailed information on over 20,000 users, 13.4 percent of whom get married through the service. In particular, the data includes not only information about whom each user dated and ultimately married, but also information about proposed dates that were turned down. A second important feature of this data set is that users can search for a spouse from a wide spectrum of potential spouses in terms of age, education, geographic location, and many other dimensions. Sorting among users is thus more likely to reflect users' preferences as opposed to constraints on their choice sets. These features of the data allow me to identify people's marital preferences over a wide variety of characteristics.

I develop a model of dating and marriage choices based on a random-effects probit specification, which I extend to allow for the possibility that people have multiple dates with the

same dating partner. Within my model, multiple dates result from a desire to learn more about one's dating partners. In order to estimate the model, I use a Laplace-type estimator, which relies on Markov Chain Monte Carlo methods.

I find that people consider a large number of traits when choosing a spouse, and they value similarity to themselves for many of these traits. However, overall preference rankings are determined by weighing the value of similarity against the benefit from having a partner of a "better" type. For specific characteristics, the latter effect is dominant. For example, people value having a partner with similar physical attractiveness or education, but all men and women unanimously prefer a partner with better appearance. In some cases, this offsetting effect is gender-specific. Male high school graduates prefer female high school graduates, while male college graduates prefer female college graduates; on the other hand, all women prefer male college graduates regardless of their own educational attainment.

I then examine how strongly first-date outcomes reflect marital preferences. To do so, I re-estimate my model using only first-date outcomes, and then using only first- and second- date outcomes. In both cases, sorting predicted by these two models is very close to that predicted by the model using all match outcomes, including the marriage decision. This suggests that in a setting where people are seriously searching for a spouse, analyzing first-date outcomes can be sufficient to identify their marital preferences.

Next, I use the estimated marital preferences to address the question: In the general population, how do changes in people's choice sets affect marital sorting? Preferences estimated using the matchmaking data set allow me to address this question correctly if there is no selection in terms of who uses the matchmaking services. While it is not possible to rule out all potential types of selection bias, I address the most likely sources in detail. In particular, people who use the service may be more or less motivated to marry than non-users, even after controlling for observables. I find evidence that bias resulting from such a scenario is unlikely to significantly affect the results. To simulate marital sorting, I use the Gale-Shapley algorithm (1962) on a random sample of users, weighted such that the distribution of characteristics matches the general Korean population.

I examine the importance of choice-set constraints by comparing marital sorting observed in the general population to sorting in simulated marriages under a fully integrated marriage market, in which people see all singles in the population. I consider six dimensions: age, marital history, education, industry, region and hometown. Sorting by age and marital history in the fully-integrated market is similar to the general population, but significantly less sorting along the remaining four dimensions is observed. The fraction of married couples with the same

education is reduced from 79 to 62 percent, while the fraction of married couples in the same industry falls from 36 to 13 percent. To understand what generates such differences, I allow the market to be segregated along the six dimensions and calibrate the degree of segregation such that the simulation results match the marital sorting observed in the population data. I find that the observed marital sorting can be generated in a marriage market that is partially segregated along four dimensions: education, industry, region, and hometown. This suggests that although preferences contribute to overall marital sorting, constraints on people's choice sets do account for observed sorting along these dimensions. New search technologies may therefore significantly reduce sorting along the latter dimensions. As a result, less sorting along education may increase intergenerational mobility, while less sorting along industry may reduce households' vulnerability to industry-level income shocks.

Finally, I examine some of the broader implications of marital sorting. In particular, the relationship between marital sorting and income inequality has been the focus of much previous discussion. In theory, greater marital sorting by income may multiply the effect of increases in individual-level income inequality, leading to even higher household-level income inequality. On the other hand, it is plausible to expect very little response in terms of sorting if income is negatively correlated with other important positively-valued traits. The overall strength and direction of this relationship is therefore ultimately an empirical question.

To address this question, I perform two experiments. In the first experiment, all people have the same income and parental wealth, effectively removing all individual-level income inequality. In the second, income-inequality is increased via a large increase in the returns to college education. In either case, I find that marital sorting along education, industry, age and other dimensions changes very little relative to sorting under the actual individual-level income inequality. This result holds regardless of whether or not marriage markets are assumed to be segregated. This finding suggests that changes in individual-level income inequality are unlikely to be amplified at the household-level by endogenous marital sorting.

This paper ties together several strands of research related to marriage. It is closely related to several recent empirical studies that estimate marital preferences based on speed dating experiments (Kurzban and Weeden, 2005; Fisman et al., 2006, 2007; Belot and Francesconi, 2006) or records from online dating services (Hitsch et al., 2006). In general, my work empirically supports this literature by suggesting that first-date outcomes may in fact be sufficiently informative for marital preferences, provided that individuals are genuinely interested in finding a spouse when dating. The overall analytical framework of this paper is most closely related to Hitsch et al. (2006) who use data from an online dating service to recover people's preferences for a first date. They then compare match outcomes simulated by the Gale-Shapley

algorithm to sorting in actual marriages in the United States. In this paper, I build upon the original contribution by Hitsch et al. (2006) in three important ways. Firstly, my analysis uses actual dating histories and realized marriages. Secondly, I extend the theoretical framework to include learning about types of partners over multiple dates. Finally, I add new counterfactual analyses based on differences in individual choice sets.

Recently, Banerjee et al. (2008) study the role of caste and other attributes in arranged marriages in India using data on a sample of parents of prospective brides and bridegrooms who placed matrimonial advertisements in a newspaper. They find strong preferences for the same caste among their sample. While there is no direct counterpart of caste in Korea, I find vertical preferences (preferences for better types) for some characteristics, such as income (for both) and education (for women), that may proxy for caste. On the other hand, I find strong preferences for similar characteristics along some dimensions, namely age and martial history.

A parallel literature analyzes the marriage market as an equilibrium model.[4] This work generally relies on population data. Since such data provides little information about people's actual choice sets, these papers are often forced to make strong assumptions about preferences and market segregation. My work uses estimates of marital preferences to support several of these findings, without imposing such assumptions. For example, Choo and Siow (2006) find that in the United States the gains from marriage generally decrease the further a couple deviates from a preferred age gap, a result that also follows from my analysis. Studies such as Angrist (2002) and Abramitzky et al. (2007) exploit exogenous shocks in the sex-ratio due to migration or war to examine changes in marital sorting. They find that a higher ratio of single men to single women raises the probability of women marrying men of higher social status and vice versa. I find that women in general prefer better educated men, which is consistent with their finding that an exogenous increase in the supply of men will lead to more women marrying such men.

The empirical estimates of people's marriage utility functions in this paper also complement both quantitative studies of household inequality and theoretical studies of matching markets.[5] For example, Pencavel (2006) studies the relationship between individual- and household-level

[4]Examples of such an approach include Wong (2003), Bisin et al. (2004), Choo and Siow (2006), Angrist (2002) and Abramitzky et al. (2007). One potential limitation of my work as compared to an equilibrium modeling approach is that my model assumes people do not care about rejection, ruling out strategic dating behavior. Suppose all men prefer beautiful women and all women prefer handsome men. Then, an average-looking man may reject a potential date with a beautiful woman if he expects and fears her rejection. Then, we cannot infer his preference rankings from his rejection of a proposed date. While my data do not allow me to directly test for strategic behavior, I find evidence suggesting that such behavior may not matter greatly.

[5]Examples of the literature studying household inequality include Kremer (1997), Greenwood et al. (2003), Fernández (2001, 2005), and Pencavel (1998, 2006). The matching literature includes Becker (1973, 1974), Burdett and Coles (1997), Shimer and Smith (2000), Gould and Paserman (2003), and Legros and Newman (2006).

income inequality, while assuming that changes in individual-level income inequality do not affect marital sorting. Results from simulations that I present below confirm the appropriateness of this assumption. Studies in the matching literature assume that a single-dimensional index adequately summarizes individuals' characteristics. However, my results show that individuals consider multiple dimensions of characteristics, suggesting the need to extend the theoretical analysis of matching to include this empirically important feature.

A brief overview of the remainder of this paper is as follows. Section 2 describes the matchmaking services industry in Korea and the data. Section 3 presents an empirical framework for estimation, identification and the estimation method. Section 4 provides the estimates of the model. I then discuss several potential issues in my analysis, such as selection bias, in Section 5. Section 6 provides results of counterfactual analyses. Section 7 concludes.

2 Industry and Data

2.1 Industry

The matchmaking industry consists of two types of providers: traditional matchmakers and corporations. Traditional matchmakers are individuals who act as sole proprietors. They typically find and match couples based on their personal connections. Individual matchmakers charge a fixed fee in advance and receive a bonus in the event that their services result in marriage. On the other hand, matchmaking companies emerged in the late 1980s and rapidly expanded their market. These matchmaking companies provide access to an internet database where users can browse one another's profiles and use a computerized algorithm to introduce singles to each other. These users are recruited through advertisements and pay a fixed advance fee for a pre-specified period, usually a year.[6]

The use of matchmaking services is common in Korea. According to the Korea Marriage Culture Institute (KMCI), 7.6 percent of couples who married in 2005 met through matchmaking companies.[7] Although use of matchmakers increases with age, the use of matchmaking companies is non-negligible even among the younger singles (see Table 1). Parents or relatives introduced an additional 12.6 percent of couples to their future spouses, often with the help of

[6] According to the Korea Consumer Association, in 2000, matchmaking companies charged a user 300,000 to 500,000 won, whereas individual matchmakers charged a fixed fee of about 300,000 won fee and an additional 2 million won upon marriage. Currently, 1,000 won is approximately equivalent to one U.S. dollar.

[7] The magnitude of using matchmaking services in Korea is high compared to the United States. According to Madden and Lenhart (2006), three percent of the sample of U.S. internet users met their spouse through the internet, including online dating services, and one percent of people met on a blind date or through a dating service.

traditional individual matchmakers. Similar results are found in a study of unmarried internet users conducted by a local research organization *Pollever*.

The number of people who use matchmaking companies is large and increasing. The Korea Consumer Association estimates that 60,000 people used matchmaking companies in 2000. Total sales of the four largest matchmaking companies was 24.3 billion Korean won in 2002 (approximately 24.3 million U.S. dollars), with average sales growth of 25.6 percent per year between 2000 and 2002 (Fair Trade Commission, 2004).

2.2 Data

The data for this study is obtained from one of the four main companies mentioned above. It contains $20,689$ individuals who used the company's services from January 2002 to June 2006 and provides information about each user's individual characteristics, stated marital preferences, and match history. These individual characteristics include socioeconomic and demographic characteristics, physical traits, and family background, described in Table 2. Stated marital preferences include his or her rating of the three most important traits for a prospective spouse and also the user's dislikes in terms of religion, hometown or region.[8] Finally, each user's match history consists of the set of all his or her proposed matches and the outcomes for each match. The match outcome for any pair of a man and a woman is characterized by up to three stages: whether or not each side of a pair wished to go on a first date, whether each of them wanted to have a second date conditional on having had a first date and finally, whether the users eventually married.

2.2.1 Motivation for Using the Matchmaking Service

It is reasonable to assume users are primarily motivated to seek marriage rather than casual dating. A membership contract, which guarantees service for one year, costs 900,000 won (as of July 2007).[9] This annual fee is about 3.5 percent of the average annual income in Korea. The fraction of users who married as a result of the matchmaking service is 13.4 percent, which is non-trivial.

[8]By region, I mean an area where a user currently lives, and by hometown I mean an area where the user grew up.

[9]This is approximately 900 US dollars. In contrast, online dating services in the United States, such as Yahoo Personals and eHarmony, currently cost about 160 to 250 dollars for a comparable one-year contract.

2.2.2 The Reliability of Information

The information about user characteristics is subject to several tests by the company itself. As far as possible, key information is legally verified or independently evaluated by the matchmaking company (see Table 2). The matchmaking company requires each user to submit legal documents in order to confirm primary information such as age, education, employment, marital status and the marital status of the user's parents. Staff members of the matchmaking company use submitted photographs to assign each user a facial grade, intended to represent attractiveness of a user's facial appearance to the opposite gender. It ranges from A (the most attractive) to F (the least attractive), and the majority of users are rated either B or C.

Self-reported user attributes that cannot be formally verified are monitored via user feedback. The company routinely surveys its users about their experiences and asks them to verify the correctness of other users' information. The matchmaking company's contract specifies that the service will be terminated if a user is found to provide incorrect information. As a reasonable test, I compare the self-reported user information to the Korean population at large in the next section and find that these attributes are comparable.

2.2.3 Comparison between Users and the General Population

As no single population-based data set captures all the features observed in my data, I use four separate nationally-representative data sets. The official marriage register (MR) is the closest analogue to the matchmaking data set. The annual MR lists all couples who report their marriage to the Korean government during that year. The MR provides a married couple's demographic information such as age, educational attainment and marital history. However, it does not contain economic information such as income and industry of employment. Thus I use the MR as the baseline data for drawing comparisons to the general population, and supplement the analysis with three other data sets: the Basic Statistics Survey of Wage Structure (WS) for industries and income, the National Household Income and Expenditure Survey (HIS) for income of husbands and wives and the Survey of Physical Traits of Koreans (PT) for height and weight (see Table 3 for a comparison of information available in these four data sets). The matchmaking data set contains a wide spectrum of Koreans. I classify the general population in the MR into 170 cells defined by gender, age, education, region and hometown. The users in my data set are distributed across 48 percent of those cells which includes 70 percent of the Korean population.

The distribution of the users' characteristics is noticeably different from the general population in three ways. First, users are, on average, older than the couples in the MR. The

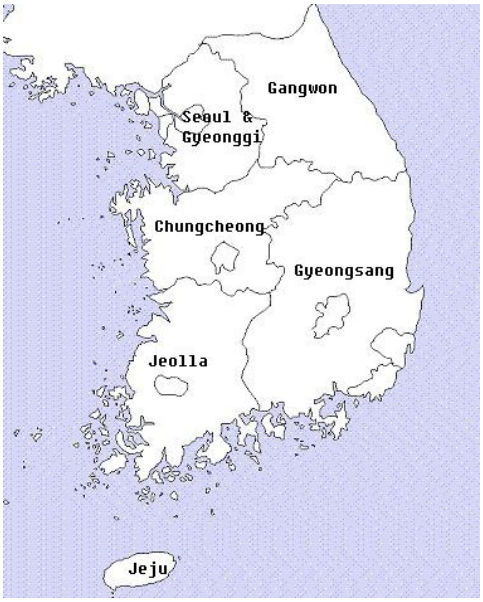

Figure 1: Regions of South Korea

average age of users is 33.4 for men and 29.9 for women, whereas the average age in the MR is 30.9 for men and 27.8 for women. Second, users are better educated. 92.5 percent are college graduates, whereas 56.6 percent of people declaring marriage are college graduates. Third, the user group is overrepresentative of people who currently live in or are originally from Seoul and its surroundings.[10] Over 75.9 percent of users live in Seoul and its surroundings while 45.1 are originally from this area. This is compared to 51.4 and 27.4 percent of people in the MR respectively. In order to examine which characteristics most distinguish users from the general population, I run a linear probability model predicting the use of the matchmaking service with all characteristics in the MR. I found that education and regional affiliation are highly predictive, accounting for 76.8 percent of the R-squared of the model.

To examine industry and income, I apply weights on the WS so that the weighted distribution of people's characteristics is comparable to that of the users in terms of age, gender and educational attainment. The top panel of Table 5 compares users to the weighted WS sample, in terms of their industries and income. I find that the average (self-reported) income of the users excluding outliers[11] is only 14 percent higher than that observed in the general population. Users tend to be more concentrated in manufacturing and education services, while wholesale and retail trade, consumer goods, hotels/restaurants and real estate/ business services are significantly underrepresented.

[10]By "surroundings", I mean Gyeonggi Province, which surrounds Seoul and is the primary region where people live in order to commute to Seoul.

[11]Outliers refer to users whose income is larger than the 99th percentile of income among all users.

The bottom panel of Table 5 compares self-reported physical traits of the users with those of the general population. The average height and weight of the matchmaking company's users are remarkably similar to those in the PT.[12] The difference in average height is one inch, and the difference in average weight is four pounds.

2.2.4 Stated Marital Preferences

Three types of information in the matchmaking data set indicate users' marital preferences (see Table 6). Each user gives a ranking of the three most important traits for their prospective spouse, as well as any religion or geographic location that they wish to avoid.

Male users' top priority is appearance (44.6 percent), which is chosen most often, followed by personality (33.7 percent) and occupation and income (11.0 percent). On the other hand, female users choose occupation and income (55.6 percent) most often, followed by personality (26.8 percent) and appearance (5.1 percent). A Kolmogorov-Smirnov test shows that the distribution of female users' top priority is statistically different from that of male users'. This gender difference in stated marital preferences is consistent with the findings in Fisman et al. (2006) and Hitsch et al. (2006), both of which find that women put greater weight on income while men respond more to physical attractiveness. Another pattern shown in Table 6 is that people consider multiple dimensions of spousal characteristics. While education, age, religion and other dimensions are not often ranked as the top priority, they are nonetheless sufficiently important that they appear regularly among individuals' top three priorities. The majority of users are open to all religions or regions.

2.2.5 The Matching System

Each user can find a partner in two ways. He/she can search the company's database independently or allow the company to suggest a partner. In the first case, the user accesses the complete company database via a website. Having found a suitable match profile, the user can then send an electronic note to propose a first date (a user-initiated first date proposal). Note that the profiles available on the website include the users' photograph, education level, names of schools attended, occupation, region, birth order and number of siblings. For online security and privacy reasons, however, the company does not immediately reveal income, weight, parental marital status, and parental wealth. This information is available and can be

[12]The only relatively large discrepancy between the matchmaking data set and the PT is for women older than 33. This may come from the fact that women of the PT are more likely to have given birth than women in the matchmaking data set.

obtained prior to a first date by asking a staff member. The data does not provide information about the exact range of a dating partner's characteristics obtained by a user prior to a first date. I thus consider multiple possibilities in my empirical analysis, which will be discussed in Section 3.

On the other hand, the company may introduce two users based on their characteristics and stated preferences (a company-initiated first date proposal). In order to match two users, the matchmaking company employs the following sequential algorithm. Suppose the company finds a female user to match with a male user m. The company first selects a set of women who the company expects m would like and also who the company expects would like m. In order to predict the extent to which a user should be attractive to the opposite gender, the company calculates an index for the user by aggregating various attributes of the user: physical attractiveness (height, weight and facial grade), socioeconomic attractiveness (education, income, occupation and wealth) and family background (marital status of a user's parents, parental education level, and parental wealth). The company aggregates these attributes by assigning a weight to each based on a survey of its staff members who are experienced in assisting users. Choosing women who m will probably like and also who will probably like m is done by selecting women whose index is close to m's index. Second, the company further reduces the set of women by considering m's age, height and preferences for avoiding any region or religion. Finally, the company ranks women within the set by m's top three priorities and then selects the top candidate. Each of m and the top candidate receives a proposal from the company, which provides information about both users and asks whether or not they want to meet their partner in person. If either party declines the first date, the company generally provides another user's information almost immediately. The median waiting time for a new proposal is four days. Company-initiated first date proposals constitute 87 percent of all first date proposals.

If both users agree to have a first date, the company contacts each user after the first date and asks whether or not they would like to meet the partner again for a second date. This response is recorded. The company does not, however, examine the results of any subsequent dates in the same automatic fashion. However, each user is assigned a staff member, who regularly follows up to inquire whether or not the match eventually resulted in marriage.

Table 7 presents match outcomes in the data set. The median male user has 28 first date proposals. Among them, he has a first date with five women (i.e., five first dates). Out of those five first dates, he is likely to meet two women for a subsequent date (i.e., two second dates). The median female user has 27 first date proposals and has four first dates. She is likely to meet two men for a second date. After dating, 14.4 percent of men and 12.6 percent

of women get married to a person they found through the matchmaking services.[13]

2.2.6 Patterns of Sorting

In this section, I present empirical facts about users' sorting patterns across different stages of relationship, and compare the sorting patterns to that observed in the general population.

I first examine the degree of sorting at different stages of a relationship. I use three measures to examine the degree of sorting. The first is the fraction of pairs who share the same level of a particular trait, such as education level. The second is the difference between a man's age and a woman's age. The third is the correlation between a man's trait and a woman's trait. I calculate these measures for four groups: pairs who both wanted to have a first date (sorting at the first date level), pairs who both wanted to have a second date (sorting at the second date level), couples who married (marital sorting), and pairs who are randomly matched (See Table 8).

There are two main patterns revealed in Table 8. First, users positively sort on all dimensions with the exception of hometown conflicts. The difference in sorting between random matching (column 4 in Table 8) and the matchmaking data set (columns 1 to 3) reveals the degree of sorting. For example, if users randomly agree to have a first date, the fraction of pairs with the same education level would be 36 percent. However, in the data, this figure is much higher, about 53 percent (column 1). This implies that people prefer a partner with a similar educational background. Second, comparing columns 1 to 3 shows that the degree of sorting across various dimensions is generally similar at different relationship stages.

Next, I examine whether marital sorting in the matchmaking data is similar to sorting observed in the general population. I use both the MR and the HIS since each data set provides a different set of information about married couples. First, I examine the subsample of married couples who are college graduates and live in Seoul or its surroundings, from both matchmaking data and data from the general population. This is because as discussed earlier, this type of people are overrepresented in the user group. Table 9 shows the observed sorting among the user groups and that in the general population. The user group shows a larger age difference within couples, less sorting by industry, and less sorting by income than the corresponding general population. Second, instead of using only a subset of the population data, I use weights to make the distribution of users in the matchmaking data set comparable to that of people in the general population. I compute two sets of weights. The first set makes

[13]These numbers are not equal because the number of female users is greater than that of male users and also because some people married users who joined the company prior to 2002.

the distribution of husbands' characteristics the same across the matchmaking data set and the general population. The second set makes the distribution of wives' characteristics the same across the matchmaking data set and the general population.[14] Appendix B provides a detailed explanation of how the weights are constructed. Columns 3 and 4 in Table 9 show that the findings using only the previous subsample are generally confirmed, although the difference in sorting by industry is higher. Sorting along education in the matchmaking data is lower than the general population if weights based on wives' characteristics are used. However, if weights based on husbands' characteristics are used, sorting along education in the matchmaking data set is not statistically different from that observed in the general population.

Interpretation of the findings described above requires caution; using weights can make the matchmaking data set resemble the general population only in terms of the gender for which the weights are constructed. However, it does not make two data sets comparable in terms of the distribution of characteristics of the other gender. To address this limitation, we need a marriage model with which we can simulate marriage sorting while controlling for the distribution of people's characteristics.

3 The Empirical Framework

The previous section depicts two patterns arising in the data. First, users consider various spousal attributes, but men and women value given attributes differently. Second, users have multiple dates with the same partner. This section provides an empirical framework to examine underlying preferences that may generate such patterns. I then discuss identification of the model and the estimation method.

3.1 The Model

I allow the possibility that men and women have different preferences by using gender-specific marriage utility functions. In order to infer the marriage utility functions, I develop a model based on a random-effects probit specification, which I extend to allow for multiple stages with the same partner.

There are three key elements of the model: a threshold crossing rule, an idiosyncratic reservation utility, and learning processes. An individual uses a **threshold crossing rule**

[14] Note that we cannot construct a weight that simultaneously controls for both husbands' and wives' characteristics because if we construct such a weight, then the category used for the weight entirely describes marital sorting.

in order to decide whether or not to continue a relationship with a partner. The individual continues a relationship with a partner if and only if the individual expects the utility from marrying the partner to be greater than his or her reservation utility. Such a threshold crossing rule is implied by search models. **Idiosyncratic reservation utilities** using random-effects allow the possibility that individuals may have differing reservation utility levels even after controlling for their observable characteristics.

Introducing **learning processes** allows for the possibility that people can acquire information about their partners over successive dates. Within the model, multiple dates with the same partner result from a desire to learn more about their potential spouse. I model two types of learning processes. In Type 1 learning process, people require more information about a partner's characteristics that are unobservable to researchers (e.g., personality). In Type 2 learning process, people require more information about a partner's characteristics that are not revealed in the database but observable to researchers (e.g., parental wealth discussed in Section 2.2.5). I present each element in detail in subsequent sections.

3.1.1 A Threshold Crossing Rule

I begin by introducing some terminology and notation. A *pair* (m, w) refers to a specific combination of man m and woman w. The set of pairs is all possible combinations of men and women who are registered users of the matchmaking service at a given time. Subscript $s \in \{1, 2, 3\}$ indicates the stage of relationship for two individuals in a match. Stage 1 represents the decision to have a first date. Stage 2 represents the decision to have a second date, and finally stage 3 contains the marriage decision.

Superscript M or W indicates the gender of the decision maker in the pair. $U_s^M(m, w)$ is m's expected utility from marrying w at stage s whereas $U_s^W(m, w)$ is w's expected utility from marrying m at stage s. As the notation is symmetric, for convenience I refer to the model from now on using a man's point of view. $v_s^M(m, w)$ is m's reservation utility at stage s, from staying single and continuing the search for a spouse. In each stage, m determines whether to continue a relationship with w. m will want to continue a relationship with w if and only if the expected utility from marrying w is higher than m's reservation utility. $Y_s^{M*}(m, w)$ represents m's expected surplus from marrying w, or the expected utility from marrying w net of m's reservation utility. The binary variable $Y_s^M(m, w)$ is one if m wants to continue a relationship with w at stage s and zero otherwise. $Y_s^M(m, w)$ is observed in the data and can be defined

as:

$$\begin{aligned} Y_s^M(m,w) &= 1\left(Y_s^{M*}(m,w) > 0\right) \\ \text{where } Y_s^{M*}(m,w) &= U_s^M(m,w) - v_s^M(m,w) \end{aligned} \tag{1}$$

The data does not provide information about who rejects a marriage proposal at $s = 3$. I therefore define $Y_3(m,w)$ as the product of two users' responses at $s = 3$:

$$Y_3(m,w) = Y_3^M(m,w) \times Y_3^W(m,w). \tag{2}$$

A match outcome for m and w then can be expressed as a sequence $\{Y_1^M(m,w), Y_1^W(m,w), Y_2^M(m,w), Y_2^W(m,w), Y_3(m,w)\}$ where $Y_2^M(m,w)$ and $Y_2^W(m,w)$ are observable only if $Y_1^M(m,w) = Y_1^W(m,w) = 1$, and $Y_3(m,w)$ is observable only if $Y_2^M(m,w) = Y_2^W(m,w) = 1$.

3.1.2 Information Revelation and Utility from Marriage

As discussed earlier in Section 2.4, some (observable) characteristics of a partner may not be revealed prior to a first date. In order to allow for such a possibility, I assume that some traits of a dating partner are revealed prior to a first date (i.e., stage 1) and the rest after a first date (i.e, stage 2).[15]

Let X^m be user m's attributes, a column vector partitioned into two parts: X_1^m and X_2^m. X_1^m and X_2^m are column vectors of attributes revealed in stage 1 and in stage 2 respectively. $A(i)$ is the i^{th} row of A. The utility that m receives from marrying w is a function of observable attributes of m and w and a pair-specific random utility $\epsilon_{m,w}^M$:

$$\begin{aligned} U^M(m,w) &= \sum_i \left\{\alpha_i^M X_1^m(i) + \beta_i^M X_1^w(i) + \gamma_i^M h\left(X_1^m(i), X_1^w(i)\right)\right\} \\ &+ \sum_j \left\{\theta_j^M X_2^m(j) + \kappa_j^M X_2^w(j) + \lambda_j^M h\left(X_2^m(j), X_2^w(j)\right)\right\} + \epsilon_{m,w}^M \end{aligned} \tag{3}$$

where $h(x,y) = (x-y)^2$ if x and y are continuous, and $h(x,y) = 1(x \neq y)$ otherwise. $\epsilon_{m,w}^M$ is characteristics of w that m cares about but that are unobservable to researchers (e.g., personality). $\epsilon_{m,w}^M$ is independent across pairs and normally distributed with mean zero and variance $(\sigma_\epsilon^M)^2$. $U_s^M(m,w)$, m's expected utility from marrying w at stage s, is $E(U^M(m,w)|\Omega_{m,w,s}^M)$ where $\Omega_{m,w,s}^M$ is the information set of m for a pair (m,w) at stage s.

[15] In theory, I can assume that some observable traits can be observable only at stage 3. However, in that case, estimation is more difficult since at stage 3, only the joint marriage decisions is observable, rather than each user's response for marriage.

This utility function has two key features. First, it allows men and women to put different weights on each trait of a spouse. $\{\alpha^M, \beta^M, \gamma^M, \theta^M, \kappa^M, \lambda^M\}$ determine the quantitative importance of each spousal trait for men. These are not necessarily the same as $\{\alpha^W, \beta^W, \gamma^W, \theta^W, \kappa^W, \lambda^W\}$ for women. Whether these two sets of parameters are the same or not will be empirically determined. Second, the utility function also explicitly allows for the possibility that, depending on their own characteristics, different people may have a different preference ranking across their partners: If any of the parameters $\{\gamma^M, \lambda^M\}$ is not zero, the utility from marriage depends on the interaction between one's own attributes and a partner's attributes. The sign of these parameters determines whether men have preferences for traits similar to, or different from their own.

As a final remark, parameters $\{\alpha^M, \theta^M\}$ determine the "net" contribution of decision maker m's attributes to m's marriage utility. Due to collinearity, I cannot separately identify the "gross" contribution of X^m to m's marriage utility and the contribution of X^m to m's reservation utility. Thus I omit X^m in m's reservation utility function which will be described in the next section. Estimates of $\{\alpha^M, \theta^M\}$ thus quantify the contribution of X^m to m's marriage utility net of the changes of m's reservation utility due to X^m.

3.1.3 Learning Processes

Type 1 Learning Process: Bayesian Updating

I assume that a user receives a noisy signal of a partner's unobservable characteristics when they meet in person. Let $\epsilon^M_{m,w}$ be the true value of unobserved w's characteristics that m values. $\epsilon^M_{m,w}$ is normally distributed with mean zero and variance $(\sigma^M_\epsilon)^2$. When m meets w (i.e., stage s with $s \geq 2$), m receives a noisy signal $\zeta^M_{m,w,s}$ that is a sum of the true value $\epsilon^M_{m,w}$ and noise $\nu^M_{m,w,s}$. The noise is normally distributed with mean zero and variance $(\sigma^M_\nu)^2$. m's information set at stage s, $\Omega^M_{m,w,s}$, then includes $\left\{\zeta^M_{m,w,s'}\right\}$ with $2 \leq s' \leq s$ that have been revealed up to stage s. m uses Bayes' Rule to update the expectation of $\epsilon^M_{m,w}$ from the observed signals.[16]

[16] Examples of papers that employ a Bayesian learning process include Parent (2002), Gibbons et al. (2005), and Brien et al. (2006).

The distribution of $\epsilon^M_{m,w}$ given signals can be written as:

$$\begin{aligned}
\epsilon^M_{m,w}|\Omega^M_{m,w,1} &\sim N\left(0,(\sigma^M_\epsilon)^2\right) && (4)\\
\epsilon^M_{m,w}|\Omega^M_{m,w,s} &\sim N\left[\frac{(\sigma^M_\nu)^{-2}\left(\sum_{i=2}^{s}\zeta^M_{m,w,i}\right)}{(\sigma^M_\epsilon)^{-2}+(s-1)(\sigma^M_\nu)^{-2}},\frac{1}{(\sigma^M_\epsilon)^{-2}+(s-1)(\sigma^M_\nu)^{-2}}\right] && (5)\\
\text{for } s &= 2,3
\end{aligned}$$

Having multiple dates with w improves the precision of m's prediction on $\epsilon^M_{m,w}$ since the conditional variance of $w's$ unobserved attributes ($Var(\epsilon^M_{m,w}|\Omega^M_{m,w,s})$) decreases in s.

$\Omega^M_{m,w,s}$, the information set of m for a match (m,w) at stage s, is then

$$\begin{aligned}
\Omega^M_{m,w,1} &= \{X^m, X^w_1\}\\
\Omega^M_{m,w,2} &= \{X^m, X^w_1, X^w_2, \zeta^M_{m,w,2}\} && (6)\\
\Omega^M_{m,w,3} &= \{X^m, X^w_1, X^w_2, \zeta^M_{m,w,2}, \zeta^M_{m,w,3}\}
\end{aligned}$$

Type 2 Learning Process: Linear Projection

As discussed in Section 3.1.2, a subset of a dating partner's attributes are revealed in the first stage (X^w_1), and the rest is revealed in the second stage (X^w_2). If a variable $X^w_2(k) \in X^w_2$ is correlated with some variable $X^w_1(j) \in X^w_1$, then a user m can use $X^w_1(j)$ in order to predict $X^w_2(k)$. I assume that individuals use a linear projection rule to predict $X^w_2(k)$:

$$E(X^w_2(k)|\{X^w_1(j)\}^J_{j=1}) = \rho_{k0} + \sum_{j=1}^{J}\rho_{kj}X^w_1(j) \quad (7)$$

3.1.4 Specification of Utility Functions

For my estimation, I use users' stated preferences to select attributes for the marriage utility functions. Table 9 presents the attributes that I assume affect a user's utility from marriage. Some variables in Table 9 require additional explanation. First, I use hours worked as a proxy for a user's industry. The underlying assumption is that when simultaneously controlling for income, this variable captures most of industry-level variation. I adopt this approach for reasons of parsimony in order to reduce the computational burden of estimation. The variable is constructed from the WS conditional on the user's gender, age group, educational attainment and industry. Second, Body Mass Index (BMI) is a height-adjusted measure of weight and

ranges between 18.5 and 24.9 for normal-weight adults 20 years old and older.[17] Third, I assume that the marriage utility function depends on either the logarithm of current income or the logarithm of present discounted value of expected future income (PDV). PDV is the product of the logarithm of current income, average income growth rate, and the job retention rate conditional on gender, age, and industry. Detailed explanation is provided in Appendix B.2 and B.3. Primary care-provider is a binary variable that is one if a man is the eldest son or if a woman is the eldest daughter and has no male siblings. This indicates whether a user is likely to be the primary care provider for his or her parents. If a person is the primary care provider of his or her parents, the burden is likely to be shared with his/her spouse. Marital status of parents is a binary variable that is zero if the biological parents of a user are alive and still married to each other. Finally, in the 1970s and 1980s, Korean government leaders from Gyeongsang discriminated against people from Jeolla in social and economic policies. This political history resulted in regional conflicts. I define conflict between hometowns as a binary variable that is one if a user from Jeolla meets a partner from Gyeongsang or vice versa.

I use four specifications for my estimation (see Table 10). Specifications A and B assume that no Type 2 learning occurs. On the other hand, Specifications C and D allow for the learning over a partner's characteristics that are observable to researchers but may not be to other people (Type 2 learning). Specifications A and C use the current income of users whereas Specification B and D use PDV.

In specifications that allow for Type 2 learning (i.e., C and D), I assume that a user only observes information available in the online database and then receives additional information at stage 2 (e.g., income and parental wealth). I assume that people use only education and hours worked to predict a partner's income and father's education to predict parental wealth. This assumption both allows identification and reduces computational burden. If all of X_1^m are assumed to be used to predict $X_2^m(k)$, the model implies that any coefficient in front of regressors in stage 1 is not the same as the corresponding coefficient in the second stage. For identification of the variance of the stage 2 errors, we need at least one restriction on coefficients across stages. In order to identify characteristics that are the least informative in the prediction of income or parental wealth, I regress income and parental wealth individually on the entire set of characteristics of a user as well as on subsets of them. I find that income is mainly accounted for by education and hours worked, and parental wealth is for by father's education. In an OLS regression of income on the entire set of characteristics, education and hours worked account for over 93 percent of R-squared. In an OLS regression of parental wealth on the entire set of characteristics, father's education accounts for over 50 percent of

[17]Source: U.S. Center for Disease Control and Prevention, Department of Health and Human Services

R-squared. Since education, hours worked, and father's education mainly account for own income and parental wealth, I assume that people use education and hours worked to predict income and father's education to predict parental wealth in order to reduce computational burden.

3.1.5 Reservation Utility

A user m's reservation utility depends on four components. A gender-stage specific component μ_s^M allows for the possibility that burden of commitment of a relationship may differ by gender and stage. The second component is the number of singles of the opposite gender per km^2 in the region where m lives L^m. It captures the option value of finding a spouse outside the matchmaking service.[18] A user-specific random utility η_m incorporates unobserved user's characteristics such as willingness to marry. Finally, a pair-and-stage specific random component $\omega_{m,w,s}^M$ is a random utility shock realized to m at stage s in a match with w. $\omega_{m,w,1}^M$ for example contains whether or not m had a bad day when m considers a first date with w.

$$\begin{aligned} v_s^M(m,w) &= \mu_s^M + \chi^M L^m + \eta_m + \omega_{m,w,s}^M \\ \text{with } \eta_m &\sim N(0,(\sigma_\eta^M)^2) \\ \omega_{m,w,s}^M &\overset{iid}{\sim} N(0,1) \end{aligned} \tag{8}$$

3.2 Missing Data Problems

Missing data problems arise in two cases. The first is when a pair receives no first-date proposals. In other words, neither side of the pair proposes a first date, nor does the company. In this case, a user's response for a first date is missing. In the second case, a pair agreed to have a first date but the data does not have information about their response for a second date or marriage. Alternatively, the pair had a first date and agreed to have a second date, but the data does not have information about whether or not they married. The second case arises if the data is collected while a pair is continuing their relationship, or alternatively if the pair "disappeared" from the data. This second case leads to the censoring problem.

The event that a pair (m,w) does not receive a first date proposal may not be random. This can be because it is immediately obvious to all parties that they are not a good match.

[18] The assumption that only L^m, not L^w, is included in the reservation utility or the marriage utility is for identification and will be discussed in Section 4.3. I also examined an alternative specification using both L^m and the sex-ratio. I find that the sex-ratio is not statistically significant at 10 percent level, after controlling for L^m.

For example, if m lives in Seoul and w lives in Jeju, an island far from Seoul, both m and w may not consider each other as a good match. The company also will not suggest they have a first date. Since there is no information about how often such an event occurs in a non-random manner, the potential bias must be examined empirically. In order to handle this potential selection issue, I introduce the following two assumptions:

- **Assumption 1:** A user exhaustively searches for other users' profiles in the online database.
- **Assumption 2:** A user sends a proposal to another user that he/she sees in the online database if and only if the expected utility from marrying the other user exceeds his/her reservation utility.

Assumption 1 may be plausible since users can easily eliminate other users who the users are not interested in using keyword searches in the online database.[19] Assumption 2 and an alternative modeling approach will be discussed in Section 4.2. Assumptions 1 and 2 imply that a match outcome for a pair with no first date proposal may be treated in the same manner as a pair in which both declined a first date proposal. A practical issue of estimation is that the number of such pairs with no first date proposals is so large that it is infeasible to use all such pairs in my estimation. Currently I randomly select 65,489 pairs with no first date proposals, constituting 24 percent of the pairs used for estimation.

Second, the censoring issue does not cause severe problems for estimation since the fraction of pairs censored is only 2.6 percent of all pairs with first-date proposals. The estimation results change very little regardless of whether I assume the pair censored eventually got married or vice versa.

3.3 Identification

Examining users' responses for a first and a second date can identify all parameters except the following six: the gender specific component for marriage decision in reservation utility (μ_3^M, μ_3^W), the gender-specific variance of unobserved types of partners $((\sigma_\epsilon^M)^2, (\sigma_\epsilon^W)^2)$, the gender-specific variance of noises for unobserved types of partners $((\sigma_\nu^M)^2, (\sigma_\nu^W)$. Identification of all other parameters comes from three features: (1) the normalization of the variance of pair-stage specific shocks in the reservation utility (i.e., the variance of $\omega_{m,w,s}^M$ and $\omega_{m,w,s}^W$ in

[19] Keywords cover eight dimensions: age, education, marital history, location, occupation, industry, religion and height. As of October 2007, there are 12,230 male users. However, the number of never-married college-educated male users aged between 30 and 35 and living in Seoul reduces to 1,760.

Eq.(8)), (2) a full rank condition of regressors, and (3) the constraints on coefficients across stages described in Section 3.1.4.

Next consider identification of the remaining six parameters. At stage 3 (marriage decision), the expected surplus from marriage can be written as

$$Y_3^{M*}(m,w) = \mu_3^M + g_3^M(m,w) + \eta_m + \xi_{m,w,3}^M. \tag{9}$$

$g_3^M(m,w)$ includes all components which are identified by analyzing first and second dates. η_m is m's willingness to marry. $\xi_{m,w,3}^M$ is a random component that is a sum of expectation about a partner's unobserved type $\epsilon_{m,w}^M$ and a pair-specific shock $\omega_{m,w,3}^M$. I simplify a user's decision rule at the third stage as:

$$Y_3(m,w) = Y_3^M(m,w) \times Y_3^W(m,w) \tag{10}$$

$$Y_3^M(m,w) = 1\left(Y_3^{M*}(m,w) > 0\right) \tag{11}$$

where

$$\begin{aligned} \xi_{m,w,3}|\Omega_{m,w,3}^M &= E(\epsilon_{m,w}^M|\Omega_{m,w,3}^M) + \omega_{m,w,3}^M \\ (\sigma_3^M)^2 &\equiv Var(\xi_{m,w,3}^M) = \frac{2(\sigma_\epsilon^M)^4}{2(\sigma_\epsilon^M)^2 + (\sigma_\nu^M)^2} + 1. \end{aligned}$$

Identification of μ_3^M and $(\sigma_3^M)^2$ comes from two features: (1) the coefficient of $g_3^M(m,w)$ is one, and (2) $g_3^M(m,w)$ is non-degenerate and is different from $g_3^W(w,m)$ since X^m and X^w vary across people and the density of singles of the opposite gender is assumed not to affect a partner's marriage utility.[20] Finally, $\{(\sigma_\epsilon^M)^2, (\sigma_\epsilon^W)^2, (\sigma_\nu^M)^2, (\sigma_\nu^W)^2\}$ are identified from the variances of the composite error terms at stages 2 and 3 $((\sigma_2^M)^2, (\sigma_3^M)^2, (\sigma_2^W)^2, (\sigma_3^W)^2)$ (see Appendix A.3.2 for further explanation).

3.4 Estimation Method

I use a Laplace type estimator (LTE) as suggested by Chernozhukov and Hong (2003). The LTEs are defined similarly to Bayesian estimators but use more general objective functions in place of the likelihood function in Bayesian estimators. For the LTE, I define my objective function to minimize the distance between an actual match outcome and the predicted proba-

[20] In addition, if the estimated coefficients in the men's marriage utility function are different from those in the women's marriage utility function, $g_3^M(m,w)$ is not the same as $g_3^W(m,w)$ without relying on the exclusion restriction of L^m or L^w.

bility that this outcome occurs, in a similar fashion as in a simulated nonlinear least squares. Appendix A provides further explanation of the estimation method.

Compared to other alternative estimators, the LTE provides a feasible and computationally attractive solution for estimation. In my data, each user has a different number of first date proposals, and the number of stages that a pair survives is different across pairs. This feature makes the use of Bayesian estimators computationally costly, since an update of the posterior distribution of the likelihood function is complicated. Simulated maximum likelihood estimator is not feasible for my estimation. This is because the model allows a random reservation utility η_m in Eq.(8) and marriage decisions at stage 3 are the product of both users' binary responses. Due to these two features, the (log) likelihood function to be used in estimation involves a high-dimensional integration of cumulative probability densities, leading to the likelihood indistinguishable from zero for computation. LTEs are also found to perform better than simulated method of moments estimators when the objective functions have many local optima and the parameter dimension is high. (see Appendix A.4 for further discussion of infeasibility of Maximum likelihood estimators, and Chernozhukov and Hong (2003) for further discussion about LTEs).

4 Findings

In this section, I discuss the estimated model and examine its goodness of fit.

4.1 Marital Preferences and Reservation Utilities

Tables 11 and 12 report the estimated model for men and women respectively, using the four specifications described in Section 3.1.4.

Marriage Utility from a Partner's Characteristics

Tables 11 and 12 show that people consider a large number of partner traits when they make their decisions on dating and marriage. In all four specifications, the estimated parameters governing how much people value others with different traits (i.e., γ^M and γ^W in Eq.(3)) are negative for almost all traits. The estimates imply that men prefer women who are younger than them but not by too much while women prefer men who are older than themselves but not by too much. For example, *ceteris paribus*, average men (33 year old) consider 28 year old women as the best whereas 28 year old women consider 31 year old men the best. However, for specific characteristics, these preferences for similarity can be dominated by preferences for a "better" type. For example, all men and women strictly prefer partners of better appearance

(i.e., higher facial grade). In some cases, this offsetting effect is gender- as well as trait-specific. Male high school graduates prefer female high school graduates, while male college graduates prefer female college graduates except Specification C; On the other hand, in Specification A and B, all women prefer men who hold masters or Ph.D. degrees regardless of their own educational attainment. The finding that people have unanimous revealed preference rankings for "better" types for some traits suggests that strategic behavior resulting from concerns about rejection may not be severe.

The estimation results also show that parental socioeconomic status, such as father's education and parental wealth, still affects people's decision even after controlling for a large number of individual characteristics.[21] Both men and women prefer a spouse from the same hometown. Marriage between a person from Jeolla and a person from Gyeongsang (hometown conflict) is more difficult than marriage between other counterparts. Both men and women avoid a partner who is likely to be the primary care-provider for his or her parents.

Marriage Utility from One's Own Characteristics

The estimation results show that people who have more desirable characteristics receive less utility from marriage, all else being equal. For example, a man with facial grade A has less utility from marriage than a man with facial grade C. This arises if people with more desirable characteristics have a higher reservation utility. Recall that the coefficient on one's own attribute reflects the attribute's contribution to the utility of marriage *net* of its contribution to reservation utility.

Reservation Utility

People who live in a region where there are many singles of the opposite gender have a higher reservation utility. This may reflect that a high density of available singles increases the opportunity of finding a more attractive spouse than the current partner.

Learning Processes

The estimated variances of pair-specific errors $\{(\sigma_\epsilon^M)^2, (\sigma_\epsilon^W)^2\}$ and noise terms $\{(\sigma_\nu^M)^2, (\sigma_\nu^W)^2\}$ determine how fast people can improve their prediction of their partner's unobserved characteristics over multiple dates. The estimates suggest that for men, the variance of the prediction based on a second date (see Eq.(5)) is 50 percent of that based on the first date alone. The equivalent number for women is 57 percent. This implies that over multiple dates, men update their beliefs about their dating partner's type faster than women. One interpretation of this is that women may take a wider variety of unobservable characteristics into account when making decisions.

[21] Charles et al. (2006) also find positive marital sorting by parental wealth even after controlling for individual characteristics among married couples in the United States.

4.2 Goodness of Fit

I perform two sets of tests for goodness of fit. The first test uses a table of hits and misses: it compares actual binary responses in the matchmaking data to a simulated response using the point estimates of parameters. In my data, 16 percent of all pairs jointly want to have a first date, 4 percent of all pairs jointly want to have a second date, and 0.3 percent of them result in marriage. Since the sample is unbalanced in the sense that the number of zeros for having a date or marriage is much higher than ones, there is no natural threshold value for computing such a "hit-miss" table (see Greene (2007) for further discussion). Here I use a threshold value that maximizes the percent correctly predicted (see Table 13).

Another test which is a less direct but more important measure of goodness of fit, is to compare the sorting among users observed in the data to predictions of the model. I randomly select 5,000 users and use the estimated model to compute the expected marriage utility prior to a first date, $U_1^M(m,w)$ and $U_1^W(m,w)$ for each possible match among the users. Applying the Gale-Shapley algorithm to the preference ranking yields a stable matching.[22]

Since there are multiple equilibria, I compute both the male-optimal stable marriage equilibrium and the female-optimal stable marriage equilibrium. If the estimates are unbiased and the search cost is negligible, the simulated marriage outcome will be close to actual sorting in the data. The left panel in Table 14 shows the sorting observed in the matchmaking in terms of married couples, pairs who had a second date, and pairs who had a first date. The center panel shows sorting when men and women are randomly matched. The right panel shows sorting in the simulated matching using the estimated model with the four previous specifications. I compare three types of statistics from the data and the simulation results: the fraction of couples with an identical trait (top panel in Table 14), the age gap, and the correlation between a husband's and a wife's traits (bottom panel in Table 14). For the first two types of statistics, the model with the male-optimal stable matching, matches the observed sorting in the data well and performs much better than a random match. For example, 55 percent of couples in the actual data have the same level of education.[23] The fraction of such couples predicted by random matching is only 36 percent, whereas the model prediction is between 52 and 55

[22]The empirical model in Section 3.1 remains agnostic about people's search algorithm. Therefore, in order to simulate marriage, I use a simplifying assumption: people's preference rankings at the first-date stage are the same as that at the marriage stage. This assumption is reasonable considering the finding in Section 2.2.6 that sorting at the first-date stage and at marriage are similar to each other. Since users continually see numerous potential spouses in the online database, the set of equilibrium in this case coincides with the set of stable matches generated by the Gale-Shapley algorithm (see Adachi (2003) for the relationship between matching outcomes and search outcomes. Examples of papers that employ the Gale-Shapley algorithm to simulate marriage include Hitsch et al. (2006), Del Boca and Flinn (2006), and Mobarak et al. (2007)).

[23]The education here is classified into four categories: high school or less, technical college, college, and master's degree or Ph.D.

percent. The model shows some weakness in matching the observed correlation between a man and a woman such as height and parental wealth. However, even for those dimensions, it significantly outperforms random matches.

Finally, in Section 3.2, I discuss the issue of pairs with no first date proposals. In column 11 of Table 14, I show the model prediction estimated excluding all such pairs. The fit of the model is poor. The magnitude of marital sorting is considerably lower than that in the data. More strikingly, the predicted correlation between traits is much lower and even reverses the sign. This suggests that analyzing pairs with no first date proposals is important to recover marital preferences.

Among all the specifications, Specification A with the male-optimal stable equilibrium generates overall sorting that best fits the data. I thus use it for my counterfactual analysis in the subsequent section.

4.3 Preferences Revealed at Early Stages

This section studies whether preferences revealed at early stages of a relationship can reasonably predict marriage decisions. This question is motivated by the observation that most data sources rarely have information about both dating partners and spouses. Even if the data contains information about dating partners, it often only describes first dates. This exercise gives us an important understanding of how much studies based only on dating outcomes can teach us about marital preferences.

I mimic "dating-only" studies by re-estimating my model first using only first-date outcomes and then using only first- and second-date outcomes. I then compare sorting based on estimates from the first two analyses to sorting in the model estimated using all match outcomes, including the marriage decisions. As Table 15 shows, sorting along various dimensions remains similar across all three. This suggests that in a setting where people are seriously searching for a spouse, analyzing first-date outcomes is sufficient to identify their marital preferences.

5 Further Discussion

In this section, I discuss assumptions of the empirical model and the issue of selection bias.

5.1 Assumptions of The Model

In this section, I will discuss assumptions of the empirical model regarding the functional form of utility from marriage and the current assumption regarding pairs with no first-date proposals.

5.1.1 Marriage Utility Functions

I assume that people within a gender are homogeneous in terms of the marriage utility function. This assumption can be relaxed to account for the possibility that a subset of people may have a different utility function. I divide men (or women) into two groups: college graduates (say C) and non-college graduates. I estimate the following model which relaxes the homogeneity assumption:

$$\begin{aligned} Y_1^M(m,w) &= \alpha^M X^M + \beta^M X^W + \gamma^M h(X^M, X^W) \\ &\quad +1(m \in C)(\beta_C^M X^W + \gamma_C^M h(X^M, X^W)) + \eta_m + \omega_{m,w,1}^M \end{aligned} \tag{12}$$

Almost all of the 70 parameters included in β_C^M or γ_C^M are not statistically different from zero at a five percent level. I find similar results when I divide people based on their father's educational attainment. I thus conclude that the assumption of within-gender homogeneity of the utility function is reasonable.

5.1.2 Pairs with No First-Date Proposals

Suppose a pair (m, w) does not receive a first date proposal either from each other or from the company. I currently assume that such an event happens if for m and w, the expected utility from marrying each other is lower than their reservation utility. Realistically, (m, w) may not receive a first date proposal for many other reasons. Consider the following case: the period of m's using the service briefly overlapped the period that w used the service. During the time w is available, suppose that m was engaged in a relationship with another female user, and thus m had stopped searching the online database and so did w. Note that the company has a policy that it stops initiating a first-date proposal to users who have an ongoing relationship. Another example is that m thinks that utility from marrying w is higher than his reservation utility but lower than that from marrying another woman w' who he is considering and vice versa. While these are legitimate counter examples, Section 4.2 shows that the estimated model with the current assumption predicts sorting patterns in the data sufficiently well to serve as a reasonable starting point.

5.2 Selection Bias

In the counterfactual analysis, I use the estimated model to understand marital sorting in the general population. This approach will be valid if people choose to use the matchmaking service in a random manner. Below I discuss two important sources of potential selection bias.

The first is when people who are more (or less) willing to marry use the matchmaking service, even after controlling for their observable characteristics. I maintain the assumption that individuals of a given gender have the same utility function. For simplicity, consider only first-date decisions of male users and suppose that the users' behavior is modeled as a random-effects linear probability model:

$$Y_1^M(m,w) = \alpha^M X^M + \beta^M X^W + \gamma^M h(X^M, X^W) + \eta_m + \omega^M_{m,w,1} \tag{13}$$

Selection bias arises if $E(\eta_m | m \in \{\text{users}\})$ is not zero. Since the source of bias is the individual-specific willingness to marry (η_m), using individual-fixed effects denoted as a_m in Eq.(14) will yield unbiased estimates of β^M and γ^M.

$$Y_1^M(m,w) = a_m + \beta^M X^W + \gamma^M h(X^M, X^W) + \omega^M_{m,w,1} \tag{14}$$

Note that in a fixed-effects model, α^M is not identified since m's time-invariant characteristics (X^M) are subsumed in the fixed effect (a_m).

Comparing the estimates using the random-effects model with that using the fixed effects can inform us about the possibility of selection bias. Table 16 presents the estimates from these two models and the results of testing whether the two sets of estimates are statistically different from each other. Among the 70 estimated parameters, I find that only 9 of them are statistically different at a five percent significance level. Even for those 9 parameters, the preference ranking generated by the two models is the same. For example, even though the coefficient of an indicator of whether a dating partner has the same marital history is different across the models, both models predict that never-married men prefer never-married women to divorced women. Both models predict that average men (33 year old) prefer 28 year old women to the others. Note that for simulating marriages, only preference rankings matter and the estimated coefficients are relevant only for the construction of these rankings. I conclude that this form of selection bias is unlikely to be severe. A caveat to this conclusion is that I use a linear probability model instead of the probit model used for my model estimation. This is due to the fact that fixed effect probit model do not yield consistent estimates.

The second possible source of selection bias is when people have heterogeneous marriage

utility functions and the user group does not reflect the population distribution. This type of selection bias could be directly examined if we observed those who do not use the matchmaking service and have some exogenous shock affecting participation. Unfortunately, no such data are available to examine this issue. As discussed in Section 2, matchmaking services are widely used in Korea and such services have been well-established. This may be suggestive that such a selection problem may not be severe.

6 Counterfactual Analysis

In this section, I examine the importance of choice-set constraints and the relationship between marital sorting and income inequality. I use a random sample of 15,000 users of the matchmaking company, weighted such that the distribution of characteristics matches the 2005 marriage register. I use the estimated model in Section 4 to predict people's preference rankings and simulate marriages using the Gale-Shapley algorithm (1962). I use Specification A and male-optimal stable equilibrium.

6.1 The Importance of Choice-Set Constraints

I compare marital sorting observed in the general population to sorting in simulated marriages under a fully integrated marriage market, in which people see all singles in the population. Column 2 in Table 17 shows marital sorting in the fully-integrated market. Sorting by age and marital history is similar to the general population, but significantly less sorting along education, industry, region and hometown is observed. The fraction of married couples with the same education is reduced from 79 to 62 percent,[24] while the fraction of married couples in the same industry falls from 36 to 13 percent.

To understand what generates such differences, I allow the market to be segregated along the six dimensions: age, marital history, education, region and hometown. I then calibrate the degree of segregation such that the simulation results match the marital sorting observed in the population data. I find that the observed marital sorting can be generated in a marriage market that is partially segregated along four dimensions: education, industry, region, and hometown. Even though these variables are correlated to each other, a marriage market

[24]Despite a decrease in sorting along education, preferences for education in the fully integrated market still generates strong sorting along education (i.e., 78 percent of the observed marital sorting along education). On the other hand, using speed dating outcomes Belot and Francesconi (2006) conclude that preferences for education account for less than 6 percent of sorting along education. The difference in our results may be due to differences in the U.K. and Korean marriage markets. However, we cannot rule out the possibility that speed-daters may not be sufficiently committed to searching for their spouse in that environment.

partially segregated along a subset of these four (columns 3 to 5) does not generate observed sorting in the data set. A somewhat surprising finding is that although a marriage market is segregated by region, additional segregation along hometown is required to match the observed sorting. This may be because a large fraction of Koreans originally from other parts of Korea move into Seoul or its surroundings where they are likely to form social networks based on their hometown.

This suggests that although preferences contribute to overall marital sorting, constraints on people's choice sets do account for observed sorting along these dimensions. As a result, new search technologies may significantly reduce sorting along the latter dimensions. Less sorting along education may increase intergenerational mobility and less sorting along industry may cause households to be less vulnerable to industry-level income shocks.

6.2 Marital Sorting and Household Income Inequality

The relationship between marital sorting and income inequality, has been the focus of much previous discussion (for example, Kremer, 1997; Greenwood et al., 2003; Fernandez, 2001, 2005; and Pencavel, 1998, 2006). In theory, greater marital sorting by income may multiply the effect of increases in individual-level income inequality, leading to even higher household-level differences in income. On the other hand, it is plausible to expect very little response in terms of sorting if income is negatively correlated with other important positively-valued traits. The overall strength and direction of this relationship is therefore ultimately an empirical question.

To address this question, I perform two experiments. In the first experiment, all people have the same income and parental wealth, effectively removing all individual-level income inequality. In the second, income-inequality is increased by raising incomes of college graduates by 10 percent and 50 percent.

I examine how marital sorting responds to such changes in individual-level income inequality in the fully-integrated market as well as the market that matches sorting in the general population discussed previously. Columns 1 and 5 in Table 18 show marital sorting in the simulation presented earlier. Columns 2 and 6 give the results of the first experiment. Columns 3 and 7 show the results for the second experiment given a 10 percent increase of college graduates' income, while columns 4 and 8 show the results of a 50 percent increase. In all cases, I find that marital sorting along various dimensions changes very little relative to sorting under actual individual-level income inequality, regardless of whether or not marriage markets are assumed to be segregated. This suggests that increasing income inequality does not lead to

greater marital sorting, implying that changes in individual-level income inequality are unlikely to be amplified at the household-level by endogenous marital sorting.

7 Conclusion

This paper studies marital preferences and disentangles the mechanisms underlying observed marital sorting. I identify people's marital preferences using a novel data set from a major Korean matchmaking company. People consider a large number of attributes when choosing a spouse. Men and women value given attributes differently, but in general people prefer partners who are similar to themselves. I also find that constraints on people's choice sets may account for a substantial fraction of observed sorting along education, industry and geographic location in the general population. In addition, changes in individual-level income inequality are unlikely to be amplified at the household-level by endogenous marital sorting.

This paper suggests several directions for future research. One limitation of this paper is that it does not allow for the possibility that people change their reservation utility depending on their match outcomes. I leave the task of extending my model to a dynamic framework for future research. Next, my estimation results show that parental socioeconomic status directly affects a person's marriage decisions, even after controlling for the person's socioeconomic status. It thus may be useful to examine intergenerational mobility in an environment in which parental socioeconomic status partially determines a child's educational attainment, as well as the child's marriage. Thirdly, I also find that multiple dimensions are an empirically important feature of matching, suggesting the need to extend the theoretical analysis of matching to include multi-dimensional traits. Finally, I find empirical evidence that in reality people's marriage choices are constrained by the fact that they are likely to meet others sharing similar traits. It may be beneficial to study the implications of search frictions resulting from such constraints in other two-sided search markets, including job search.

References

Abramitzky, Ran, Adeline Delavande, and Luís Vasconcelos, "Marriage and War." 2007, Working Paper, Stanford University.

Adachi, Hiroyuki, "A Search Model of Two-Sided Matching under Nontransferable Utility," *Journal of Economic Theory*, 2003, *113*, 182–198.

Angrist, Joshua, "How Do Sex Ratios Affect Marriage And Labor Markets? Evidence From America's Second Generation," *The Quarterly Journal of Economics*, August 2002, *117* (3), 997–1038.

Banerjee, Abhijit, Esther Duflo, Jeanne Lafortune, and Maitreesh Ghatak, "Marry for What? Mate Selection in Modern India," 2008. Working Paper, MIT.

Becker, Gary, "A Theory of Marriage: Part I.," *Journal of Political Economy*, 1973, *81*, 813–849.

___ , "A Theory of Marriage: Part II.," *Journal of Political Economy*, 1974, *82* (2), 511–526.

Belot, Michèle and Marco Francesconi, "Can Anyone be 'The' One? Evidence on Mate Selection from Speed Dating," IZA Discussion Papers 2377, Institute for the Study of Labor (IZA) October 2006.

Bisin, Alberto, Giorgio Topa, and Thierry Verdier, "Religious Intermarriage and Socialization in the United States," *Journal of Political Economy*, 2004, *112*, 615–664.

Blossfeld, Hans-Peter and Andreas Timm, eds, *Who Marries Whom? Educational Systems as Marriage Markets in Modern Societies*, Kluwer Academic Publisher, 2003.

Boca, Daniela Del and Christopher Flinn, "Household Time Allocation and Modes of Behavior: A Theory of Sorts," 2006. New York University.

Brien, Michael J., Lee A. Lillard, and Steven Stern, "Cohabitation, Marriage, and Divorce in a Model of Match Quality," *International Economic Review*, 2006, *47*, 451–494.

Burdett, Ken and Melvyn G. Coles, "Marriage and Class," *The Quarterly Journal of Economics*, February 1997, *112* (1), 141–168.

Charles, Kerwin, Liqian Ren, and Erik Hurst, "The Nature and Consequences of Marital Sorting by Parental Wealth." 2006, Working Paper, University of Chicago.

Chernozhukov, Victor and Han Hong, "An MCMC Approach to Classical Estimation," *Journal of Econometrics*, 2003, *115*, 293–346.

Choo, Eugene and Aloysius Siow, "Who Marries Whom and Why," *The Journal of Political Economy*, February 2006, *114* (1), 175–202.

Fernández, Raquel and Richard Rogerson, "Sorting and Long-Run Inequality," *The Quarterly Journal of Economics*, November 2001, *116* (4), 1305–1341.

___ , Nezih Guner, and John Knowles, "Love and Money: A Theoretical and Empirical Analysis of Household Sorting and Inequality," *The Quarterly Journal of Economics*, January 2005, *120* (1), 273–344.

Fisman, Raymond, Sheena S. Iyengar, Emir Kamenica, and Itamar Simonson, "Gender Differences in Mate Selections: Evidence from a Speed Dating Experiment," *The Quarterly Journal of Economics*, May 2006, pp. 673–679.

___ , ___ , ___ , and ___ , "Racial Preferences in Dating: Evidence from a Speed Dating Experiment," 2007. forthcoming, Review of Economic Studies.

Fox, Jeremy, "Estimating Matching Games with Transfers," February 2007. University of Chicago.

Gale, David and Lloyd S. Shapley, "College Admissions and the Stability of Marriage," *The American Mathematical Monthly*, January 1962, *69* (1), 9–15.

Gibbons, Robert, Lawrence F. Katz, Thomas Lemieux, and Daniel Parent, "Comparative Advantage, Learning, and Sectoral Wage Determination," *Journal of Labor Economics*, October 2005, *23* (4), 681–723.

Gould, Eric D. and M. Daniele Paserman, "Waiting for Mr. Right: Rising Inequality and Declining Marriage Rates," *Journal of Urban Economics*, March 2003, *53* (2), 257–281.

Greene, William, *Econometric Analysis*, Prentice Hall, 2007.

Greenwood, Jeremy, Nezih Guner, and John A. Knowles, "More on Marriage, Fertility, and the Distribution of Income," *International Economic Review*, August 2003, *44* (3), 827–862.

Hajivassiliou, Vassilis, "Some Practical Issues in Maximum Simulated Likelihood," in Roberto Mariano, Til Schuermann, and Melvyn J. Weeks, eds., *Simulation-Based Inference in Economics*, Cambridge University Press, 2000, chapter 3, pp. 71–99.

___ , Daniel McFadden, and Paul Ruud, "Simulation of Multivariate Normal Rectangle Probabilities and Their Derivatives Theoretical and Computational Results," *Journal of Econometrics*, 1996, *72* (1-2), 85–134.

Hitsch, Günter J., Ali Hortaçsu, and Dan Ariely, "What Makes You Click? - Mate Preferences and Matching Outcomes in Online Dating," April 2006. University of Chicago.

Judd, Kenneth, *Numerical Methods in Economics*, The MIT Press, 1999.

Kalmijn, Matthijs, "Intermarriage and Homogamy: Causes, Patterns, Trends," *Annual Review of Sociology*, 1998, *24*, 395–421.

Korea Consumer Association, *Survey of Matchmaking Services Providers* 2000.

Korea Labor Institute, *Labor Statistics*, The Korean Labor Institute, 2007.

Korea Marriage Culture Institute, *Survey of the Korean Marriage Culture* 2005.

Korean Agency of Techonology and Standards, *Survey of Physical Traits of Koreans* 2004.

Kremer, Michael, "How Much Does Sorting Increase Inequality?," *The Quarterly Journal of Economics*, February 1997, *112* (1), 115–139.

Legros, Patrick and Andrew F. Newman, "Beauty is a Beast, Frog is a Prince: Assortative Matching with Nontransferabilites," December 2006.

Madden, Mary and Amanda Lenhart, "Online Dating," Technical Report, PEW/INTERNET March 2006.

Mobarak, A. Mushfiq, Randall Kuhn, and Christina Peters, "Marriage Market Effects of a Wealth Shock in Bangladesh," 2007. Yale School of Management.

Parent, Daniel, "Matching, Human Capital, and the Covariance Structure of Earnings," *Labour Economics*, 2002, *9* (3), 375–404.

Pencavel, John, "Assortative Mating by Schooling and the Work Behavior of Wives and Husbands," *The American Economic Review*, May 1998, *88* (2), 326–329.

___ , "Earnings Inequality and Market Work in Husband-Wife Families," *IZA Discussion Paper*, 2006.

Pollever, *Survey of Korean Marriage* 2004.

Republic of Korea. Fair Trade Comission, *Press Release* March 2004.

Republic of Korea. Ministry of Labor, *Basic Statistical Survey of Wage Structure* 1994 – 2006.

____ , *Labor Demand Survey* 1994 – 2006.

Republic of Korea. National Statistical Office, *National Population and Fertility Survey* 1998 – 2005.

____ , *National Household Income and Expenditure Survey* 2002 – 2005.

Sándor, Zsolt and Kenneth Train, "Quasi-Random Simulation of Discrete Choice Models," 2002.

____ and P. Pèter András, "Alternative Sampling Methods for Estimating Multivariate Normal Probabilities," *Journal of Econometrics*, 2004, *120* (2), 207–234.

Shimer, Robert and Lones Smith, "Assortative Matching and Search," *Econometrica*, March 2000, *68* (2), 343–369.

Stevenson, Betsey and Justin Wolfers, "Marriage and Divorce: Changes and Their Driving Forces," *Journal of Economic Perspectives*, 2007, *21* (2), 27–52.

Train, Kenneth, "Halton Sequences for Mixed Logit," 1999. Working paper No. E00-278, Department of Economics, University of California, Berkeley.

____ , *Discrete Choice Methods with Simulation*, Cambridge University Press, New York, 2002.

Wang, Xiaoqun and Fred J. Hickernell, "Randomized Halton Sequences," *Mathematical and Computer Modeling*, 2000, *32*, 887–899.

Wong, Linda Y., "Structural Estimation of Marriage Models," *Journal of Labor Economics*, July 2003, *21* (3), 699–727.

Appendix

A Identification and Estimation Method

Here I present the gender-specific marriage utility functions of Section 3 in detail. To do so, I consider a man m's decisions for a match with a woman w. w's decision is characterized in the same way as m's.

A.1 The Surplus from Marriage

In the first stage, m can only observe F^w. However, m can predict w's income based on her educational attainment and her hours worked, and her parental wealth from her father's educational attainment. Let w^w be the log of w's income, e^w be the level of w's education, h^w be w's hours worked, pw^w be the log of w's parental wealth, and pe^w be the level of w's father's education. Let Γ^w be a set of variables revealed in the first stage and correlated with some variables revealed in the second stage (i.e., $\Gamma^w = \{e^w, h^w, pe^w\}$).

The expected utility from the log income ($S_{k'}^{Cm}$) given w's hours worked (F_1^{Cw}) and education (F_2^{Dw}) can be expressed as below:

$$\begin{aligned} & E\left\{\theta_{k'}^{MS} S_{k'}^{Cm} + \kappa_{k'}^{MS} S_{k'}^{Cw} + \lambda_{k'}^{MS}\left(S_{k'}^{Cm} - S_{k'}^{Cw}\right)^2 | F_1^{Cw}, F_2^{Dw}\right\} \\ = \; & a_1 + a_2 S_{k'}^{Cm} + \lambda_{k'}^{MS}\left(S_{k'}^{Cm}\right)^2 \\ & + a_3 F_1^{Cw} + a_4 F_2^{Dw} + a_5\left[S_{k'}^{Cm} F_1^{Cw}\right] + a_6\left[S_{k'}^{Cm} F_2^{Dw}\right] \\ & + a_7\left(F_1^{Cw}\right)^2 + a_8\left(F_2^{Dw}\right)^2 + a_9\left[F_1^{Cw} F_2^{Dw}\right] \end{aligned}$$

where a_i with $i \in 1, .., 9$ is a nonlinear function of parameters. The expected utility from the log parental wealth given the education level of w's father can be expressed similarly. Therefore, the surplus from marriage in the first stage is summarized as

$$\begin{aligned} & Y_1^{M*}(m, w) \\ = \; & \sum_{i \notin \Gamma}\left\{\alpha_i^M F_i^m + \beta_i^M F_i^w + \gamma_i^M h\left(F_i^m, F_i^{Cw}\right)\right\} \\ & + \sum_{j \in \Gamma}\left\{\alpha_j^{MC} F_j^{Cm} + \left(\beta_j^{MC} + a_{3j}\right) F_j^{Cw} + \gamma_j^{MC}\left(F_j^{Cm} - F_j^{Cw}\right)^2 + a_{7j}\left(F_j^{Cw}\right)^2\right\} \end{aligned}$$

$$+\sum_{k\in\Gamma}\left\{\alpha_k^{MD}F_k^{Dm}+\left(\beta_k^{MD}+a_{4k}+a_{8k}\right)F_k^{Dw}+\gamma_k^{MD}1\left(F_k^{Dm}\neq F_k^{Dw}\right)\right\}$$
$$+\sum_{l\in\Gamma}\left\{a_{2l}S_l^{Cm}+\lambda_l^{MS}\left(S_l^{Cm}\right)^2+a_{5l}\left(S_l^{Cm}F_{1l}^{Cw}\right)+a_{6l}\left(S_l^{Cm}F_{2l}^{Dw}\right)+a_{9l}\left(F_{1l}^{Cw}F_{2l}^{Dw}\right)\right\}$$
$$+\chi^M L^m+\left(\mu_1^M+c_1^M\right)+\eta_m+\ \xi_{m,w,1}^M$$

It is worth noting that coefficients of regressors which are not correlated with stage 2 regressors are the same in stage 1 and stage 2. Since there is no Type 2 learning between stage 2 and stage 3, the coefficients of all regressors except the constant in stage 2 are the same as those in stage 3. This feature leads to constraints on coefficients across stages.

A.2 The LPE Method for Joint Estimation of All Stages

The empirical model is characterized by Eq.(1) and (2) where the latent index $Y_s^{M*}(m,w)$ can be simplified as below:

$$\begin{aligned}
Y_s^{M*}(m,w) &= f_s^M\left(X^m,X^w\right)+\eta_m+\xi_{m,w,s}^M \\
\xi_{m,w,s}^M &\equiv E(\epsilon_{m,w}^M|\Omega_{m,w,s}^M)+\omega_{m,w,s}^M \\
(\sigma_2^M)^2 &\equiv Var(\xi_{m,w,2}^M)=\frac{(\sigma_\epsilon^M)^4}{(\sigma_\epsilon^M)^2+(\sigma_\nu^M)^2}+1 \\
(\sigma_3^M)^2 &\equiv Var(\xi_{m,w,3}^M)=\frac{2(\sigma_\epsilon^M)^4}{2(\sigma_\epsilon^M)^2+(\sigma_\nu^M)^2}+1
\end{aligned}$$

Let $R_{m,w}$ be a discrete variable to indicate the match outcome out of all possible events:

Table A.1 Possible Match Outcomes

$R_{m,w}$	$Y_1^M(m,w)$	$Y_1^W(m,w)$	$Y_2^M(m,w)$	$Y_2^W(m,w)$	$Y_3(m,w)$
1	0	0	.	.	.
2	0	1	.	.	.
3	1	0	.	.	.
4	1	1	.	.	.
5	1	1	0	0	.
6	1	1	0	1	.
7	1	1	1	0	.
8	1	1	1	1	.
9	1	1	1	1	0
10	1	1	1	1	1

Similar to a simulated nonlinear least squares, I model the objective function L as a weighted average distance between a dummy variable of the realized match outcome and the corresponding probability that the outcome may be realized given parameters Θ:

$$\begin{aligned}\max\ L &= -\sum_{(m,w)} L_{m,w} \\ &= -\sum_{(m,w)}\sum_{r=1}^{10}\left(1(R_{m,w}=r) - \frac{1}{S}\sum_{s=1}^{S}\Pr\left(R_{m,w}=r|\Theta, X^m, X^w, \eta_m^s, \eta_w^s\right)\right)^2\end{aligned}$$

where S is the number of simulations used for computing conditional probabilities.

I estimate parameters Θ using the following algorithm: First, I estimate the model with a two-step simulated maximum likelihood and use the estimates Θ^0 as the starting value for the LTE estimates. Second, I use a normal distribution as a proposal density and tune a scaling matrix V such that the acceptance ratio of proposed parameters is between 0.4 and 0.6. Third, I generate a set of parameters φ from the proposed density $q(\varphi|\Theta^j) \sim N(\Theta^j, V)$ and update Θ using the following rule:

$$\Theta^{j+1} = \begin{cases} \varphi \text{ with probability } \rho(\Theta^j, \varphi) \\ \Theta^j \text{ with probability } 1-\rho(\Theta^j, \varphi) \end{cases}$$

where $\rho(x,y) = \inf\left(\frac{\exp(L(y))}{\exp(L(x))}, 1\right)$. I repeat the third step for $10,000$ times and take the results' mean value as $\widehat{\Theta}$.

By Theorem 4 in Chernozhukov and Hong (2003), $\widehat{\Theta}$ follows normal distribution with mean Θ and variance-covariance matrix $V\left(\widehat{\Theta}\right)$ where

$$\begin{aligned} V\left(\widehat{\Theta}\right) &= \frac{1}{n}G_n\left(\widehat{\Theta}\right)' \times W_n\left(\widehat{\Theta}\right) \times G_n\left(\widehat{\Theta}\right) \\ G_n\left(\widehat{\Theta}\right) &\equiv n \times Cov(\Theta^1, ..., \Theta^I) \\ W_n\left(\widehat{\Theta}\right) &= \frac{1}{n}\sum_{(m,w)}\left[\frac{\partial L_{m,w}}{\partial\widehat{\Theta}} \times \frac{\partial L_{m,w}}{\partial\widehat{\Theta}'}\right] \end{aligned}$$

A.3 A Two-Step Simulated Maximum Likelihood Estimation

I classify parameters in the model into two parts: parameters that determine only a response at the third stage $\Theta_2 \equiv \{\mu_3^M, \mu_3^W, (\sigma_\epsilon^M)^2, (\sigma_\nu^M)^2, (\sigma_\epsilon^W)^2, (\sigma_\nu^W)^2\}$ and the remaining parameters Θ_1. Θ_1 is identified by users' responses in the first and the second stages (the first step). I then estimate Θ_2 given parameter estimates of Θ_1 as the second step.

A.3.1 The Likelihood Function of the First Step

Let $R'_{m,w}$ be a discrete variable that is the same as $R_{m,w}$ in Table A.1 if $R_{m,w} < 9$ and is 8 if $R'_{m,w} > 8$. The probability that an event $R'_{m,w}$ occurs given individuals' random effects is as follows:

$$\begin{aligned} & \Pr(R'_{m,w} = r|X^m, X^w, \eta_m, \eta_w) \\ = \quad & P^M(R'_{m,w} = r|X^m, X^w, \eta_m) \times P^W(R'_{m,w} = r|X^w, \eta_w) \end{aligned}$$

where

$$P^M(R'_{m,w} = r|X^m, X^w, \eta_m) = \begin{cases} \Phi\left(q^M_{m,w,1}\left(f^M_1\left(X^m, X^w\right) + \eta_m\right)\right) & \text{for } r \leq 4 \\ \Phi\left(q^M_{m,w,1}\left(f^M_1\left(X^m, X^w\right) + \eta_m\right)\right) & \\ \times\Phi\left(q^M_{m,w,2}\left(f^M_2\left(X^m, X^w\right) + \eta_m\right)/\sigma^M_2\right) & \text{otherwise} \end{cases}$$

with $q^M_{m,w,s} = 2Y^M_s(m,w) - 1$. Let Θ_M be the set of all men and $W(m)$ be the set of all women with whom m was matched. Θ_W and $M(w)$ are likewise defined. The log likelihood function is then the sum of the log likelihood of men and the log likelihood of women where the log likelihood function for men is

$$\sum_{m=1}^{M} ln\left\{\int_{\eta_m} \prod_{(m,w)\in m\times W(m)} \left[\sum_{s=1}^{8} 1\left(R'_{m,w} = r\right) P^M(R'_{m,w} = r|X^m, X^w, \eta_m)\right] \phi\left(\frac{\eta_m}{\sigma^M_\eta}\right) d\eta_m\right\}$$

and the log likelihood function for women is

$$\sum_{w=1}^{W} ln\left\{\int_{\eta_w} \prod_{(m,w)\in M(w)\times w} \left[\sum_{r=1}^{8} 1\left(R'_{m,w} = r\right) P^W(R'_{m,w} = r|X^m, X^w, \eta_w)\right] \phi\left(\frac{\eta_w}{\sigma^W_\eta}\right) d\eta_w\right\}.$$

Since there is no restriction of parameters across genders, I estimate parameters for men and those for women by separately maximizing the log likelihood of each gender. I use Gauss-Hermite quadrature to compute the log likelihood function that requires a one-dimensional integration over η_m or η_w. I estimate parameters using NPSOL, an optimization algorithm that was developed by Stanford Business Software Inc. with tolerance level e^{-7}.

A.3.2 The Likelihood Function of the Second Step

A third stage error term given the information set is expressed as below:

$$\begin{aligned}\xi^M_{m,w,3}|\Omega^M_{m,w,3} &= \rho_M \xi^M_{m,w,2} + c_M \varsigma^M_{m,w} \\ \xi^W_{m,w,3}|\Omega^W_{m,w,3} &= \rho_W \xi^W_{m,w,2} + c_W \varsigma^W_{m,w}\end{aligned}$$

where

$$\begin{aligned}\rho_M &= \frac{(\sigma^M_2)^2 - 1}{(\sigma^M_2)^2}, c_M = \sqrt{(\sigma^M_3)^2 - \frac{\left((\sigma^M_2)^2 - 1\right)^2}{(\sigma^M_2)^2}} \\ \rho_W &= \frac{(\sigma^W_2)^2 - 1}{(\sigma^W_2)^2}, c_W = \sqrt{(\sigma^W_3)^2 - \frac{\left((\sigma^W_2)^2 - 1\right)^2}{(\sigma^W_2)^2}} \\ \xi^M_{m,w,2}|\ Y^M_2(m,w) &= 1 \sim N\left(0, (\sigma^M_2)^2\right) 1\left(\xi^M_{m,w,2} > -\left(f^M_2\left(X^m, X^w\right) + \eta_m\right)\right) \\ \xi^W_{m,w,2}|\ Y^W_2(m,w) &= 1 \sim N\left(0, (\sigma^W_2)^2\right) 1\left(\xi^W_{m,w,2} > -\left(f^W_2\left(X^m, X^w\right) + \eta_w\right)\right) \\ \varsigma^W_{m,w}, \varsigma^W_{m,w} &\sim N(0,1)\end{aligned}$$

ρ_M and ρ_W allow the possibility that $\xi^M_{m,w,3}$ and $\xi^W_{m,w,3}$, are correlated with $\xi^M_{m,w,2}$ and $\xi^W_{m,w,2}$ respectively due to pair-specific error terms $\epsilon^M_{m,w}$ and $\epsilon^W_{m,w}$.

Given individual random effects and $Y^M_2(m,w) = Y^W_2(m,w) = 1$, the probability of realizing $Y_3(m,w) = 1$ is

$$\begin{aligned}& P\left(Y_3(m,w) = 1 | X^m, X^w, \eta_m, \eta_w\right) \\ = \;& P\left(Y^{M*}_3(m,w) > 0 | Y^{M*}_2(m,w) > 0, X^m, X^w, \{\eta_m | Y^M_s(m,w') \text{ with } w' \in W(m), s \in \{1,2\}\}\right) \\ & \times P\left(Y^{W*}_3(w,m) > 0 | Y^{W*}_2(w,m) > 0, X^m, X^w, \{\eta_w | Y^W_s(w,m') \text{ with } m' \in M(w), s \in \{1,2\}\}\right)\end{aligned}$$

which I calculate using a Halton sequence with ten points per pair. I use Halton sequences instead of a Monte-Carlo simulation such as the GHK in order to reduce computation time. (see Judd, 1999; Sándor and András, 2004; Sándor and Train, 2002; Train, 1999; Train, 2002 for further discussion)

The second step estimation problem is thus

$$
\begin{aligned}
&\left(\mu_3^M, \mu_3^W, (\sigma_3^M)^2, (\sigma_3^W)^2\right) = \\
&\text{argmin} \sum_{(m,w)} \sum_{j=1}^{J} \frac{1}{J} \left\{ \begin{array}{c} (1 - Y_3(m,w)) \left(P\left(Y_3(m,w) = 0 | X^m, X^w, \eta_m^j, \eta_w^j, Y_2^M, Y_2^W \right) \right) \\ +Y_3(m,w) \left(P\left(Y_3(m,w) = 1 | X^m, X^w, \eta_m^j, \eta_w^j, Y_2^M, Y_2^W \right) \right) \end{array} \right\} \\
s.t. \quad & (\sigma_2^M)^2 \leq (\sigma_3^M)^2 < 2(\sigma_2^M)^2 - 1 \\
& (\sigma_2^W)^2 \leq (\sigma_3^W)^2 < 2(\sigma_2^W)^2 - 1
\end{aligned}
$$

$\left\{ \eta_m^j, \eta_w^j \right\}_{j=1}^{J}$ are simulated random effects drawn from a normal distribution conditional on match outcomes of m and w up to the first and the second stage, and J is set 20.

Finally, from $\left\{ (\sigma_2^M)^2, (\sigma_3^M)^2, (\sigma_2^W)^2, (\sigma_3^W)^2 \right\}$, I identify $\{ (\sigma_\epsilon^M)^2, (\sigma_\nu^M)^2, (\sigma_\epsilon^W)^2, (\sigma_\nu^W)^2 \}$ using the following formula:

$$
\begin{aligned}
(\sigma_\epsilon^M)^2 &= \frac{\left((\sigma_2^M)^2 - 1 \right) \left((\sigma_3^M)^2 - 1 \right)}{\left(2(\sigma_2^M)^2 - (\sigma_3^M)^2 - 1 \right)} \\
(\sigma_\nu^M)^2 &= \frac{2 \left((\sigma_3^M)^2 - (\sigma_2^M)^2 \right) \left((\sigma_2^M)^2 - 1 \right) \left((\sigma_3^M)^2 - 1 \right)}{\left(2(\sigma_2^M)^2 - (\sigma_3^M)^2 - 1 \right)^2}
\end{aligned}
$$

A.4 Infeasibility of MLE

For illustration of the need for high dimension integration, suppose that there are two men and two women characterized by the four pairs $(m_1, w_1), (m_1, w_2), (m_2, w_1), (m_2, w_2)$. For simplification, suppose that all four matches yield two successful dates but no marriage (i.e., $Y_3(m,w) = 0$ for $\forall$ m and w). The probability that the pair (m_1, w_1) reaches to the third stage depends on the users' random utilities (i.e., η_{m1} and η_{w1}). Since we can only observe the product of m_1 and w_1's responses at the third stage, the probability that the match outcome of (m_1, w_1) may arise, should be jointly computed with the probability of other match outcomes, which include m_1 or w_1. As Eq.(15) shows, the likelihood function is not separable across matches nor across genders, and thus it involves integration with dimension 4, the number of people. Since the number of users in my data set is over 20,000, computing the likelihood function is not feasible.

$$
\begin{aligned}
L \equiv \quad & \int_{\eta_{w2}} \int_{\eta_{w1}} \int_{\eta_{m2}} \int_{\eta_{m1}} \left(\begin{array}{c} P(Y^3(m_1, w_1) = 0 | X^{m1}, X^{w1}, \eta_{m1}, \eta_{w1}) \\ \times P(Y^3(m_1, w_2) = 0 | X^{m1}, X^{w2}, \eta_{m1}, \eta_{w2}) \\ \times P(Y^3(m_2, w_1) = 0 | X^{m2}, X^{w1}, \eta_{m2}, \eta_{w1}) \\ \times P(Y^3(m_2, w_2) = 0 | X^{m2}, X^{w2}, \eta_{m2}, \eta_{w2}) \end{array} \right) \\
& \times \phi(\frac{\eta_{m1}}{\sigma_\eta^M}) \phi(\frac{\eta_{m2}}{\sigma_\eta^M}) \phi(\frac{\eta_{w1}}{\sigma_\eta^W}) \phi(\frac{\eta_{w2}}{\sigma_\eta^W}) d\eta_{m1} d\eta_{m2} d\eta_{w1} d\eta_{w2}.
\end{aligned} \tag{15}
$$

B Construction of Variables

B.1 Weights for Comparison of Marital Sorting

I classify people in the marriage register (MR) into 1,176 categories based on eight age groups, three education levels, seven regions and seven hometowns. I then calculate $\{p_c\}_{c=1}^{1176}$ which is the frequency of observations in each category. I likewise calculate the frequency of observations $\{m_c\}_{c=1}^{1176}$ using the matchmaking data set. I compute weights $w_c = (p_c/P)/(m_c/M)$ where P is the total number of individuals in the MR and M is the number in the matchmaking data set. I then apply the weights to the matchmaking data set to adjust the importance of a married couple in which a wife belongs to a category c.

I similarly compute weights for comparison between the matchmaking data set and the HIS, except I use the WS instead of the MR as the baseline population data. Following the classification used in the WS, I classify people into 316 categories based on eight age groups, three education levels, and thirteen industries.

B.2 Hours Worked, Income Growth, and Job Retention Rate

I construct a user's hours worked and income growth using the WS. A user's hours worked is constructed as the average hours worked of workers in the WS who are the same as the user in terms of gender, industry, and education level. Likewise, the annual income growth rate of a user is constructed as the average growth rate of the annual income of corresponding workers in the WS. I use the WS from 1994 to 2006, excluding 1997 and 1998 when the Korean economy was seriously influenced by financial crisis. I construct a user's job retention rate from the annual Labor Demand Survey (LDS) conducted by the Korean Ministry of Labor. I take the LDSs from 1994 to 2006 (excluding 1997 and 1998) and calculate the job separation rate of workers who are the same as a user in terms of gender, industry, and education level. The job retention rate for a user is then one minus the job separation rate.

B.3 Present Value of Expected Utility of Future Income

Let g_{Mei} be the average annual income growth rate of m whose education level is e and who works in the industry i. Let θ_{mi} be m's job retention rate. m's income at time t is then

$$E\left(\ln(w_t^m)|w^m, e, i\right) = \ln(w^m) + g_{Mei}t + \alpha_1(a_t^m - a_0^m) + \alpha_2(a_t^m - a_0^m)^2$$

where w^m is m's current income, and a_t^m is m's age at time t. g_{Mei} allows industry-specific time trends and α_1 and α_2 incorporate gender-specific returns to experience. I estimate α_1 and α_2 using the HIS from 2002 to 2005.

The present discount value of utility from expected future income is the present discount value of expected future income multiplied by the job retention rate.

$$I^m = \sum_{t=1}^{T} \frac{\theta_{Mi}^t}{(1+r)^t} E\left(\ln(w_t^m)\right)$$

I assume the discount rate r to be 7.61 percent, the average interest rate of a one-year bond of the Bank of Korea from 1994 to 2005, excepting 1997 and 1998. I assume that all people retire at age 60 since the average retirement age is 56 in companies with 300 or more employees (the Korean Ministry of Labor, 2007).

Table 1: Route of Finding a Spouse

Survey Conductor	KMCI				Pollever
Survey year	2005				2004
Sample	305 couples married in 2005				1,941 unmarried internet users
Fraction of men	50				67.2
Age Groups					
- younger than 30	29.3				63.9
- 30~33	49.8				25.9
- 34~	20.9				10.2
Fraction of survey participants who are in college, college graduates, or more*	93.8				69.7
Route of Finding a Spouse by age groups**	all	(1)	(2)	(3)	all
Friends	31.8	28.8	37.6	30.4	68.6
College or Work Place	29.5	33.7	25.8	21.7	
Family/Relatives/Matchmakers	12.6	11.7	11.8	17.4	8.0
Matchmaking companies	7.6	3.7	4.3	28.3	2.5
Club/Internet	7.9	8.0	10.8	2.2	2.7
Others	10.6	14.1	9.7	0.0	18.2

* In the 2005 marriage register, the fraction of people with tertiary education was 52.28 percent.
** Definition of age groups: (1) younger than 30, (2) between 30 and 33, and (3) older than 34.
Sources: KMCI Survey of the Korean Marriage Culture 2005, Pollever Survey of Korean Marriage 2004
Route of finding a spouse refers to how a married person met his or her spouse (for the KMCI) or to how an unmarried person wants to find his or her spouse (for Pollever)

Table 2: Contents of the Matchmaking Data Set

Variables	Source
1. User's Demographic and Socioeconomic Information	
Age	Legal documents
User's birth order	Legal documents
Marital history	Legal documents
Region	Legal documents
Educational background	College diploma
Occupation and industry	Proof of employment
Annual income	Self-reported
Wealth	Self-reported
2. Family Background Information	
Father's educational background	Self-reported
Parental wealth	Self-reported
Father's occupation	Self-reported
Parent's marital status	Legal documents
3. Physical Traits	
Facial Grade (A to F)*	Evaluated by the matchmaking company
Height	Self-reported
Weight	Self-reported

* A facial grade A is the most attractive to the opposite gender whereas F is the least attractive. In the data, the distribution of facial grades is as follows: A (7.1 percent), B (38.3 percent), C (42.71 percent), D (9.56 percent), E (2.27 percent), F (0.06 percent).

Table 3: Comparison Across National-Level Data

This table compares information that is available in four data sets from the general population. MR refers to the official marriage register. WS refers to the Basic Statistics Survey of Wage Structure. HIS refers to the National Household Income and Expenditure Survey. Finally, PT refers to the Survey of Physical Traits of Koreans.

	MR	WS	HIS	PT
Survey Conductor	Ministry of Government Administration and Home Affairs	Ministry of Labor	National Statistical Office	Korean Agency of Technology and Standards
Level of Data	Micro level	Statistics	Micro level	Statistics
Spousal Information	Yes	N.A	Yes	N.A
Classification of Education				
- None	(1)	(1)	(1)	N.A
- Primary School	(2)	(1)	(2)	
- Middle School	(3)	(1)	(3)	
- Technical College (2 year)	(4)	(2)	(4)	
- University (4 year)	(4)	(3)	(5)	
- Master or Ph.D.	(4)	(3)	(6)	
Region	Detailed level	N.A.	Seoul or Non-Seoul	N.A.
Hometown	Provincial Level	N.A.	N.A.	N.A.
Industry	N.A.	Yes	Yes	N.A.
Income	N.A	Average income given gender, age, education, and industry	Income of husbands and wives	
Occupation	Yes*	Yes	Yes	N.A.
Physical Trait	N.A.	N.A.	N.A.	Yes

* The classification of occupations in the MR is not consistent with a standard classification used by the Korean National Statistical Office.

Table 4: Users' Characteristics 1

This table compares characteristics of users in the matchmaking data set (MM) with the official marriage register (MR).

Year	MM		MR
	January, 2002 ~ June, 2006		2002~2005
	All	Married	
Number of individuals	20,689	1,594	2,477,648
Composition (percentage)			
Women	53.90	50.00	50.00
Divorced	10.70	12.57	18.82
Non-Korean	0.00	0.00	4.87
Age			
younger than 27	9.01	5.83	28.79
27~29	25.28	24.76	28.08
30~33	40.05	43.61	21.84
Older than 33	25.66	25.8	21.31
Educational attainment			
Middle School or less	0.87	0.09	5.14
High School	6.63	8.06	38.27
College or more	92.50	91.86	56.59
Technical College	13.65	12.70	-
University	61.25	64.83	-
Master and Ph.D	17.60	14.33	-
Region			
Seoul or Gyeonggi	75.92	77.65	51.44
Gangwon	0.55	0.57	2.79
Chungcheong	4.44	5.00	9.59
Jeolla	3.34	3.46	9.63
Gyeongsang	11.39	13.25	25.15
Jeju and others	4.35	0.06	1.40
Hometown			
Seoul or Gyeonggi	45.12	42.48	27.36
Gangwon	3.26	3.79	4.86
Chungcheong	10.65	11.76	15.47
Jeolla	13.60	14.58	19.32
Gyeongsang	25.86	26.11	31.61
Jeju and others	1.51	1.29	1.38

Table 5: Users' Characteristics 2

This table compares users of the matchmaking service with the general population. For population data, the top panel uses the WS (2002-2006), and the bottom panel uses the PT (2004).

		MM	General Population
Year		2002~June, 2006	
Distribution across industries (Percentage)			
Agriculture, forestry, fishing, Mining		0.04	7.92
Manufacturing		20.37	16.36
Public, electric power, gas, water supply		9.23	6.27
Construction		4.26	10.54
Wholesales & retail trade, consumer goods, restaurants & hotels		4.74	19.32
Transportation, storage, communication		9.41	5.49
Finance & insurance		10.19	5.17
Real estate rent & business service		0.76	12.69
Education service		20.32	11.01
Health & social welfare		9.55	3.02
Entertainment, housekeeping, personal service		5.6	2.2
International & other foreign institution		2.41	-
Others or unemployed		3.12	-
Average annual income (10,000 won)			
Average		4054.63	3046.49
Average excluding 99th percentile and above		3468.76	N.A
Median		3137.05	N.A.
Gender-specific Physical Traits			
Height (foot, inch)			
younger than or equal to 34:	Men	5' 9"	5' 8"
	Women	5' 4"	5' 3"
Older than 34	Men	5' 8"	[5' 4", 5' 7"]
	Women	5' 4"	5' 2"
Weight (lb)			
younger than or equal to 34	Men	153.7	[153.2 , 157.0]
	Women	111.4	[116 , 120.4]
Older than 34	Men	153.2	[151.9 , 158.3]
	Women	112.0	[123.9 , 131.0]
Body Mass Index*			
younger than or equal to 34	Men	22.8	[22.6 , 24.0]
	Women	19.0	[20.3 , 21.7]
Older than 34	Men	23.0	[24.7 , 25.0]
	Women	19.4	[22.8 , 25.1]

* BMI = 703 * weight (pounds) / (height (inches))^2

Table 6: Users' Stated Marital Preferences

- **The Three Most Important Characteristics for a Prospective Spouse**

	Number of observations	Distribution across prospective spouse's characteristics (Percentage)						
		Appearance	Personality	Occupation and Income	Education	Religion	Age	Others
Men								
1st priority	6,334	44.57	33.71	11.02	2.01	1.97	2.78	3.95
2nd priority	6,334	34.13	25.51	16.47	5.00	1.36	6.71	10.82
3rd priority	5,991	20.35	15.31	23.21	6.96	2.65	8.31	23.21
Women								
1st priority	7,539	5.07	26.82	55.64	4.42	3.32	0.90	3.83
2nd priority	7,421	8.56	24.40	44.19	11.44	1.82	2.12	7.47
3rd priority	7,156	23.30	16.62	21.03	8.06	3.14	3.44	24.41

- **A Prospective Spouse's Religion that a User Avoids**

	Number of Observations	Avoiding religion					
		None	Protestant	Catholic	Buddhist	No religion	Other religions
Men	9,458	50.9	0.3	1.8	0.0	25.0	21.9
Women	11,052	50.7	0.6	2.1	0.0	24.1	22.5

- **A Prospective Spouse's Residential Area or Hometown that a User Avoids**

	None	Seoul	Gyeonggi	Gangwon	Chungcheong	Jeolla	Gyeongsang	Jeju and Others
Men	62.21	0.12	1.49	2.03	0.26	0.23	1.27	32.39
Women	63.80	0.29	1.52	0.98	0.13	0.16	0.73	32.39

Table 7: Description of Match Outcomes

	First Date Proposals	First Date	Second Date	Marriage**
Men				
All users	9,538	9,538	9,538	9,538
Median	28	4	1	
Mean	42.94	5.66	1.8	
Standard Deviation	45.81	5.29	2.15	
Users with obs.* >0	9,538	8,911	6,690	1,370
[Percentage out of all users]	[100]	[93.43]	[70.14]	[14.37]
Median	28	5	2	
Mean	42.94	6.06	2.56	
Standard Deviation	45.81	5.24	2.15	
Women				
All users	11,151	11,151	11,151	11,151
Median	27	3	1	
Mean	38.28	4.46	1.58	
Standard Deviation	36.72	4.38	1.91	
Users with obs.* >0	11,151	10,006	7,351	1,409
[Percentage out of all users]	[100]	[89.73]	[65.92]	[12.64]
Median	27	4	2	
Mean	38.28	4.97	2.4	
Standard Deviation	36.72	4.34	1.89	

* The unit of observation is a pair which reaches each stage. For example, users with obs.>0 for a second date means the number of users who have at least one match which reaches up to the second date.

** There is a discrepancy between the number of male users who found their spouse and the number of female users who found their spouse because 185 male users and 224 female users married persons who joined the matchmaking company prior to 2002.

Table 8: Degree of Sorting

This table shows the similarity between a man and a woman in the matchmaking data set, who both agreed to have a first date (Column 1), who both agreed to have a second date (Column 2), or who got married each other (Column 3). Column 4 shows similarity between a man and a woman if users are randomly matched.

	Matchmaking Data			Random
	1st date	2nd date	married	Matches
Type	(1)	(2)	(3)	(4)
Number of couples	58,833	14,884	1,594	
[Obs. used for statistics]	[32,334]	[8,394]	[1,594]	
Percentage of couples with the same				
- Education	0.529	0.535	0.549	0.361
- Industry	0.130	0.129	0.127	0.108
- Facial grade	0.394	0.407	0.456	0.310
- Marital status	0.985	0.985	0.986	0.710
- Care provider	0.592	0.589	0.605	0.586
- Region	0.927	0.928	0.926	0.352
- Religion	0.561	0.558	0.588	0.471
- Hometown	0.475	0.484	0.531	0.220
- Hometown conflicts	0.031	0.032	0.035	0.137
- Father's education	0.536	0.533	0.553	0.437
- Parental marital status	0.749	0.752	0.772	0.580
Difference in age	3.374	3.331	3.343	5.040
Correlation				
- Age	0.878	0.878	0.878	0.000
- Income	0.193	0.172	0.261	-0.009
- Hours worked	0.142	0.137	0.178	0.013
- Parental wealth	0.202	0.191	0.257	-0.030
- Height	0.333	0.313	0.339	-0.014
- BMI	0.015	0.027	0.040	-0.003

Table 9: Marriage Sorting

This table shows the similarity between husbands and wives across the matchmaking data set (MM) and data from the general population. The marriage register (MR) is used to compute statistics for the general population, except industry and income which are computed using the household income survey (HIS). Statistics in column (3) are computed using weights based on men in the population data whereas those in column (4) are computed using weights based on women in the population data. In column (5), measures of sorting along industry and income are computed using weights based on men and women in shown in the Basic Statistical Survey of Wage Structure (WS) since the HIS is not a representative sample of workers. When the HIS is used, I write the statistics using weights on husbands and then using weights based on wives.

Data	MM				Population	
Sample	All	Seoul and college graduates	Weighted (Men)	Weighted (Women)	All	Seoul and college graduates
	(1)	(2)	(3)	(4)	(5)	(6)
Number of couples	1,594	1,036	1,594	1,594	680,134	243,223
Mean difference of age	3.34	3.27	2.94	3.74	2.45	2.31
		[0.00]	[0.02]	[0.00]		
Percentage of couples with						
- Same education	89.28	100	79.40	58.73	79.13	100
			[0.29]	[0.00]		
- Same location	92.60	100	86.99	87.80	89.05	100
			[0.62]	[0.93]		
- Same hometown	53.14	28.51	38.01	66.87	55.01	30.99
		[0.09]	[0.00]	[0.60]		
- Hometown conflicts	3.51	3.30	8.51	3.74	4.65	4.15
		[0.17]	[0.12]	[0.68]		
- Same industry	12.70	12.51	13.07	10.64	36.51, 40.04	15.51
		[0.01]	[0.00]	[0.00]		
Mean income difference	1,827.37	1,550.31	1,662.32	1,599.66	1,622.28,1,548.18	1,577.64
(10,000 won)		[0.00]	[0.79]	[0.69]		
Income correlation	0.36	0.13	0.29	0.42	0.22, 0.15	0.26

Table 10: Specification of Marriage Utility Functions

This table presents the characteristics that are included in the marriage utility function in four specifications. Among a partner's characteristics, variables with F are included from the first stage whereas variables with S are included from the second stage. Specifications with "current" income use the reported income of users whereas Specifications with "PDV" use the expected present discount value of income given each user's gender, education, industry, and current income level. Appendix B.3 explains how to construct such measures in detail.

		Specification			
		A	B	C	D
	Learning of observables	No	No	Yes	Yes
	Type of income used	Current	PDV	Current	PDV
Variables	Unit				
Age	0.1 years	F	F	F	F
Educational attainment	High school or less, Technical college, and University*, and Master or Ph.D.	F	F	F	F
Hours worked per year	100 hours	F	F	F	F
Father's educational attainment	High school or less, Technical college, and University*, and Master or Ph.D.	F	F	F	F
Facial grade	Dummy: A, B, C*, and D or F	F	F	F	F
Height	Meter	F	F	F	F
Marital History	0= Never-married, 1=ever-divorced	F	F	F	F
Primary care provider	0=No, 1=Yes	F	F	F	F
Region	6 regions	F	F	F	F
Religion	6 religions	F	F	F	F
Hometown	6 regions	F	F	F	F
Conflicts between hometowns	0=No , 1=Yes	F	F	F	F
The log of current income	The unit of income is 10,000 won. The log of income is divided by 10 for scaling.	F	-	S	-
The log of expected future income	The unit of income is 10,000 won The log of income is divided by 100 for scaling.	-	F	-	S
The log of parental wealth	The unit of income is 10,000 won. The log of income is divided by 10 for scaling.	F	F	S	S
Body Mass Index	1/10	F	F	S	S
Parental marital status	0=Both biological parents are alive and not divorced, 1= otherwise	F	F	S	S

* Variables with asterisk serve as a baseline variable and thus are omitted from the regression.

Table 11: Estimated Men's Marital Preferences

This table presents the estimation results of men's surplus from marriage. The dependant variable is whether or not a decision maker wants to continue a relationship with a partner at each stage.

		Specification A		Specification B		Specification C		Specification D	
	Learning of Observables	No		No		Yes		Yes	
	Type of Income Used	Current		PDV		Current		PDV	
		Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Regressors									
Age	own	1.678	0.038	1.768	0.038	1.616	0.038	1.593	0.038
	partner	-2.259	0.034	-2.276	0.034	-2.211	0.033	-2.247	0.034
	squared difference	-2.495	0.038	-2.374	0.039	-2.482	0.039	-2.416	0.039
Educational	own = high school or less	0.293	0.045	0.184	0.041	0.222	0.050	0.229	0.046
attainment	own = technical college	0.093	0.024	0.020	0.024	0.047	0.026	0.026	0.024
	own = master of Ph.D.	0.003	0.017	0.009	0.017	-0.001	0.017	0.006	0.017
	partner = high school or less	0.087	0.090	0.080	0.013	0.243	0.088	-0.072	0.013
	partner = technical college	0.088	0.069	0.081	0.010	0.094	0.068	-0.073	0.010
	partner = master or Ph.D.	-0.030	0.074	-0.046	0.009	-0.214	0.074	-0.175	0.009
	own ≠ partner	-0.115	0.007	-0.104	0.007	-0.105	0.007	-0.113	0.007
Industry	own ≠ partner	-0.048	0.009	-0.043	0.009	-0.044	0.009	-0.047	0.009
Hours worked	Own	0.004	0.076	0.212	0.052	0.189	0.075	0.305	0.046
	partner	-0.205	0.216	-0.167	0.032	-1.139	0.123	-0.332	0.026
	squared difference	-0.128	0.023	-0.027	0.020	-0.134	0.020	-0.071	0.017
log of income	Own	-0.191	0.355	-2.905	0.111	-2.546	2.191	-13.269	0.691
	partner	0.417	0.494	3.062	0.060	2.021	2.690	8.707	0.369
	squared difference	0.257	0.233	-26.251	0.089	0.908	17.616	27.523	6.448
Father's	own = high school or less	0.099	0.016	0.120	0.015	0.098	0.016	0.126	0.015
educational	own = technical college	0.118	0.075	0.188	0.084	0.088	0.151	0.243	0.139
attainment	own = master of Ph.D.	-0.034	0.032	-0.072	0.031	-0.062	0.031	-0.027	0.031
	partner = high school or less	-0.027	0.054	-0.032	0.007	-0.088	0.054	-0.106	0.007
	partner = technical college	0.092	0.258	0.053	0.031	0.279	0.260	0.079	0.031
	partner = master or Ph.D.	0.023	0.089	0.025	0.010	0.000	0.090	0.022	0.010
	own ≠ partner	-0.032	0.007	-0.039	0.007	-0.041	0.007	-0.046	0.007
log of parental	own	-0.338	0.166	-0.347	0.045	-0.032	0.168	-0.201	0.046
wealth	partner	-0.207	0.171	-0.184	0.022	0.807	0.172	0.752	0.022
	squared difference	-0.274	0.191	-0.333	0.065	-0.511	0.193	-0.651	0.065

(continue)

Facial grade	own = A	-0.171	0.030	-0.193	0.029	-0.168	0.031	-0.176	0.030
	own = B	-0.143	0.015	-0.129	0.015	-0.133	0.016	-0.140	0.015
	own = D or F	0.122	0.022	0.152	0.021	0.130	0.022	0.143	0.021
	partner = A	0.399	0.011	0.407	0.011	0.418	0.011	0.424	0.011
	partner = B	0.179	0.006	0.183	0.007	0.192	0.006	0.192	0.007
	partner = D or F	-0.117	0.010	-0.152	0.010	-0.134	0.010	-0.146	0.010
	own ≠ partner	-0.033	0.006	-0.042	0.006	-0.031	0.006	-0.032	0.006
Height	own	1.743	0.284	0.696	0.283	1.697	0.280	1.329	0.281
	partner	-2.460	0.222	-1.538	0.222	-2.353	0.220	-2.131	0.220
	squared difference	-15.572	0.942	-14.060	0.941	-14.718	0.932	-14.501	0.935
Body Mass	own	0.086	0.156	0.094	0.049	0.415	0.154	0.358	0.049
Index	partner	-0.426	0.225	-0.492	0.036	-1.983	0.229	-2.068	0.036
	squared difference	-0.103	0.111	-0.158	0.042	-0.096	0.106	-0.064	0.042
Marital history	own = ever divorced	0.574	0.034	0.555	0.034	0.684	0.034	0.680	0.034
	partner = ever divorced	0.110	0.025	0.078	0.025	0.125	0.025	0.101	0.025
	own ≠ partner	-0.738	0.024	-0.782	0.024	-0.757	0.024	-0.730	0.024
Primary care	own = yes	-0.030	0.015	-0.032	0.014	-0.022	0.015	-0.038	0.015
provider	partner = yes	-0.030	0.007	-0.025	0.007	-0.021	0.007	-0.023	0.007
	own ≠ partner	0.010	0.007	0.036	0.007	0.015	0.007	0.024	0.007
Region	own ≠ partner	-0.515	0.011	-0.513	0.011	-0.510	0.011	-0.516	0.011
Religion	own ≠ partner	-0.081	0.006	-0.087	0.006	-0.095	0.006	-0.091	0.006
Hometown	own ≠ partner	-0.084	0.007	-0.065	0.007	-0.086	0.007	-0.086	0.007
	hometown conflict= yes	-0.158	0.016	-0.116	0.016	-0.133	0.016	-0.131	0.016
Parental	own	0.126	0.084	0.089	0.019	0.302	0.084	0.315	0.019
marital status	partner	-0.008	0.081	-0.025	0.010	-0.046	0.082	0.048	0.010
	own ≠ partner	-0.026	0.081	-0.030	0.010	-0.143	0.082	-0.046	0.010
Reservation Utility									
Density	own	1.498	0.110	1.321	0.109	1.457	0.109	1.413	0.109
Random effects	σ^{M}_{η}	0.158	0.030	0.033	0.007	0.429	0.018	0.371	0.127
Bayesian	σ^{M}_{ε}	3.318	0.017	4.638	0.014	2.347	0.003	2.805	0.107
learning	σ^{M}_{v}	0.032	0.008	0.032	0.017	0.032	0.003	0.032	0.001
No. of pairs		270,699		270,699		270,699		270,699	
No. of people		6,727		6,727		6,727		6,727	

Table 12: Estimated Women's Marital Utility Function

This table presents the estimation results of women's surplus from marriage. The dependant variable is whether or not a decision maker wants to continue a relationship with a partner at each stage.

		Specification A		Specification B		Specification C		Specification D	
	Learning of Observables	No		No		Yes		Yes	
	Type of Income Used	Current		PDV		Current		PDV	
		Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Regressors									
Age	own	-1.662	0.037	-1.750	0.037	-1.642	0.036	-1.666	0.037
	partner	1.437	0.034	1.526	0.034	1.419	0.034	1.438	0.034
	squared difference	-1.662	0.044	-1.750	0.044	-2.524	0.044	-2.520	0.044
Educational	own = high school or less	0.193	0.165	0.185	0.020	-0.674	0.166	0.338	0.020
Attainment	own = technical college	0.030	0.122	0.037	0.016	-0.540	0.121	-0.254	0.016
	own = master of Ph.D.	0.083	0.118	0.072	0.015	0.085	0.117	-0.042	0.015
	partner = high school or less	-0.108	0.084	-0.075	0.021	0.120	0.155	0.338	0.021
	partner = technical college	-0.298	0.064	-0.310	0.014	-0.306	0.137	-0.254	0.014
	partner = master or Ph.D.	0.151	0.034	0.149	0.008	0.154	0.066	-0.042	0.008
	own ≠ partner	-0.068	0.008	-0.069	0.008	-0.061	0.008	-0.069	0.008
Industry	own ≠ partner	-0.052	0.009	-0.057	0.009	-0.058	0.009	-0.051	0.009
Hours worked	own	0.132	0.326	-0.044	0.046	-0.394	0.194	-0.647	0.033
	partner	-0.282	0.092	-0.224	0.029	-1.553	0.080	-0.576	0.027
	squared difference	0.074	0.043	0.034	0.022	-0.039	0.033	-0.043	0.018
log of income	own	-0.148	0.762	-0.415	0.091	0.111	4.354	13.166	0.569
	partner	1.023	0.513	5.357	0.057	3.044	3.038	4.903	0.382
	squared difference	-0.146	0.209	-24.084	0.091	-0.225	18.906	-29.086	6.541
Father's	own = high school or less	0.010	0.089	0.079	0.011	-0.080	0.089	-0.211	0.011
Educational	own = technical college	0.061	0.415	0.109	0.053	-0.018	0.409	-0.048	0.053
Attainment	own = master of Ph.D.	0.007	0.150	-0.031	0.018	0.239	0.148	0.294	0.018
	partner = high school or less	-0.107	0.044	-0.122	0.007	-0.027	0.044	-0.032	0.007
	partner = technical college	-0.118	0.212	-0.215	0.034	0.090	0.212	0.102	0.034
	partner = master or Ph.D.	0.055	0.075	0.064	0.012	0.022	0.074	-0.046	0.012
	own ≠ partner	-0.060	0.007	-0.081	0.007	-0.060	0.007	-0.051	0.007
log of parental	own	-0.203	0.271	-0.195	0.034	0.533	0.267	0.506	0.034
Wealth	partner	-0.045	0.269	-0.058	0.022	0.764	0.268	0.942	0.022
	squared difference	-0.030	0.022	0.006	0.064	-0.116	0.022	-0.103	0.063

(continue)

Facial grade	own = A	-0.088	0.019	-0.043	0.019	-0.104	0.019	-0.077	0.019
	own = B	-0.063	0.011	-0.078	0.011	-0.055	0.011	-0.059	0.011
	own = D or F	0.045	0.016	0.078	0.016	0.074	0.016	0.089	0.016
	partner = A	0.252	0.013	0.254	0.013	0.257	0.013	0.263	0.013
	partner = B	0.134	0.007	0.133	0.007	0.126	0.007	0.142	0.007
	partner = D or F	-0.114	0.009	-0.087	0.009	-0.108	0.009	-0.114	0.009
	own ≠ partner	-0.051	0.006	-0.048	0.006	-0.045	0.006	-0.046	0.006
Height	own	-4.275	0.264	-4.860	0.258	-4.777	0.263	-4.768	0.263
	partner	4.672	0.251	5.216	0.246	5.407	0.251	5.569	0.251
	squared difference	-13.790	0.959	-17.176	0.936	-16.555	0.957	-17.101	0.957
Body Mass	own	0.087	0.282	0.194	0.044	-0.077	0.279	0.138	0.044
Index	partner	0.124	0.220	0.157	0.035	1.150	0.222	0.818	0.035
	squared difference	-0.159	0.009	-0.241	0.035	0.061	0.009	0.044	0.035
Marital history	own = ever divorced	0.145	0.031	0.032	0.031	0.179	0.031	0.158	0.031
	partner = ever divorced	0.589	0.028	0.664	0.027	0.629	0.028	0.614	0.027
	own ≠ partner	-0.747	0.027	-0.805	0.027	-0.727	0.027	-0.786	0.027
Primary care	own = yes	-0.020	0.013	0.011	0.013	0.005	0.013	-0.008	0.013
Provider	partner = yes	-0.031	0.008	-0.034	0.008	-0.022	0.008	-0.026	0.008
	own ≠ partner	-0.014	0.008	-0.019	0.008	0.006	0.008	-0.011	0.008
Region	own ≠ partner	-0.616	0.012	-0.599	0.012	-0.606	0.012	-0.583	0.012
Religion	own ≠ partner	-0.118	0.007	-0.111	0.006	-0.108	0.006	-0.118	0.006
Hometown	own ≠ partner	-0.068	0.006	-0.058	0.006	-0.075	0.006	-0.067	0.006
	hometown conflict= yes	-0.115	0.017	-0.099	0.017	-0.097	0.017	-0.114	0.017
Parental	own	0.102	0.129	0.122	0.016	-0.135	0.127	0.093	0.016
marital status	partner	-0.050	0.128	-0.013	0.011	-0.160	0.127	0.034	0.011
	own ≠ partner	-0.039	0.131	-0.012	0.010	-0.326	0.129	0.004	0.010
Reservation Utility									
Density	own	2.179	0.075	2.084	0.074	2.045	0.075	1.392	0.075
Random effects	σ^W_η	0.018	0.036	0.149	0.069	0.306	0.049	0.250	0.063
Bayesian	σ^W_ε	63.371	0.399	62.041	0.496	62.878	0.119	62.627	0.113
Learning	σ^W_v	32.337	27.748	7.454	4.314	32.095	7.735	17.058	11.826
No. of pairs		270,699		270,699		270,699		270,699	
No. of people		8,093		8,093		8,093		8,093	

Table 13: Model Fit (1)

	Specification A		Specification B		Specification C		Specification D	
Learning of observables	No		No		Yes		Yes	
Types of income used	Current		PDV		Current		PDV	
Actual Outcome	Model Prediction							
Men								
First Date	0	1	0	1	0	1	0	1
0	0.598	0.402	0.602	0.398	0.583	0.417	0.591	0.409
1	0.374	0.626	0.377	0.623	0.386	0.614	0.397	0.603
Second Date								
0	0.550	0.450	0.536	0.464	0.558	0.442	0.568	0.432
1	0.481	0.519	0.465	0.535	0.452	0.548	0.469	0.531
Women								
First Date								
0	0.603	0.397	0.607	0.393	0.506	0.494	0.496	0.504
1	0.363	0.637	0.367	0.633	0.419	0.581	0.433	0.567
Second Date								
0	0.507	0.493	0.501	0.499	0.540	0.460	0.546	0.454
1	0.453	0.547	0.446	0.554	0.445	0.555	0.454	0.546
Marriage								
0	0.513	0.487	0.540	0.460	0.494	0.506	0.489	0.511
1	0.408	0.592	0.420	0.580	0.409	0.591	0.400	0.600

Table 14: Model Fit (2)

This table compares sorting observed in the matchmaking data set (Column 1), to simulated marriages (Columns 2 to 11). Column 2 shows the predicted sorting generated by random matching. Columns 3 to 10 show the predicted marriage using the four specifications described in Section 3.1.4. The Gale-Shapley algorithm is used to simulate marriages. M refers to the male-optimal stable matching equilibrium whereas W refers to the female-optimal stable matching equilibrium. Column 11 is computed by estimating the model with excluding all pairs with no first date proposals. Specification A and male-optimal equilibrium is used to compute statistics in Column 11.

	Data	Random	Model Prediction Specification								Excl. pairs with no first
	Married	Matches	A		B		C		D		date proposals
			M	W	M	W	M	W	M	W	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Fraction of couples with same											
- Education	0.549	0.361	0.554	0.458	0.543	0.458	0.522	0.462	0.520	0.460	0.440
Father's education	0.553	0.437	0.533	0.469	0.526	0.480	0.532	0.494	0.512	0.504	0.477
- Facial grade	0.456	0.310	0.358	0.332	0.377	0.360	0.357	0.356	0.373	0.352	0.333
- Marital history	0.986	0.710	0.967	0.881	0.966	0.879	0.962	0.895	0.953	0.888	0.819
- Primary caretaker	0.605	0.586	0.600	0.585	0.576	0.575	0.587	0.591	0.566	0.590	0.576
- Region	0.943	0.352	0.831	0.692	0.838	0.701	0.818	0.704	0.797	0.701	0.637
- Religion	0.588	0.471	0.523	0.452	0.511	0.456	0.493	0.470	0.502	0.454	0.460
- Hometown	0.462	0.220	0.444	0.343	0.452	0.353	0.428	0.351	0.422	0.345	0.319
- Hometown conflict	0.031	0.137	0.034	0.061	0.030	0.064	0.035	0.064	0.041	0.061	0.057
- Industry	0.138	0.108	0.124	0.096	0.122	0.101	0.123	0.103	0.122	0.108	0.102
- Parental marital status	0.772	0.580	0.739	0.745	0.742	0.730	0.743	0.731	0.748	0.737	0.723
Age gap	3.287	5.04	3.101	3.101	3.101	3.101	3.101	3.101	3.101	3.101	3.101
Correlation											
- Height	0.339	0.000	0.238	0.044	0.221	0.056	0.233	0.078	0.141	0.081	-0.019
- Age	0.878	-0.009	0.753	0.242	0.747	0.212	0.753	0.294	0.712	0.310	-0.368
- Hours worked	0.178	0.013	0.143	-0.007	0.121	0.022	0.137	-0.016	0.020	-0.002	-0.032
- BMI	0.040	-0.030	0.090	0.022	0.035	-0.030	0.038	0.014	0.045	0.038	0.013
- Log income	0.055	-0.014	0.115	0.010	0.120	0.016	0.119	-0.015	0.049	-0.011	-0.027
- Log parental wealth	0.429	-0.003	0.114	0.025	0.114	0.029	0.104	0.015	0.128	0.012	-0.066

Table 15: Goodness of Fit for Models Using a Subset of Information

This table compares marital sorting in the data and sorting predicted by the model. Column 2 shows the predicted sorting by the model estimated only with the first-date outcomes, Column 3 shows the case for model estimated with both first- and second- date outcomes. Column 4 shows the prediction of the model with all information including marriage.

	Data	Prediction of Model Estimated By Using Match Outcomes of		
	Married	First, Second dates and Marriage	Only First and Second Dates	Only First
	(1)	(2)	(3)	(4)
Fraction of couples with same				
- Education	0.549	0.554	0.524	0.525
- Father's education	0.553	0.533	0.504	0.527
- Industry	0.138	0.124	0.115	0.111
- Facial grade	0.456	0.358	0.354	0.371
- Marital history	0.986	0.967	0.961	0.968
- Primary caretaker	0.605	0.600	0.581	0.579
- Region	0.943	0.831	0.809	0.833
- Religion	0.588	0.523	0.513	0.506
- Hometown	0.462	0.444	0.434	0.424
- Hometown conflict	0.031	0.034	0.042	0.025
- Parental marital status	0.772	0.739	0.746	0.754
Age gap	3.287	3.101	3.101	3.101
Correlation				
- Height	0.339	0.238	0.203	0.240
- Age	0.878	0.753	0.734	0.746
- Hours worked	0.178	0.143	0.092	0.077
- BMI	0.040	0.090	0.014	0.008
- Log income	0.055	0.115	0.099	0.081
- Log parental wealth	0.429	0.114	0.105	0.139

Table 16: Sample Selection

		Men				Women			
		Random	Fixed	Abs. Diff	P-value	Random	Fixed	Abs. Diff	P-value
Facial Grade	partner = A	0.109	0.110	0.001	0.705	0.066	0.067	0.001	0.668
	partner = B	0.049	0.049	0.000	0.888	0.032	0.032	0.000	0.865
	partner = D or F	-0.030	-0.031	0.002	0.508	-0.021	-0.026	0.005	0.029
	own = partner	0.007	0.007	0.000	0.944	0.012	0.011	0.001	0.678
Marital History	partner = ever divorced	0.092	0.101	0.010	0.070	0.132	0.130	0.002	0.639
	own = partner	0.157	0.140	0.016	0.002	0.144	0.149	0.006	0.263
Caretaking burden	partner = yes	-0.007	-0.006	0.001	0.789	-0.008	-0.009	0.000	0.836
	own = partner	-0.003	-0.003	0.000	0.968	0.001	0.001	0.000	0.867
Region	own = partner	0.146	0.144	0.002	0.565	0.139	0.139	0.000	0.877
Religion	own = partner	0.028	0.028	0.000	0.809	0.030	0.033	0.004	0.033
Hometown	own = partner	0.025	0.027	0.002	0.337	0.020	0.020	0.001	0.756
	hometown conflict= yes	-0.040	-0.041	0.000	0.943	-0.025	-0.030	0.005	0.243
Height	partner	-0.694	-0.782	0.088	0.162	1.231	1.287	0.056	0.313
	squared difference	-4.259	-4.702	0.443	0.091	-3.815	-4.006	0.191	0.375
Age	partner	-0.198	-0.236	0.037	0.000	0.058	0.084	0.025	0.000
	squared difference	-0.164	-0.201	0.037	0.000	-0.137	-0.160	0.022	0.000
Industry	own = partner	0.013	0.013	0.000	0.933	0.012	0.012	0.000	0.968
Educational attainment	partner = high school or less	0.026	0.024	0.002	0.618	-0.026	-0.028	0.002	0.682
	partner = technical college	0.029	0.025	0.004	0.204	-0.057	-0.057	0.000	0.963
	partner = master or Ph.D.	-0.010	-0.009	0.001	0.704	0.039	0.039	0.000	0.839
	own = partner	0.035	0.035	0.000	0.877	0.017	0.018	0.001	0.699
Hours worked	partner	-0.051	-0.058	0.007	0.408	-0.049	-0.045	0.004	0.591
	squared difference	-0.022	-0.025	0.003	0.508	0.007	0.007	0.000	0.964
Father's educational attainment	partner = high school or less	-0.005	-0.006	0.001	0.688	-0.028	-0.027	0.000	0.957
	partner = technical college	0.035	0.037	0.002	0.794	-0.020	-0.023	0.002	0.799
	partner = master or Ph.D.	0.013	0.013	0.000	0.937	0.013	0.013	0.001	0.874
	own = partner	0.015	0.015	0.000	0.974	0.016	0.017	0.000	0.825
log of income	partner	0.587	0.662	0.076	0.480	1.051	1.038	0.013	0.892
	squared difference	-0.323	0.320	0.643	0.729	-4.497	-4.351	0.146	0.931
log of parental wealth	partner	-0.050	-0.032	0.018	0.006	-0.015	0.016	0.031	0.000
	squared difference	-0.101	-0.122	0.020	0.301	-0.007	-0.059	0.053	0.001
Body Mass Index	partner	-0.117	-0.118	0.001	0.914	0.042	0.043	0.001	0.899
	squared difference	-0.012	-0.010	0.001	0.918	-0.041	-0.045	0.004	0.683
Parental Marital Status	partner	-0.001	-0.002	0.001	0.715	-0.008	-0.008	0.000	0.981
	own = partner	0.004	0.001	0.003	0.310	0.009	0.006	0.003	0.210

Table 17: Counterfactual Exercises I (Market Segregation)

Column 1 is observed sorting in the general population. Column 2 is sorting in simulated marriages in a fully integrated marriage market. Column 3 is sorting in simulated marriages in a market partially segregated by region; Column 4 in a market partially segregated by education; Column 5 in a market partially segregated by both region and education; and Column 6 in a market partially segregated by region, education, industry and hometown.

	Population Data	Simulation				
Market segregation		None	Region	Education	Region and Education	Region, Education, Industry and Hometown
	(1)	(2)	(3)	(4)	(5)	(6)
Percentage of couples with same						
- Education	0.791	0.622	0.586	0.786	0.785	0.792
- Father's education	-	0.609	0.610	0.612	0.618	0.624
- Marital history	0.921	0.969	0.921	0.929	0.927	0.933
- Region	0.890	0.735	0.882	0.693	0.906	0.883
- Hometown	0.550	0.433	0.507	0.393	0.499	0.557
- Hometown conflict	0.047	0.061	0.055	0.068	0.048	0.042
- Industry	0.365,0.400	0.127	0.127	0.137	0.138	0.382
- Facial grade	-	0.384	0.390	0.393	0.389	0.387
- Primary care-provider	-	0.565	0.580	0.572	0.575	0.560
- Religion	-	0.506	0.507	0.495	0.505	0.499
- Parental marital status	-	0.707	0.725	0.725	0.734	0.722
Age gap	2.485	2.799	2.923	2.699	2.974	2.834
Correlation						
- Height	-	0.162	0.155	0.126	0.159	0.136
- Age	0.859	0.833	0.791	0.754	0.740	0.659
- Hours worked	-	0.095	0.040	0.323	0.337	0.398
- Income	0.189,0.150	0.070	0.039	0.028	0.028	0.039
- Parental wealth	-	0.040	0.037	0.034	0.017	0.084
- Body Mass Index	-	0.025	0.047	0.046	0.074	0.040

Table 18: Counterfactual Exercises II (Changes in Income Distribution)

The fully integrated market is one which people see all singles prior to marriage. The partially-segregated market is a market partially segregated by education, industry, region, and hometown and matches the observed sorting in the general population data. Columns 1 and 5 maintain the individual-level income distribution observed in the general population data. Columns 2 and 6 assume that people have the same level of their own income and parental wealth. Columns 3 and 7 assume that college graduates earn 10 percent more income than their current income level. Columns 4 and 8 assume that college graduates earn 50 percent more income than their current income level.

Market	Fully Integrated Market				Partially Segregated Market			
Income Distribution	Baseline	No Inequality	10%↑	50%↑	Baseline	No Inequality	10%↑	50%↑
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Percentage of couples with same								
- Education	0.622	0.618	0.620	0.621	0.792	0.783	0.784	0.785
- Father's education	0.609	0.619	0.610	0.610	0.624	0.624	0.625	0.622
- Marital history	0.969	0.977	0.976	0.976	0.933	0.933	0.932	0.933
- Region	0.735	0.756	0.744	0.764	0.883	0.888	0.885	0.879
- Hometown	0.433	0.445	0.434	0.460	0.557	0.560	0.548	0.548
- Hometown conflict	0.061	0.055	0.062	0.052	0.042	0.038	0.040	0.041
- Industry	0.127	0.132	0.130	0.129	0.382	0.384	0.383	0.384
- Facial grade	0.384	0.395	0.381	0.389	0.387	0.392	0.392	0.387
- Primary care-provider	0.565	0.568	0.571	0.571	0.560	0.563	0.566	0.566
- Religion	0.506	0.511	0.505	0.512	0.499	0.505	0.501	0.493
- Parental marital status	0.707	0.716	0.715	0.714	0.722	0.715	0.719	0.725
Age gap	2.799	2.799	2.799	2.799	2.834	2.834	2.834	2.834
Household-level Income Inequality								
- Gini coefficients	0.246	0.000	0.241	0.240	0.252	0.000	0.251	0.251
- 90th percentitle/10th percentile	2.463	1.000	2.371	2.462	2.536	1.000	2.485	2.593
- Coefficient of Variation	2781.547	0.000	2821.598	3240.824	3092.175	0.000	3115.482	3538.261
Individual-level Income Inequality								
- Gini coefficients	0.329	0.000	0.327	0.331	0.329	0.000	0.327	0.331
- 90th percentitle/10th percentile	3.552	1.000	3.375	3.750	3.552	1.000	3.375	3.750
- Coefficient of Variation	3211.253	0.000	3257.837	3702.548	3211.253	0.000	3257.837	3702.548