Partner Choice and the Marital College Premium

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

December 13, 2010

Abstract

Several theoretical contributions have argued that the returns to schooling within marriage play a crucial role for human capital investments. Our paper empirical investigates the evolution of these returns over the last decades. We consider a frictionless matching framework à la Becker-Shapley-Shubik, in which the gain generated by a match between two individuals 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; following Choo and Siow (2006), we assume the latter component has an additively separable structure. We derive a complete, theoretical characterization of the model. We show that under the null that supermodularity on the surplus function is invariant over time and errors have extreme value distributions, the model is overidentified even if the surplus function varies over time. We apply our method to US data on individuals born between 1943 and 1972. The overidentification tests do not reject; moreover, we find that the deterministic part of the surplus is indeed supermodular and that, in line with theoretical predictions, the "marital college premium" has increased more for women than for men over the period.

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

*Columbia University

†Columbia University.

‡Tel Aviv University.

[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.

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, one can identify a more general structure, in which the systematic component of the surplus involves class-specific temporal drifts; moreover, this generalized model still generates strong overidentification restrictions.[3]

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,

[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.

[3]We show that an even more general structure, in which the scale of individual heterogeneity may vary by education class, could also be identified. In practice, however, the assumption of identical homogeneity is not rejected by the data.

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 a change in 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. We find strong evidence that for educated women, the additive gains from marriage have shifted over time. 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). However, 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 it 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, the following result. Assume we take our structural framework under the assumption that the forces driving assortative matching (i.e., the matrix of marital

gains by education classes) remain constant, and we use it to generate marriage data. If the log-linear regression adopted in the sociological literature were applied to such data, they would (spuriously) conclude 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 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 homoskedasticity 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 dynamic context, provided that the structures driving assortative matching remains constant. In particular, our identification strategy is original.

Another related approach is due to Galichon and Salanié (2010), who provide a theoretical and econometric analysis of multicriterion 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 are out of school; the resulting sample at this stage has 1,307,465 observations and is 49.5% male. We used the "detailed education variable" of the ACS to define three subcategories:

1. High School Dropouts (HSD)
2. High School Graduates (HSG)
3. Some College (SC).

Our category "some college" aggregates all individuals who at least started college. The drawback is that 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; the results are qualitatively similar. A finer classification would be desirable, but cell sizes shrink fast.

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?

It is notoriously hard to model divorce and remarriage in an empirically credible manner. Since this is not the object of this paper, we chose instead 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 individuals who are separated from their spouse
  - 218,839 who are divorced
  - 143,963 who are 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

Outcomes are truncated in our data, since 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. To examine marriage patterns, we dropped the small number of couples where one partner married before age 16 or after age 35 (recall that these are first unions.)

This leaves us with 179,353 couples, 44,344 single men, and 32,985 single women. The increasing level of education of women is shown on Figure 1: in cohorts born after 1955 women graduate more from college. Not coincidentally, the proportion of marriages in which the husband is more educated than the wife has fallen quite dramatically. Figure 2 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.)

Figures 3 and 4 describe changes in the level of education of the partners of married men (resp. women) between the earlier cohorts (born in the early 40s) and the most recent cohorts in our sample (born in the early 70s.) Figure 3 shows that college-educated men now find a college-educated wife much more easily; and in fact even less-educated men are now more likely to marry a college-educated woman—if they marry at all. On the other hand, the marriage patterns of women are remarkably stable, as evidenced in Figure 4.

We illustrate the decline in marriages by plotting the percentage of individuals of a given cohort who never married in Figures 5 and 6. They show that a higher education has tempered the decline in marriage, especially for women; and that high-school dropouts on the other hand have faced a very steep decline in marriage rates.

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 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, Z^{IJ} denotes the common component and ε_{ij}^{IJ} is a random shock with mean zero.

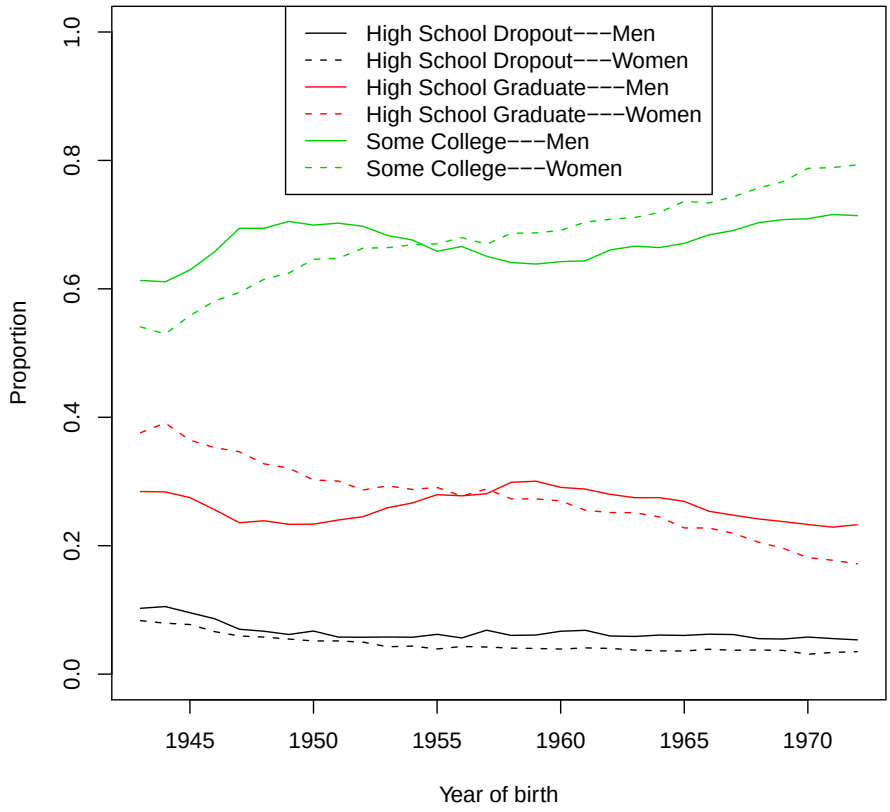

Figure 1: Education levels of men and women

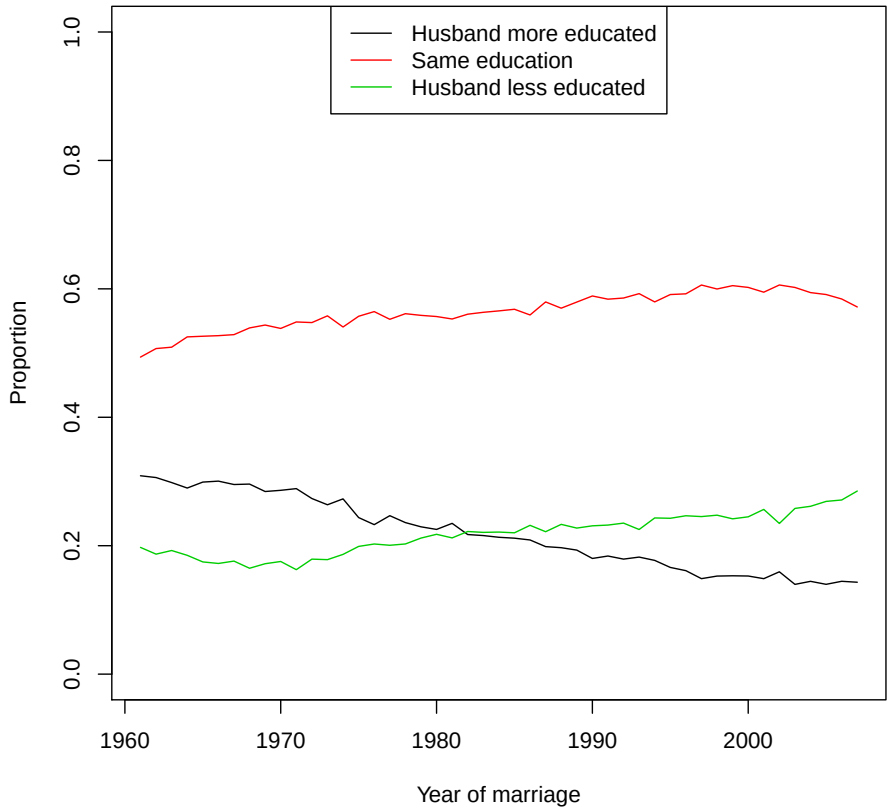

Figure 2: Relative education of partners

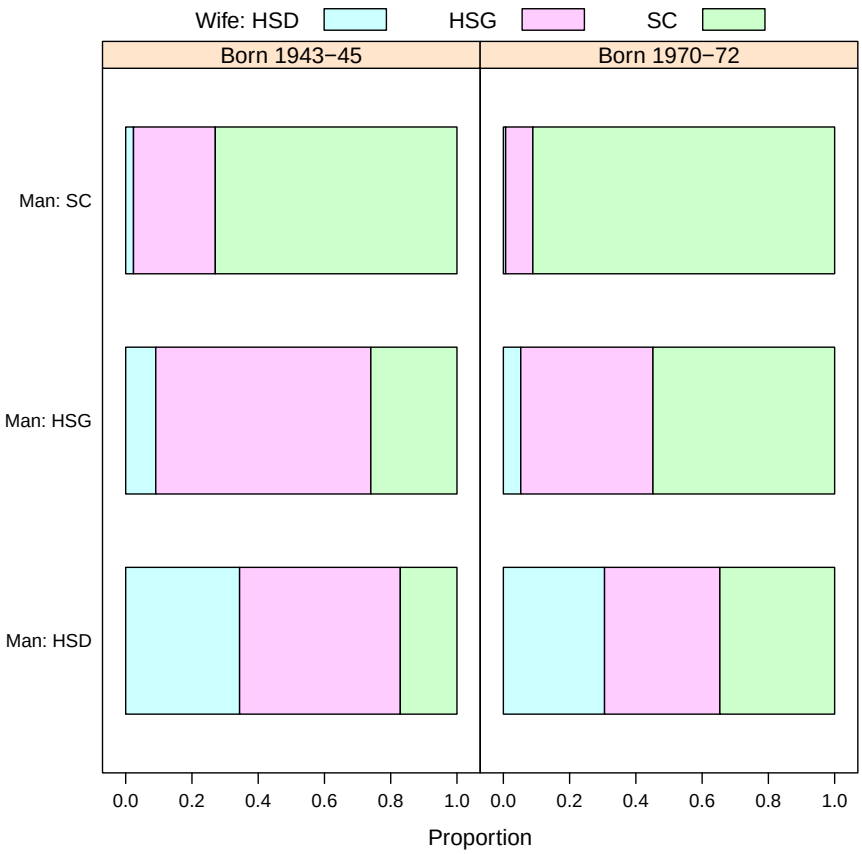

Figure 3: Marriage patterns of men who marry

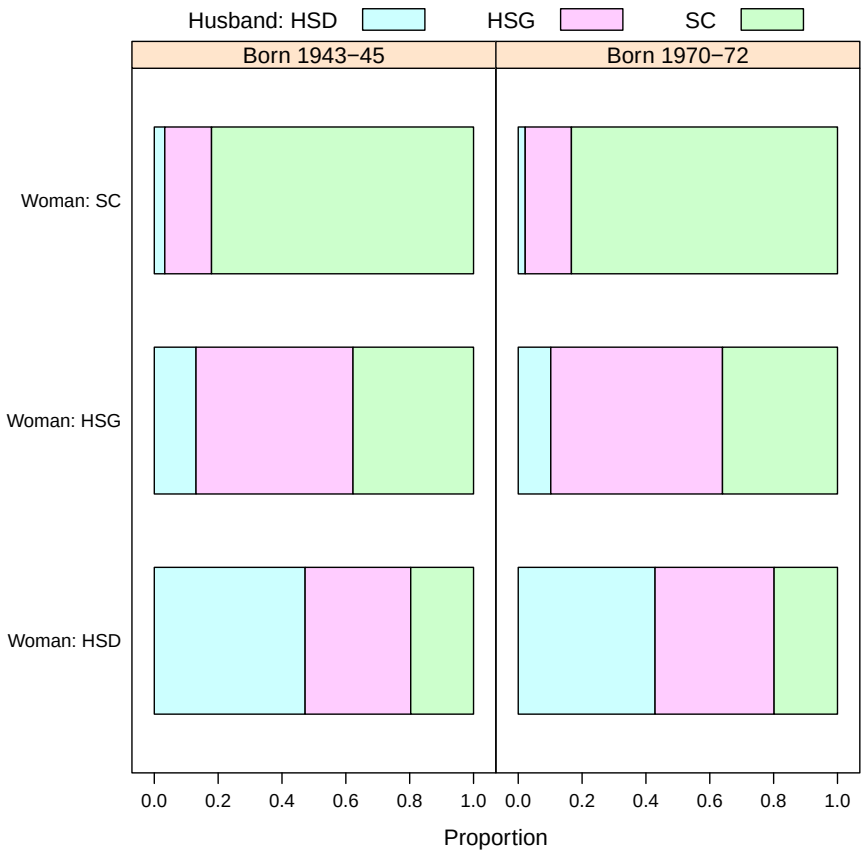

Figure 4: Marriage patterns of women who marry

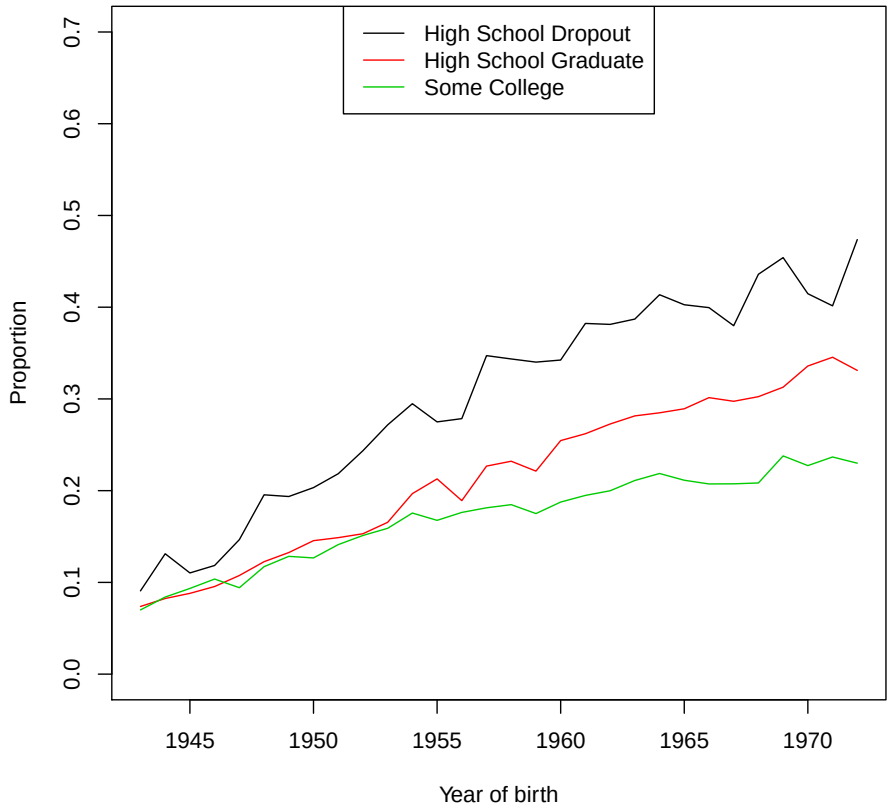

Figure 5: Proportion of men who never married

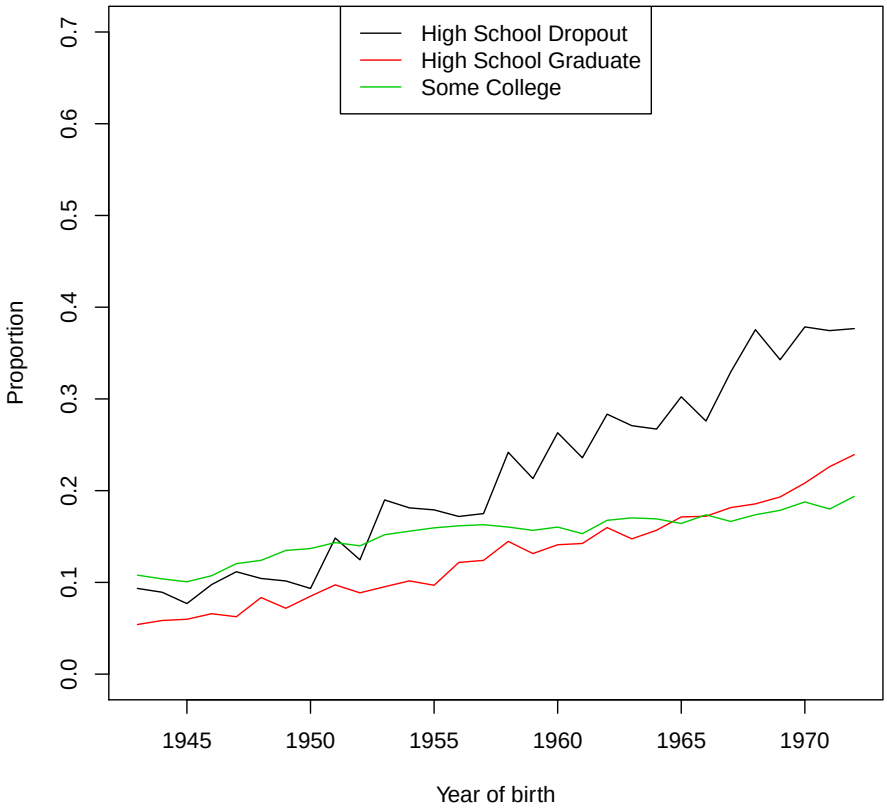

Figure 6: Proportion of women who never married

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

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

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:

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

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 property implies that within each pair of classes, (I, J), any matching would be stable. In practice, this means that 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*

$$\begin{aligned} \alpha_i^{IJ} - \alpha_i^{IK} &\geq U^{IK} - U^{IJ} \quad \textit{for all } K \qquad (3) \\ \alpha_i^{IJ} - \alpha_i^{I0} &\geq U^{I0} - U^{IJ} \qquad (4) \end{aligned}$$

and

$$\begin{aligned} \beta_j^{IJ} - \beta_j^{KJ} &\geq V^{KJ} - V^{IJ} \quad \textit{for all } K \qquad (5) \\ \beta_j^{IJ} - \beta_j^{0J} &\geq V^{0J} - V^{IJ} \qquad (6) \end{aligned}$$

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

$$\alpha_i^{IJ} - \alpha_i^{I0} \leq U^{I0} - U^{IJ} \quad \textit{for all } J \qquad (7)$$

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

$$\beta_j^{IJ} - \beta_j^{0J} \leq V^{0J} - V^{IJ} \quad \textit{for all } J \qquad (8)$$

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:*

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

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[4]:

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*

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

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

and

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

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

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

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

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

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

[4]Gumbel distributions are better known to economists under the clumsier name of "type-I extreme value distributions.

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

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

and

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

where U^{I0} has been normalized to 0. Similarly, for any J and any female $j \in J$:

$$b^{IJ} = P\left(j \text{ matched with a male in } I\right) \tag{15}$$

$$= \frac{\exp\left(V^{IJ}/\mu^J\right)}{\sum_K \exp\left(V^{KJ}/\mu^J\right) + \exp\left(V^{0J}/\mu^J\right)} \text{ and} \tag{16}$$

$$b^{0J} = P\left(j \text{ single}\right) = \frac{\exp\left(V^{0J}/\mu^J\right)}{\sum_K \exp\left(V^{KJ}/\mu^J\right) + \exp\left(V^{0J}/\mu^J\right)}$$

where $V^{0J} = 0$.

These formulas can be inverted to give:

$$\exp\left(U^{IJ}/\sigma^I\right) = \frac{a^{IJ}}{1 - \sum_K a^{IK}} \tag{17}$$

and

$$\exp\left(V^{IJ}/\mu^J\right) = \frac{b^{IJ}}{1 - \sum b^{KJ}} \tag{18}$$

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. Note that a direct consequence of these results is that, knowing the Z^{IJ} and the population sizes, we can *algebraically* compute U^{IJ}/σ^I and V^{IJ}/μ^J for all (I, J).

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:

$$\begin{aligned}\bar{u}^I &= \sigma^I E\left[\max_J \left(U^{IJ}/\sigma^I + \tilde{\alpha}_i^{IJ}\right)\right] \\ &= \sigma^I \ln\left(\sum_J \exp\left(U^{IJ}/\sigma^I\right) + 1\right) = -\sigma^I \ln\left(a^{I0}\right) \end{aligned} \tag{19}$$

and similarly

$$\bar{v}^J = \mu^J \ln\left(\sum_I \exp\left(V^{IJ}/\mu^J\right) + 1\right) = -\mu^J \ln\left(b^{0J}\right) \tag{20}$$

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 (19), 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 (19) 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 7). For instance, if one sees that college graduate are more likely to remain single than high school graduates ($a^{I0} > a^{K0}$, implying that $\ln\left(a^{K0}/a^{I0}\right) < 0$), we can conclude that the expected marital gain is smaller for college graduates ($\bar{u}^I < \bar{u}^K$), therefore that the marital college premium is negative.

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

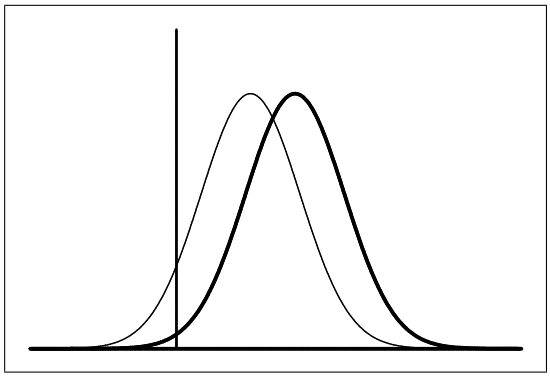

Figure 7: Homoskedastic random gains

Figure 8. 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. Technically, we now have that:

$$\bar{u}^I - \bar{u}^K = \sigma^K \ln\left(a^{K0}\right) - \sigma^I \ln\left(a^{I0}\right) \tag{21}$$

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, education operates on marital prospects through three different channels: it increases marriage probabilities; it changes the potential 'quality' (here education) of the future spouses; and it affects the distribution of surplus within the household. In the basic, homoskedastic version of the model, due to the assumptions made on the distributions of the random terms, these three channels are intrinsically mixed, and the expected utility of each spouse is fully determined by the percentage of persons in the same education class that remains single. The heteroskedastic version is much richer; welfare impacts go beyond the sole probability of marriage, and involve other considerations. Clearly, 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.[5]

[5]Note, however, that if the variances are assumed constant across time, then the *variations* in singlehood

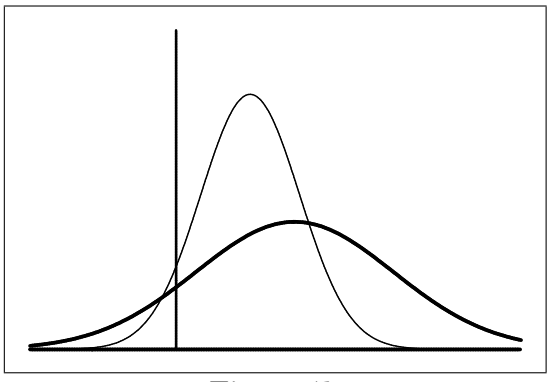

Figure 8: Heteroskedastic random gains

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):

$$\begin{aligned}
\alpha_i^{IJ} &= X_i.\zeta_m^{IJ} + \tilde{\alpha}_i^{IJ} \\
\alpha_i^{I0} &= X_i.\zeta_m^{I0} + \tilde{\alpha}_i^{I0} \\
\beta_j^{IJ} &= Y_j.\zeta_f^{IJ} + \tilde{\beta}_j^{IJ} \\
\beta_j^{0J} &= Y_j.\zeta_f^{0J} + \tilde{\beta}_j^{0J}
\end{aligned}$$

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, type 1 extreme values distributions $G(-k, 1)$. Then the computations are as above. In other words, we can estimate for $i \in I$:

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

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 cohorts c and c', we can unambiguously conclude that the gains from marriage have diminished for these women over the period.

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

and the conclusions follow. In particular, these models can be estimated running standard (multinomial) 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*

$$U^{IJ} = \sigma^I \log \frac{a^{IJ}}{1 - \sum_K a^{IK}} \tag{22}$$

and

$$V^{IJ} = \mu^J \log \frac{b^{IJ}}{1 - \sum_K b^{KJ}} \tag{23}$$

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*

$$\frac{\bar{\sigma}^I}{\sigma^I} U^{IJ} + \frac{\bar{\mu}^J}{\mu^J} V^{IJ} = \bar{Z}^{IJ} \tag{24}$$

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

$$\bar{U}^{IJ} = \frac{\bar{\sigma}^I}{\sigma^I} U^{IJ} \tag{25}$$

$$\bar{V}^{IJ} = \frac{\bar{\mu}^J}{\mu^J} V^{IJ} \tag{26}$$

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 (24), (25) and (26) must hold.*
Proof. *From the previous calculations, there is a one-to-one relationship between the* a^{IJ} *and the* v^{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 utility scales. The unit in which the Us and Vs 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 $\mathcal{M} = \left(Z^{IJ}, \sigma^I, \mu^J\right)$ holds for different cohorts of agents, $c = 1, ..., T$, with varying class compositions. The basic structure becomes:

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

with

$$\varepsilon^{IJ}_{ij,c} = \sigma^I \alpha^{IJ}_{i,c} + \mu^J \beta^{IJ}_{j,c} \tag{S}$$

Also, assume for the time being that each man marries a woman within his cohort.[6] 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}. Then

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

therefore

$$\exp\left(U_c^{IJ}/\sigma_I\right) = \frac{a_c^{IJ}}{1-\sum_K a_c^{IK}} \tag{27}$$

and similarly:

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

implying that

$$\exp\left(V_c^{IJ}/\mu_J\right) = \frac{b_c^{IJ}}{1-\sum_K b_c^{IK}} \tag{28}$$

Moreover, we have

$$U_c^{IJ} + V_c^{IJ} = Z^{IJ} \tag{29}$$

Now, let $p_c^{IJ} = U_c^{IJ}/\sigma_I$ and $q_c^{IJ} = V_c^{IJ}/\mu_J$. The crucial remark is that from (27) and (28), the p_c^{IJ} and q_c^{IJ} *are directly observable from the data.* It follows that (29) 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:

$$\sigma^I\ \mathbf{p}^{IJ} + \mu^J\ \mathbf{q}^{IJ} - Z^{IJ}\mathbf{1} = 0 \tag{30}$$

[6] 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.

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 (30) 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 (30) 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 $\sigma = 1$):

$$\mathbf{p} = -\mu\ \mathbf{q} + \sum_{I,J} Z^{IJ}\mathbf{1}_{IJ} \tag{31}$$

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:

$$Z_c^{IJ} = \zeta_c^I + \xi_c^J + Z^{IJ} \tag{32}$$

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 rules out. Under (32), 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* (or supermodularity) between education classes - the second differences $\left(Z^{IJ}-Z^{IL}\right)-\left(Z^{KJ}-Z^{KL}\right)$ - 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 test whether this hypothesis is compatible with the evolutions in marital patterns observed over the last decades.

Normalizations The form (32) 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 (32). We can therefore normalize ξ_c^1 to be zero for all c.

Testing the framework Under (32), equation (29) becomes:

$$\sigma^I\ p_c^{IJ}+\mu^J\ q_c^{IJ}=\zeta_c^I+\xi_c^J+Z^{IJ}\quad\forall I,J,c \tag{33}$$

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

$$\sigma^I\ \left(\ p_c^{IJ}-p_c^{I1}\right)+\mu^J\ q_c^{IJ}-\mu^1\ q_c^{I1}=\xi_c^J+Z^{IJ}-Z^{I1} \tag{34}$$

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

$$\sigma^I\ \left(p_c^{IJ}-p_c^{I1}\right)-\sigma_1\ \left(p_c^{1J}-p_c^{11}\right)+\mu^J\ \left(q_c^{IJ}-q_c^{1J}\right)-\mu^1\left(q_c^{I1}-q_c^{11}\right)=Z^{IJ}-Z^{I1}-Z^{1J}+Z^{11}. \tag{35}$$

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. To simplify notation, denote

$$D_2Z^{IJ}=Z^{IJ}-Z^{I1}-Z^{1J}+Z^{11}$$

the second difference of the mean surplus; and 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 $(I>1,J>1)$:

$$\mathbf{R}^{IJ}=\sigma^I\ \mathbf{P}^{IJ}+\mu^J\ \mathbf{Q}^{IJ}-\mu^1\mathbf{Q}^{I1}-D_2Z^{IJ}\mathbf{1} \tag{36}$$

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 first estimate the probabilities of the various marital outcomes directly from the data, and we use them to construct estimates of the vectors $\mathbf{P}, \mathbf{Q}$ and $\mathbf{R}$; then we choose the heterogeneity parameters $((\sigma^I), (\mu^J))$ and the second differences $(D_2 Z^{IJ})$ so as to minimize the deviations from the conditions in (36). This minimum distance estimation technique also allows us to test the model by evaluating the distance function at its minimum. In our application there are 116 conditions in (36), and only 9 free parameters; this is quite a stringent test since the probabilities of the various matches are estimated from a large sample and thus very precisely.

Once we have estimated the heterogeneity parameters σ^I and μ^J we can also reconstruct the left-hand side of equation (33):

$$\hat{A}_c^{IJ} = \hat{\sigma}^I \; p_c^{IJ} + \hat{\mu}^J \; q_c^{IJ}.$$

Our theory states that in an ANOVA regression of this $\hat{A}_c^{IJ}$, only 1-way and 2-way effects should appear. To put this in terms more familiar to applied econometricians: a regression of $\hat{A}_c^{IJ}$ on fixed effects for I, for J, and for c (the 1-way effects) and on fixed effects for the interactions (I, J), (I, c) and (J, c) (the 2-way effects) should have an R^2 of one. This is an alternative way of evaluating departures from the theory, based more on economic significance than on statistical significance.

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 (33) 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 (33) identifies ζ_c^I for all I since we set $\xi_c^1 \equiv 0$; and (34) 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 (35) 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:

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

In other words, the ratio $\Delta_3 a_c^{IJ}/\Delta_3 b_c^{IJ}$ should not depend on the classes I and J nor on the cohort - and the ratio μ/σ has then a natural interpretation in terms of minus this ratio (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.[7]

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 Results

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 $(c+1)$ (the one-year gap is both the mode and the median age difference at marriage.) We ran the analysis for cohorts of men born between 1943 to 1971.

Then we reconstitute the p and q terms and we run the minimum distance procedure, 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

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

[7] This property could in principle be used to construct both a specification test and a non parametric estimator of the ratio. In our data, however, the power of the test is quite weak, due to insufficient variations in the second difference across cohorts.

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.

Minimum distance estimation amounts to choosing the heterogeneity parameters and the second difference so as to minimize the length of the residuals in (36). As usual, the optimal choice of a norm is the inverse of the variance-covariance matrix of the residuals. Since we use 29 cohorts and we have three categories, the vector of residuals has dimension $29*(3-1)*(3-1) = 116$, and its variance-covariance matrix is rather unwieldy. To avoid relying too much on imprecise estimates of some off-diagonal elements of the variance matrix, we only used its diagonal elements[8]. Using the full matrix does not materially alter our results.

6.1 Tests

The hypotheses implied by our model is very roundly rejected. While this sounds like a disappointing outcome, the ANOVA procedure described in section 5.2.2 gives much more positive results. When we reconstructed the $\hat{A}_c^{IJ}$ factor, we found that in the 2-way ANOVA regression[9] the main effects were the 1-way effects on I and J (for a total of 46.2% of the variance), the 1-way cohort effect (for 13.8%), and the 2-way (I, J) effect (for 37.4%). As it turns out, the residual, which measures the deviation from our theory, accounts for only... 0.5% of the variance of $\hat{A}_c^{IJ}$.

These apparently divergent results are a striking illustration of the difference between statistical significance and economic significance. Since we use rather large samples of men and women, the odds ratios p_c^{IJ} and q_c^{IJ} are very precisely estimated, and any small deviation from the theory (the 0.5% of the variance above) results in a very large value of the test statistic, and thus a spectacular statistical rejection. Thus the statistical rejection of our theory is a minor distraction, and we pursue our analysis of the 99.5% of the variance in marriage patterns that we manage to explain.

6.2 Estimated Heterogeneities

Table 2 gives our estimates of the σ^I and μ^J terms. The model in Choo-Siow (2006) imposes that they all be equal; on the contrary, we find clear and significant variations across our estimates. In particular, each estimated μ is larger than the corresponding σ; and the hypothesis that each σ equals the corresponding μ is strongly rejected. There also appears to be much less heterogeneity among high-school graduates than for the other two categories; given the discussion of section 4.2, this will play an important role in what follows.

[8] Our estimator of these diagonal elements relies on a first-step minimum distance estimator based on weighting the residuals by the observed number of marriages. In computing it, we neglect the correlation between the estimated $\mathbf{P}$ and $\mathbf{Q}$.

[9] We weighted each (I, J, c) observation by the corresponding number of marriages in the data.

Group	σ^I	μ^J
HSD	0.089	0.148
	(0.017)	(0.027)
HSG	0.060	0.071
	(0.017)	(0.018)
COLL	0.087	0.137
	(0.014)	(0.017)

Table 2: σ^I in rows, μ^J in columns

6.3 Estimated Surpluses

The reconstructed values[10] 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	0.331	0.193	−0.128
HSG	0.195	0.272	0.098
COLL	−0.028	0.233	0.468

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 9 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.

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 10 (for men) and 11 (for women).

6.4 Interpretation

All these estimates have immediate structural interpretations. In practice, the marital college premium can be decomposed into several components. First, education

[10] The estimated standard errors are between 0.01 and 0.04.

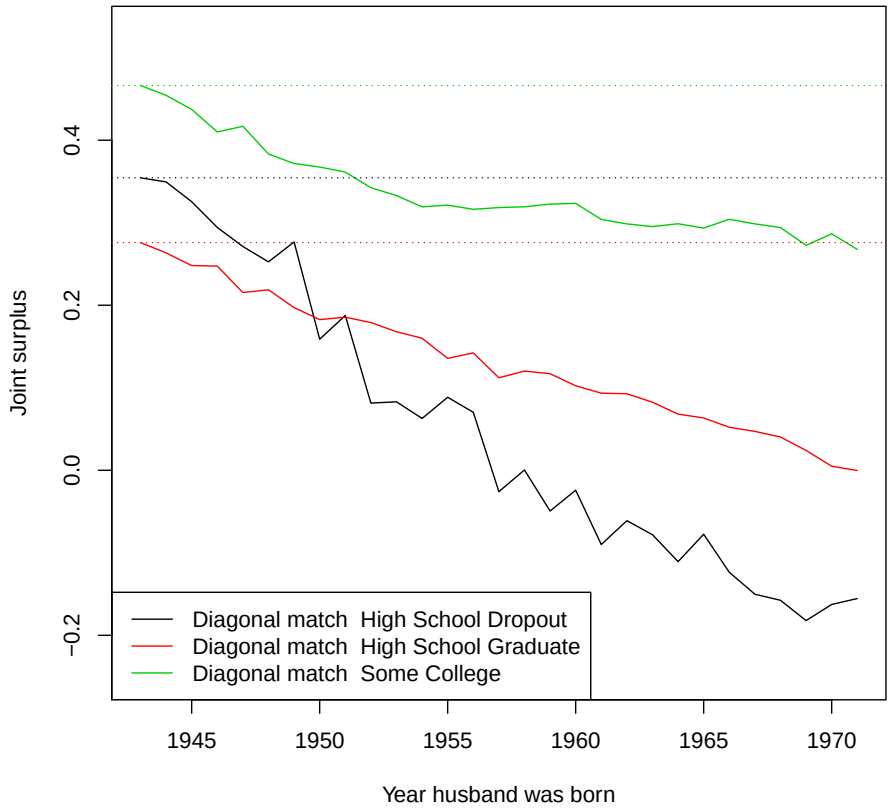

Figure 9: Joint surplus of diagonal matches

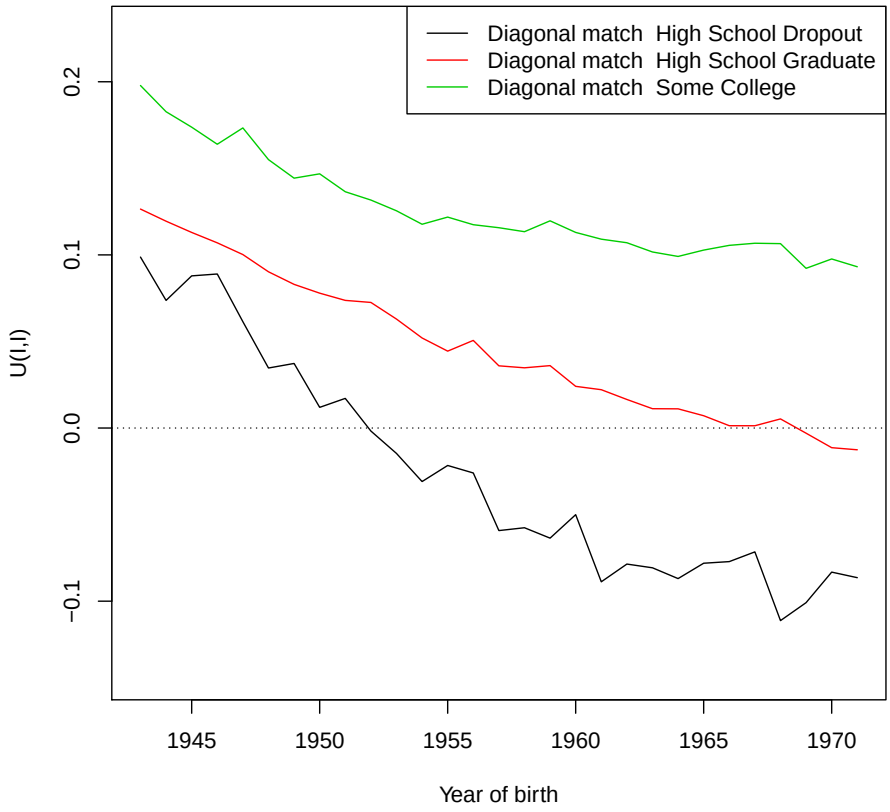

Figure 10: Gain from diagonal matching for men

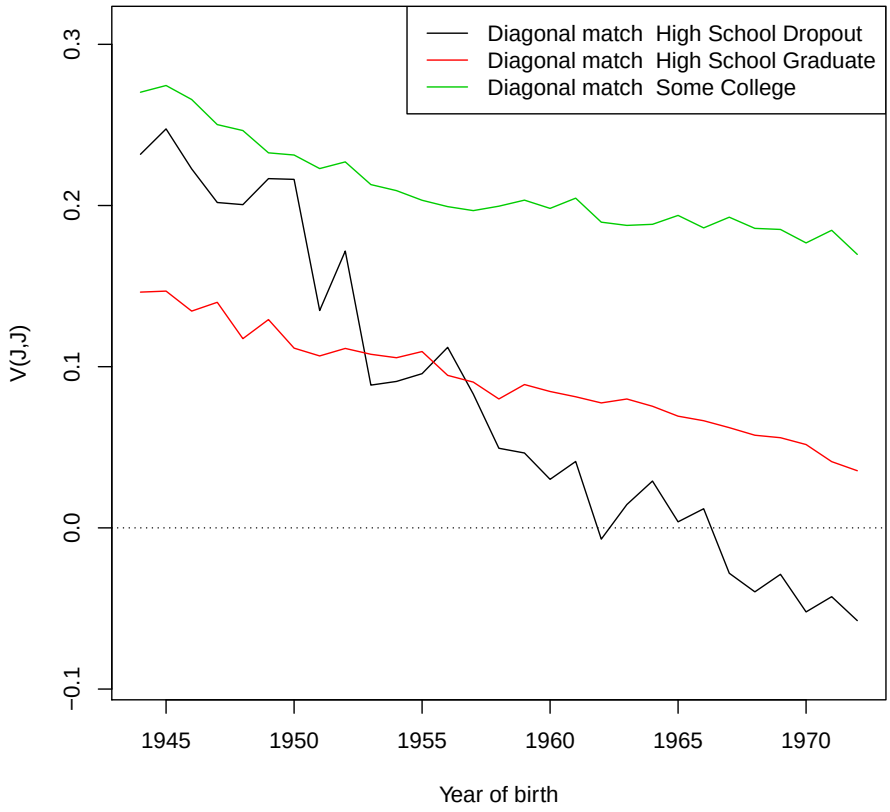

Figure 11: Gain from diagonal matching for women

affects the probability of being married. Second, conditional on being married, it also affects the education of the spouse (or more exactly its distribution); intuitively, we expect educated women to find a "better" husband, at least in terms of education, and conversely. Third, the impact on the *total* surplus generated by marriage is twofold. Take women for instance. A wife's education has a direct impact on the surplus; this impact can be measured, for college education, by the difference $(Z_c^{I3} - Z_c^{I2})$, where I denotes the husband's education. In addition, since a more-educated woman is more likely to marry a more educated husband, the husband's higher expected education further boosts the surplus, by the average of these $(Z_c^{I3} - Z_c^{I2})$ terms weighted by the difference in probability of marrying a college-educated husband instead of a high school graduate.

Finally, the share of the surplus going to the wife in any given match is also affected by her education. Consider the average surplus form a match between an I-man of cohort c and a J-woman of cohort $(c+1)$—recall that we assumed a fixed age difference. This average surplus is the expected value of

$$E\left(Z_c^{IJ} + \sigma^I \alpha_{i,c}^{IJ} + \mu^J \beta_{j,c}^{IJ}\right),$$

conditional on i and j marrying each other in equilibrium. Given the additive structure of our theory, it can also be rewritten as the sum of

$$E \max_K (U_c^{IK} + \sigma^I \alpha_{i,c}^{IK})$$

and

$$E \max_K (V_c^{KJ} + \mu^J \beta_{j,c}^{KJ})$$

where the first expectation is conditional on i marrying a J-woman, and the second one is conditional on j marrying a I-woman. But given the peculiar nature of type-I extreme value errors, the first expectation is $\bar{u}^I$, independently of the value of J; and the second one is $\bar{v}^J$, independently of the value of I. Therefore the ratio

$$\frac{\bar{v}^J}{\bar{u}^I + \bar{v}^J}$$

measures the share of the surplus that goes to the wife in an (I, J) marriage, in expected terms.

All these components can readily be computed from our estimates. For instance, start with the early cohorts of women, born between 1944 and 1946. For these women:

1. College education reduced the probability of marrying: it was 93.9% for a high school graduate, but only 89.6% after college.
2. It allowed women who did marry to get a better-educated partner: for instance, the conditional probability of marrying a college-educated man jumped from 38.0% for a high school graduate. to 83.3% for a college-educated woman.
3. The marriage of a college-educated husband with a college-educated wife generated a total surplus that was 0.464 on average, as opposed to only 0.191 if the wife did not attend college.

4. Finally, still in the case of a college-educated husband, the wife's share of total surplus was 57.0% on average if she was college-educated, while a high school graduate received only 41.9% of the smaller surplus.

Note that items 1 and 2 above can be evaluated directly from the data (and our model reproduces them very closely); but on the other hand, items 3 and 4 cannot be computed without a model, such as the one we use in this paper.

The same analysis can be made for the most recent cohorts of women, born between 1970 and 1972. Here:

1. College education now *increases* the probability of marrying (it is 79.1% for a high school graduate and 81.8% for a college graduate)
2. Its impact on the husband's education is pretty much unchanged: the conditional probabilities of marrying a college-educated man are 37.6% for a high school graduate and 84.1% for a college graduate.
3. Regarding the direct impact of female education on total surplus, the marriage of a high school graduate wife with a college-educated man generates negative total surplus on average (-0.041); if the wife attended college, the total surplus is 0.289. At 0.330, the difference is much larger than it was for early cohorts (0.273).
4. The wife's share of the total surplus in a marriage with a college-educated man has decreased for high school graduates, at 40.4% now; and it has markedly increased for college-educated women—it is now 62.5%.

All in all, the impact of education on a person's marital situation is quite complex: it involves changes in the marriage probabilities, but also in the "quality" of the spouse, in the size of the surplus generated by marriage and ultimately in the distribution of this surplus between spouses. These various components may not evolve in the same direction. A spouse's expected gain, on which the definition of the marital college premium is based, must take all these elements into account; as a result, even the direction of its evolution may in principle be quite difficult to figure out.

An obvious advantage of our structural model, though, is that the value of this expected gain can be directly computed from the data. For the main concepts at stake, the model actually provides explicit expressions that can readily be evaluated from our estimates. For instance, the marital college premium for any generation c is given by equation (21) above:

$$MCP_c^m = \bar{u}_c^3 - \bar{u}_c^2 = \sigma^2 \ln\left(a_c^{20}\right) - \sigma^3 \ln\left(a_c^{30}\right)$$

for men and

$$MCP_c^w = \bar{v}_c^3 - \bar{v}_c^2 = \mu^2 \ln\left(b_c^{20}\right) - \mu^3 \ln\left(b_c^{30}\right)$$

for women.

Figures 12 and 13 plot the evolution over cohorts of the expected gains $\bar{u}_c^I$ and $\bar{v}_c^J$ for the various education classes under consideration. One sees, in particular, that the fate of high-school dropouts has deteriorated for both genders, while that of college-educated women has improved.

The latter point is confirmed on Figure 14, which plots the evolution of the "marital college premium" $(\bar{u}_c^3 - \bar{u}_c^2)$ and $(\bar{v}_c^3 - \bar{v}_c^2)$ over cohorts for both genders. Beyond the year-to-year changes, the nonparametric smoothers in dashed lines tell a clear story: the marital college premium of women started to increase sharply for cohorts born around 1955, who graduated from college around 1980; and it has crept upwards ever since. No such change can be seen for men: their marital college premium has remained remarkably flat over the period.

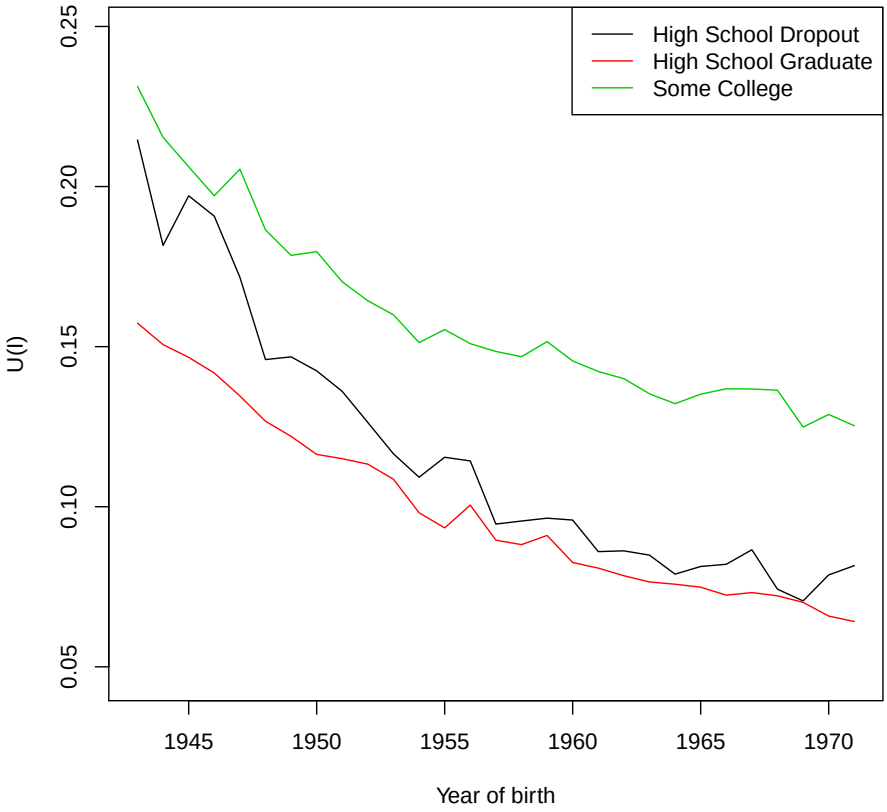

Figure 12: Gain from marriage for men

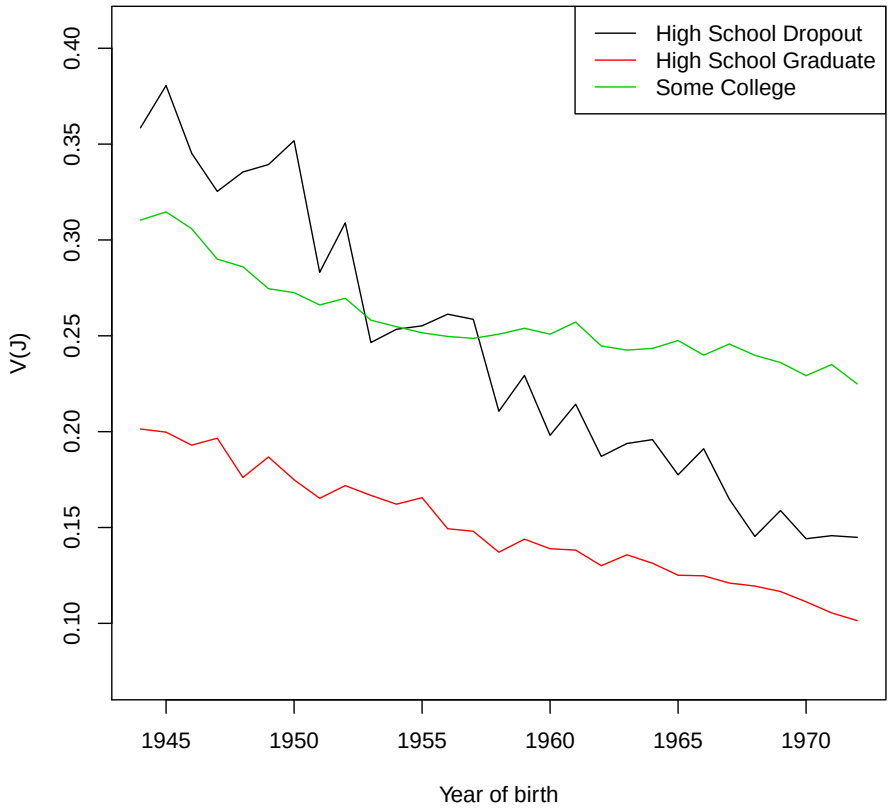

Figure 13: Gain from marriage for women

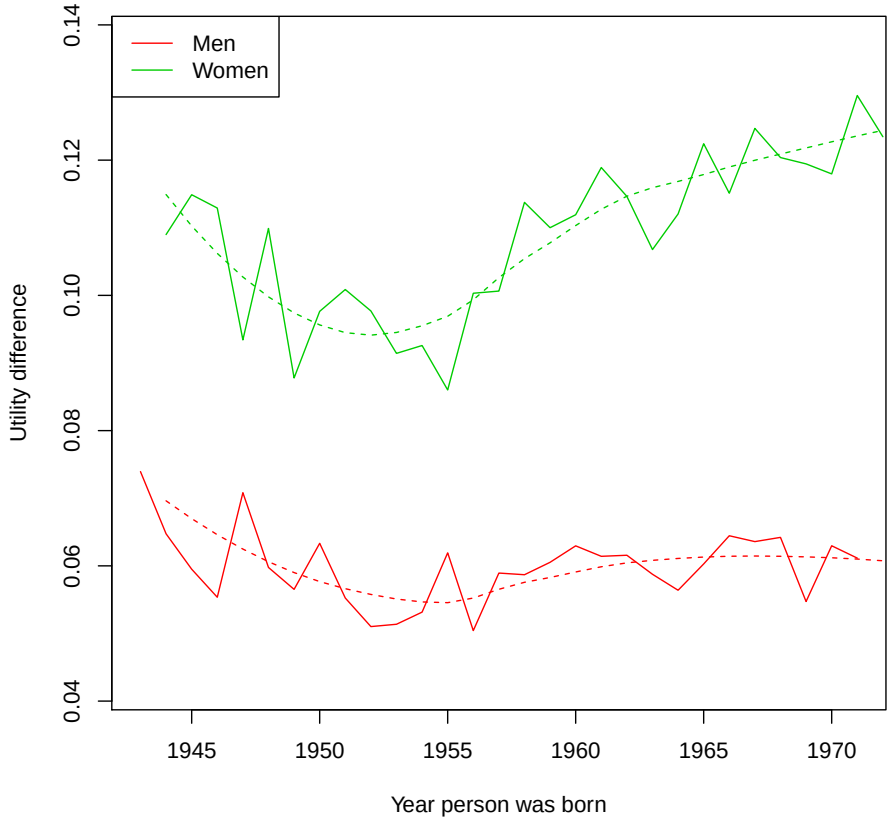

Figure 14: Marital college premium

7 Conclusion

It has long been recognized (at least since Becker's 1991 seminal contributions) that the division of the surplus generated by marriage should be analyzed as an equilibrium phenomenon. As such, it responds to changes in the economic environment; conversely, investments made before marriage are partly driven by agents' current expectations about the division of surplus that will prevail after marriage. Theory shows that such considerations may explain the considerable differences in male and female demand for higher education. In a nutshell, when deciding whether to go to college, agents take into acount not only the market college premium (i.e., the wage differential resulting from a college education) but also the "marital college premium"' which represents the impact of education on marital prospects; the later includes not only marriage probabilities, but also the expected "quality" of the future spouse and the resulting distribution of marital surplus. Our first contribution is to provide a simple but rich model in which these components can be econometrically identified. Our framework generalizes a previous contribution by Choo and Siow (2006); we show, in particular, that it can be (over)identified using temporal variations in the compositions of the populations at stake. Applying the model to US data, we first find that the main identifying assumption—that the gains from assortative matching, as measured by level of supermodularity in the marital surplus, remain constant over time— is rejected by the data. On the other hand,

[.....]

We can then fully identify the structural model. While the gains from mariage have declined over the period, the decline has been smaller for educated agents. In particular, the "marital college premium" has markedly increased for women while remaining stable for men, which confirms the theoretical predictions discussed above.

References

[1] Becker, Gary S. (1991), *Treatise on the family*, Cambridge Mass: Harvard University Press.

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

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

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

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

[6] 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

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

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

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

[10] 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.

[11] 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.

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