barcelona gse
graduate school of economics

Marry Your Like: Assortative Mating and Income Inequality

Jeremy Greenwood
Nezih Guner
Georgi Kocharkov
Cezar Santos

January 2014

Barcelona GSE Working Paper Series

Working Paper nº 748

January 2014

Marry Your Like: Assortative Mating and Income Inequality*

by

Jeremy Greenwood, Nezih Guner, Georgi Kocharkov and Cezar Santos†

Abstract

Has there been an increase in positive assortative mating? Does assortative mating contribute to household income inequality? Data from the United States Census Bureau suggests there has been a rise in assortative mating. Additionally, assortative mating affects household income inequality. In particular, if matching in 2005 between husbands and wives had been random, instead of the pattern observed in the data, then the Gini coefficient would have fallen from the observed 0.43 to 0.34, so that income inequality would be smaller. Thus, assortative mating is important for income inequality. The high level of married female labor-force participation in 2005 is important for this result.

Keywords: Assortative mating, married female labor supply, inequality.

JEL Classifications: D31, J11, J12, J22.

*This paper is prepared for the 2014 American Economic Review Papers and Proceedings.

†*Affiliations*: University of Pennsylvania; ICREA-MOVE, Universitat Autonoma de Barcelona and Barcelona GSE; University of Konstanz; and University of Mannheim. Guner acknowledges support from European Research Council (ERC) Grant 263600.

1 Introduction

Has there been an increase in positive assortative mating on the marriage market since 1960? How does positive assortative mating in the marriage market contribute to income inequality across households? These two questions are addressed here. To answer them, samples of hundreds of thousands of households from the United States Census Bureau are analyzed for the period 1960 to 2005.

2 The Rise in Positive Assortative Mating

Start with the first question: Has there been a rise in positive assortative mating in marriage across U.S. households? To address this question, consider a regression between a wife's educational level and her husband's. In particular, a regression of the following form is run for the years $y = 1960, 1970, 1980, 1990, 2000, 2005$:

$$\text{EDU}^w_{my} = \alpha + \beta \text{EDU}^h_{my} + \sum_{t \in \mathcal{T}} \gamma_t \times \text{EDU}^h_{my} \times \text{YEAR}_{ty} + \sum_{t \in \mathcal{T}} \theta_t \times \text{YEAR}_{ty} + \varepsilon_{my}, \text{ with } \varepsilon_{my} \sim N(0, \sigma).$$

Here EDU^h_{my} and EDU^w_{my} represent the years of education for the husband and wife in marriage m for year y. The variable YEAR_{ty} is a time dummy. It is set up so that $\text{YEAR}_{ty} = 1$, if $t = y$, and $\text{YEAR}_{ty} = 0$, if $t \neq y$, where $t \in \mathcal{T} \equiv \{1970, 1980, 1990, 2000, 2005\}$. The coefficient β measures the impact of a husband's education on his wife's for the baseline year 1960, since $\text{YEAR}_{ty} = 0$, for all t, when $y = 1960$. The coefficient γ_t gives the additional impact of a husband's education on his wife's relative to the baseline year, 1960. The evolution of γ_t over time speaks to changes in the degree of assortative mating. The regression also includes a fixed effect for each year as measured by the constants α and θ_t. The θ_t's control for the secular rise in the educational levels for the married population. The lefthand side panel of Figure 1 plots the upshot of the regression analysis. As can be seen, γ_t rises over time, implying that the degree of assortative mating has increased.

The levels of education are collapsed into five categories for the rest of the analysis: less than high school ($HS-$), high school (HS), some college ($C-$), college (C), and post college ($C+$). Kendall's τ rank correlation is computed between a husband's and wife's education for each year. The changes in Kendall's τ over time are illustrated in Figure 1, righthand side panel. While the series displays some nonmonotonicity, Kendall's τ is clearly higher in 2005 relative to 1960.

Last, the pattern of assortative mating between a husband's and wife's educational levels can be examined in a contingency table, as the upper panel in Table 1 does. Each cell in the contingency table has two entries. The first entry gives the observed fraction of married households in the cell, while the second number displays the fraction that would occur if matching was random. The diagonal of the contingency table describes the matches that occur when husband and wife have identical educational levels, both for the observed matches and when they are random. Take sum along the diagonal for each of these two types of matches, actual and random. Next, compute the ratio of the actual to random matches and denote it by δ. This ratio is also plotted in Figure 1, righthand side panel, for the years $1960, 1970, \cdots, 2005$. First, as can be seen, there is positive assortative mating. That is, the ratios are larger than one, implying that the number of matches between husband and wife with the identical education level is larger than what would occur if matching was random. Second, they increase over time. Greenwood et al. (2013) present a structural model of this rise in positive assortative mating.

3 Assortative Mating and Income Inequality

Turn to the second question: How does marital sorting affect household income inequality? Interest in this question is not without precedence. For example, Cancian and Reed (1998) and Schwartz (2010) both conclude that an increase in assortative mating has led to a rise in income inequality. The current research addresses this question using an accounting-based methodology, which is quite different from other studies. Some income statistics for married households by educational class are presented in the lower panel of Table 1. Again, each cell has two entries. The first number gives the married household's income *relative* to mean

income in the economy across *all* households, married and/or single. The second number is the share of the wife's labor income in household labor income. In 1960 if a woman with a less-than-high-school education ($HS-$) married a similarly educated man their household income would be 77 percent of mean household income. If that same woman married a man with a college education (C) then household income would be 124 percent of the mean. Alternatively, in 2005 if a woman with post-college education ($C+$) marries a man with a less-than-high-school education their income would be 92 percent of mean household income. This rises to 219 percent if her husband also has a post-college education. So, at some level, sorting matters for household income.

3.1 Constructing Lorenz Curves and Gini Coefficients

Let f_{ij} denote the fraction of households in the U.S. that are of type i in income percentile j and r_{ij} represent the income of such a household relative to mean household income. The percentile index j is expressed in terms of fractions (e.g., 0.10 instead of 10). The types are classified as follows: There are married and single households. In a married household each person is indexed by one of the above educational levels. The wife is also categorized by whether she works or not. There are ten income percentiles (deciles) so $j \in \{0.1, 0.2, \cdots, 1.0\}$. This leads to 550 (i, j)-combinations of households in total for each year. The share of aggregate income that percentile j accounts for, s_j, is given by $s_j = \sum_i f_{ij} r_{ij}$. The cumulative share of income at percentile p is thus $l_p \equiv \sum_j^p s_j = \sum_j^p \sum_i f_{ij} r_{ij}$. A Lorenz curve plots l_p against $p = \sum_j^p \sum_i f_{ij}$. The Gini coefficient, g, is twice the area between the Lorenz curve and the 45^0 line. If p moves continuously then the Gini coefficient is defined by $g = 2\int_0^1 |l_p - p| dp$, where $0 \leq g \leq 1$. A higher value for g implies a greater degree of income inequality. The Lorenz curve and Gini coefficient are clearly functions of the f_{ij}'s and the r_{ij}'s, for all i and j, so write $l_p = \textsc{Lorenz}_p(\{f_{ij}\}, \{r_{ij}\})$ and $g = \textsc{Gini}(\{f_{ij}\}, \{r_{ij}\})$.

The Lorenz curves for 1960 and 2005 are pictured in the lefthand side panel of Figure 2. They show a rise in inequality. The Gini coefficient rises from 0.34 to 0.43. The righthand side panel shows the relative income for each percentile. In 1960 a household at the 10th percentile earned 16 percent of mean income. This dropped to 8 percent in 2005. A household in the 90th percentile earned 251 percent of mean income in 1960 versus 317 percent in 2005.

Incomes are more polarized in 2005. The change in wages across individuals is the primary driver of this increase in income inequality.

3.2 Assortative versus Random Matching

Suppose that matching was random instead of assortative. What would have happened to the income distributions in 1960 and 2005? To do this experiment the observed pattern of matching for married couples shown in the contingency table is replaced by the pattern that would occur if matching was random. Let $\mathcal{M}$ represent that set of indices for married couples and $\mathcal{S}$ be the set for singles. The experiment involves replacing the observed $\{f_{ij}\}$'s for $(i,j) \in \mathcal{M}$ with the set that would obtain if matching was random, denoted by $\{\widetilde{f}_{ij}\}$ for $(i,j) \in \mathcal{M}$. The counterfactual Lorenz curve and Gini coefficient are given by $\textsc{Lorenz}_p(\{f'_{ij}\}, \{r_{ij}\})$ and $\textsc{Gini}(\{f'_{ij}\}, \{r_{ij}\})$, where $\{f'_{ij}\} \equiv \{\widetilde{f}_{ij}\}_{\mathcal{M}} \cup \{f_{ij}\}_{\mathcal{S}}$.

The results of the counterfactual experiment are interesting. Moving from the observed pattern of assortative matching in 1960 to a random pattern has little discernible impact on income inequality. The Gini coefficient drops only slightly to 0.33. Repeating the experiment for 2005 has a marked impact on the income distribution, that is shown in the lefthand side panel of Figure 3. As can be seen, the Lorenz curve shifts in and the Gini drops from 0.43 to 0.34. (The analogous diagram for 1960 is not shown since the shift in the Lorenz curve is not noticeable.) Why does this experiment affect the Lorenz curve for 2005 but not 1960? This question will be addressed now.

3.3 Matching and Married Female Labor-Force Participation

For positive assortative matching to have an impact on income inequality married females must work. Married females worked more in 2005 than 1960. The righthand side panel of Figure 4 shows married female labor-force participation by percentile. As can be seen, across all income percentiles labor-force participation was higher in 2005 versus 1960, but the increase is most precipitous at the highest percentiles. For example, at the 80th percentile 42 percent of married women worked in 1960. This rose to 77 percent in 2005. At the 20th percentile the numbers are 25 and 34 percent. The lefthand side panel of Figure 4 shows the

contribution of the wife's labor income to household labor income, again by percentile. The wife's contribution to household labor income is significantly larger in 2005 relative to 1960. This share rises with the income percentile. At the 80th percentile the share that married woman provided to household income rose from 16 to 34 percent, and from 13 to 25 percent at the 20th percentile.

To examine the impact of married female labor-force participation (MFLP) and sorting on income inequality, undertake this thought experiment. Assume that matching is random in the years 1960 and 2005 with one twist: assume that in 1960 married woman participate in the labor force at their 2005 levels and that in 2005 they work at their 1960 levels. The resulting Gini coefficients are 0.32 and 0.45. When matching is random, married female labor-force participation has a significant dampening effect on income inequality for the year 2005. Random sorting works to equalize incomes across households in 2005 because it diversifies income across husbands and wives. But, this effect is only operational to the extent that married women work. (That is, for 2005 compare 0.34 with 0.45.) Random matching has less of an effect in 1960 than in 2005. Incomes are less polarized in 1960, as Figure 1 and Table 1 both show.

Another interesting question to ask is what would have happened to income inequality if couples in 2005 matched as in 1960. That is, replace the 2005 contingency table with the 1960 one. This experiment is somewhat tricky to operationalize. In 2005 people were much more educated than in 1960. The fractions of wives (husbands) in the various educational groups can be obtained by summing each column (row) across the rows (columns). In other words, the marginal distributions for husbands and wives linked with the contingency tables have changed across 1960 and 2005. The marginal distributions for females are shown in Table 1, upper panel. The rise in educational attainment for females is readily apparent.

A *standardized* contingency table for the years 1960 and 2005 can be constructed to control for this. The essential idea is that shifts in the marginal distributions across non-standardized contingency tables can distort the comparison of the core patterns of association between the variables in the tables. Using the iterative procedure outlined in Mosteller (1968), a contingency table for 1960 can be computed using the 2005 marginal distributions over educational categories for husbands and wives. Another one can be built for 2005 using

the 1960 marginal distributions. (These standardized contingency tables, and the method for generating them, are presented in the appendix.) A comparison of the 1960 contingency table from the data with the standardized one for 2005 shows an increase in assortative mating. The (straight) sum of the diagonals rises from 0.54 to 0.60. (A comparison of 1960 standardized contingency table with the one in the data for 2005 also shows an increase along the diagonal from 0.44 to 0.48.) The Gini coefficients associated with these two standardized tables are 0.34 and 0.35. Therefore, if people matched in 2005 according to the 1960 standardized mating pattern there would be a significant reduction in income inequality; i.e., the Gini drops from 0.43 to 0.35. The inward shift in the Lorenz curve is shown in the righthand side panel of Figure 3.

Last, take the 1960 standardized table and additionally impose the 2005 levels of married female labor-force participation. Likewise, force the 1960 levels of married female labor-force participation on the 2005 standardized contingency table. Now, the Gini coefficients are 0.33 and 0.44. Income inequality rises for 2005 (from 0.35 to 0.44). By shutting down married female-labor participation for 2005 income inequality worsens. The Lorenz curve for this experiment virtually lies on top of the one from the data for 2005 (but shifts very slightly outward), so it is not shown in Figure 3. This illustrates the importance of married female labor-force participation for understanding income inequality. The results of the experiments are catalogued in Table 2. So, if people matched in 2005 according to the standardized mating pattern observed in 1960, which showed less positive assortative matching, then income inequality would drop because income is more diversified across husband and wife. For this effect to function females need to work in 2005, as they did, or diversification in household income can't operate.

REFERENCES

Cancian, Maria and Deborah Reed. 1998. "Assessing the Effects of Wives' Earnings on Family Income Inequality." *The Review of Economics and Statistics*, 80 (1): 73-79.

Greenwood, Jeremy, Nezih Guner, Georgi Kocharkov and Cezar Santos. 2013. "Technology and the Changing Family: A Unified Model of Marriage, Divorce, Educational Attainment and Married Female Labor-Force Participation." NBER Working Paper 17735.

Mosteller, Frederick. 1968. "Association and Estimation in Contingency Tables." *Journal of the American Statistical Association* 63 (321): 1-28.

Schwartz, Christine R. 2010. "Earnings Inequality and the Changing Association between Spouses' Earnings." *American Journal of Sociology* 115 (5): 1524-1557.

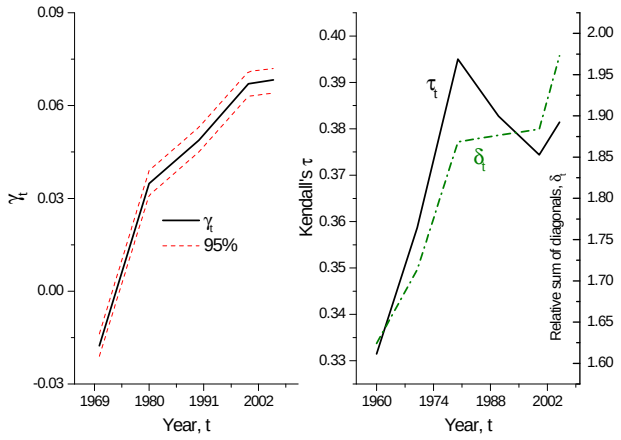

Figure 1: The Rise in Positive Assortative Mating, 1960-2005

Note: The variables γ_t, τ_t and δ_t are measures of assortative mating for the years $t = 1960, 1970, .., 2000, 2005$. A higher value for a variable shows a higher degree of positive assortative mating. See the text for a description of the variables. Source: See the appendix for a description of the data used in all figures and tables.

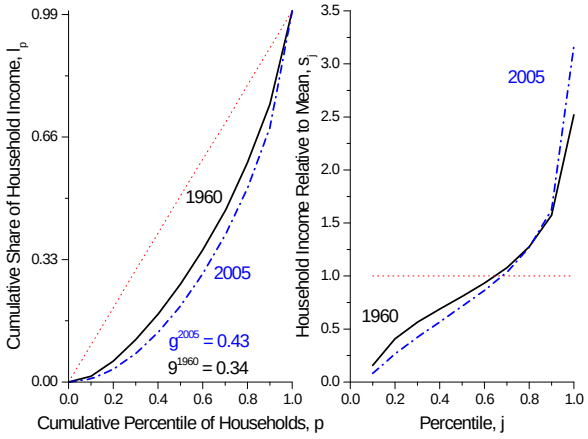

Figure 2: Income Inequality, 1960 amd 2005

Note: The righthand side panel shows average income for a household in the j-th percentile relative to mean household income in the economy. The lefthand side panel shows the Lorenz curves for 1960 and 2005. See the appendix for more detail on how the Lorenz curves are constructed.

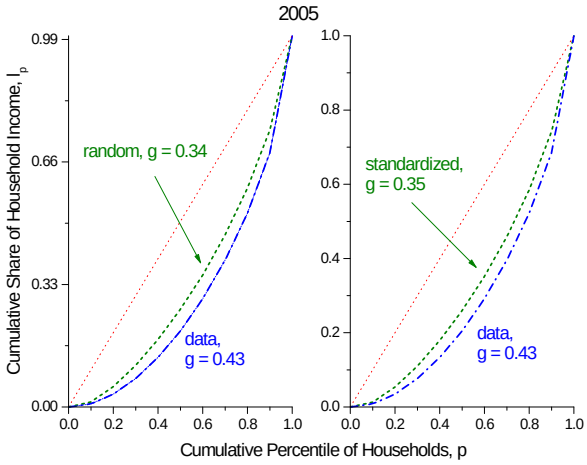

Figure 3: Assortative versus Random Mating, 2005

Note: The lefthand side panel shows the Lorenz curves in 2005 both for the data and when matching is random. The righthand side panel shows the Lorenz curves in 2005 both for the data and when matching is done according to a contingency table for 1960 that is standardized using the 2005 marginal distributions over education for men and women.

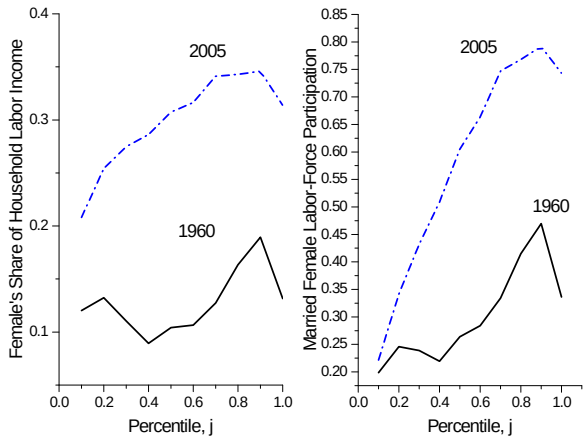

Figure 4: Married Female Labor-Force Participation, 1960 and 2005

Note: The righthand side panel shows married female labor-force participation by income percentile for 1960 and 2005. The lefthand side panel illustrates the share of the wife's labor income in household labor income.

Table 1: Contingency Table: Assortative Mating and Income by Educational Class

Marital Sorting by Education

1960

Husband	Wife									
	HS-		HS		C-		C		C+	
HS-	**0.323**[1]	**0.207**[2]	0.138	*0.192*	0.019	*0.053*	0.004	*0.026*	0.001	*0.008*
HS	0.076	*0.118*	**0.165**	**0.110**	0.028	*0.031*	0.008	*0.015*	0.002	*0.004*
C-	0.018	*0.045*	0.051	*0.042*	**0.027**	**0.012**	0.008	*0.006*	0.002	*0.002*
C	0.005	*0.030*	0.027	*0.028*	0.019	*0.008*	**0.018**	**0.004**	0.003	*0.001*
C+	0.003	*0.025*	0.016	*0.024*	0.017	*0.007*	0.016	*0.003*	**0.008**	**0.001**
Marginal	0.425		0.396		0.110		0.054		0.016	

2005

HS-	**0.039**	**0.006**	0.031	*0.027*	0.010	*0.020*	0.003	*0.020*	0.001	*0.010*
HS	0.023	*0.024*	**0.192**	**0.114**	0.082	*0.084*	0.037	*0.084*	0.012	*0.041*
C-	0.005	*0.015*	0.065	*0.073*	**0.088**	**0.054**	0.047	*0.053*	0.016	*0.026*
C	0.002	*0.015*	0.030	*0.072*	0.045	*0.053*	**0.104**	**0.053**	0.037	*0.026*
C+	0.001	*0.009*	0.010	*0.043*	0.018	*0.032*	0.050	*0.032*	**0.053**	**0.015**
Marginal	0.070		0.329		0.242		0.241		0.118	

Marital Income by Education

1960

HS-	0.765[3]	*0.131*[4]	0.918	*0.147*	1.040	*0.204*	1.243	*0.356*	1.395	*0.415*
HS	0.935	*0.101*	1.031	*0.119*	1.148	*0.168*	1.344	*0.263*	1.581	*0.358*
C-	1.071	*0.106*	1.185	*0.113*	1.278	*0.139*	1.442	*0.196*	1.593	*0.328*
C	1.234	*0.080*	1.349	*0.076*	1.420	*0.080*	1.529	*0.121*	1.673	*0.222*
C+	1.357	*0.087*	1.476	*0.083*	1.568	*0.090*	1.631	*0.126*	1.764	*0.215*

2005

HS-	0.409	*0.219*	0.586	*0.346*	0.692	*0.415*	0.904	*0.462*	0.918	*0.522*
HS	0.554	*0.221*	0.827	*0.319*	0.932	*0.376*	1.166	*0.447*	1.327	*0.503*
C-	0.661	*0.190*	0.958	*0.278*	1.042	*0.337*	1.255	*0.402*	1.434	*0.485*
C	0.852	*0.195*	1.250	*0.229*	1.335	*0.256*	1.600	*0.308*	1.793	*0.389*
C+	1.303	*0.165*	1.495	*0.199*	1.666	*0.202*	1.896	*0.224*	2.193	*0.333*

Note: Each cell in the contingency table has two entries. In the top panel they refer to 1) the observed matching pattern between husband and wife and 2) what would happen if matching was random matching. In the bottom panel they denote 3) household income relative to mean income across all households and 4) the share of the wife's labor income in total household labor income. Household income is adjusted by an equivalence scale to account for the differences in household size (including children) in each cell. The row marked marginal gives the fraction of females in each educational category; i.e., the marginal distribution over education for females.

Table 2: Gini Coefficients, Data and Experiments

Basis for Gini Coefficient	1960	2005
Data	0.34	0.43
Random Matching	0.33	0.34
Random + 2005 MFLP	0.32	
Random + 1960 MFLP		0.45
Standardized Table	0.34	0.35
Standardized Table + 2005 MFLP	0.33	
Standardized Table + 1960 MFLP		0.44

Note: The appendix contains additional information on the methodology used to generate this table.

4 Appendix

4.1 Data

The data used for this paper is freely available from the Integrated Public Use Microdata Series (IPUMS) website. The samples used in this study are taken from the 1 percent sample of the Census for the years 1960, 1970, 1980, 1990, 2000 and the American Community Survey (ACS) for the year 2005. The following variables were included for every year: year or the survey (variable name: year), spouse location flag (sploc), number of family members in the household (famsize), number of children in the household (nchild), age (age), sex (sex), marital status (marst), educational attainment (educ), employment status (empstat), total family income (ftotinc), wage and salary income (incwage). Only singles and married couples that are 25 to 54 years old are considered. The adults in these households either live by themselves or with their children, who are less than 19 years old. Households in which there are other members such as grandparents, uncles/aunts, or other unrelated individuals are excluded. Households with subfamilies of any other type are also excluded from the analysis. Finally, widows, widowers and married individuals whose spouses are absent are excluded as well. Income variables are restricted to be non-negative.

There are 560 types of households used in the analysis. Households are broken down into finer categories than are reported in the text. In principle, this doesn't affect the analysis, since the finer classifications can be combined to attain the more aggregated ones. Following a counterfactual experiment, some households are moved into new income percentiles. So, in practice the finer classification allows more accurate re-sorting into the various income percentile when conducting the counterfactual experiments. Households are classified into different types as follows:

1. Marital status: married, never married males, never married females, divorced males, divorced females.

2. Education: less than high school, high school, some college, college, more than college. For married households, both the husband and wife will have one of these educational levels.

3. Market work: work, does not work. For married households both the husband and wife will have one of these levels of labor market activity.

4. Children: no children, 1 child, 2 children, more than 2 children.

Finally, households are divided into 10 deciles. So, for every year, there are 5,600 (i, j)-combinations of household types/deciles.

4.2 The Lorenz Curve and Gini Coefficient

Think of a sample of different household types, $i \in \{1, 2, ..., m\}$, situated in different percentiles, $j \in \mathcal{J}$, of the income distribution. Again, j is expressed as a fraction. Define f_{ij} as the fraction of households that are of type-i in income percentile j. Let r_{ij} represent household (i, j)'s income, y_{ij}, *relative* to mean income, y. Each household's income is adjusted to a per-adult-equivalent basis using the OECD modified equivalence scale, which counts the first adult as 1, the second adult as 0.5 and each child as 0.3 adults. Equivalized households incomes are then divided by mean household income across the whole sample.

The share of income earned by percentile j is

$$s_j = \sum_i f_{ij} r_{ij}.$$

The Lorenz curve is derived by plotting the cumulative shares of the population indexed by percentile p,

$$p = \sum_j^p \sum_i^m f_{ij}$$

on the x-axis, against the cumulative share of income indexed by percentile p,

$$l_p = \sum_j^p s_j,$$

on the y-axis. Suppose that the unit interval is split up into n equally sized segments. Then, $j \in \mathcal{J} = \{1/n, \cdots, 1 - 1/n, 1\}$.

Take the example of $n = 4$ (quartiles). The Lorenz curve described above is plotted in Figure 1. The Gini coefficient associated with the Lorenz curve equals twice the area between the Lorenz curve and the 45-degree line. Alternatively, the coefficient can be calculated as equaling $1 - 2\Delta$, where Δ is the area below the Lorenz Curve. In the case of quartiles the area Δ is the summation of the areas of the right triangle A, the right trapezoids B, C, and D. The coordinates on the x-axis are given by 0, $p_1 = 0.25$, $p_2 = 0.5$, $p_3 = 0.75$, and 1.0. The y-axis coordinates of the Lorenz curve are given by 0, $l_{0.25}$, $l_{0.5}$, $l_{0.75}$, and 1.0.

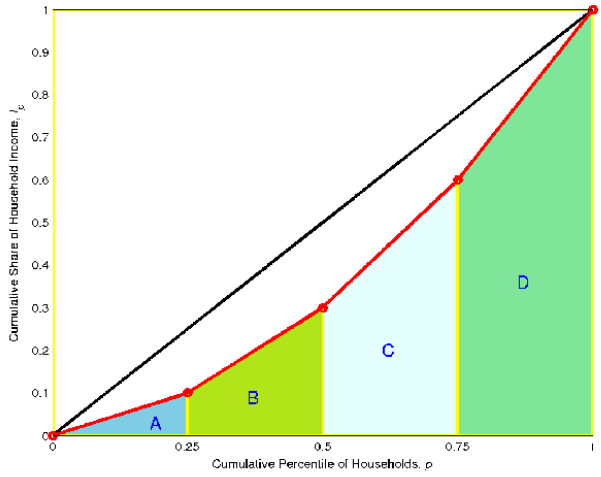

Figure A1: Lorenz Curve and the Gini Coefficient

Note: The figure shows the construction of a Lorenz curve when there are four percentiles (quartiles). The Gini coefficient is twice the area between the 45 degree line and the Lorenz curve.

Then, using the formulas for the geometric areas A, B, C, and D, the Gini coefficient, g, can be derived as

$$g = 1 - 2\left(\underbrace{\frac{p_1 l_1}{2}}_{\text{Area } A} + \underbrace{\frac{(l_1 + l_2)(p_2 - p_1)}{2}}_{\text{Area } B} + \underbrace{\frac{(l_2 + l_3)(p_3 - p_2)}{2}}_{\text{Area } C} + \underbrace{\frac{(l_3 + 1)(1 - p_3)}{2}}_{\text{Area } D}\right).$$

After rearranging and canceling out terms, the expression for the Gini coefficient can be simplified to

$$g = (p_1 l_2 - p_2 l_1) + (p_2 l_3 - p_3 l_2) + (p_3 - l_3).$$

The cumulative shares of the population, the p's, are based on quartiles; i.e., $p_1 = 1/4, p_2 = 2/4, \cdots$. Thus, above expression can be rewritten as

$$g = \frac{1}{4}\left[(l_2 - 2l_1) + (2l_3 - 3l_2) + (3 - 4l_3)\right].$$

In the more general case of n percentiles, the Gini coefficient equals

$$g = \sum_{p=1/n}^{1-1/n} [pl_{p+1} - (p + 1/n)l_p].$$

The version of this formula for an arbitrary number of income groups of any size and an arbitrary number of sub-populations (types) is presented in Rao (1969).

4.3 Counterfactual Experiments

4.3.1 Imposing Random Matching

Random matching can be imposed on the demographic structure of the U.S. population for each of these years in the sample. Counterfactual Gini coefficients can then be computed. How is this done?

First a bit of notation. Take the distribution of household, $\{f_{ij}\}$. Recall that married households are indexed by the education of the husband, the education of the wife, their labor-force participation, and the number of children in the household. Let the sets $\mathcal{M}_{E_H}$ contain the indices of all married households with a husband who has the educational level, $E_H \in \{HS-, HS, C-, C, C+\}$, where $HS-$ refers to a less-than-high-school educated person, HS refers to someone with a high-school education, $C-$ is some college, C is college, and $C+$ is more-than-college educated. Similarly, the sets $\mathcal{M}_{E_W}$ contain married households with different educational levels for wives, E_W. Furthermore, $\mathcal{M}_{LFP_H}$ ($\mathcal{M}_{LFP_W}$) contain all the married households with a husband's (wife's) labor-force participation status $LFP_{H(W)} \in \{WORK_{H(W)}, \tilde{}WORK_{H(W)}\}$. Finally, the set $\mathcal{M}_{KIDS}$ contain married households with a particular number of children $KIDS \in \{0, 1, 2, 2+\}$. The set of all married households with a particular mix of the education, $\mathcal{M}_{E_H,E_W}$, for the husband and the wife

reads

$$\mathcal{M}_{E_H,E_W} = \mathcal{M}_{E_H} \cap \mathcal{M}_{E_W}.$$

Let $\mathcal{M}$ represent the set containing all of the different types of married households. Clearly,

$$\mathcal{M} = \bigcup\nolimits_{E_H,E_W,LFP_H,LFP_W,KIDS} (\mathcal{M}_{E_H} \cap \mathcal{M}_{E_W} \cap \mathcal{M}_{LFP_H} \cap \mathcal{M}_{LFP_W} \cap \mathcal{M}_{KIDS}),$$

where the term in parenthesis is the set of all married households of type (E_H, E_W, LFP_H, $LFP_W, KIDS$).

Here is an example illustrating how the random matching experiment is performed. Take the first element of the matching table in 1960–see Table 1 of the main text. These are the marriages where both the husband and the wife are less-than-high-school educated. In 1960, the fraction of such marriages was 0.32. In terms of the current notation,

$$\frac{\sum_{i \in \mathcal{M}_{HS-,HS-}} \sum_{j=0.1}^{1} f_{ij}^{1960}}{\sum_{i \in \mathcal{M}} \sum_{j=0.1}^{1} f_{ij}^{1960}} = 0.32.$$

Now impose the random matching table entry for these marriages–again, see Table 1 of the main text. The fraction of such marriages, if matching in 1960 was random, is 0.21. Denote the counterfactual distribution to be imposed in 1960 by $\widetilde{f}_{ij}^{1960}$. The following equation must hold for the particular marriage group being discussed

$$\frac{\sum_{i \in \mathcal{M}_{HS-,HS-}} \sum_{j=0.1}^{1} \widetilde{f}_{ij}^{1960}}{\sum_{i \in \mathcal{M}} \sum_{j=0.1}^{1} f_{ij}^{1960}} = 0.21.$$

The elements in the contingency table refer to the fraction of all married households that a particular type of match between husbands' and wives' educational levels constitutes. The elements in the cells are totals across all income percentiles. The f_{ij}'s refer to the fraction of all households, married and single, that are of type i in income percentile j. Therefore, the cells in the contingency table are aggregated over income percentiles (as well as the other non-educational traits characterizing married households). The ratio of the total number of

type-$(HS-, HS-)$ marriages under random matching to what occurs in the data is

$$\frac{\sum_{i\in\mathcal{M}_{HS-,HS-}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}}{\sum_{i\in\mathcal{M}_{HS-,HS-}}\sum_{j=0.1}^{1}f_{ij}^{1960}}=\frac{0.21}{0.32}=0.66.$$

So, under random matching the number of type-$(HS-, HS-)$ marriages is reduced by factor of $0.66 = 0.21/0.31$. Assume that this reduction is spread out evenly across all of the income percentiles, or across all of the j's. Therefore, when undertaking the random matching experiment, $\widetilde{f}_{ij}^{1960}$should be constructed as follows:

$$\widetilde{f}_{ij}^{1960}=\frac{0.21}{0.31}f_{ij}^{1960}, \text{ for } i\in\mathcal{M}_{HS-,HS-} \text{ and all } j.$$

A similar scaling operation is performed for each of the other 24 possible matches. Thus, there is a scaling factor specific to each type of marriage (in the contingency table). For all single and divorced people, keep the original fractions; i.e., $\widetilde{f}_{ij}^{1960}=f_{ij}^{1960}$.

4.3.2 Imposing Random Matching While Holding Fixed Married Female Labor-Force Participation

The impact of random matching on inequality can be interacted with changes in the labor-force participation decisions of married females. The procedure for imposing random matching in 1960 is outlined in the previous section. Suppose that in addition to imposing random matching in 1960, married female labor-force particpation is fixed at its 2005 level. How can this be implemented?

The married female labor-force participation rate in 1960 when random matching is imposed is

$$\frac{\sum_{i\in\mathcal{M}_{WORK_W}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}}{\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}}=0.33,$$

while the labor-force participation rate in 2005 is

$$\frac{\sum_{i\in\mathcal{M}_{WORK_W}}\sum_{j=0.1}^{1}f_{ij}^{2005}}{\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}f_{ij}^{2005}}=0.68.$$

Denote the desired new counterfactual distribution for married households in 1960 by $\widehat{f}_{ij}^{1960}$, for $i \in \mathcal{M}$ and all j. This new counterfactual distribution for 1960 must give the 2005 married female labor-force participation rate so

$$\frac{\sum_{i\in\mathcal{M}_{WORK_W}}\sum_{j=0.1}^{1}\widehat{f}_{ij}^{1960}}{\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widehat{f}_{ij}^{1960}} = 0.68.$$

Bear in mind that the fraction of married people in 1960 does not change in the counterfactual experiments; i.e.,

$$\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1} f_{ij}^{1960} = \sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960} = \sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widehat{f}_{ij}^{1960}.$$

Consequently,

$$\frac{\sum_{i\in\mathcal{M}_{WORK_W}}\sum_{j=0.1}^{1}\widehat{f}_{ij}^{1960}}{\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}} = \frac{0.68}{0.33}\frac{\sum_{i\in\mathcal{M}_{WORK_W}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}}{\sum_{i\in\mathcal{M}}\sum_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}} = 0.68.$$

Imposing a labor-force participation rate from 2005 onto the 1960 counterfactual distribution of random matching amounts to scaling up all (i,j)-combinations of married households in which women work. On the other hand, the married households in which women do not work should be scaled down so that the total fraction of married households does not change.

Therefore, the counterfactual distribution, $\{\widehat{f}_{ij}^{1960}\}$, should be constructed in the following way:

$$\widehat{f}_{ij}^{1960} = \frac{0.68}{0.33}\widetilde{f}_{ij}^{1960}, \text{ for } i \in \mathcal{M}_{WORK_W} \text{ and all } j,$$

and

$$\widehat{f}_{ij}^{1960} = \frac{1-0.68}{1-0.33}\widetilde{f}_{ij}^{1960}, \text{ for } i \in \mathcal{M}\text{-}_{WORK_W} \text{ and all } j.$$

This way the total fraction of married households stays constant,

$$\begin{aligned}
\sum_{i\in\mathcal{M}}\sum\nolimits_{j=0.1}^{1}\widehat{f}_{ij}^{1960} &= \sum_{i\in\mathcal{M}_{WORK_W}}\sum\nolimits_{j=0.1}^{1}\widehat{f}_{ij}^{1960} + \sum_{i\in\mathcal{M}_{\text{-}WORK_W}}\sum\nolimits_{j=0.1}^{1}\widehat{f}_{ij}^{1960} \\
&= \frac{0.68}{0.33}\sum_{i\in\mathcal{M}_{WORK_W}}\sum\nolimits_{j=0.1}^{1}\widetilde{f}_{ij}^{1960} + \frac{1-0.68}{1-0.33}\sum_{i\in\mathcal{M}_{\text{-}WORK_W}}\sum\nolimits_{j=0.1}^{1}\widetilde{f}_{ij}^{1960} \\
&= 0.68\sum_{i\in\mathcal{M}}\sum\nolimits_{j=0.1}^{1}\widetilde{f}_{ij}^{1960} + (1-0.68)\sum_{i\in\mathcal{M}}\sum\nolimits_{j=0.1}^{1}\widetilde{f}_{ij}^{1960} \\
&= \sum_{i\in\mathcal{M}}\sum\nolimits_{j=0.1}^{1}\widetilde{f}_{ij}^{1960}.
\end{aligned}$$

As with the previous counterfactual distribution adjustment, keep the original fractions, $\widehat{f}_{ij}^{1960} = f_{ij}^{1960}$, for all single and divorced people.

4.4 Standardizing Contingency Tables

Mosteller (1968) suggests that when comparing two contingency tables they should first be standardized so that they both have the same marginal distributions associated with the rows and columns. Take a 5×5 table. It can be standardized so that each element of the two marginal distributions is $1/5$. This can be done by employing the Sinkhorn-Knopp (1967) algorithm, which iteratively scales each row and column. Standardization preserves the core pattern of association in a contingency table. For example, Tan, Kumar and Srivastava (2004) note that such standardization does not affect the odds ratios in a contingency table, a typical measure used to gauge the pattern of association between variables.

4.4.1 Sinkhorn-Knopp (1967) Algorithm

1. Enter an iteration with a contingency table.

2. This contingency table has a marginal distribution associated with the rows (for men) obtained by summing each row along its columns to obtain a total for that row. Divide each row through by 5 times its total. The marginal distribution associated with the rows is now $(1/5, 1/5, 1/5, 1/5, 1/5)$.

3. Compute the marginal distribution associated with the columns (for women) by summing each column along its rows to obtain a total for that column. Divide each column through by its 5 times its total.

4. Recompute the marginal distribution associated with the rows. It has changed following the previous two steps. Check its distance from the desired marginal distribution $(1/5, 1/5, 1/5, 1/5, 1/5)$. If it has reached the desired level of closeness then stop. If not, go back to Step 1.

4.4.2 The Standardized Tables

The two resulting standardized tables for 1960 and 2005 are shown in Table A1. The diagonal elements in the 2005 table are larger than in the 1960 one. Assortative mating has increased.

There is no need to standardize the tables so that each element of the marginal distributions is 1/5. One can standardize the 1960 table so that its marginal distributions coincide with those in the data for 2005, or vice versa. This way the standardized table for 1960 (2005) can be compared with the one from the data for 2005 (1960). Both tables will have the same 2005 (1960) marginal distributions. This results are shown in Table A2. By comparing the standardized table for 1960 with the one in the data for 2005 (see Table 1 in the text) it can be seen that assortative mating has increased. Once again, the diagonal elements are larger in the table for 2005. Likewise, a comparison of the standardized table for 2005 with the one in the data for 1960 shows an increase in assortative mating (again, see Table 1 in the text).

Table A1: Standardized Contingency Table: Assortative Mating by Educational Class

Marital Sorting by Education

1960

Marginal Distributions $= (1/5, \cdots, 1/5)$

Husband	Wife				
	HS-	HS	C-	C	C+
HS-	**0.126**	0.043	0.017	0.007	0.007
HS	0.046	**0.079**	0.038	0.019	0.017
C-	0.020	0.045	**0.067**	0.037	0.032
C	0.005	0.023	0.047	**0.081**	0.043
C+	0.002	0.010	0.031	0.055	**0.102**
Marginal, Wives	1/5	1/5	1/5	1/5	1/5

2005

Marginal Distributions $= (1/5, \cdots, 1/5)$

HS-	**0.146**	0.035	0.014	0.004	0.002
HS	0.035	**0.088**	0.047	0.019	0.011
C-	0.013	0.047	**0.079**	0.038	0.023
C	0.004	0.021	0.039	**0.082**	0.054
C+	0.002	0.010	0.022	0.057	**0.109**
Marginal, Wives	1/5	1/5	1/5	1/5	1/5

Note: The upper panel shows the contingency table for 1960 when it has been normalized using the Sinkhorn-Knopp algorithm so that each element of marginal distributions over education for men and women equals $1/5$. The lower panel shows the same thing for 2005.

Table A2: Standardized Contingency Table: Assortative Mating by Educational Class

Marital Sorting by Education

1960

Using the 2005 Marginal Distributions

Husband	Wife				
	HS-	HS	C-	C	C+
HS-	**0.029**	0.035	0.011	0.005	0.003
HS	0.030	**0.186**	0.072	0.040	0.019
C-	0.008	0.065	**0.079**	0.048	0.022
C	0.002	0.032	0.055	**0.101**	0.028
C+	0.001	0.010	0.025	0.048	**0.047**
Marginal, Wives	0.070	0.329	0.242	0.241	0.118

2005

Using the 1960 Marginal Distributions

HS-	**0.354**	0.114	0.015	0.002	0.000
HS	0.054	**0.183**	0.033	0.007	0.001
C-	0.011	0.054	**0.031**	0.008	0.001
C	0.004	0.027	0.017	**0.019**	0.003
C+	0.002	0.017	0.013	0.017	**0.009**
Marginal, Wives	0.425	0.396	0.110	0.054	0.016

Note: The upper panel shows the contingency table for 1960 when it has been normalized using the Sinkhorn-Knopp algorithm so that the marginal distributions for men and women over education equal what there are in the data for 2005. The lower panel shows the contingency table for 2005 when it has been normalized so that the marginal distributions for men and women equal what there are in the data for 1960.

4.5 A Brief Literature Review

The increase in assortative mating in the U.S. has also been examined by Hou and Myles (2008), Lam (1997), Qian and Preston (1993), and Schwartz and Mare (2005) to name a few papers. Siow (2013) documents an increase in educational homogamy, but not a general increase in positive assortative matching. Lam (1997) and Schwartz (2010) discuss the relationship between assortative mating and income inequality. Cancian and Reed (1998, 1999) also focus on the role that married female-labor force participation plays in the relationship between assortative mating and income inequality.

ADDITIONAL REFERENCES FOR THE APPENDIX

Cancian, Maria and Deborah Reed. 1999. "The Impact of Wives' Earnings on Income Inequality: Issues and Estimates." *Demography*, 36, (2): 173–84.

Hou, Feng and John Myles. 2008. "The Changing Role of Education in the Marriage market: Assortative Marriage in Canada and the United States since the 1970s." *Canadian Journal of Sociology*, 33 (2): 337-366.

Lam, David. 1997. "Demographic Variables and Income Inequality." In *Handbook of Population and Family Economics*, edited by Mark R. Rosenzweig and Oded Stark, 1015–1059. Amsterdam, Elsevier North Holland.

Mosteller, Frederick. 1968. "Association and Estimation in Contingency Tables." *Journal of the American Statistical Association*, 63 (321): 1-28.

Qian, Zhenchao and Samuel H. Preston. 1993. "Changes in American Marriage, 1972 to 1987: Availability and Forces of Attraction by Age and Education." *American Sociological Review*, 58 (4): 482-495.

Rao, V. M. 1969. "Two Decompositions of Concentration Ratio." *Journal of the Royal Statistical Society*, Series A (General), 132 (3): 418-425.

Schwartz, Christine R. and Robert D. Mare. 2005. "Trends in Educational Assortative Marriage from 1940 to 2003." *Demography*, 42 (4): 621-646.

Sinkhorn, Richard and Paul Knopp. 1967. "Concerning Nonnegative Matrices and Doubly Stochastic Matrices." *Pacific Journal of Mathematics*, 21 (2): 343-348.

Siow, Aloysius. 2013. "Testing Becker's Theory of Positive Assortative Matching." *Journal of Labor Economics*, forthcoming.

Tan, Pang-Ning, Vipin Kumar, Jaideep Srivastava. 2004. "Selecting the Right Objective Measure for Association Analysis." *Information Systems*, 29 (4): 293–313.