Partner Choice and the Marital College Premium

Pierre-André Chiappori* **Bernard Salanié†** **Yoram Weiss‡**

June 6, 2010

Abstract

We propose a method to identify the returns to schooling on the marriage market from observed marriages. Under a separability assumption similar to Chow and Sioo (2006), our model is overidentified; and the data fail to reject our identifying restrictions. Our estimates show that the observed changes in marriage patterns in the US can be explained by changes in the make-up of the populations of men and women, without resorting to changes in preferences for assorted matches. Educated women now marry at higher rates than less-educated women, and they receive a higher share of the marital surplus. This may help explain why more women than men graduate from college now.

*Columbia University
†Columbia University.
‡Tel Aviv University.

1 Introduction

Marital college premium and the demand for higher education The market rate of return to schooling has increased for both men and women in recent decades. It was to be expected, therefore, that both men and women should increase their investment in schooling. However the data shows that women increased their schooling substantially more than men; in many countries, women are now more educated than men. Chiappori et al (2009) argue that this difference can be attributed to gender differences in the returns to schooling *within marriage*[1]. This hypothesis is supported by the substantial improvements in household and birth control technology, as well as by the changing roles of women within the household. Still, it is hard to prove empirically (see Greenwood et al, 2004). In contrast to the returns to schooling in the market that can be estimated from observed wages data, the returns to schooling within marriage are not directly observed and can only be inferred indirectly from the marriage patterns of individuals with different levels of schooling.

In this paper, we provide such estimates. Our empirical approach is based on a structural model of matching on the marriage market that is close, in spirit, to that adopted by Chiappori et al. (2009). Specifically, we consider a frictionless matching framework a la Becker-Shapley-Shubik, in which the gain generated by the match of male i and female j is the sum of a systematic effect, that only depends on the spouses' education classes, and a match-specific term that we treat as random. Our crucial identifying assumption, similar to that in Choo and Siow (2006), is that the latter term is additively separable into a male-specific and a female-specific components. A natural interpretation is that the complementarity properties of the model, which drive the assortativeness of the stable matching, operate *only between classes*, and are not affected by the unobservable variables. While undoubtedly restrictive, this assumption allows us to focus on our main topic of interest, namely matching between education classes; in that sense, our model is essentially motivated by a concern for parsimony. Moreover, our separability assumption generates testable restrictions that the data markedly fails to reject.[2]

Under this separability assumption, we derive a set of necessary and sufficient conditions for stability, and show that these conditions can be interpreted in terms of a standard, discrete choice framework. We then discuss the identifiability of our theoretical setup. In a cross-sectional context, the simplest version, which relies on a strong homoskedasticity assumptions, is exactly identified; so that we cannot identify any pattern of heteroskedasticity. If, however, the same structure (as summarized by the matrix of economic gains by spouses' education classes) is observed for subpopulations with different compositions, then the basic model is (vastly) overidentified. In fact,

[1] Another, largely complementary explanation proposed by Becker, Hubbard and Murphy (2009) relies on the differences between male and female distributions of unobserved ability. Still, these authors also emphasize that educated women must have received some additional, intrahousehold return to their education. It is precisely that additional term that our approach allows to evaluate.

[2] While the frictionless nature of our model would be a strong assumption in many contexts, we believe that it is probably more acceptable in our framework, precisely because of the separability assumption. In our separable world, the absence of frictions only means that any agent can meet at least one potential mate from each of the education classes under consideration at (almost) no cost.

one can even identify a much more general structure, in which the scale of individual heterogeneity may vary by education class and the systematic surplus may involve class-specific temporal drifts; moreover, this generalized model still generates strong overidentification restrictions.

We apply our model to the US population, for cohorts born between 1935 and 1975 and married between ages 18 and 35. We show that the returns from schooling received within marriage (the "marital education premium") have increased sharply for women over the period, while they have not changed much for men. Educated women have gained relative to uneducated women in two ways: by marrying at higher rates and by receiving a higher share of the marital surplus. We also find that the gains generated by marriages with equally educated partners have declined for all types of marriage, reflecting the general reduction in marriage over time. However, the smallest decline is in matches when one or both partners have college education. This finding can be related to empirical work showing that such marriages are also less likely to break (see Weiss and Willis 1997 and Bruze, Svarer and Weiss 2010).

The evolution of assortative matching A related issue is what happened to assortative mating. The observed patterns are quite complex. Overall, the percentage of couples in which both spouses have a college degree has significantly increased over the period; however, as women with college degree became more abundant, the *proportion of educated women* who marry educated men has declined (because some educated women had to marry downwards with less educated men), while men with high school degree shifted upwards from marrying women with high school degree to marrying more often women with college degree. All in all, many observers have nevertheless concluded that assortative matching was stronger now than four decades ago, and that this evolution had a deep impact on intrahousehold inequality (see for instance Burtless 1999).

An old question is whether this phenomenon is entirely due to the mechanical effect of the increase in female education, or whether it also reflects a shift in preferences towards assortative matching (as would be the case, for instance, if the share of public goods in households rises with time - or income - and similar education facilitates agreements on the composition and level of these public goods). An important advantage of our structural approach is that it allows to formally disentangle the two aspects. In this respect, our conclusions are clear-cut. We do not find any evidence for an evolution of preferences for assortative matching. In fact, we do not reject the null that the interaction in marital gains by level of schooling (as summarized by the supermodularity of the matrix of systematic gains) has remained stable over time. To the extent that we find an increasing proportion of couples in which both partners are educated, this is not because the gains from having a college degree for both partners (compared with only one partner having a college degree) have risen over time. Instead, we explain the data by shifts over time in the additive gains from marriage of educated women. One possible interpretation is that it became less costly for educated women to marry mainly because household chores have been reduced, so that married women can participate more in the labor market (see Greenwood, Seshadri and Yorukoglu 2005), and also because birth control technologies have drastically improved over the period, allowing for better planning of family fertility (see Michael, 2000, and Goldin

and Katz 2002). Our findings suggest that these "liberating effects" are more or less independent of the schooling of the husband. As a by-product of our investigation, we can identify the matrix of systematic gains; we find that ti is, indeed, significantly supermodular.

This finding seems to contradict results in the sociological literature that have shown, using log linear models, that even after accounting for changes in the relative number of men and women in each skill group, homogamy has increased in the US and several other countries (see Mare, 1991, 2008). However, these conclusions were drawn from reduced-form models with no direct economic interpretation. We show, in particular, that if the reduced form regression adopted in this literature were applied to data generated, in our setting, under the assumption that the forces driving assortative matching (i.e., the matrix of marital gains by education classes) remain constant, log-linear models would spuriously suggest that preferences for homogamy have changed. These findings further outline the importance of a structural approach to guide the interpretation of the empirical results.

Finally, another outcome of our structural approach is the identification of the group specific "prices" that determine the division of the gains from marriage between husbands and wives of different types. We find that in couples in which both spouses have a college degree, the share of the wife in the gains from marriage is larger than that of the husband. Surprisingly, the gap in favor of the wife has increased over time, despite the rise in the number of educated women relative to educated men. This happened because the marginal contribution of educated women to the surplus with educated men has risen over time. We find that the increase is mainly due to the variable component: educated women became more productive relative to less educated women in all marriages, irrespective of the type of the husband. This finding confirms the analysis of Chiappori et al. (2009), according to which the increase in the marital component of the education premium for women could explain the spectacular increase in female demand for higher education.

Related literature The analysis of the marriage market as a matching process, which dates back to Becker's seminal contributions (see Becker 1981), has recently attracted renewed attention. Probably the most important empirical work is due to Choo and Siow (2006), who propose one of the first implementations of a Becker-Shapley-Shubik model based on a discrete choice model. Our paper extends their contribution in three directions. First, we clarify the underlying theoretical structure, in particular by working out the assumptions needed on the *fundamentals* of the model (i.e., the matrix of systematic gain) and their implications for the *endogenous* variables (individual utilities at the stable match). Secondly, we consider a model that allows for interclass *heteroskedasticity* of the random components. Thirdly, we study the *evolution* of matching patterns throughout time, in a framework that also allows the gains for marriage to evolve in a class-specific way. Our ultimate goal is to study matching on education, and more specifically to provide a dynamic perspective on the evolution of the corresponding patterns over several decades.

These various extensions are necessary for our purpose. Evaluating how the 'marital college premium' evolves over time obviously necessitates a dynamic analysis. It also requires comparing (expected) utilities between classes, a task for which the ho-

moskedasticity assumption of the standard version is potentially inappropriate. In Choo and Siow's approach, the basic, homoskedastic version is exactly identified, implying that any generalization will face severe identifiability problems. We show that these problems can however be overcome in a more dynmic context (provided that the structural framework that drives assortative matching remains constant). In particular, our identification strategy is largely original.

Another related approach is due to Galichon and Salanié (2010), who provide a theoretical and econometric analysis of multidimensional matching under the same separability assumption. Their focus is different: while our paper considers a small number of classes, they seek to provide a general method to estimate and test flexible parametric specifications of the gains from marriage when many covariates are available.

Section 2 presents some stylized facts. Then we introduce our theoretical framework in Section 3, and section 4 describes the basic principles underlying its empirical implementation. In Section 5, we discuss identification issues and present our main theoretical results on that topic. Section 6 describes the matching patterns in the data, and our empirical findings are presented in Section 7.

2 Some stylized facts

We first briefly describe some raw facts about the evolution of matching by education over the last decades. To do this, we use the American Community Survey, a representative extract of the Census, which we downloaded from IPUMS (see Ruggles et al (2008).) Unlike earlier waves of the survey, the 2008 survey has information on current marriage status, number of marriages, and year of current marriage. Of the 3,000,057 observations in our original sample, we only keep white adults (aged 18 to 70) who our out of school; the resulting sample has 1,307,465 observations and is 49.5% male. We used the "detailed education variable" of the ACS to define four subcategories:

1. High School Dropouts (HSD): $\texttt{educd} < 61$
2. High School Graduates (HSG): $62 \leq \texttt{educd} < 65$
3. Some College (SC): $65 \leq \texttt{educd} < 101$
4. College-educated (COLL): $\texttt{educd} \geq 101$.

Outcomes are truncated in that young men and women who are single in 2008 may still marry; in our figures (and later in our estimates) we circumvent this difficulty by stopping at the cohort born in 1972—the first union occurs before age 35 for most men and women.

The increasing level of education of women is shown on Figure 1: in cohorts born after 1955 women graduate more often from high school, and from college a few years after that.

We illustrate the decline in marriages by plotting the percentage of individuals of a given cohort who never married in Figures 2 and 3. Figures 4 and 5 take this further by plotting the logarithm of the ratio of the proportion of a given education and gender who never married, relative to the proportion for that gender. They show that a higher

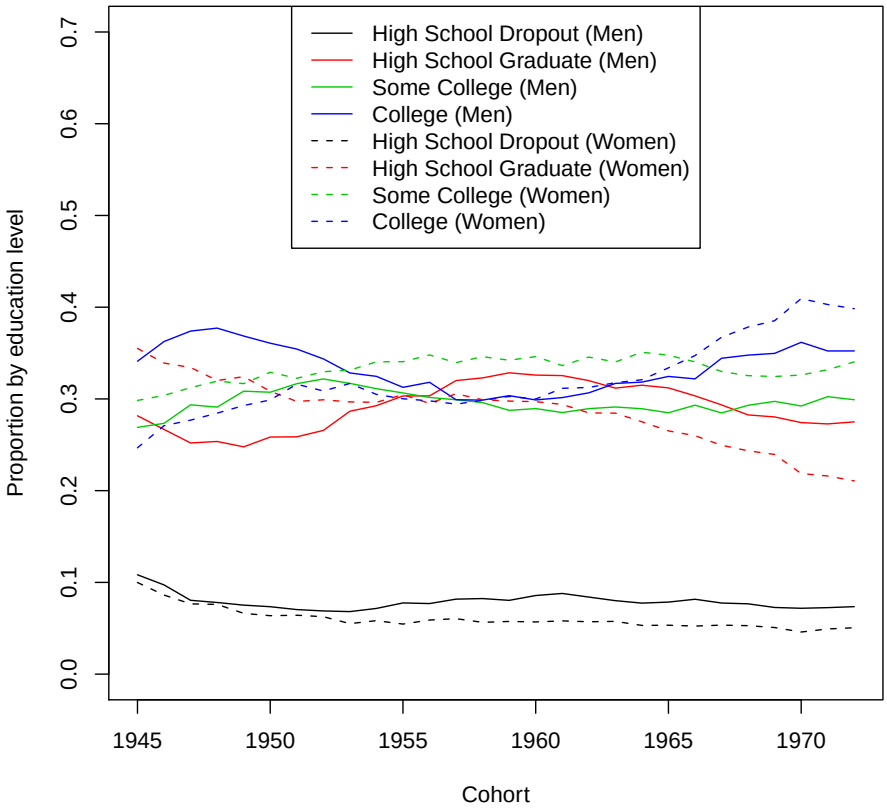

Figure 1: Education levels of men and women

education has tempered the decline in marriage, especially for women; and that women high-school dropouts on the other hand have faced a very steep decline in marriage.

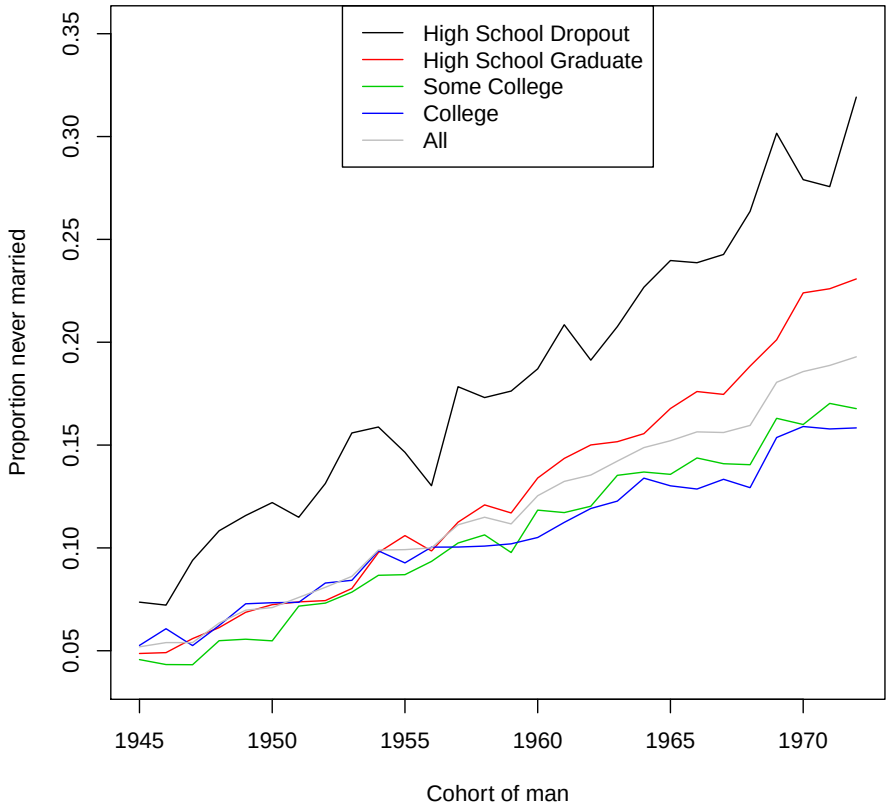

Figure 2: Proportion of men who never married

We will give more information on observed matching patterns after we restrict the sample further in section 6.

3 Theoretical framework

The basic structure We consider a frictionless, Becker-Shapley-Shubik matching game between a male population M, endowed with some measure $d\mu_M$, and a female population F, endowed with some measure $d\mu_F$. Each population is partitioned into a finite number of classes, $I = 1, ..., N$ for men and $J = 1, ..., M$ for women. The gain

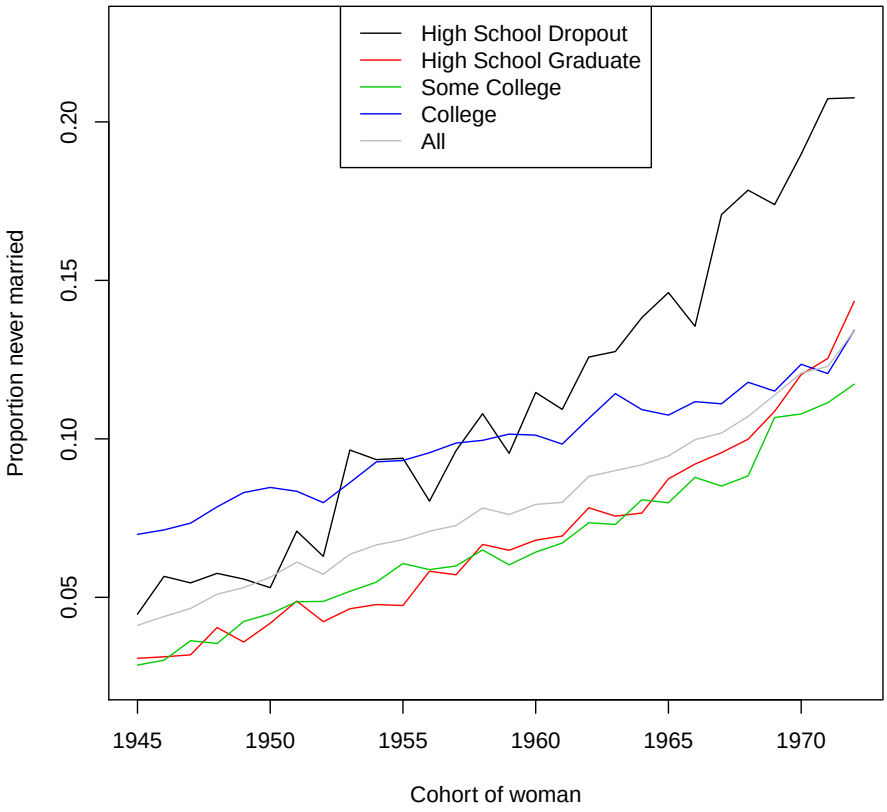

Figure 3: Proportion of women who never married

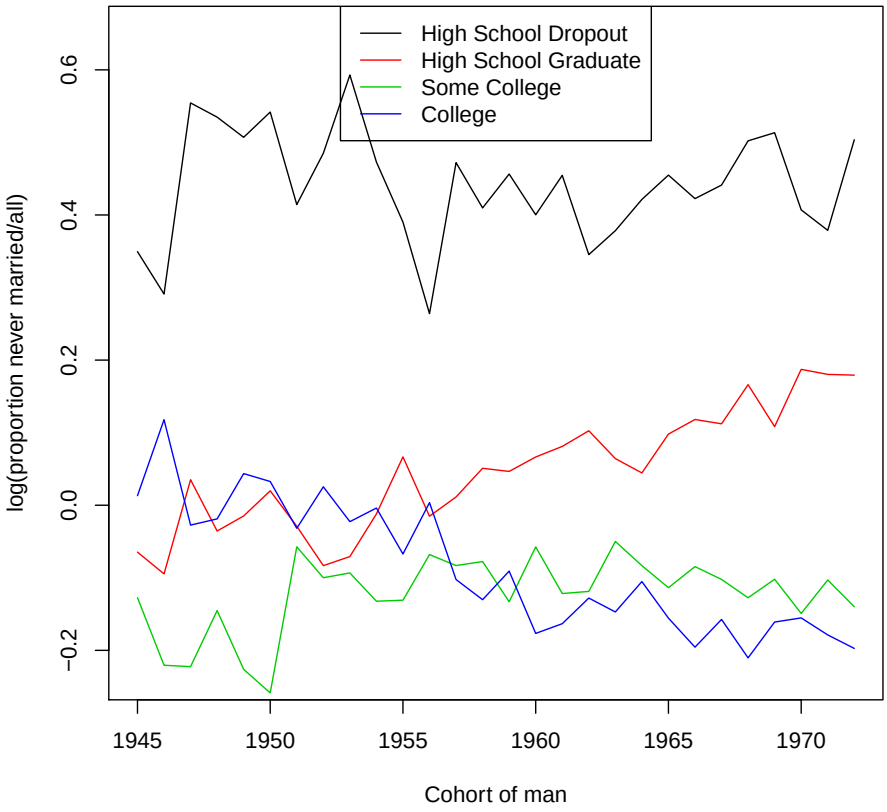

Figure 4: Log-odds of men never marrying

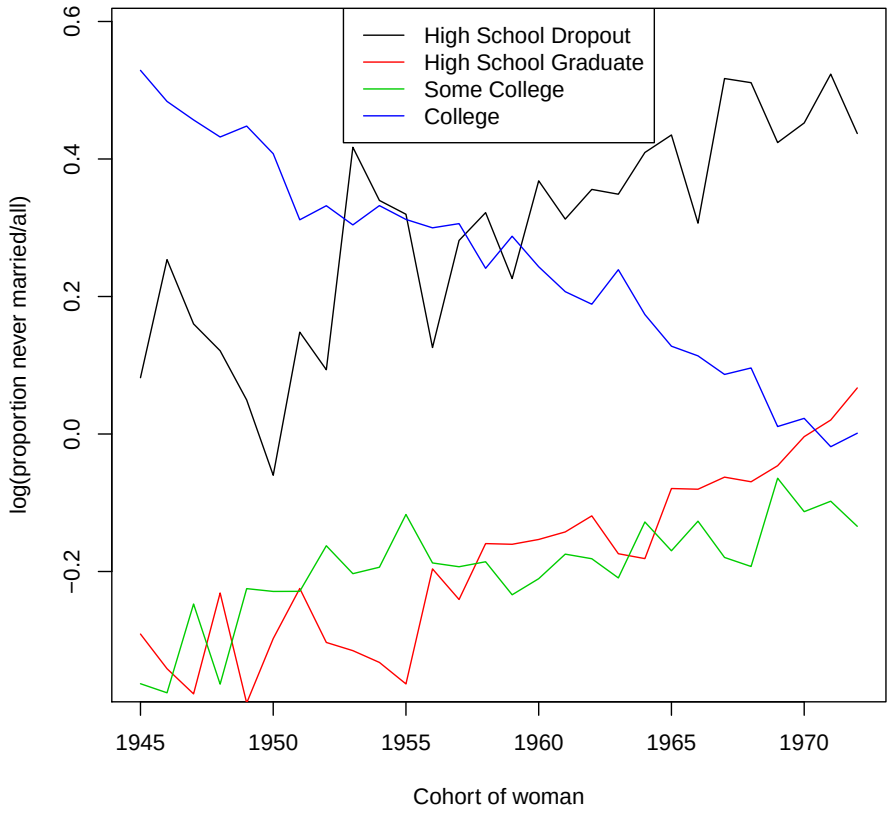

Figure 5: Log-odds of women never marrying

generated by the match of Mr. i, belonging to class I, and Mrs. j, belonging to class J, is the sum of two components, one common to all individuals in the same class, the other match specific:

$$g_{ij} = Z^{IJ} + \varepsilon_{ij}^{IJ}$$

with the notation $I = 0$, $J = 0$ for singles. Here ε_{ij}^{IJ} is a random shock with mean zero.

A *matching* consists of (i) a measure $d\mu$ on the set $M \times F$, such that the marginal of $\delta\mu$ over M (resp. F) is $d\mu_M$ ($d\mu_F$), and (ii) a set of payoffs (or imputations) $\{u_i, i \in M\}$ and $\{v_j, j \in F\}$ such that

$$u_i + v_j = g_{ij} \quad \text{for any } (i,j) \in Supp\,(\mu)$$

In words, a matching indicates who marries whom (note that the allocation may be random, hence the measure), and how any married couple shares the gain generated by their match.

A matching is stable if one can find neither a man i who is currently married but would rather be single, nor a woman j who is currently married but would rather be single, nor a woman j and a man i who are not currently married together but would both rather be married together than remain in their current situation. Formally, we must have that:

$$u_i + v_j \geq g_{ij} \quad \text{for any } (i,j) \in M \times F \tag{1}$$

which translate the fact that for any possible match (i,j), the realized gain g_{ij} cannot exceed the sum of utilities respectively reached by i and j in their current situation.

As is well known, a matching model of this type is equivalent to a maximization problem; specifically, a match is stable if and only if it maximizes total gain, $\int g d\mu$, over the set of measures whose marginal over M (resp. F) is $d\mu_M$ ($d\mu_F$). A first consequence is that existence is guaranteed under mild assumptions. Moreover, the dual of this maximization problem generates, for each male i (resp. female j), a 'shadow price' u^i (resp. v^j), and the dual constraints these variables must satisfy are exactly (1); in other words, the dual variables exactly coincide with payoffs associated to the matching problem.

Finally, is the stable matching unique? With finite populations, the answer is no; in general, the payoffs u_i and v_j can be marginally altered without violating the (finite) set of inequalities (1). However, when the populations become large, the intervals within which u_i and v_j may vary typically shrink; in the limit of continuous populations, (the distributions of) individual payoffs are exactly determined. On all these issues, the reader is referred to Chiappori, McCann and Nesheim (2009) for precise statements.

The main empirical assumption We now introduce a simplifying assumption that will be crucial in what follows:

Assumption S (separability): *the idiosyncratic component ε_{ij} is additively separable:*

$$\varepsilon_{ij}^{IJ} = \alpha_i^{IJ} + \beta_j^{IJ} \tag{S}$$

where $E\left[\alpha_i^{IJ}\right] = E\left[\beta_j^{IJ}\right] = 0$.

In words, the match specific term is the sum of two contributions. The male contribution is individual specific and may depend on both his and his spouse's class - but it does not depend on the precise identity of i's spouse; and the same property holds for the female contribution. Note that this assumption is equivalent to the following property: for any $i, i' \in I$ and any $j, j' \in J$,

$$g_{ij} + g_{i'j'} = g_{ij'} + g_{i'j}$$

This implies that within each pair of classes, (I, J), all matches are equivalent; in practice, we exclusively concentrate on the marital patterns between classes (although this can be relaxed by the introduction of covariates, see below).

Each male i is thus fully characterized by the realization of the vector $\alpha_i = \left(\alpha_i^{11}, ..., \alpha_i^{MN}\right)$. For notational consistency, we define

$$\alpha_i^{I0} = \varepsilon_{i0}^{I0} \text{ and } \beta_j^{0J} = \varepsilon_{0j}^{0J}$$

(and similarly for women).

Then we have the following Lemma:

Lemma 1 *Assume the ε satisfy the separability property (S). For any stable matching, there exist numbers U^{IJ} and $V^{IJ}, I = 1, ..., M, J = 1, ..., N$, with*

$$U^{IJ} + V^{IJ} = Z^{IJ} \tag{2}$$

satisfying the following property: for any matched couple (i, j) such that $i \in I$ and $j \in J$,

$$\begin{aligned} u_i &= U^{IJ} + \alpha_i^{IJ} \\ &\quad \textit{and} \qquad\qquad\qquad\qquad\qquad\qquad (L) \\ v_j &= V^{IJ} + \beta_j^{IJ} \end{aligned}$$

Proof. *Assume that i and i' both belong to I, and are both matched with a spouse (resp. j and j') belonging to J. Stability requires that:*

$$\begin{aligned} u_i + v_j &= Z^{IJ} + \alpha_i^{IJ} + \beta_j^{IJ} && (1) \\ u_i + v_{j'} &\geq Z^{IJ} + \alpha_i^{IJ} + \beta_{j'}^{IJ} && (2) \\ u_{i'} + v_{j'} &= Z^{IJ} + \alpha_{i'}^{IJ} + \beta_{j'}^{IJ} && (3) \\ u_{i'} + v_j &\geq Z^{IJ} + \alpha_{i'}^{IJ} + \beta_j^{IJ} && (4) \end{aligned}$$

Subtracting (1) from (2) and (4) from (3) gives

$$\beta_{j'}^{IJ} - \beta_j^{IJ} \leq v_{j'} - v_j \leq \beta_{j'}^{IJ} - \beta_j^{IJ}$$

hence

$$v_{j'} - v_j = \beta_{j'}^{IJ} - \beta_j^{IJ}$$

It follows that the difference $v_j - \beta_j^{IJ}$ does not depend on j, i.e.:

$$v_j - \beta_j^{IJ} = V^{IJ} \quad \textit{for all } i \in I, j \in J$$

The proof for u_i is identical. ■

In words: the differences $u_i - \alpha_i^{IJ}$ and $v_j - \beta_j^{IJ}$ only depend on the spouses' classes, not on who they are. The U^{IJ} and V^{IJ} denote how the common component of the gain is divided between spouses; then a spouse's utility is the sum of their share of the common component and their own, idiosyncratic contribution. Note, incidentally, that (L) is also valid for singles if we set $u^{I0} = z^{I0}$ and $v^{0J} = z^{0J}$.

An intuitive interpretation of U^{IJ} (or equivalently of V^{IJ}) would be the following. Assume that a man randomly picked in class I is *forced* to marry a woman belonging to class J (assuming that the populations are large, so that this small deviation from stability does not affect the equilibrium payoffs). Then his expected utility is exactly U^{IJ} (the expectation being taken over the random choice of the individual within the class). Note, however, that this value does *not* coincide with the average utility of men in class I married to women J at a stable matching. The latter value is larger than U^{IJ} (reflecting the fact that an agent *chooses* his wife's class), and will be computed below.

Stable matchings: a characterization Under this separability assumption, the empirical characterization of the stable match becomes much easier. We first provide a simple translation of the stability properties:

Lemma 2 *A set of necessary and sufficient conditions for stability is that*

1. *for any matched couple* $(i \in I, j \in J)$ *one has*

$$\alpha_i^{IJ} - \alpha_i^{IK} \geq U^{IK} - U^{IJ} \quad \text{for all } K \tag{3}$$
$$\alpha_i^{IJ} - \alpha_i^{I0} \geq U^{I0} - U^{IJ} \tag{4}$$

and

$$\beta_j^{IJ} - \beta_j^{KJ} \geq V^{KJ} - V^{IJ} \quad \text{for all } K \tag{5}$$
$$\beta_j^{IJ} - \beta_j^{0J} \geq V^{0J} - V^{IJ} \tag{6}$$

2. *for any single male* $i \in I$ *one has*

$$\alpha_i^{IJ} - \alpha_i^{I0} \leq U^{I0} - U^{IJ} \quad \text{for all } J \tag{7}$$

3. *for any single female* $j \in J$ *one has*

$$\beta_j^{IJ} - \beta_j^{0J} \leq V^{0J} - V^{IJ} \quad \text{for all } J \tag{8}$$

Proof. *The proof is in several steps. Let* $(i \in I, j \in J)$ *be a matched couple. Then:*

1. *First, male* i *must better off than being single, which gives:*

$$\begin{aligned} U^{IJ} + \alpha_i^{IJ} &\geq U^{I0} + \alpha_i^{I0} \\ &\textit{hence} \\ \alpha_i^{IJ} - \alpha_i^{I0} &\geq U^{I0} - U^{IJ} \end{aligned}$$

and the same must hold with female j*. This shows that 4, 6, 7 and 8 are necessary.*

2. *Take some female j' in J, currently married to some i' in I. Then i must be better off matched with j than j', which gives:*

$$U^{IJ} + \alpha_i^{IJ} \geq z_{ij'} - v_{j'} = z^{IJ} + \alpha_i^{IJ} + \beta_j^{IJ} - \left(V^{IJ} + \beta_{j'}^{IJ}\right)$$

and one can readily check that this inequality is always satisfied as an equality, reflecting the fact that i is indifferent between j and j', and symmetrically j is indifferent between i and i'.

3. *Take some female k in $K \neq J$, currently married to some i' in I. Then 'i is better off matched with j than k' gives:*

$$U^{IJ} + \alpha_i^{IJ} \geq z_{ik} - v_k = z^{IK} + \alpha_i^{IK} + \beta_k^{IK} - \left(V^{IK} + \beta_k^{IK}\right)$$

which is equivalent to

$$\alpha_i^{IJ} - \alpha_i^{IK} \geq U^{IK} - U^{IJ}$$

and we have proved that the conditions 3 are necessary. The proof is identical for 5.

4. *We now show that these conditions are sufficient. Assume, indeed, that they are satisfied. We want to show two properties. First, take some female j' in J, currently married to some l in $L \neq I$. Then i is better off matched with j than j'. Indeed,*

$$U^{IJ} + \alpha_i^{IJ} \geq z_{ij'} - v_{j'} = z^{IJ} + \alpha_i^{IJ} + \beta_j^{IJ} - \left(V^{LJ} + \beta_{j'}^{LJ}\right)$$

is a direct consequence of 5 applied to l. Finally, take some female k in $K \neq J$, currently married to some l in $L \neq I$. Then i is better off matched with j than j'. Indeed, it is sufficient to show that

$$U^{IJ} + \alpha_i^{IJ} \geq z_{ik} - v_k = z^{IK} + \alpha_i^{IK} + \beta_j^{IK} - \left(V^{LK} + \beta_k^{LK}\right)$$

But from 5 applied to k we have that:

$$\beta_k^{LK} - \beta_k^{IK} \geq V^{IK} - V^{LK}$$

and from 3 applied to i:

$$\alpha_i^{IJ} - \alpha_i^{IK} \geq U^{IK} - U^{IJ}$$

and the required inequality is just the sum of the previous two.

■

In summary, under our separability assumption, stability can readily be translated into a set of inequalities, each of which *relates to one agent only*. This property is crucial, because it implies that the model can be estimated using standard statistical procedures applied at the individual level, *without considering conditions on couples*. We now see how these insights can be implemented in practice.

4 Empirical implementation

4.1 Probabilities

Assume, first, that the classes are large, so that while the α and β are random the U^{IJ} and V^{IJ} are not.

Given the computations above, it is natural to make the following assumption:

Assumption HG (Heteroskedastic Gumbel): *The random terms α and β are such that*

$$\begin{aligned} \alpha_i^{IJ} &= \sigma^I.\tilde{\alpha}_i^{IJ} \\ \beta_i^{IJ} &= \mu^J.\tilde{\beta}_i^{IJ} \end{aligned}$$

where the $\tilde{\alpha}_i^{IJ}$ and $\tilde{\beta}_j^{IJ}$ follow independent Gumbel distributions $G(-k,1)$.

In particular, the $\tilde{\alpha}_i^{IJ}$ and $\tilde{\beta}_j^{IJ}$ have mean zero and variance $\frac{\pi^2}{6}$, therefore the α_i^{IJ} and β_j^{IJ} have mean zero and respective variance $\frac{\pi^2}{6}\left(\sigma^I\right)^2$ and $\frac{\pi^2}{6}\left(\mu^J\right)^2$. The previous Lemma then implies:

Lemma 3 *A set of necessary and sufficient conditions for stability is that*

1. *for all matched couple $(i \in I, j \in J)$ one has*

$$\alpha_i^{IJ} - \alpha_i^{IK} \geq \frac{U^{IK} - U^{IJ}}{\sigma^I} \quad \text{for all } K \tag{9}$$

$$\alpha_i^{IJ} - \alpha_i^{I0} \geq \frac{U^{I0} - U^{IJ}}{\sigma^I} \tag{10}$$

and

$$\beta_j^{IJ} - \beta_j^{KJ} \geq \frac{V^{KJ} - V^{IJ}}{\mu^J} \quad \text{for all } K \tag{11}$$

$$\beta_j^{IJ} - \beta_j^{0J} \geq \frac{V^{0J} - V^{IJ}}{\mu^J} \tag{12}$$

2. *for all single male $i \in I$ one has*

$$\alpha_i^{IJ} - \alpha_i^{I0} \leq \frac{U^{I0} - U^{IJ}}{\sigma^I} \quad \text{for all } J \tag{13}$$

3. *for all single female $j \in J$ one has*

$$\beta_j^{IJ} - \beta_j^{0J} \leq \frac{V^{0J} - V^{IJ}}{\mu^J} \quad \text{for all } J \tag{14}$$

Therefore, for any I and any $i \in I$:

$$\begin{aligned} a^{IJ} &= \Pr\left(i \text{ matched with a female in } J\right) \\ &= \frac{\exp\left(U^{IJ}/\sigma^I\right)}{\sum_K \exp\left(U^{IK}/\sigma^I\right) + 1} \\ &= \frac{\upsilon^{IJ}}{1 + \sum_K \upsilon^{IK}} \end{aligned}$$

and

$$\begin{aligned} a^{I0} &= \Pr(i \text{ single}) \\ &= \frac{1}{\sum_K \exp(U^{IK}/\sigma^I) + 1} \\ &= \frac{1}{1 + \sum_K \upsilon^{IK}} \end{aligned}$$

where $\upsilon^{IJ} = \exp\left(U^{IJ}/\sigma^I\right)$ and U^{I0} has been normalized to 0. Similarly, for any J and any female $j \in J$:

$$\begin{aligned} b^{IJ} &= P(j \text{ matched with a male in } I) && (15) \\ &= \frac{\exp\left(V^{IJ}/\mu^J\right)}{\sum_K \exp(V^{KJ}/\mu^J) + \exp(V^{0J}/\mu^J)} && (16) \\ &= \frac{\omega^{IJ}}{1 + \sum_K \omega^{KJ}} && (17) \\ b_{0J} &= P(j \text{ single}) = \frac{\exp\left(V^{0J}/\mu^J\right)}{\sum_K \exp(V^{KJ}/\mu^J) + \exp(V^{0J}/\mu^J)} \\ &= \frac{1}{1 + \sum_K \omega^{KJ}} && (18) \end{aligned}$$

where $\omega^{IJ} = \exp\left(V^{IJ}/\mu^J\right)$ and $V^{0J} = 0$.

These formulas can be inverted to give:

$$\upsilon^{IJ} = \frac{a^{IJ}}{1 - \sum_K a^{IK}} \tag{19}$$

and

$$\omega^{IJ} = \frac{b^{IJ}}{1 - \sum b^{KJ}} \tag{20}$$

therefore:

$$\begin{aligned} U^{IJ} &= \sigma^I \ln\left(\frac{a^{IJ}}{1 - \sum_K a^{IK}}\right) \\ V^{IJ} &= \mu^J \ln\left(\frac{b^{IJ}}{1 - \sum b^{KJ}}\right) \end{aligned}$$

In what follows, we assume that there are singles in each class: $a_{I0} > 0$ and $b_{0J} > 0$ for each I, J, implying that $\sum_K a^{IK} < 1$ and $\sum_K b^{KJ} < 1$ for all I, J.

A direct consequence of these results is that knowing the Z^{IJ} and the population sizes, we can *algebraically* compute the υs and the ωs. Indeed, let N_m^I (resp. N_f^J) be the number of males in class I (females in class J). At any match, the number of (I, J) couples can be computed in two ways that must be compatible, implying that:

$$N_m^I \frac{\upsilon^{IJ}}{1 + \sum_K \upsilon^{IK}} = N_f^J \frac{\omega^{IJ}}{1 + \sum_K \omega^{KJ}}$$

In addition, we know from (2) that:

$$\left(\upsilon^{IJ}\right)^{\sigma^I}.\left(\omega^{IJ}\right)^{\mu^J}=z^{IJ}$$

Therefore the υ^{IJ} satisfy the equations:

$$N_m^I\frac{\upsilon^{IJ}}{1+\sum_K\upsilon^{IK}}=N_f^J\frac{\left(z^{IJ}\right)^{1/\mu^J}\left(\upsilon^{IJ}\right)^{-\sigma^I/\mu^J}}{1+\sum_K\left(z^{KJ}\right)^{1/\mu^J}\left(\upsilon^{KJ}\right)^{-\sigma^K/\mu^J}} \tag{21}$$

This provides a system of M^2 equations in M^2 unknowns, that can be solved numerically. We conclude that, under the extreme value distribution assumption, one can directly compute the shadow prices associated with given population sizes and gain matrices.

Finally, define:

$$\bar{u}^I=E\left[\max_J\left(U^{IJ}+\sigma^I\tilde{\alpha}_i^{IJ}\right)\right]$$

In words, $\bar{u}^I$ is the expected utility of an agent in class I, given that this agent will chose a spouse in his preferred class. From the properties of Gumbel distributions, we have that:

$$\begin{aligned}\bar{u}^I &= \sigma^I E\left[\max_J\left(U^{IJ}/\sigma^I+\tilde{\alpha}_i^{IJ}\right)\right]\\ &= \sigma^I\ln\left(\sum_J\exp U^{IJ}/\sigma^I+1\right)=-\sigma^I\ln\left(a^{I0}\right)\end{aligned} \tag{22}$$

and similarly

$$\bar{v}^J=\mu^J\ln\left(\sum_I\exp V^{IJ}/\mu^J+1\right)=-\mu^J\ln\left(b^{0J}\right)$$

4.2 Why does heteroskedasticity matter?

An important property of the model just presented is heteroskedasticity: the variance of the unobserved heterogeneity parameters is class-specific. This property may in principle matter for various reasons. For one thing, the expected utility of an arbitrary agent in class I, as given by (22), is directly proportional to the standard deviation of the random shock. Indeed, remember that the agent chooses the class of his spouses so as to maximize his utility; and the expectation of the max of i.i.d variables increases with the variance. It follows that the utility generated by the access to the marriage market cannot be exclusively measured by the probability of remaining single (reflected in the $-\ln\left(a^{I0}\right)$ term).

This remark, in turn, has important consequences for measuring the marital college premium. To see how, start from a model in which the random component of the marital gain is homoskedastically distributed (i.e., the variance is the same across categories: $\sigma^I=\mu^J=1$ for all I,J). The marital college premium is measured by the

difference $\bar{u}^I - \bar{u}^K$, where I is the college education class whereas K is the high school graduate one. Condition (22) then implies that

$$\bar{u}^I - \bar{u}^K = \ln\left(\frac{a^{K0}}{a^{I0}}\right)$$

In words, the gain can directly be measured by the (log) ratio of singlehood probabilities in the two classes. The intuition is that people marry if and only if their (idiosyncratic) gain is larger than some threshold. If these random gains are homoskedastically distributed, then there is a one-to-one correspondence between the mean of the distribution for a particular class and the percentage of that class that is below the threshold, i.e. that remains single: the higher the mean, the smaller the proportion (see Figure 6). For instance, if one sees that college graduate are more likely to remain single than high school graduate ($a^{I0} > a^{K0}$, implying that $\ln\left(a^{K0}/a^{I0}\right) < 0$), we can conclude that the expected marital gain is smaller for college graduate ($\bar{u}^I < \bar{u}^K$), therefore that the marital college premium is negative.

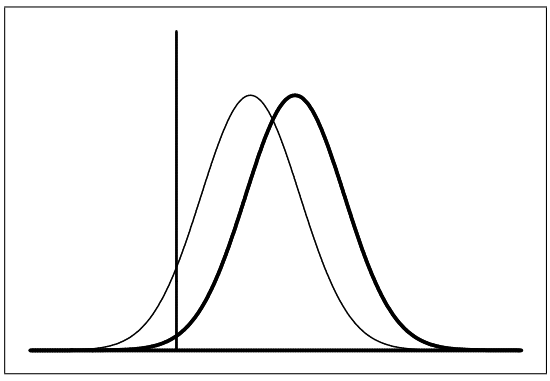

Figure 6: Homoskedastic random gains

Consider, now, the heteroskedastic version. Things are different here, because the percentage of single depends on both the mean and the variance. If educated women are more likely to remain single, it may be because the gain is on average smaller, but it may also be that the variance is larger (even with a higher mean), as illustrated in Figure 7. The one-to-one relationship needs not hold, and a higher percentage does not *necessarily* imply a smaller mean. One has to compute the respective variances - which, in turn, may affect the computation of the marital college premium.

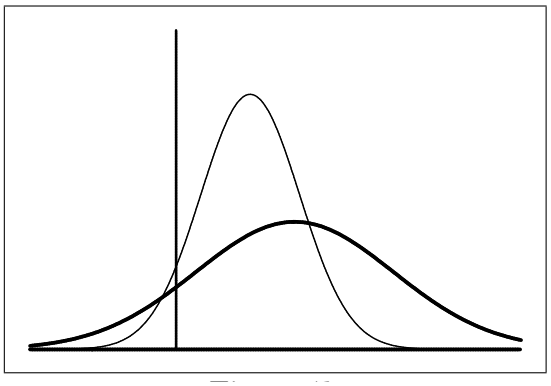

Figure 7: Heteroskedastic random gains

Technically, we now have that:

$$\bar{u}^I - \bar{u}^K = \sigma^K \ln\left(a^{K0}\right) - \sigma^I \ln\left(a^{I0}\right)$$

If $a^{I0} > a^{K0}$ *and* $\sigma^I \leq \sigma^K$, one can conclude that $\bar{u}^I - \bar{u}^K < 0$; but whenever $\sigma^I > \sigma^K$ the conclusion is not granted, and depends on the precise estimates. In other words, the conclusions drawn from the model may significantly depend on the assumptions made regarding its homoskedasticity properties. It is therefore important that these assumptions be testable rather than ad hoc - i.e., that homoskedasticity be imposed by the data (or at least compatible with them) rather than assumed a priori. In that sense, the estimation of the variances is a crucial part of the identification process.[3]

4.3 Extension: Covariates

The basic framework just described can be extended to the presence of covariates; i.e., we may specify the ε_{ik} (hence the α and β) as a function of individual characteristics (*other* than the matching ones). Let X_i be a vector of such characteristics of man i, and Y_j of woman j. We may use the following stochastic structure (where, for simplicity, we disregard heteroskedasticity):

$$\begin{aligned} \alpha_i^{IJ} &= X_i.\zeta_m^{IJ} + \tilde{\alpha}_i^{IJ} \\ \alpha_i^{I0} &= X_i.\zeta_m^{I0} + \tilde{\alpha}_i^{I0} \end{aligned}$$

[3]Note, however, that if the variances are assumed constant across time, then the *variations* in singlehood probability must still reflect similar changes in the expected gains from marriage. In other words, if we find that the percentage of, say, unskilled women remaining single has increased between two moments t and t', we can unambiguously conclude that the gains from marriage have diminished for these women over the period.

$$\begin{aligned}\beta_j^{IJ} &= Y_j.\zeta_f^{IJ} + \tilde{\beta}_j^{IJ} \\ \beta_j^{0J} &= Y_j.\zeta_f^{0J} + \tilde{\beta}_j^{0J}\end{aligned}$$

where $\zeta_m^{IJ}, \zeta_f^{IJ}$ are vector parameters, with the normalization $U^{I0} = \zeta_m^{I0} = 0$ and $V^{0J} = \zeta_f^{0J} = 0$, and where as above the $\tilde{\alpha}_i^{IJ}$ (resp. $\tilde{\beta}_j^{IJ}$) follow independent Gumbel distributions $G(-k, 1)$. Define $v^{IJ} = \exp\left(U^{IJ} + X_i.\zeta_m^{IJ}\right)$ and $\omega^{IJ} = \exp\left(V^{IJ} + Y_j.\zeta_f^{IJ}\right)$; then the computations are as above. In other words, we can estimate for $i \in I$:

$$a^{IJ} = \Pr(i \text{ matched with a female in } J) = \frac{\exp\left(U^{IJ} + X_i.\zeta_m^{IJ}\right)}{\sum_K \exp\left(U^{IK} + X_i.\zeta_m^{IK}\right) + \exp\left(U^{I0} + X_i.\zeta_m^{I0}\right)}$$

$$a^{I0} = \Pr(i \text{ single}) = \frac{\exp\left(U^{I0} + X_i.\zeta_m^{I0}\right)}{\sum_K \exp\left(U^{IK} + X_i.\zeta_m^{IK}\right) + \exp\left(U^{I0} + X_i.\zeta_m^{I0}\right)}$$

and the conclusions follow. In particular, these models can be estimated running standard logits.

5 Identification

We now consider the identification problem. In practice, we observe realized matchings - i.e., populations in each classes and the corresponding marital patterns. To what extend can one recover the fundamentals - i.e., the surplus matrix Z and the heteroskedasticity parameters σ and μ - crucially depends on the type of data available.

We first consider a static context, in which population sizes are fixed. We show that in that case, the model is exactly identified if we assume complete homoskedasticity, and not identified otherwise. Much more interesting is the situation in which population sizes vary over time while (some of) the structural parameters remain constant. Then one can identify both the surplus matrix Z and the heteroskedasticity parameters σ and μ, provided that they remain constant over time; actually, one can even introduce either time varying heteroskedasticity or a drift in the surplus matrix without losing identifiability; and finally, the model generates strong overidentifying restrictions. We consider the two cases successively.

5.1 The static framework

We start with a purely static framework. Define a model $\mathcal{M}$ as a set $\left(Z^{IJ}, \sigma^I, \mu^J\right)$ such that

$$g_{ij} = Z^{IJ} + \varepsilon_{ij}^{IJ}$$

with

$$\varepsilon_{ij}^{IJ} = \sigma^I \alpha_i^{IJ} + \mu^J \beta_j^{IJ} \tag{S}$$

and where the α_i^{IJ} and β_j^{IJ} follow independent Gumbel distributions $G(-k,1)$. Note that the model is clearly invariant when the $\left(Z^{IJ},\sigma^I,\mu^J\right)$ are all multiplied by a common, positive constant; for that reason, in what follows we normalize σ^1 to be 1.

The following result is valid for static (cross-sectional) data:

Proposition 4 *Assume that a model $\mathcal{M}=\left(Z^{IJ},\sigma^I,\mu^J\right)$ generates some matching probabilities $\left(a^{IJ},b^{IJ}\right)$, and let U^{IJ},V^{IJ} denote the corresponding dual variables. Then*

$$U^{IJ}=\sigma^I\log\frac{a^{IJ}}{1-\sum_K a^{IK}} \tag{23}$$

and

$$V^{IJ}=\mu^J\log\frac{b^{IJ}}{1-\sum_K b^{KJ}} \tag{24}$$

therefore

$$Z^{IJ}=\sigma^I\log\frac{a^{IJ}}{1-\sum_K a^{IK}}+\mu^J\log\frac{b^{IJ}}{1-\sum_K b^{KJ}}$$

Moreover, for any $\left(\bar{\sigma}^I,\bar{\mu}^J\right)\in\mathbb{R}^+$, the model $\mathcal{N}=\left(\bar{Z}^{IJ},\bar{\sigma}^I,\bar{\mu}^J\right)$ where

$$\frac{\bar{\sigma}^I}{\sigma^I}U^{IJ}+\frac{\bar{\mu}^J}{\mu^J}V^{IJ}=\bar{Z}^{IJ} \tag{25}$$

generates the same matching probabilities, and the corresponding, dual variables are

$$\bar{U}^{IJ} = \frac{\bar{\sigma}^I}{\sigma^I}U^{IJ} \tag{26}$$

$$\bar{V}^{IJ} = \frac{\bar{\mu}^J}{\mu^J}V^{IJ} \tag{27}$$

Conversely, if two models $\mathcal{M}=\left(Z^{IJ},\sigma^I,\mu^J\right)$ and $\mathcal{N}=\left(\bar{Z}^{IJ},\bar{\sigma}^I,\bar{\mu}^J\right)$ generate the same matching probabilities, then the conditions (25), (26) and (27) must hold.

Proof. *From the previous calculations, there is a one-to-one relationship between the a^{IJ} and the υ^{IJ}; the result follows.* ■

The previous result is essentially negative; it states that in a static context, the heteroskedastic version of the model is not identified. The heteroskedasticity parameters $\left(\sigma^I,\mu^J\right)$ can be chosen arbitrarily; for any value of these parameters, one can find values $\left\{Z^{IJ},I=1,...,N,J=1,...,M\right\}$ that exactly rationalize the data. An interpretation of the non identifiability result is in terms of units; i.e., the unit in which the υ and ωs are measured is not determined unless we make assumptions on the variances of the αs and βs. This negative result is important, in particular, for welfare comparisons. In a cross-sectional setting, comparing welfare between males and females or between individuals belonging to different classes is highly problematic, since it can only rely on arbitrary choices of the units.

5.2 Changes in population sizes

Much more promising is a situation in which one can observe the market over different periods (or for different cohorts), when the various populations change in respective sizes over the periods. Then a richer model can actually be estimated. We start with the benchmark case, then consider the generalized version that will be taken to data later.

5.2.1 The benchmark version

Let us now assume that the previous, heteroskedastic structural model $\mathcal{M} = \left(Z^{IJ}, \sigma^I, \mu^J\right)$ is faced with different cohorts of agents, $c = 1, ..., T$, with varying class compositions. The basic structure becomes:

$$g_{ij,c} = Z^{IJ} + \varepsilon_{ij,c}^{IJ}$$

with

$$\varepsilon_{ij,c}^{IJ} = \sigma^I \alpha_{i,c}^{IJ} + \mu^J \beta_{j,c}^{IJ} \tag{S}$$

Also, assume for the moment that each man marries a woman within his cohort. Empirically, this is not exactly right; women tend to marry slightly older men, so that in the application the wife of a man in cohort c typically belongs to cohort $(c+2)$ - a fact that will be taken into account in the empirical application, but can be ignored for the time being. As before, the matching model defines, for each cohort, a matching problem associated to shadow prices; the latter are now cohort specific. Under the same assumptions as above, the previous construct applies for each cohort, leading to the definition of U_c^{IJ} and V_c^{IJ}. Next, define $\upsilon_c^{IJ} = \exp\left(U_c^{IJ}/\sigma_I\right)$; then

$$\begin{aligned} a_c^{IJ} &= \Pr\left(i \in I \text{ matched with a female in } J \text{ in cohort } c\right) = \frac{\upsilon_c^{IJ}}{1 + \sum_K \upsilon_c^{IJ}} \\ a_c^{I0} &= \Pr\left(i \in I \text{ single}\right) = \frac{1}{1 + \sum_K \upsilon_c^{IJ}} \end{aligned}$$

therefore

$$\upsilon_c^{IJ} = \frac{a_c^{IJ}}{1 - \sum_K a_c^{IK}} \tag{28}$$

and similarly if $\omega_c^{IJ} = \exp\left(V_c^{IJ}/\mu_J\right)$:

$$\begin{aligned} b_c^{IJ} &= \Pr\left(j \in J \text{ matched with a female in } I \text{ in cohort } c\right) = \frac{\omega_c^{IJ}}{1 + \sum_K \omega_c^{IJ}} \\ b_c^{I0} &= \Pr\left(j \in J \text{ single}\right) = \frac{1}{1 + \sum_K \omega_c^{IJ}} \end{aligned}$$

implying that

$$\omega_c^{IJ} = \frac{b_c^{IJ}}{1 - \sum_K b_c^{IK}} \tag{29}$$

Moreover, we have

$$\sigma^I \log \upsilon_c^{IJ} + \mu^J \log \omega_c^{IJ} = Z^{IJ} \tag{30}$$

Now, let $p_c^{IJ} = \log \upsilon_c^{IJ} \left(= U_c^{IJ}/\sigma_I\right)$ and $q_c^{IJ} = \log \omega_c^{IJ} \left(= V_c^{IJ}/\mu_J\right)$. The crucial remark is that from (28) and (29), the υ and the ω, therefore the p and q, *are directly observable from the data*. It follows that (30) has a direct, testable implications. Indeed, define the vectors:

$$\begin{aligned} \mathbf{p}^{IJ} &= \left(p_1^{IJ}, ..., p_T^{IJ}\right) \\ \mathbf{q}^{IJ} &= \left(q_1^{IJ}, ..., q_T^{IJ}\right) \\ &\text{and} \\ \mathbf{1} &= (1, ..., 1) \end{aligned}$$

Then for each pair (I, J), the vectors $\mathbf{p}^{IJ}, \mathbf{q}^{IJ}$ and $\mathbf{1}$ must be colinear:

$$\sigma^I \ \mathbf{p}^{IJ} + \mu^J \ \mathbf{q}^{IJ} - Z^{IJ}\mathbf{1} = 0 \tag{31}$$

which generates a first testable restriction - namely that for each (I, J), the determinant

$$D_{IJ} = \left|\mathbf{p}^{IJ}, \mathbf{q}^{IJ}, \mathbf{1}\right|$$

must be zero.

If that restriction is satisfied, assume that either p^{IJ} or q^{IJ} is not constant over the cohorts. Then the vectors $\mathbf{p}^{IJ}$ and $\mathbf{1}$ (or $\mathbf{q}^{IJ}$ and $\mathbf{1}$) are linearly independent, so that the linear combination in (31) is unique up to a common multiplicative constant. Since, in our case, the constant is pinned down by the normalization $\sigma^1 = 1$, we conclude that *for each pair (I, J), the regression exactly identifies σ^I, μ^J and Z^{IJ}*. Finally, since each σ^I but σ^1 (resp. each μ^J) is identified from N (M) different regressions, the model generates a second set of overidentifying restrictions.

Finally, a more parsimonious version of the model obtains by imposing that the σs and the μs are identical *across classes* (i.e., $\sigma^I = \sigma$ for all I and $\mu^J = \mu$ for all J), although these values may be different between gender (i.e., we do *not* impose that $\sigma = \mu$). Condition (31) is then strengthened: if we define the vectors $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{1}_{IJ}$ in $R^{N\times M\times T}$ by:

$$\mathbf{p} = \left(\mathbf{p}^{11}, ..., \mathbf{p}^{NM}\right), \mathbf{q} = \left(\mathbf{q}^{11}, ..., \mathbf{q}^{NM}\right) \text{ and } \mathbf{1}_{IJ} = (0, ...0, 1, ..., 1, 0, ...0)$$

then (keeping the normalization $\sigma = 1$):

$$\mathbf{p} = -\mu \ \mathbf{q} + \sum_{I,J} Z^{IJ}\mathbf{1}_{IJ} \tag{32}$$

This requires that $(2 + NM)$ vectors be colinear in a space of dimension NMT, a strong restriction as soon as $T \geq 2$; moreover, if this property is satisfied, then μ and the Z^{IJ} are identified.

We conclude that *whenever the populations are not constant across cohorts, both the homoskedastic and the heteroskedastic versions of the benchmark structural model are (vastly) overidentified.*

5.2.2 Extension: category-specific drifts

The previous, overidentification result suggest that a more general version of the model may actually be identifiable. We now proceed to show that this is indeed the case. Specifically, we now relax the assumption that the Z_c^{IJ} are constant across cohorts; we therefore introduce category-specific drifts, whereby the Z^{IJ}s vary according to:

$$Z_c^{IJ} = \zeta_c^I + \xi_c^J + Z^{IJ} \tag{33}$$

This is equivalent to assuming that, for all (I, J) and (K, L), the second difference:

$$Z_c^{IJ} - Z_c^{IL} - Z_c^{KJ} + Z_c^{KL} = Z^{IJ} - Z^{IL} - Z^{KJ} + Z^{KL}$$

is independent of c. Clearly, what we are assuming is therefore that the supermodularity properties of the marital gains are constant over time

It is important to stress what this extension allows and what it leaves aside. Under (33), the benefits of marriage may evolve over time (although the variances do not); and these evolutions may be both gender- and education- specific. In words, we allow, for instance, the gains generated by marriage to decrease less for an educated women than for an unskilled man. However, the components reflecting *complementarity* between education classes - the Z^{IJ} - are left invariant. In particular, the forces driving the assortativeness of the match are supposed to be constant for the various cohorts. Our challenge is precisely to see - and test - whether such a null is compatible with the evolutions in marital patterns observed over the last decades.

Normalizations The form (33) requires additional normalizations. We normalize $\zeta_1^I = \xi_1^J = 0$ so that $Z^{IJ} = Z_1^{IJ}$. Also, note that for any $c > 1$, the ζ_c^I and ξ_c^J are only defined up to a (common) additive constant; i.e. for any given scalar k, one can replace $\left(\zeta_c^I, \xi_c^J\right)$ with $\left(\zeta_c^I + k, \xi_c^J - k\right)$ for all (I, J) without changing (33). We can therefore normalize ξ_c^1 to be zero for all c.

Testing the framework Under (33), equation (30) becomes:

$$\sigma^I \; p_c^{IJ} + \mu^J \; q_c^{IJ} = \zeta_c^I + \xi_c^J + Z^{IJ} \quad \forall I, J, c \tag{34}$$

This implies that for all I and all $J \geq 2$, we have:

$$\sigma^I \left(\; p_c^{IJ} - p_c^{I1}\right) + \mu^J \; q_c^{IJ} - \mu^1 \; q_c^{I1} = \xi_c^J + Z^{IJ} - Z^{I1} \tag{35}$$

Computing this expression for $I = 1$ and differencing:

$$\sigma^I \; \left(p_c^{IJ} - p_c^{I1}\right) - \sigma_1 \; \left(p_c^{1J} - p_c^{11}\right) + \mu^J \; \left(q_c^{IJ} - q_c^{1J}\right) - \mu^1 \left(q_c^{I1} - q_c^{11}\right) = Z^{IJ} - Z^{I1} - Z^{1J} + Z^{11}. \tag{36}$$

This requires a normalization since all terms can be multiplied by the same factor. We could choose for instance $\sigma_1 = 1$, so that

$$p_c^{1J} - p_c^{11} = \sigma^I \; \left(p_c^{IJ} - p_c^{I1}\right) + \mu^J \; \left(q_c^{IJ} - q_c^{1J}\right) - \mu^1 \left(q_c^{I1} - q_c^{11}\right) - \left(Z^{IJ} - Z^{I1} - Z^{1J} + Z^{11}\right)$$

From this, we derive a first testable restriction. Define the vectors:

$$\begin{aligned}\mathbf{P}^{IJ} &= \left(p_1^{IJ} - p_1^{I1}, ..., p_T^{IJ} - p_T^{I1}\right)\\ \mathbf{Q}^{IJ} &= \left(q_1^{IJ} - q_1^{1J}, ..., q_T^{IJ} - q_T^{1J}\right)\\ \mathbf{R}^{IJ} &= \left(p_1^{1J} - p_1^{11}, ..., p_T^{1J} - p_T^{11}\right)\\ &\text{and}\\ \mathbf{1} &= (1, ..., 1)\end{aligned}$$

Then for each pair $(I > 1, J > 1)$:

$$\mathbf{R}^{IJ} = \sigma^I\ \mathbf{P}^{IJ} + \mu^J\ \mathbf{Q}^{IJ} - \mu^1 \mathbf{Q}^{I1} - \left(Z^{IJ} - Z^{I1} - Z^{1J} + Z^{11}\right)\mathbf{1} \tag{37}$$

and $\mathbf{R}^{IJ}$ belongs to the subspace generated by $\left\{\mathbf{P}^{IJ}, \mathbf{Q}^{IJ}, \mathbf{Q}^{I1}, \mathbf{1}\right\}$, a first testable restriction for each $(I > 1, J > 1)$. A second set of testable restrictions comes from the fact that when we decompose $\mathbf{R}^{IJ}$ over the basis $\left\{\mathbf{P}^{IJ}, \mathbf{Q}^{IJ}, \mathbf{Q}^{I1}, \mathbf{1}\right\}$, the coefficient of $\mathbf{P}^{IJ}$ (resp. $\mathbf{Q}^{IJ}$,resp. $\mathbf{Q}^{I1}$) does not depend on J (resp. I, resp. is constant).

In practice, we will test both restrictions simultaneously by 'stacking' the various subvectors into 'large' vectors and running the regression in (37). The main implication of our model is that the fit should be *exact*, resulting in an R^2 non significantly different from 1. In practice, we will be projecting a vector with 132 components over a 9-dimensional space, and we expect the projection to coincide with the initial vector—a strong test indeed.

Identification: the main result Finally, should we fail to reject, the model is identified. To see why, note that the decomposition of $\mathbf{R}^{IJ}$ over $\left\{\mathbf{P}^{IJ}, \mathbf{Q}^{IJ}, \mathbf{Q}^{I1}, \mathbf{1}\right\}$ is generically unique; the σ^I and μ^J are therefore (over) identified as the respective coefficients of the first two vectors in the decomposition, and μ^1 as minus the coefficient of the third. Rewriting (34) for $c = 1$ gives

$$\sigma^I\ p_1^{IJ} + \mu^J\ q_1^{IJ} = Z^{IJ}$$

which shows that the Z^{IJ} are identified. Last, applying (34) identifies ζ_c^I for all I since we set $\xi_c^1 \equiv 0$; and (35) then identifies ξ_c^J for all $J \geq 2$.

A more parsimonious version Coming back to the parsimonious version introduced above ($\sigma^I = \sigma$ for all I and $\mu^J = \mu$ for all J), condition (36) becomes (with the same notations as above):

$$\sigma\ \left(\left(p_c^{IJ} - p_c^{I1}\right) - \left(p_c^{1J} - p_c^{11}\right)\right) + \mu\left(\ \left(q_c^{IJ} - q_c^{1J}\right) - \left(q_c^{I1} - q_c^{11}\right)\right) = Z^{IJ} - Z^{I1} - Z^{1J} + Z^{11}$$

In this case, the computation of μ has a simple and intuitive interpretation. For any $(I \geq 1, J \geq 1)$, let $\Delta_2 a_c^{IJ}$ denote the second difference of the log probability a_c^{IJ} that a man in I marries a woman in J, taking for instance the first category as a benchmark for both genders:

$$\Delta_2 a_c^{IJ} = \ln a_c^{IJ} - \ln a_c^{I1} - \ln a_c^{1J} + \ln a_c^{11}$$

Clearly, the use of such second differences refers to the supermodularity properties *of the (log) probabilities*. In particular, if $\ln a_c^{IJ}$ is additively separable:

$$\ln a_c^{IJ} = s_c^I + t_c^J$$

then $\Delta_2 a_c^{IJ} = 0$ for all (I, J, c).

Now, let $\Delta_3 a_c^{IJ}$ denote the *variation* of this second difference over cohorts:

$$\Delta_3 a_c^{IJ} = \Delta_2 a_{c+1}^{IJ} \ - \Delta_2 a_c^{IJ}$$

We can similarly define $\Delta_2 b_c^{IJ}$ and $\Delta_3 b_c^{IJ}$ for women. Then our model implies that:

$$\frac{\Delta_3 a_c^{IJ}}{\Delta_3 b_c^{IJ}} = -\frac{\mu}{\sigma}$$

This generates a simple test - namely, the ratio $\Delta a_c^{IJ}/\Delta b_c^{IJ}$ should not depend on the classes I and J - and a simple, non parametric estimator of the ratio σ/μ (remember that some normalization, say $\sigma = 1$, is still needed). For instance, the ratio is close to zero if the second difference Δ_2 varies much less for men than for women.

***

Actually, more complex models can in principle be tested and estimated in this framework. For instance, one may assume a uniform drift in the Zs but allow for cohort-specific variances; the model would then become:

$$g_{ij,c} = Z^{IJ} + \zeta_c + \sigma_c^I \alpha_{i,c}^{IJ} + \mu_c^J \beta_{j,c}^{IJ}$$

Again, one can show that this model (i) generates testable restrictions and (ii) is identified up to simple normalizations (a formal proof is available from the authors).

6 The Restricted Sample

When studying matching patterns, we have to decide which match we consider: the current match of a couple, or earlier unions in which the current partners entered? also, do we define a single as someone who never married, or as someone who is currently not married?

We chose to only keep first matches, and never-married singles. Given this sample selection, in each cohort we miss:

- those individuals who died before the 2008 Survey;
- those who are single in 2008 but were married before: there are
  - 36,094 separated
  - 218,839 divorced
  - 143,963 widowed.
- those who are married in 2008, but not in a first marriage—more precisely, in Table 1, we only kept the top left cell.

Number of marriages	1	2	≥ 3	Total
1	384,291	42,147	5,945	432,383
2	46,773	56,210	14,146	117,129
≥ 3	7,250	15,334	9,069	31,653
Total	438,314	113,691	29,160	581,165

Table 1: Men in rows, women in columns

6.1 Education categories

In order to simplify the analysis, we reduced the number of categories to three. We decided to aggregate "Some college" and "College Graduates". Then our highest education category includes 2-year and 4-year college graduates, along with college students who did not graduate. One may want to separate 4-year college graduates instead, and aggregate the rest of our third category with high-school graduates; we give the results in an appendix.

6.2 Marriage Patterns

To examine marriage patterns, we dropped the small number of couples where one partner married after age 35 (recall that these are first unions); and we focus on individuals who had reached the end of this "marriage window" in 2008. This leaves us with 179,353 couples, 44,344 single men, and 32,985 single women.

Figures 8 and 9 plot the proportion of diagonal matches for men and women who are married; that is,

$$\frac{a_c^{II}}{1 - a_c I 0} \quad \text{and} \quad \frac{b_c^{JJ}}{1 - b_c 0 J}.$$

The most striking change is that college-educated men now of course find it easier to find a college-educated wife; no similar change can be seen for college-educated women.

The proportion of marriages in which the husband is more educated than the wife has fallen quite dramatically. Figure 10 shows that since the early 1980s, there are now more marriages in which the wife has a higher level of education (this figure uses 4 levels of education.)

7 The Tests

We estimate the $\Pr(J|I,c)$ and $\Pr(I|J,c)$ probabilities by the obvious nonparametric technique of counting numbers of marriages in cells, assuming that a man of cohort c marries a woman of cohort $(c+2)$ (the two-year gap is roughly the observed average, with a very slow decrease over time.)

Then we reconstitute the p and q terms and we run the pooled regression, taking I and $J = 3$ rather than 1 as reference, since category 1 (high-school dropouts) becomes less numerous over time. We also found it more convenient to normalize estimates using the restriction

$$Z^{33} + Z^{11} - Z^{13} - Z^{31} = 1,$$

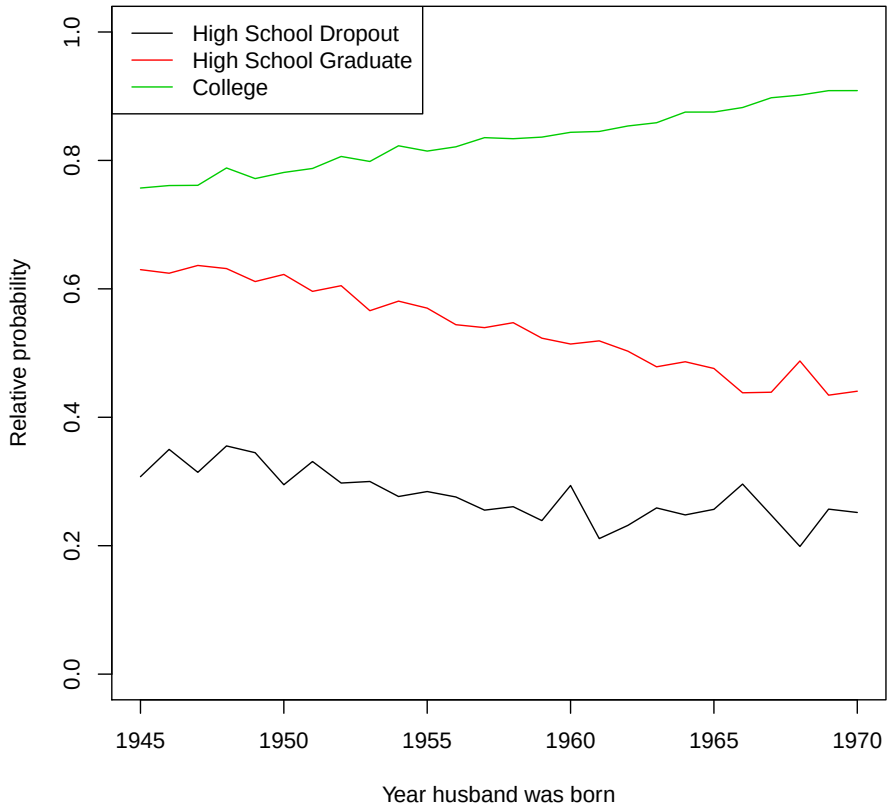

Figure 8: Relative probability of diagonal matches for men

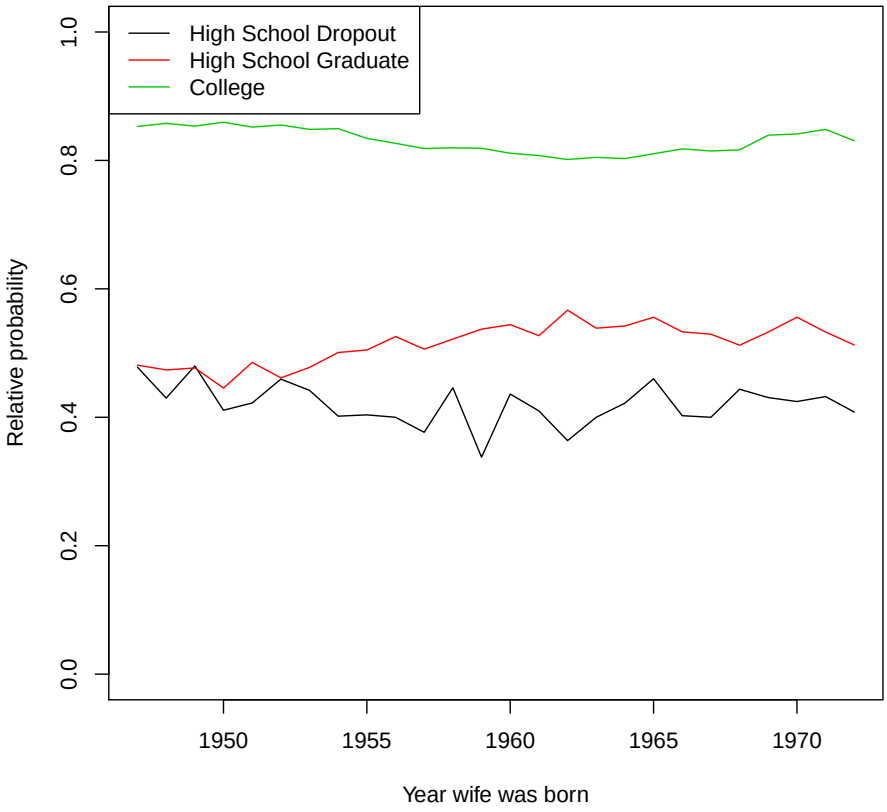

Figure 9: Relative probability of diagonal matches for women

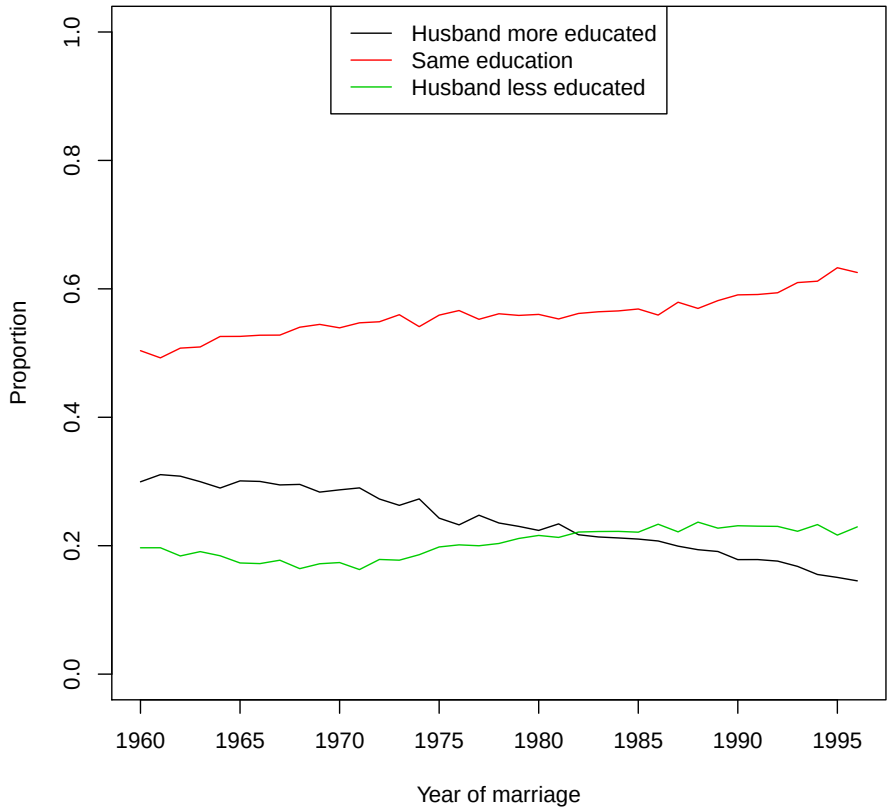

Figure 10: Relative education of partners

which scales the constant part of the joint surplus by making the largest cross-difference term equal to one. This allows us to maintain the symmetry between men and women.

We could measure the fit of our models in a variety of ways. For simplicity, we use the R^2 of equation (37), relative the to that of the model in which there is no heterogeneity at all (all σ^I and μ^J are zero.)

The model with all σ's and μ's equal has an R^2 of 0.993, which is remarkably high since it explains 132 numbers with only 4 parameters. Models in which the σ's and μ's are less constrained turn out to increase the R^2 very little; the main effect is that the estimates become less precise. As an illustration, the unconstrained estimators for the standard errors of the heterogeneity terms are in Table 2 (standard errors of the estimators are in parentheses.) For the model with all σ's and μ's equal, the common estimate is 0.115, with a very small standard error of 0.001.

Group/Gender	Men	Women
HSD	0.112	0.133
	(0.018)	(0.030)
HSG	0.069	0.094
	(0.020)	(0.022)
COLL	0.087	0.142
	(0.016)	(0.019)

Table 2: σ^I in rows, μ^J in columns

We focus from now on the model in which all σ's and μ's are equal, since the data does not reject that constraint. The reconstructed values of the Z^{IJ} (the cohort-independent part of the joint surplus) are in Table 3. We ran "supermodularity tests" by evaluating the 9 cross-difference terms

$$Z^{KL} + Z^{IJ} - Z^{IL} - Z^{KJ}$$

with $K > I$ and $L > J$. Rather strikingly, they were all positive. Since the joint surplus

$$Z^{IJ} + \xi_c^I + \zeta_c^J$$

adds to Z a part which is additive in I and J, we can conclude that the joint surplus is supermodular in educations.

Group	HSD	HSG	COLL
HSD	0.274	0.244	−0.177
HSG	0.134	0.455	0.134
COLL	−0.105	0.321	0.437

Table 3: Z values: men in rows, women in columns

Our method also yield estimates of the ξ and ζ terms, so that for any value of (I, J) we can reconstruct changes in the joint surplus across cohorts. Figure 11 focuses on

"diagonal" matches $I = J$. The dashed horizontal lines give the values of Z^{II}, and the curves add $\xi_c^I + \zeta_c^I$. The differences that prevailed for the older cohorts are dwarfed by the evolutions since then: while all categories of matches have become less attractive (relative to staying single), the fall is much steeper for high-school dropouts.

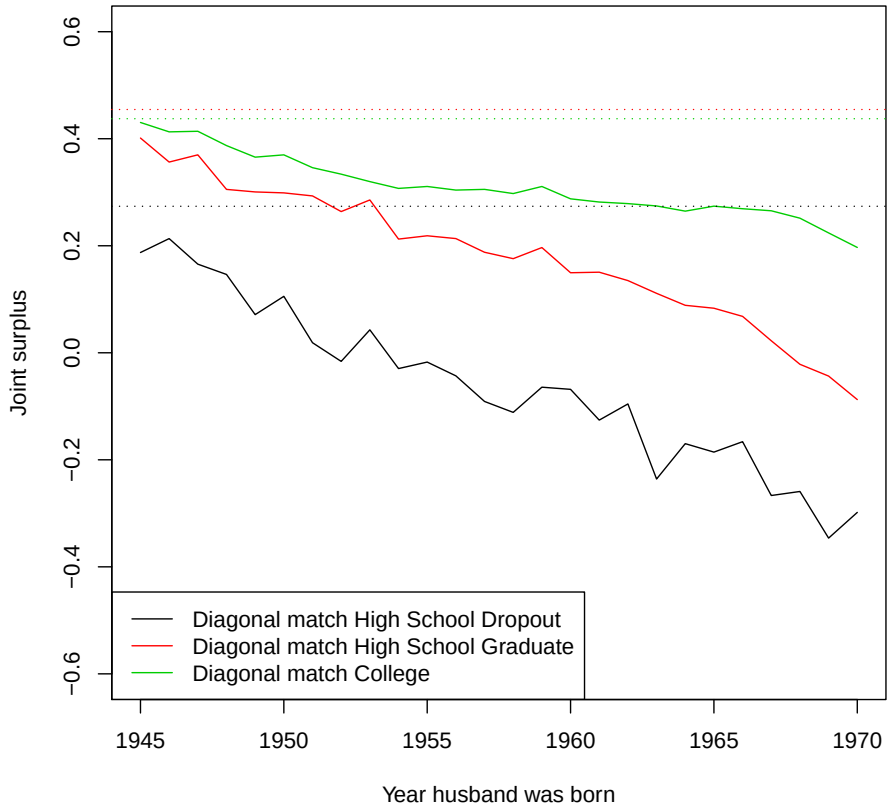

Figure 11: Joint surplus of diagonal matches

Our estimates also allow us to reconstruct changes in U_c^{IJ} and V_c^{IJ} over time. Again, we focus on diagonal terms $I = J$, which are plotted in Figures 12 (for men) and 13 (for women). The divergence is even more obvious if we look at the expected gains from marriage of the various categories U_c^I (Figure 14) and V_c^J (Figure 15).

Figure 15, in particular, shows that the fate of college-educated women has spectacularly improved relative to that of both other categories. This is confirmed on Figure 16, which plots the evolution of the "marital college premium" $(U_c^3 - U_c^2)$ and $(V_c^3 - V_c^2)$ over cohorts for both genders. It has always been positive for men, but it

used to be negative for women. The marital college premium of women turned positive for cohorts born around 1965, who married around 1990; and while for men too it has increased, the increase for women is clearly larger.

Finally, Figure 17 plots the evolution of the marital college premium for both genders for cohorts separated by 5-year intervals. If the marital college premium had changed in the same way for both genders, the red line would be on the diagonal. While the trajectory cannot be interpreted in terms of "shares of the surplus", its message is fairly clear.

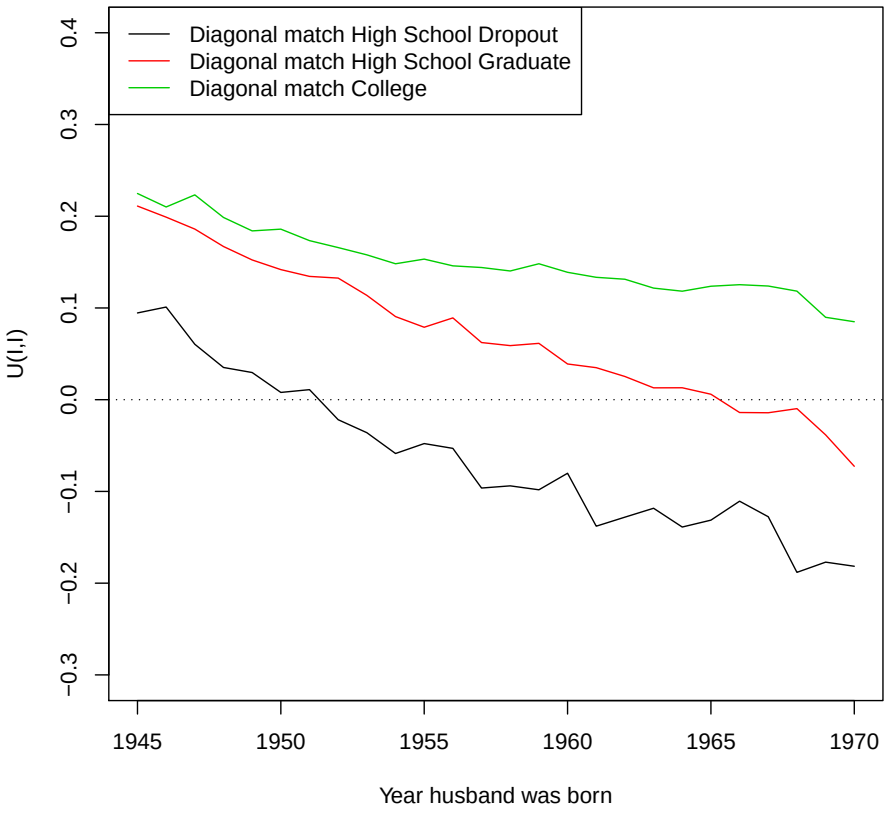

Figure 12: Gain from diagonal matching for men

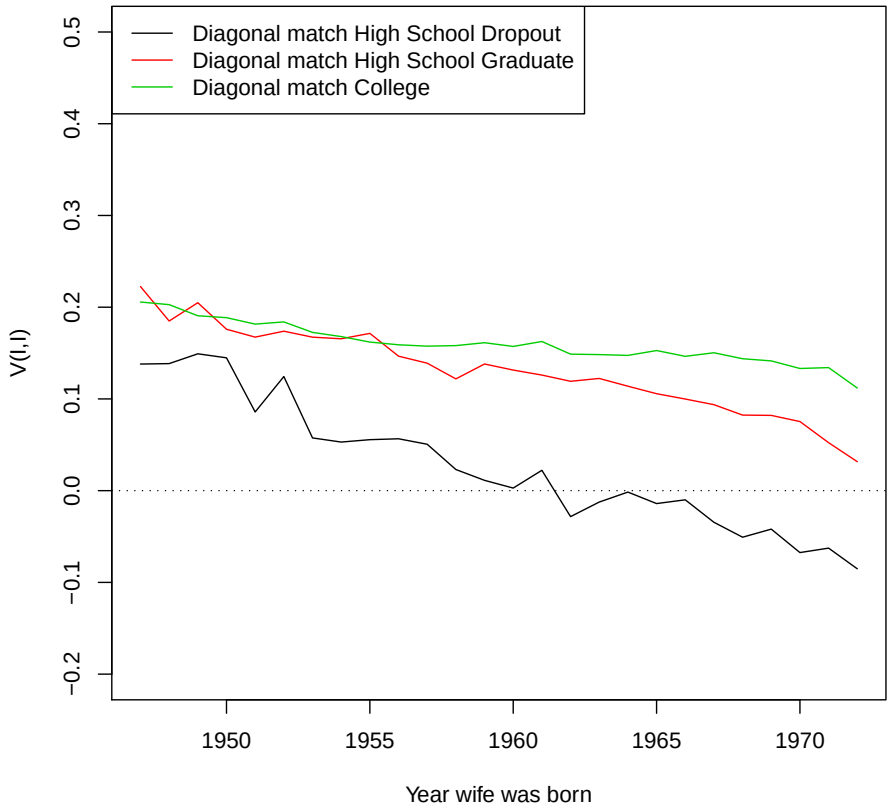

Figure 13: Gain from diagonal matching for women

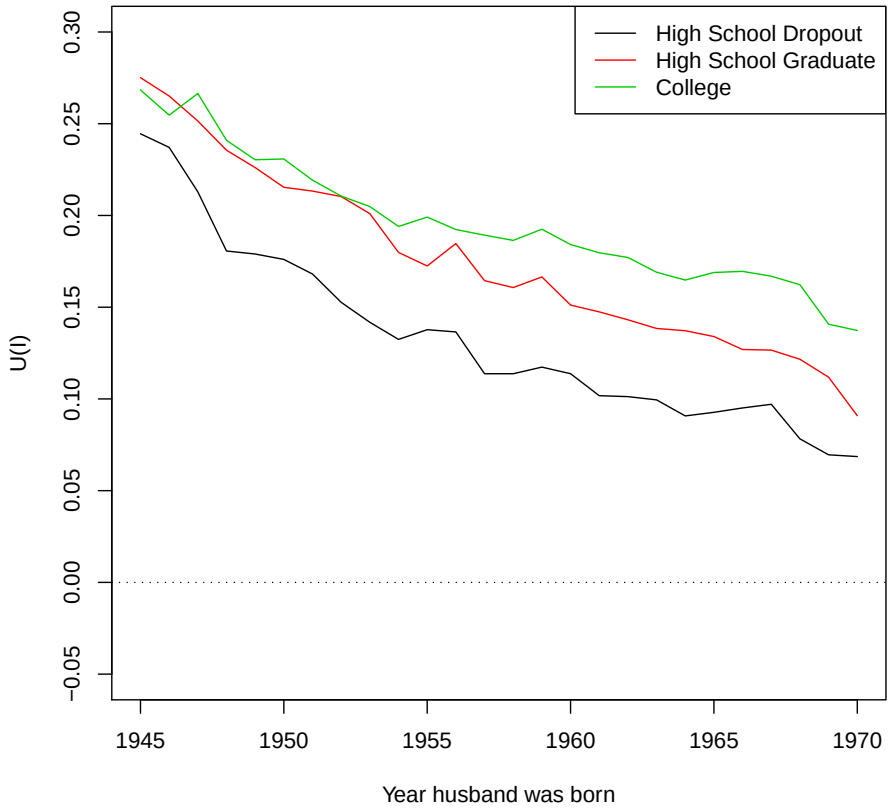

Figure 14: Gain from marriage for men

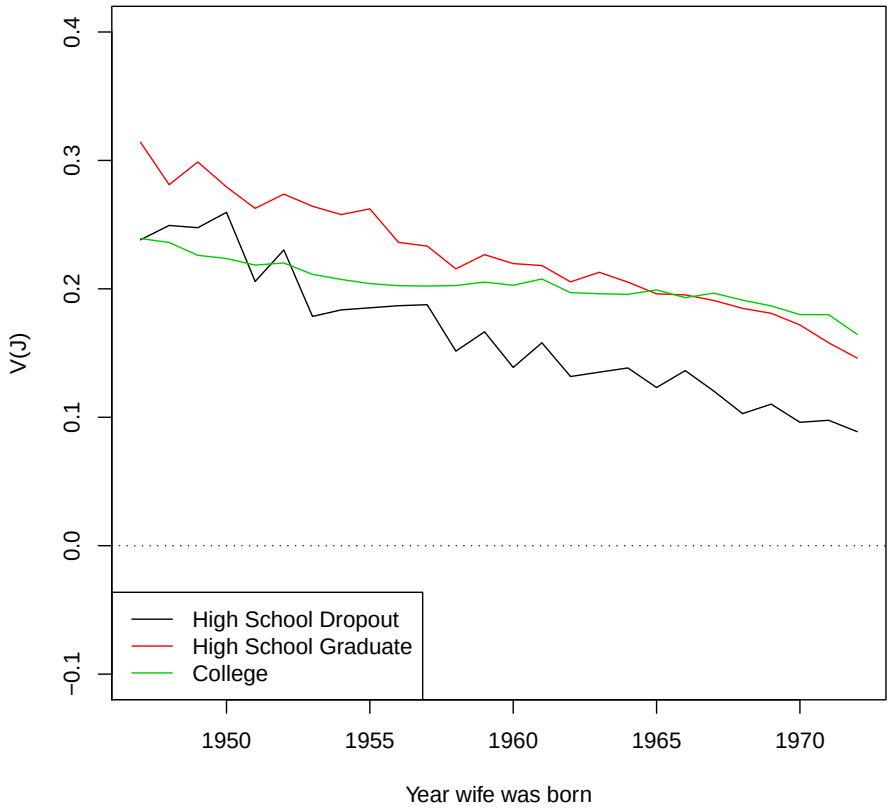

Figure 15: Gain from marriage for women

Figure 16: Marital college premium

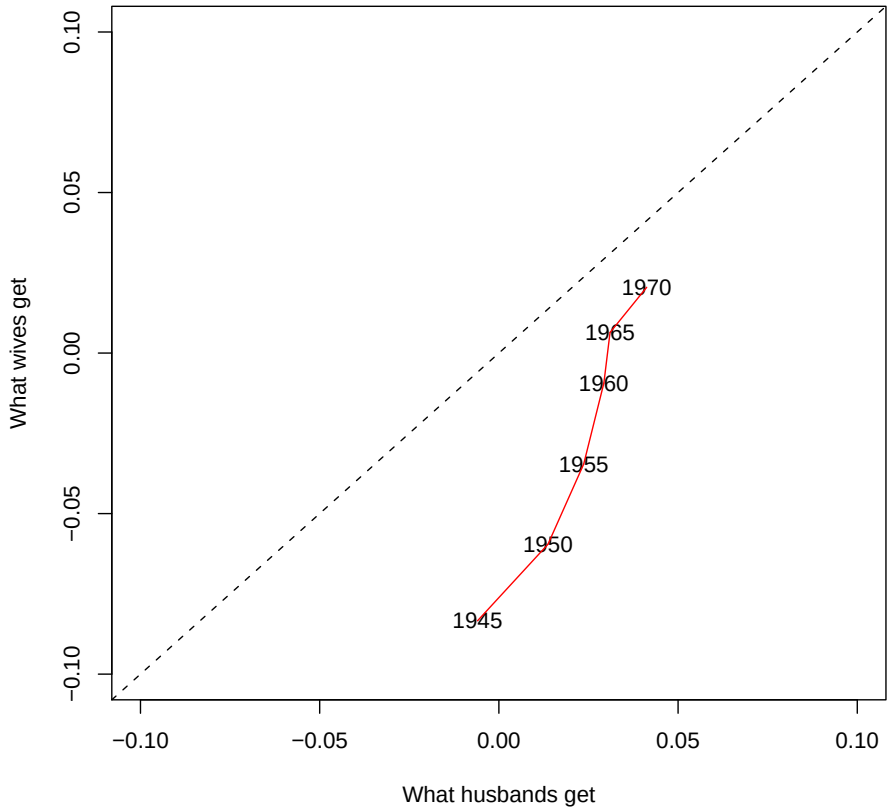

Figure 17: Marital college premium trajectories

Appendix: Estimates with Different Education Categories

We give here estimates obtained by aggregating "some college" with "high-school graduates" rather than with "college graduates". Once again, the model with all σ's and μ's equal has an excellent fit, with an R^2 of 0.988; and we present estimates for this constrained model.

Figures 18, 19, 20, 21, 22 and 23 correspond to Figures 12, 13, 14, 15, 16 and 17 in the main text.) The numbers differ of course, but the qualitative conclusions are very similar.

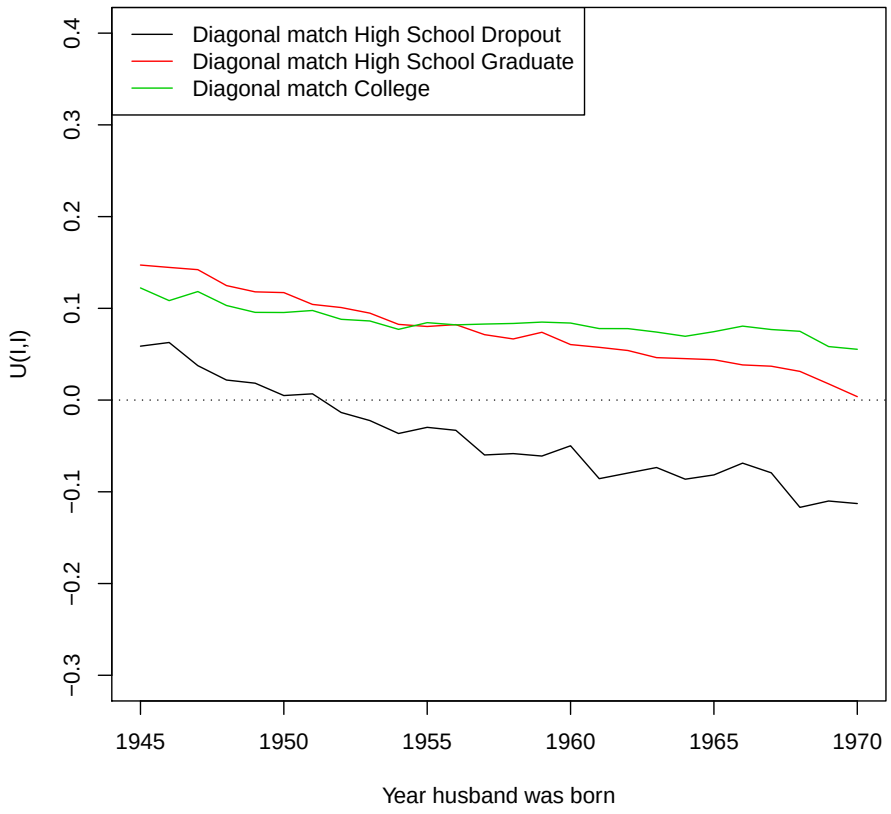

Figure 18: Gain from diagonal matching for men—different education categories

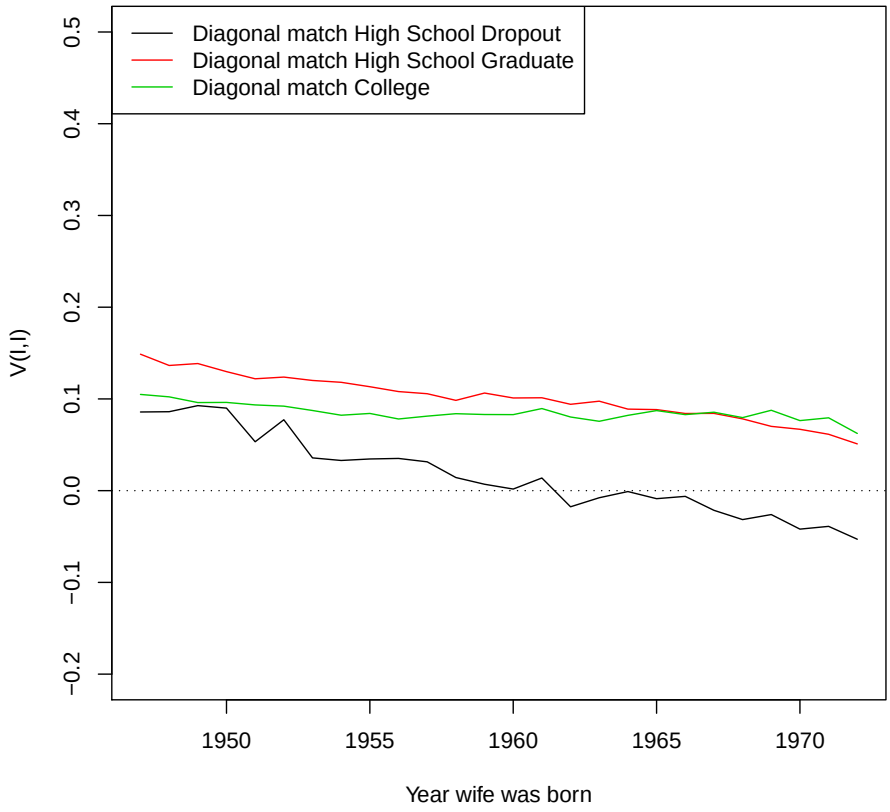

Figure 19: Gain from diagonal matching for women—different education categories

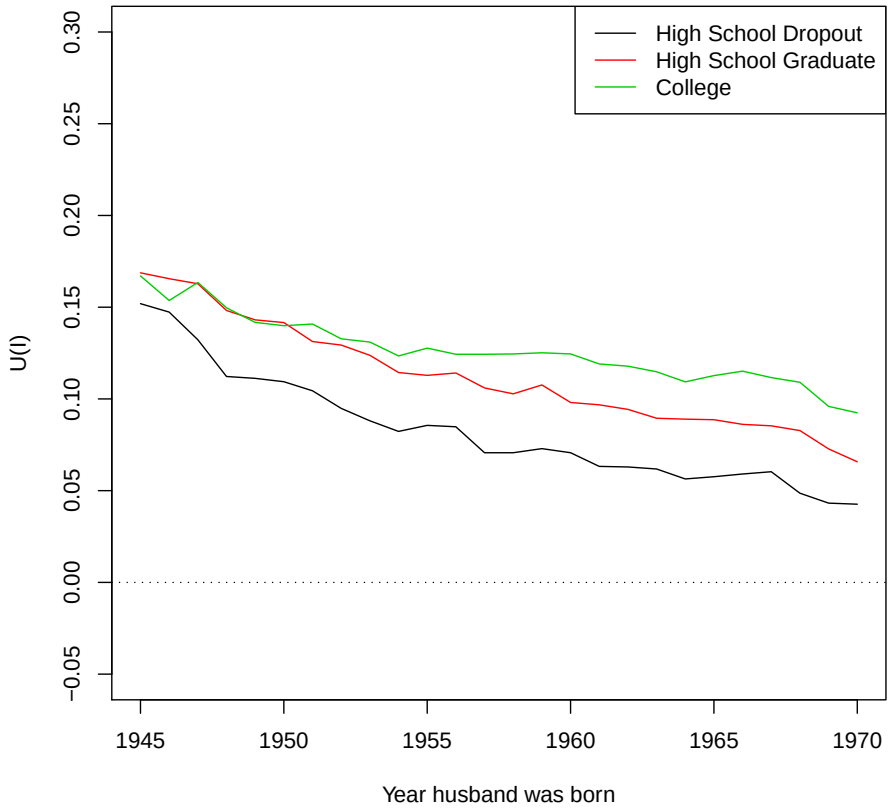

Figure 20: Gain from marriage for men—different education categories

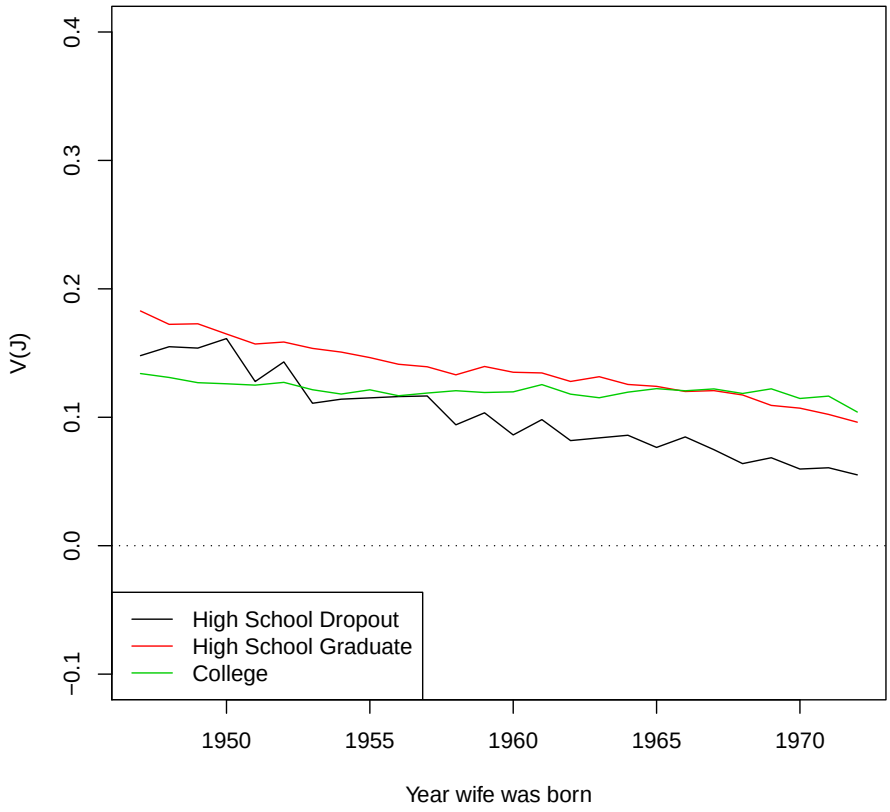

Figure 21: Gain from marriage for women—different education categories

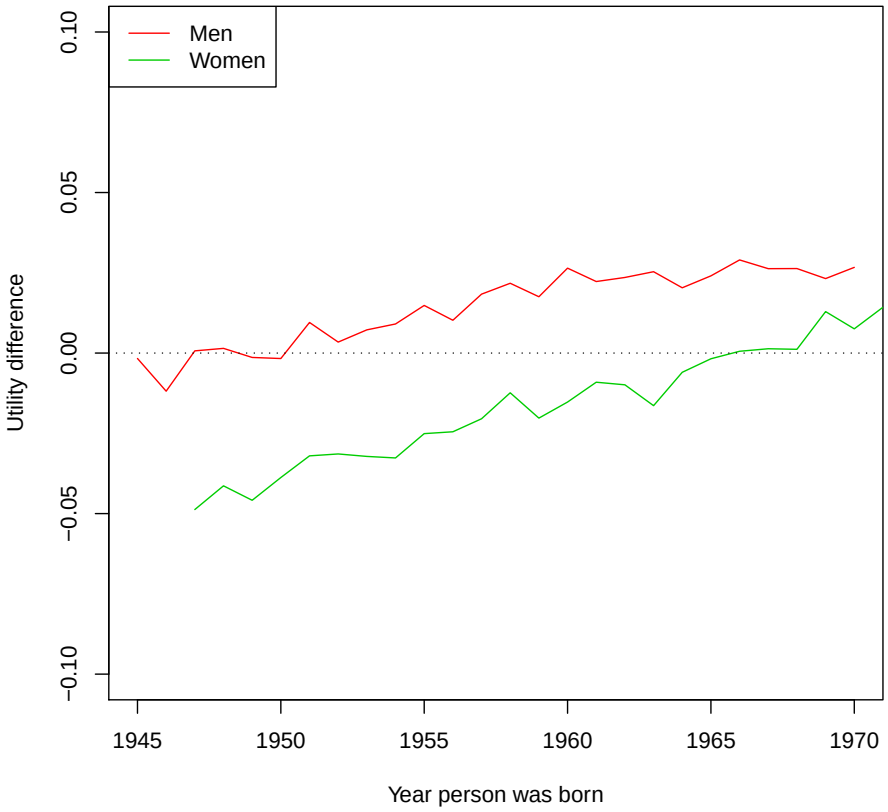

Figure 22: Marital college premium—different education categories

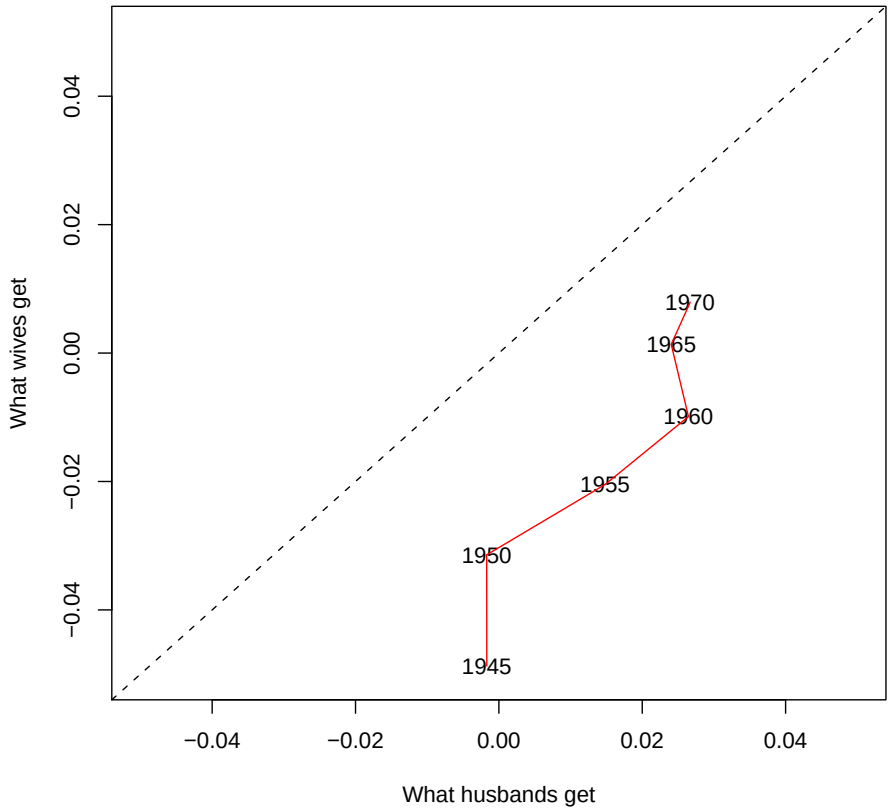

Figure 23: Marital college premium trajectories—different education categories

References

[1] Bruze, G., Svarer M. and Y. Weiss (2010), "The Dynamics of Marriage and Divorce" University of Aarhus.

[2] Chiappori, PA, Iygin, M. and Y. Weiss (2009), "Investment in Schooling and the Marriage Market," *American Economic Review,* 99, 1689-1714.

[3] Choo, E. and A. Siow, 'Who Marries Whom and Why', *Journal of Political Economy*, 114 (2006), 175-201

[4] Galichon, A., and B. Salanié (2010), "Matching with Trade-offs: Revealed Preferences over Competing Characteristics", mimeo.

[5] Goldin, C., Katz, L. (2002), "The power of the pill: oral contraceptives and women's career and marriage decisions", *Journal of Political Economy*, 110-4, 730-770

[6] Greenwood, J., Seshadri, A. and M. Yorukoglu (2005), "Engines of Liberation," *The Review of Economic Studies,* 72, 109-133

[7] Mare, R. (1991), "Five Decades of Educational Assortative Mating." *American Sociological Review* 56, 15-32.

[8] Mare, R. (2008), "Educational Assortative Mating in Two Generations," Department of Sociology, University Of California Los Angeles.

[9] Michael, R., (2000), "Abortion decisions in the U.S", in Laumann, E, Michael, R (eds.), *Sex, Love and Health: Public and Private Policy*, University of Chicago Press.

[10] Ruggles, S., M. Sobek, T. Alexander, C. Fitch, R. Goeken, P. Hall, M. King, and C.Ronnander (2008): "Integrated Public Use Microdata Series: Version 4.0," Discussion paper, Minnesota Population Center.

[11] Weiss, Y. and R. Willis (1997), Match Quality, New Information, and Marital Dissolution," *Journal of Labor Economics,* 15, 293-329, January.