*Journal of Demographic Economics* (2023), **89**, 1–27
doi:10.1017/dem.2021.1

CAMBRIDGE
UNIVERSITY PRESS

RESEARCH PAPER

# A new method for identifying the role of marital preferences at shaping marriage patterns

Anna Naszodi[1,2*] and Francisco Mendonca[1]

[1]European Commission, Joint Research Centre (JRC), Ispra, Italy and [2]Centre for Economic and Regional Studies (KRTK), Budapest, Hungary
*Corresponding author. E-mails: anna.naszodi@ec.europa.eu, anna.naszodi@gmail.com

(Received 12 December 2019; revised 28 December 2020; accepted 5 January 2021;
first published online 1 March 2021)

## Abstract

We develop a method which assumes that marital preferences are characterized either by the scalar-valued measure proposed by Liu and Lu, or by the matrix-valued generalized Liu–Lu measure. The new method transforms an observed contingency table into a counterfactual table while preserving its (generalized) Liu–Lu value. After exploring some analytical properties of the new method, we illustrate its application by decomposing changes in the prevalence of homogamy in the US between 1980 and 2010. We perform this decomposition with two alternative transformation methods as well where both methods capture preferences differently from Liu and Lu. Finally, we use survey evidence to support our claim that out of the three considered methods, the new transformation method is the most suitable for identifying the role of marital preferences at shaping marriage patterns. These data are also in favor of measuring assortativity in preferences à la Liu and Lu.

**Key words:** Altham's index; assortative mating; Choo–Siow model; counterfactual decomposition; educational homogamy; generalized Liu–Lu measure; index of marital preferences; iterative proportional fitting algorithm; Liu–Lu measure; marital sorting parameter

**JEL classification:** J12; C02

## 1. Introduction

Measuring the effect of changing marital preferences on the changing marital patterns is in the center of interest of many demographers, economists, and sociologists. This task is challenging because the observed equilibrium outcome in the marriage market depends not only on marital preferences but also on the structural availability of prospective partners with different traits as well as the interplay between preferences and availability (Kalmijn 1998).[1] Accordingly, social scientists aim at answering questions such as: what would be the share of educationally homogamous couples

[1] For instance, if there are no marriages between women with a university diploma and men without a high school degree in a hypothetical society, then it might signal that either one or both of the educational

(the proportion of couples where the spouses have the same education level) in our society today provided people nowadays had the same marital/mating preferences as people used to have in the past? What would be the share of educationally homogamous couples like in our society today if the education levels of young men and women were the same as in an older generation?

The assortative mating literature offers various ways for addressing these questions for cases when the assorted trait is a categorical variable and the marriage matching equilibrium is represented by a contingency table. The examples include the following: (i) controlling for marital preferences by the aggregate marriage matching function derived by Choo and Siow (2006); (ii) applying the iterative proportional fitting (IPF) algorithm developed by Stephan and Deming (1940) and generalized by Sinkhorn (1967); (iii) keeping a similarity coefficient fixed, e.g., the correlation between the couples' trait [e.g., Kremer (1997), Fernandez *et al.* (2005)].

In this paper, we develop a method (new method hereafter) for constructing counterfactuals.[2] The purpose of this paper is to introduce the new method. We do that by presenting its theoretical background, discussing its analytical properties and empirical properties, illustrating its empirical application, and comparing it with some alternative methods (see Table 1). Our work facilitates method selection for those researchers who would like to study changing marriage patterns by analyzing contingency tables.

The new method builds on the work by Liu and Lu (2006). They propose a new measure on the "degree of sorting". The new method transforms an observable contingency table into a contingency table under a counterfactual while preserving the Liu–Lu measure of the table. For this reason, the new method is seemingly similar to those ad hoc statistical approaches, where the value of a similarity coefficient is preserved by the transformation.

However, our choice of the Liu–Lu measure and our choice of the new method are well motivated. First, these choices are not independent of each other. If we knew that the Liu–Lu measure is more appropriate to characterize marital preferences than its alternatives, including the countless number of similarity coefficients, then we would also know that the new method is better than its alternatives. Similarly, if we knew that the new method is better than its alternatives, then we would also know that the Liu–Lu measure is more adequate to characterize marital preferences than its alternatives. Second, Liu and Lu (2006) present theoretical arguments in favor of their measure. Specifically, they claim that their measure can adequately control for changes in the trait distribution. For this reason, it is apt for separating variations in the distribution of assorted traits from changes in marital preferences. Third, in this paper, we present empirical evidence supporting the new method.

In the empirical part of this paper, we illustrate the application of the new method, while we also apply some existing methods. There, we decompose changes in the American marital patterns between 1980 and 2010 using census data. Our empirical findings are the following. First, some simple transformation methods do not even yield counterfactuals that make sense from the point of view of economics. Second, the well-known IPF algorithm and the Choo–Siow (CS hereafter) marriage matching function do. Third, we find that the results of some of our decompositions are sensitive to the choice of the method.

groups do not find the members of the other group attractive as partners. Alternatively, it can also be the consequence of all men in this society being highly educated.

[2] The new method is implemented in Excel, Visual Basic, and R. It can be downloaded from http://dx.doi.org/10.17632/x2ry7bcm95.1.

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

**Table 1.** Comparing the new method with five deterministic methods

| | New method | Iterative proportional fitting algorithm | Choo–Siow model-based method | Regression coeff.-based method | Pearson’s correlation coeff.-based method | Covariance coeff.-based method |
|---|---|---|---|---|---|---|
| Theoretical background | Statistical approach | Computational algorithm | Structural model | Statistical approach | Statistical approach | Statistical approach |
| Analytical properties | | | | | | |
| Gender symmetry♣ | Yes | Yes | Yes | No | Yes | Yes |
| Transformed table | Unique | | | | | |
| Transformation preserves | Liu–Lu matrix | Odds ratio matrix | Marital surplus matrix | Matrix of regression coeff.s◆ | Matrix of correlation coeff.s◆ | Matrix of covariance coeff.s◆ |
| Transformed table is obtained by | Closed-form formula | Algorithm | Solving a system of quadratic equations | | | |
| Transformation method | Commutes with the operation of merging categories | | | | | |
| Potential limitations/lack of limitations | Seed table can contain zeros | Zero in the seed table can be a problem | Number of singles of each type should be positive | | | |

(*Continued*)

**Table 1.** (*Continued.*)

| | New method | Iterative proportional fitting algorithm | Choo–Siow model-based method | Regression coeff.-based method | Pearson's correlation coeff.-based method | Covariance coeff.-based method |
|---|---|---|---|---|---|---|
| Theoretical background | Statistical approach | Computational algorithm | Structural model | Statistical approach | Statistical approach | Statistical approach |
| Empirical properties | | | | | | |
| US 1980–1990● | Preference component: negative♥★ | Positive | Positive | Transformed tables have negative entries | | |
| US 1990–2000● | Preference component: positive♥ | Positive | Positive | Transformed tables have negative entries | | |
| US 2000–2010● | Preference component: positive♥★ | Positive★ | Zero | Transformed tables have negative entries | | |
| US 1980–2010♠ | Preference component: positive★ | Positive★ | Positive★ | | | |
| Portugal 1981–2011● | Transformed table has a neg. entry | | | | | |

*Notes*: ♣: invariance to interchanging wives' data and husbands' data. ♦: wives' and husbands' education level are dummy variables that take the value 1 for high level of education defined in all possible ways. ♥: robust to considering a fourth educational category: either "some college", or "less than primary" (in addition to the three categories in the main specification that are "less than high school", "high school completed", "university degree"), and also to an alternative definition of young couples (men are aged 30–34 years vs. women are aged 30–34 years). ★: supported by survey evidence (Changing American Family survey conducted by the Pew Research Center in 2010). ●: decomposition of the change in the share of educationally homogamous couples is performed with the single-period decomposition scheme in equation (11). ♠: decomposition is performed with the multi-period decomposition scheme in equation (12). Empty cells: not analyzed in this paper.

The contribution of this paper is threefold. First, a minor contribution is that we generalize the Liu–Lu measure. Unlike the original scalar-valued Liu–Lu measure, the generalized Liu–Lu measure is a matrix. While the original Liu–Lu measure is defined for a dichotomous assorted trait, the generalized Liu–Lu measure is defined even when the trait variable can take more than two possible values. For instance, when we distinguish not only between low and high education levels, but more.

Second, and most importantly, we develop a novel transformation method and investigate its properties. The new method is suitable to transform not only 2-by-2 contingency tables, but also larger tables. It facilitates comparing the Liu–Lu matrix with other matrix-valued measures of assortativity in marital preferences. The discussion of this point is presented in Appendix A in the online appendix (see: https://doi.org/10.1017/dem.2021.1 ).[3]

Third, in addition to providing empirical support to the view that the choice of the transformation method is not innocuous, we also present a supplementary analysis in subsection 5.4. The aim of that analysis is to facilitate the choice of the method and also the choice of statistics for characterizing marital preferences.

Although our supplementary analysis has some limitations, its idea of using survey data for method selection is novel. Liu and Lu (2006) also illustrate that the mode of measuring the degree of sorting is crucial: they obtain diverse dynamics with three competing measures by using US census data from the period between 1940 and 2000. However, they offer exclusively theoretical arguments in favor of one of the three measures, the Liu–Lu measure.

The rest of the paper is structured as follows. Section 2 reviews the literature. Section 3 presents the Liu–Lu measure and its generalization. Section 4 introduces the methodology we apply: it presents the decomposition scheme, and the new method. Further, it discusses some analytical properties of the new method. Section 5 conducts decompositions using US census data from four census waves between 1980 and 2010. It illustrates that the sign and the magnitude of the components can be sensitive to the choice of the method. Section 5 ends with the supplementary analysis exploiting survey data. Finally, section 6 concludes the paper.

## 2. Literature

First, we take a stock of some solutions for constructing counterfactuals put forward in three strands of the assortative mating literature. Further, we position our paper with respect to the literature. Then we discuss several points of contribution of the recent paper by Eika *et al.* (2019) to the literature on measure selection, and we highlight the connection between their research and ours.

### *2.1. Three strands of the literature*

The assortative mating literature proposes three main approaches for constructing counterfactuals. First, approaches for conducting counterfactual experiments are typically dictated by *structural models* in those papers, where marital preferences are explicitly modeled. Examples include the models in the seminal paper by Choo and

[3] In Appendix A, one of the matrix-valued measures considered is the matrix of odds-ratios. It forms the basis of a scalar-valued aggregate measure, the Altham's index [see Altham (1970)]. Other examples for matrix-valued measures are the marital sorting matrix, the marital surplus matrix, and the Liu–Lu matrix. The latter forms the basis of our index of changing marital preferences.

Siow (2006), and Chiappori *et al.* (2017) developed in the transferable utility framework, and the models by Dagsvik (2000) and Menzel (2015) in the non-transferable utility framework.

Chen *et al.* (2019) show that these models, together with some other recent contributions to the matching literature, share a common structure. The key elements of the structure are a behaviorally coherent aggregate matching function, and a system of nonlinear equations. The aggregate matching function relates to the distribution of couples to the trait distributions of available men and women in the population. Whereas the system of nonlinear equations determines the number of singles of each type.

The model by Choo and Siow (2006) is probably the best known model in the "matching function—nonlinear system of equations class". Its aggregate matching function is given by a simple closed form formula. More importantly, the contingency tables under the counterfactuals are given by a system of quadratic equations [see equation (20) in Choo and Siow (2006)], which is easy to solve. This motivates us to use the Choo–Siow solution as a benchmark of the new method.

Second, the counterfactuals are often constructed by *computational algorithms*. The most commonly used algorithm in the assortative mating literature is probably the well-known *iterative proportional fitting algorithm* [e.g., Altham and Ferrie (2007), Breen and Salazar (2005), Breen and Salazar (2011), Hu and Qian (2016)].[4] Due to the popularity of the IPF algorithm among demographers, economists, and sociologists, we use it as another benchmark for the new method. As a strength of the computational algorithms, we mention that these allow us to obtain solutions outside the subset of solutions represented by closed form formulas.

Third, some studies apply *statistical approaches*. Papers in this strand of the literature use various, often ad hoc similarity coefficients for controlling for marital preferences. Examples for such coefficients include the *regression coefficient* obtained by regressing the wives' years of education on the husbands' years of education [Greenwood *et al.* (2014, 2015)]; the *proportion of homogamous couples* [Fernandez (2001), Fernandez and Rogerson (2001)]; and the *Pearson's correlation coefficient* between the couples' traits [Kremer (1997), Fernandez *et al.* (2005)]. As the last example, we mention the *Liu–Lu measure* that forms the basis of the new transformation method proposed in this paper.

The distinctive feature of these papers is that their similarity coefficients fulfill certain criteria dictated by heuristics. One criterion is symmetry. The symmetry of a similarity coefficient in the marriage matching context means this. The coefficient does not change if we interchange husbands' data and wives' data. The correlation coefficient, the proportion of homogamous couples, the marital sorting parameters, and the Liu–Lu measure are symmetric. Another criterion is the following. Statistics characterizing marital preferences should be immune to ceteris paribus changes in the trait distributions. Liu and Lu (2006) claim that their new measure fulfills this criterion, while earlier statistical measures in the literature do not.

It is important to note that our classification of the literature is based on the exposition of the methods in the papers. However, the categorization based on the methods themselves would be different. First, such a categorization would put some papers in the second group and the third group into the same box. This is because the assumption behind the application of the IPF algorithm is that marital preference

[4] Another example for a computational algorithm is given by Eika et al. (2019) (see subsection 2.2).

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

at a societal level is characterized by a simple statistic. It is the so-called cross-product ratio [or odds ratio defined below equation (16)]. The cross-product ratio corresponds to a similarity coefficient in the third group of papers.

Second, even papers with structural models can apply the IPF algorithm, if this computational method is consistent with their models. For instance, Dupuy and Galichon (2014) use the IPF method to solve their structural matching model.

Third, one can recover a micro foundation for some seemingly ad hoc approaches in the third group. An example is the harmonic marriage matching function. Although this matching function was not derived from preferences, Chen *et al.* (2019) rationalize it with a structural model.

Due to the vagueness of the categories, the following question arises. Is the transformation method put forward in this paper really new? A related question is whether the new method is better than its alternatives. The second question is addressed in the empirical part of the paper in subsection 5.4 since the criteria we choose are empirical in nature. Here, we address the former question.

First, the exposition of the new method is certainly new: we are not aware of any other paper in the literature proposing a transformation method with Liu–Lu measure invariance. Second, in this paper, we highlight the differences between the new method and its alternatives.

## 2.2. Some recent contributions to the literature on measure selection

The recent paper by Eika *et al.* (2019) decomposes changes in household income inequality in the US between 1962 and 2013. One factor considered by them is the changing educational assortative mating. They contribute to the assortative mating literature in several respects. In this section, we survey three of their contributions, which are particularly relevant for this paper with a focus on methodology.

First, Eika *et al.* (2019) illustrate the same point as Liu and Lu (2006): different measures of marital sorting can exhibit different dynamics. Eika *et al.* (2019) present the time series of six different measures of marital sorting for five decades. They find that some of the six measures hardly change after 1980 in the US, while some others declined. However, they emphasize the result that all six measures have similar dynamics over the entire period studied. Since this long period is in the focus of their study, they do not perform a systematic comparison of the analytical properties of all the six measures.

They investigate a seventh measure in detail: it is the regression coefficient used by Greenwood *et al.* (2014). Eika *et al.* (2019) point out that the regression coefficient is an asymmetric measure, i.e., its value and trend may depend on whether it is obtained by regressing wives' years of education on husbands' years of education or the other way around.

Eika *et al.* (2019) develop a stochastic matching procedure. It transforms a contingency table into a stochastic table representing a counterfactual.[5]

[5] The transformation keeps some moments of the generated table close to their deterministic counterparts characterizing the original table, while it also makes the marginal distributions of the transformed table equal to their preset distributions. The transformed table is overdetermined typically due to the high number of moment conditions. The stochastic matching procedure offers a solution for this problem (see Appendix C in Eika et al. 2019).

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

Let us see now, how our research is related to these contributions. We think that the finding by Eika *et al.* (2019) on the diverse dynamics of some scalar-valued measures after 1980 deserves special attention. Since this period breaks the empirical equivalence among different measures, it facilitates measure selection. This we do in subsection 5.4 based on empirical criteria derived from survey evidence.

The work by Eika *et al.* motivates us to narrow down our analysis to symmetric measures. Finally, the scope of the empirical part of our paper is different from the scope of the paper by Eika *et al.* (2019): our paper is limited to applying methods producing deterministic transformed tables. Additional differences between our paper and theirs include differences in the decomposition schemes applied, the age of observed individuals, the number of education categories, the way of aggregating matrix-valued measures (and thereby producing a scalar-valued composite indicator of marital preferences). While each of these may influence the outcome of the decompositions, almost none of them is crucial for theoretical considerations about measure/method/model selection.

## 3. Characterizing marital preferences à la Liu and Lu

In this section, first we introduce a characterization of marital preferences in the society based on the work by Liu and Lu (2006). According to this concept, marital preferences are unchanged in a society if the measure on the degree of assortative mating developed by Liu and Lu (2006) is constant. Then, in subsection 3.2, we propose a novel extension of the Liu–Lu measure.

Readers interested only in the intuition behind this measure may skip subsection 3.2. What subsection 3.2 adds is the mathematical formula we use in the empirical decompositions.

### *3.1. The Liu–Lu measure*

In the Liu–Lu model, there are $N$ women and $N$ men of marriageable age. It is assumed that no one will remain single: all of the $N$ women and $N$ men will marry a person of the opposite sex eventually.[6] Further, it is assumed that individuals are matched on a one-dimensional trait. We can think of this trait as the education level. It is captured by a dichotomous variable that can take two possible values, low ($L$) and high ($H$). Accordingly, couples can be of four types and the contingency table representing the matching outcome is a 2-by-2 matrix:

$$K = \begin{bmatrix} N_{L,L} & N_{L,H} \\ N_{H,L} & N_{H,H} \end{bmatrix}, \tag{1}$$

where $N_{L,L}$ ($N_{H,H}$) denotes the number of couples where both spouses are low (high) educated. $N_{L,H}$ ($N_{H,L}$) stands for the number of couples where the husband (wife) is low educated, while the wife (husband) is high educated.

If the contingency table $K$ is known, the trait distributions of both married men and married women are known as well. Specifically, the number of husbands with high education level is $N_{H,\cdot} = N_{H,L} + N_{H,H}$. The number of wives with low (high) education

[6] This assumption is relaxed in our companion paper, Naszodi and Mendonca (2019a). There, we even distinguish between voluntary singles and involuntary singles.

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

level is $N_{\cdot,L} = N_{L,L} + N_{H,L}$ $(N_{\cdot,H} = N_{L,H} + N_{H,H})$. The gender-specific educational distributions are assumed to be non-degenerate: $0 < N_{H,\cdot} < N$, and $0 < N_{\cdot,H} < N$.

The *Liu–Lu measure* is given by

$$\mathrm{LL}(K) = \begin{cases} \dfrac{N_{H,H} - Q^-}{\min(N_{H,\cdot}, N_{\cdot,H}) - Q^-}, & \text{if} \quad N_{H,H} \geq Q, \\ \dfrac{N_{H,H} - Q^+}{Q^+ - \max(0, N_{H,\cdot} - N_{\cdot,L})}, & \text{if} \quad N_{H,H} < Q, \end{cases} \tag{2}$$

where again $N$ denotes the total number of couples, while $Q = N_{H,\cdot} N_{\cdot,H}/N$ is the expected number of $H$, $H$-type couples under random matching. Furthermore, $Q^-$ is the biggest integer being smaller than or equal to $Q$, while $Q^+$ is the smallest integer being larger than or equal to $Q$.

The *statistical interpretation of the Liu–Lu measure* is this. It is the signed normalized distance between the realized matching outcome $K$ and a benchmark outcome where individuals are randomly matched. If the number of $H$, $H$-type couples equals to its (integer valued) expected value under random matching, i.e., $N_{H,H} = Q^-$, the Liu–Lu measure takes the value zero. If the number of $H$, $H$-type couples is higher, then the Liu–Lu measure is positive. Otherwise, the Liu–Lu measure is negative.

In the extreme case, when sorting maximizes the number of $H$, $H$-type couples, the Liu–Lu measure takes its maximum value, 1. In the other extreme case, when sorting minimizes the number of $H$, $H$-type couples, the Liu–Lu measure takes its minimum value, minus 1. Finally, all feasible matching outcomes of any given trait distribution is ranked uniquely by the number of $H$, $H$-type couples. This ranking principle defines the distance.

Under the assumption $N_{H,H} \geq Q$, the Liu–Lu formula simplifies to

$$\mathrm{LL}^{\mathrm{sim}}(K) = \frac{N_{H,H} - Q^-}{\min(N_{H,\cdot}, N_{\cdot,H}) - Q^-}. \tag{3}$$

Equation (3) is the empirically relevant version of the Liu–Lu measure for positively assorted traits, such as education level.

### 3.2. The generalized Liu–Lu measure

In the empirical part of the paper, we consider not only two, but three education levels, i.e., low ($L$ = no high school degree was obtained), medium ($M$ = high school was completed but no college or university degree was obtained), and high ($H$ = tertiary education was completed). Allowing the trait variable to take more than two possible values requires to *generalize the original scalar-valued Liu–Lu measure*.

We show the generalization under the assumption that the *trait variable is ordered*. First, we discuss the case where the contingency table is a 3-by-3 table, such as

$$K = \begin{bmatrix} N_{L,L} & N_{L,M} & N_{L,H} \\ N_{M,L} & N_{M,M} & N_{M,H} \\ N_{H,L} & N_{H,M} & N_{H,H} \end{bmatrix}. \tag{4}$$

Then, we also provide the formula for any contingency table larger than 2 × 2.

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

As a first step, we dichotomize the educational trait variable, i.e., transform it into a binary variable. If the trait variable can take three possible values, the dichotomization can be done in four different ways depending on whether the $M$-type husbands and the $M$-type wives are considered to be low or high educated in the dichotomous world. We note that the number of dichotomizations is not independent of our assumption that the trait variable is ordered.[7]

The four dichotomizations result in four 2-by-2 contingency tables. Let us introduce the notation $K^{H,H}$ for the 2-by-2 contingency table obtained by reclassifying all the $M$-type husbands and wives in equation (4) to $H$-type. It gives

$$K^{H,H} = \begin{bmatrix} N_{L,L} & N_{L,M} + N_{L,H} \\ N_{M,L} + N_{H,L} & N_{M,M} + N_{M,H} + N_{H,M} + N_{H,H} \end{bmatrix}. \tag{5}$$

Analogously, we use the notation $K^{H,L}$ for the 2-by-2 contingency table obtained by reclassifying the $M$-type husbands to $H$-type and the $M$-type wives to $L$-type:

$$K^{H,L} = \begin{bmatrix} N_{L,L} + N_{L,M} & N_{L,H} \\ N_{M,L} + N_{M,M} + N_{H,L} + N_{H,M} & N_{M,H} + N_{H,H} \end{bmatrix}, \tag{6}$$

The notation $K^{L,H}$ stands for the 2-by-2 contingency table obtained by reclassifying the $M$-type husbands to $L$-type and the $M$-type wives to $H$-type:

$$K^{L,H} = \begin{bmatrix} N_{L,L} + N_{M,L} & N_{L,M} + N_{L,H} + N_{M,M} + N_{M,H} \\ N_{H,L} & N_{H,M} + N_{H,H} \end{bmatrix}. \tag{7}$$

Finally, the notation $K^{L,L}$ is used for the 2-by-2 contingency table obtained by reclassifying all the $M$-type husbands and wives to $L$-type:

$$K^{L,L} = \begin{bmatrix} N_{L,L} + N_{L,M} + N_{M,L} + N_{M,M} & N_{L,H} + N_{M,H} \\ N_{H,L} + N_{H,M} & N_{H,H} \end{bmatrix}. \tag{8}$$

In the second step, we calculate the original Liu–Lu measure of $K^{H,H}$, $K^{H,L}$, $K^{L,H}$, and $K^{H,H}$ to obtain the following matrix:

$$\mathrm{LL}^{\mathrm{gen}}(K) = \begin{bmatrix} \mathrm{LL}(K^{H,H}) & \mathrm{LL}(K^{H,L}) \\ \mathrm{LL}(K^{L,H}) & \mathrm{LL}(K^{L,L}) \end{bmatrix}. \tag{9}$$

[7] In the case of unordered variables, some further dichotomizations would also be needed, including the one where $L$-type and $H$-type individuals are put in the same class, while $M$-type individuals are kept in the other class.

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

Equation (9) defines the generalized Liu–Lu measure when the contingency table is a 3-by-3 table. We refer to it as the *Liu–Lu matrix characterizing marital preferences.*

Finally, we define the Liu–Lu matrix for the general case, where $K$ is an $n \times m$ table. Its $(i, j)$th element is

$$\mathrm{LL}_{i,j}^{\mathrm{gen}} = \mathrm{LL}(V_i K W_j^T), \tag{10}$$

where $V_k$ is the $2 \times n$ matrix $V_k = \begin{bmatrix} \overbrace{\begin{matrix} 1 & \dots & 1 \\ 0 & \dots & 0 \end{matrix}}^{k} & \overbrace{\begin{matrix} 0 & \dots & 0 \\ 1 & \dots & 1 \end{matrix}}^{n-k} \end{bmatrix}$ and $W_p^T$ is the $m \times 2$ matrix $W_p = \begin{bmatrix} \overbrace{\begin{matrix} 1 & \dots & 1 \\ 0 & \dots & 0 \end{matrix}}^{p} & \overbrace{\begin{matrix} 0 & \dots & 0 \\ 1 & \dots & 1 \end{matrix}}^{m-p} \end{bmatrix}$ with $k \in \{1, \dots, n-1\}$, and $p \in \{1, \dots, m-1\}$.

So, this is how we generalize equation (9). Such a generalization is relevant when the assorted trait is a polytomous variable with more than three possible values. For instance, when education level is measured on a refined scale with four or five categories.

The next section shows how the Liu–Lu measure and the Liu–Lu matrix can be used for constructing counterfactuals essential for decompositions.

## 4. Methodology

This section describes the methodology that will be used for the empirical counterfactual decomposition analysis in section 5. First, we present the decomposition scheme. Then we introduce the new transformation method. Finally, we visit some of its properties.

### 4.1. Decomposition scheme

For the empirical decompositions, we apply the decomposition scheme promoted by Biewen (2012).[8] It works as follows with two factors $P$ and $A$ observed at time $t_0$ and $t_1$; and a function $f(A, P)$ mapping the space spanned by the two factors into R:

$$\begin{aligned} f(A_{t_1}, P_{t_1}) - f(A_{t_0}, P_{t_0}) = {} & \overbrace{[f(A_{t_0}, P_{t_1}) - f(A_{t_0}, P_{t_0})]}^{\text{due to } \Delta P_{t_1}} + \overbrace{[f(A_{t_1}, P_{t_0}) - f(A_{t_0}, P_{t_0})]}^{\text{due to } \Delta A_{t_1}} + \\ & + \underbrace{[f(A_{t_1}, P_{t_1}) - f(A_{t_1}, P_{t_0}) - f(A_{t_0}, P_{t_1}) + f(A_{t_0}, P_{t_0})]}_{\text{interaction term capturing the joint effect of } \Delta P_{t_1} \text{ and } \Delta A_{t_1}} . \end{aligned} \tag{11}$$

[8] Unlike the popular DiNardo et al. (1996) decomposition scheme, the Biewen decomposition scheme is designed to isolate not only the ceteris paribus effects of certain factors, but also their interaction(s). As it is shown by Biewen (2012), the DiNardo et al. (1996) scheme attributes the interaction term(s) to either of the factors arbitrarily.

Provided that we have observations from multiple years, e.g., $t_0 < t_1 < t_2 < t_3$, the decomposition scheme of equation (11) gives

$$\begin{aligned}
&f(A_{t_3}, P_{t_3}) - f(A_{t_0}, P_{t_0}) = f(A_{t_3}, P_{t_3}) - f(A_{t_2}, P_{t_2}) + f(A_{t_2}, P_{t_2}) - f(A_{t_1}, P_{t_1})\\
&\quad + f(A_{t_1}, P_{t_1}) - f(A_{t_0}, P_{t_0}) = \overbrace{[f(A_{t_0}, P_{t_1}) - f(A_{t_0}, P_{t_0})]}^{\text{due to } \Delta P_{t_1}} + \overbrace{[f(A_{t_1}, P_{t_0}) - f(A_{t_0}, P_{t_0})]}^{\text{due to } \Delta A_{t_1}}\\
&\quad + \overbrace{[f(A_{t_1}, P_{t_2}) - f(A_{t_1}, P_{t_1})]}^{\text{due to } \Delta P_{t_2}} + \overbrace{[f(A_{t_2}, P_{t_1}) - f(A_{t_1}, P_{t_1})]}^{\text{due to } \Delta A_{t_2}}\\
&+ \overbrace{[f(A_{t_2}, P_{t_3}) - f(A_{t_2}, P_{t_2})]}^{\text{due to } \Delta P_{t_3}} + \overbrace{[f(A_{t_3}, P_{t_2}) - f(A_{t_2}, P_{t_2})]}^{\text{due to } \Delta A_{t_3}}\\
&+ \text{interaction terms capturing the joint effects of}\Delta P_{t_1}\Delta A_{t_1},\ \Delta P_{t_2}\Delta A_{t_2},\ \text{and } \Delta P_{t_3}\Delta A_{t_3}\\
&= \overbrace{[f(A_{t_0}, P_{t_1}) - f(A_{t_0}, P_{t_0})] + [f(A_{t_1}, P_{t_2}) - f(A_{t_1}, P_{t_1})] + [f(A_{t_2}, P_{t_3}) - f(A_{t_2}, P_{t_2})]}^{\text{due to change in } P\text{from } P_{t_0} \text{to } P_{t_3}}\\
&+ \overbrace{[f(A_{t_1}, P_{t_0}) - f(A_{t_0}, P_{t_0})] + [f(A_{t_2}, P_{t_1}) - f(A_{t_1}, P_{t_1})] + [f(A_{t_3}, P_{t_2}) - f(A_{t_2}, P_{t_2})]}^{\text{due to change in } A\text{from } A_{t_0} \text{to } A_{t_3}}\\
&+ \text{interaction terms capturing the joint effects of } \Delta P_{t_1}\Delta A_{t_1},\ \Delta P_{t_2}\Delta A_{t_2},\ \text{and } \Delta P_{t_3}\Delta A_{t_3}.
\end{aligned} \tag{12}$$

In our specific empirical application, the function $f(A, P)$ tells us the *share of homogamous couples* in a society, whereas $A$ stands for *availability* (i.e., trait distribution of men and women), and $P$ captures *preferences*.

While $f(A_t, P_t)$ (the share of homogamous couples at time $t$) can be observed, $f(A_t, P_{s \neq t})$ cannot. So, to perform decompositions with equation (11), it is essential to construct contingency tables under the counterfactual scenarios such that structural availability is measured in a given year, while preferences are from another year.

## 4.2. The new method for constructing counterfactuals

In this section, we introduce a new method for constructing contingency tables under counterfactual scenarios. In consonance with the literature, we call the observed contingency table in the period from which the marital preferences are taken as the *seed table*. While the trait distributions in the other period are referred to as the *target marginals*.

Subsection 4.2.1 presents the new method in the simplest set up where the contingency table is a 2-by-2 table. Then subsection 4.2.2 introduces the new method for the general case. Reading subsection 4.2.1 without subsection 4.2.2 is sufficient to learn the logic of the new method. What subsection 4.2.2 adds is the mathematical formula applicable when the assorted trait is a polytomous variable as it is in our empirical decomposition problem.

### 4.2.1. The new method for constructing counterfactuals with dichotomous trait variable

The *problem* can be *formalized in the 2-by-2 case* as follows. Suppose that the seed table is given by $K$ in equation (1), while the target marginals are defined by the $1 \times 2$

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

vector $C$, and the $2 \times 1$ vector $R$. Our goal is to determine the elements of the transformed contingency table

$$K^* = \begin{bmatrix} N^*_{L,L} & N^*_{L,H} \\ N^*_{H,L} & N^*_{H,H} \end{bmatrix},$$

under the restrictions $R = \begin{bmatrix} N^*_{L,L} + N^*_{L,H} \\ N^*_{H,L} + N^*_{H,H} \end{bmatrix}$, $C = \begin{bmatrix} N^*_{L,L} + N^*_{H,L} & N^*_{L,H} + N^*_{H,H} \end{bmatrix}$, and

$$\mathrm{LL}(K) = \mathrm{LL}(K^*). \tag{13}$$

To solve the problem, we assume $N_{H,H} \geq Q$. It allows us to use the simplified version of the Liu–Lu measure in equation (3). By substituting it into equation (13), we get

$$\frac{N_{H,H} - \mathrm{int}(N_{H,\cdot} N_{\cdot,H}/N)}{\min(N_{H,\cdot}, N_{\cdot,H}) - \mathrm{int}(N_{H,\cdot} N_{\cdot,H}/N)} = \frac{N^*_{H,H} - \mathrm{int}(N^*_{H,\cdot} N^*_{\cdot,H}/N^*)}{\min(N^*_{H,\cdot}, N^*_{\cdot,H}) - \mathrm{int}(N^*_{H,\cdot} N^*_{\cdot,H}/N^*)}. \tag{14}$$

We obtain the *solution* by rearranging equation (14):

$$\begin{aligned} N^*_{H,H} = {} & \frac{[N_{H,H} - \mathrm{int}(N_{H,\cdot} N_{\cdot,H}/N)][\min(N^*_{H,\cdot}, N^*_{\cdot,H}) - \mathrm{int}(N^*_{H,\cdot} N^*_{\cdot,H}/N^*)]}{\min(N_{H,\cdot}, N_{\cdot,H}) - \mathrm{int}(N_{H,\cdot} N_{\cdot,H}/N)} \\ & + \mathrm{int}\left(\frac{N^*_{H,\cdot} N^*_{\cdot,H}}{N^*}\right) \end{aligned}. \tag{15}$$

The right-hand-side of equation (15) expresses $N^*_{H,H}$ as a function of known variables. Trivially, once $N^*_{H,H}$ and the target marginals are known, all the other three elements of the $K^*$ matrix are known as well.

#### 4.2.2. The new method for constructing counterfactuals with polytomous trait variable

Let us discuss the *problem in the 3-by-3 case*, i.e., where the seed table $K$ is given by equation (4), while the row-sum of the transformed table $K^*$ is the $3 \times 1$ vector $R$, and the column-sum of $K^*$ is the $1 \times 3$ vector $C$. While the problem itself is a trivial extension of the problem in the 2-by-2 case, its solution is not. We obtain the solution in the 3-by-3 case by solving four 2-by-2 problems:

$$\begin{aligned} N^*_{L,L} &= (K^{H,H})^*_{1,1} \text{ with target marginals } R^H = \begin{bmatrix} R_1 & R_2 + R_3 \end{bmatrix}^T \text{ and } C^H = \begin{bmatrix} C_1 & C_2 + C_3 \end{bmatrix} \\ N^*_{H,H} &= (K^{L,L})^*_{2,2} \text{ with target marginals } R^L = \begin{bmatrix} R_1 + R_2 & R_3 \end{bmatrix}^T \text{ and } C^L = \begin{bmatrix} C_1 + C_2 & C_3 \end{bmatrix} \\ N^*_{L,H} &= (K^{H,L})^*_{1,2} \text{ with target marginals } R^H = \begin{bmatrix} R_1 & R_2 + R_3 \end{bmatrix}^T \text{ and } C^L = \begin{bmatrix} C_1 + C_2 & C_3 \end{bmatrix} \\ N^*_{H,L} &= (K^{L,H})^*_{2,1} \text{ with target marginals } R^L = \begin{bmatrix} R_1 + R_2 & R_3 \end{bmatrix}^T \text{ and } C^H = \begin{bmatrix} C_1 & C_2 + C_3 \end{bmatrix}. \end{aligned}$$

The other five elements of the 3 × 3 K* matrix can be expressed with the target marginals.

Finally, we discuss the general case, where the seed matrix $K$ is of size $n \times m$. The *problem in the general case* can be formalized as follows. Our goal is to determine the transformed contingency table $K^*$ of size $n \times m$ under the restrictions given by

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

the target marginals $R = K^* e_m^T$, and $C = e_n K^*$, where $e_m$ and $e_n$ are all-ones row vectors of size $m$ and $n$, respectively. The additional restriction is $\mathrm{LL}^{\mathrm{gen}}(K) = \mathrm{LL}^{\mathrm{gen}}(K^*)$.

By using equation (10), we can rewrite the problem as follows. We look for $K^*$, where $V_i R = V_i K^* e_m^T$, and $C W_j^T = e_n K^* W_j^T$; and $\mathrm{LL}(V_i K W_j^T) = \mathrm{LL}(V_i K^* W_j^T)$ for all $i \in \{1, \ldots, n-1\}$ and $j \in \{1, \ldots, m-1\}$. The matrices $V_k$ and $W_p$ are the same as defined under equation (10). For each $i$, $j$ pairs, these equations define a problem that is of the 2-by-2 form. Each problem can be solved separately by applying equation (15). The solutions determine $(m-1) \times (n-1)$ entries of the $K^*$ table. The remaining $m + n - 1$ elements of the $K^*$ matrix can be determined with the help of the target marginals.

## 4.3 Analysis of the new method

Some properties of the new transformation method are worth visiting. They will be compared, in this section, with the properties of the IPF algorithm, and the CS method. In addition to analyzing and comparing some analytical properties of the three methods, we also illustrate some differences and similarities with the help of a numerical example.

### 4.3.1. Some analytical properties of the new method

First, we highlight the difference between the 2-by-2 problems solved by the new method and its two alternatives. Each of the three methods transforms a seed table $K$ defined by equation (1) into another contingency table $K^*$ so as to make the marginals of $K^*$ equal to their preset targets $R$ and $C$. However, the additional restrictions are different. As we have seen, in the case of the *new method*, it is given by equation (13).

The additional restriction with the *IPF algorithm* is

$$\mathrm{CPR}(K) = \mathrm{CPR}(K^*), \tag{16}$$

where CPR is the cross-product ratio[9] defined as $\mathrm{CPR}(K) = (N_{L,L} N_{H,H})/(N_{L,H} N_{H,L})$.

In the case of the *CS method*, the set of additional restrictions is on the marital surpluses in the CS model [see Choo and Siow (2006)]:

$$\begin{aligned}
\frac{N_{L,L}}{\sqrt{N_{L,0} N_{0,L}}} &= \frac{N^*_{L,L}}{\sqrt{N^*_{L,0} N^*_{0,L}}} \\
\frac{N_{L,H}}{\sqrt{N_{L,0} N_{0,H}}} &= \frac{N^*_{L,H}}{\sqrt{N^*_{L,0} N^*_{0,H}}} \\
\frac{N_{H,L}}{\sqrt{N_{H,0} N_{0,L}}} &= \frac{N^*_{H,L}}{\sqrt{N^*_{H,0} N^*_{0,L}}} \\
\frac{N_{H,H}}{\sqrt{N_{H,0} N_{0,H}}} &= \frac{N^*_{H,H}}{\sqrt{N^*_{H,0} N^*_{0,H}}},
\end{aligned} \tag{17}$$

[9] The cross-product ratio is identical to the odds ratio. It has the following interpretation: the odds of being matched with an $H$-type women is CPR-times higher for an $H$-type men, than for an $L$-type men.

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

where $N_{i,0}(N_{0,j})$ denote the observed number of $i$-type single men ($j$-type single women), while $N^*_{i,0}(N^*_{0,j})$ denote the number of $i$-type single men ($j$-type single women) under the counterfactual.

Second, the new method provides us a *closed-form* solution for the transformed contingency table representing the counterfactual.

Third, a solution can be obtained even if the seed table contains *zero entries*. The above two properties make the new method particularly attractive relative to the IPF algorithm from a computational perspective since the latter offers a numerical solution with its iterative procedure; and generalizing the IPF algorithm for seed tables with zero cells is still considered a problem solved only partially.[10]

Next, we discuss *three properties of the counterfactual* constructed with the new method. First, the solution offered by the new method is *unique*. For the 2-by-2 case, this is apparent from equation (15) derived in subsection 4.2.1. Although we do not provide a formal step-by-step proof of the uniqueness in the general case with a contingency table of size $n \times m$, we give the intuition of the proof. The number of additional cells to be determined relative to the 2-by-2 case is equal to the number of additional independent linear restrictions.[11]

Second, the *new transformation method commutes with* another operation by construction. This other operation is the *merging of the categories* of the assorted trait. For the point of the counterfactual, it means the following. Suppose that originally the transformed contingency table was constructed when the educational categories were low ($L$), medium ($M$), and high ($H$). If $L$ and $M$ are merged in the seed table and target marginals then the new transformed contingency table is the same as the original transformed contingency table with merged corresponding rows and columns.

The above properties of the new method do not guarantee that the various elements making up the conditions of the constructed counterfactuals can jointly happen. Apart from a few exceptions, we do not know what set of conditions is possible to happen and what is not, i.e., what counterfactuals are possible and what are impossible. Certainly, having a negative frequency of couples under a counterfactual clearly indicates that the counterfactual is impossible.[12]

Constructing *impossible counterfactuals* with the new method is not only a theoretical possibility. For instance, if one applies the new method for census data from Portugal with the seed table taken from 2011 and the target marginal distributions from the year 1981, then the transformed contingency table contains a

[10] Fienberg (1970) provides an overview of those special cases with a zero in the seed table that can be solved with existing generalizations of the IPF algorithm. Sinkhorn (1964) presents some numerical examples to illustrate the drawback of replacing zeros with some small quantities.

[11] For instance, in the 3-by-3 case, one needs to determine the value of $5( = 3 \times 3 - 2 \times 2)$ more cells of the transformed table relative to the 2-by-2 case. We have $(3 - 1) \times (3 - 1) - (2 - 1) \times (2 - 1) = 3$ more elements in the Liu–Lu matrix relative to the 2-by-2 case with scalar valued Liu–Lu measure. These three elements in the Liu–Lu matrix yield three additional independent linear restrictions and we have two additional independent linear restrictions from the target marginals.

[12] To note, it is independent of the choice of the transformation method which set of conditions is possible to happen. However, checking the possibility of a counterfactual may not be independent of the transformation method applied. Checking the non-negativity criterion of the frequencies offers a joint test of the transformation method and the possibility of the set of conditions.

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

negative element.[13] (By contrast, the transformed matrix generated by the CS method is always non-negative if the seed table is non-negative. This follows from equation (17).)

In the empirical part of this paper, we decompose changes in marital patterns in the US over a similarly long period as in the example with Portugal. There, we follow an approach different from the one in the example.[14] In particular, we use observations from four years ($t_0 < t_1 < t_2 < t_3$) instead of two. Accordingly, we apply the decomposition scheme represented by equation (12).

Our multi-period decomposition involves constructing $f(A_{t_1}, P_{t_0}), f(A_{t_0}, P_{t_1})$, $f(A_{t_2}, P_{t_1}), f(A_{t_1}, P_{t_2}), f(A_{t_3}, P_{t_2}), f(A_{t_2}, P_{t_3})$, whereas it does not require to construct counterfactual with factors measured in relatively distant years. This is a fortunate case, because over a relatively long horizon any factor may change so that the respective counterfactual is impossible. Using the multi-period decomposition scheme is a potential solution for avoiding such counterfactuals.[15] It works well in practice: at least we obtain no negative entries in any of the transformed tables for Portugal using data from 1981, 1991, 2001, and 2011. This gives us hope that the counterfactual tables for the US generated for the multi-period decomposition scheme with data from 1980, 1990, 2000, and 2010 also represent realistic scenarios.

### 4.3.2. A numerical example with the new method

Let us illustrate with a numerical example the difference between the new method and some other statistical methods. Suppose that the seed table is $K = \begin{bmatrix} 45 & 15 \\ 5 & 35 \end{bmatrix}$. While the target marginals are $R = \begin{bmatrix} 105 \\ 45 \end{bmatrix}$, and $C = \begin{bmatrix} 100 & 50 \end{bmatrix}$.

We obtain $K^*_{\text{new}} = \begin{bmatrix} 92.5 & 12.5 \\ 7.5 & 37.5 \end{bmatrix}$ with the new transformation method.[16] By contrast, we would get $K^*_{\text{conv}} = \begin{bmatrix} 90 & 15 \\ 10 & 35 \end{bmatrix}$ as a transformed table (with entries rounded to the nearest integers) by many conventional approaches including the IPF algorithm. The cross-product ratio, the trait correlation, and the regression slope coefficient are either exactly invariant or approximately invariant to transforming $K$ to $K^*_{\text{conv}}$.[17]

[13] See sheet PT_A1981_P2011_Not_meaningful of the Excel file Naszodi_Mendonca 2020_NewMethod_US_PT_1980_2010_age3034.xlsx that can be downloaded from http://dx.doi.org/10.17632/x2ry7bcm95.1.

[14] The way Eika *et al.* (2019) construct their counterfactual is similar to the one in the example with Portugal. They measure marital sorting in the year 1962 and the educational distributions in 2013.

[15] An example for an impossible counterfactual is the one under which Europeans already know how to make popcorn without having discovered the North American continent. Researchers might need to construct such a counterfactual if they aim at disentangling the effects of inventions and discoveries on growth in Europe over the history of mankind while being restricted to use observations only from the endpoints of the time period.

[16] We check the invariance of the Liu–Lu measure by using equation (3): $\text{LL}(K) = \frac{35-\text{int}(40\times 50/100)}{\min(40,\,50)-\text{int}(40\times 50/100)} = \text{LL}(K^*_{\text{new}}) = \frac{37.5-\text{int}(45\times 50/150)}{\min(45,\,50)-\text{int}(45\times 50/150)} = \frac{3}{4}$.

[17] The only reason for having approximate invariance for some statistics instead of exact invariance is the rounding of the transformed table entries. The cross-product ratio [defined below equation (16)] is exactly invariant to transforming $K$ to $K^*_{\text{conv}}$ since $\text{CPR}(K) = \text{CPR}(K^*_{\text{conv}}) = 21$. Whereas it is not invariant to transforming $K$ to $K^*_{\text{new}}$ since $\text{CPR}(K) \neq \text{CPR}(K^*_{\text{new}}) = 37$. Similarly, the Liu–Lu measure is not invariant to transforming $K$ to $K^*_{\text{conv}}$ since $\frac{3}{4} = \text{LL}(K) \neq \text{LL}(K^*_{\text{conv}}) = \frac{35-\text{int}(45\times 50/150)}{\min(45,\,50)-\text{int}(45\times 50/150)} = \frac{2}{3}$.

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

This numeric example illustrates two points. First, the *dissimilarity across the counterfactuals obtained with various conventional methods can be negligible*. This point has relevance for robustness checks of decomposition results: *one may not represent model uncertainty sufficiently well if the applied methods are all from the conventional family even if those are derived from different theoretical models.*

Second, our numerical example illustrates well the point that *the Liu–Lu measure is not from the family of conventional measures populated by the cross-product ratio, the trait correlation, and the regression slope coefficient.* And also, a counterfactual constructed with the new method can be different from the counterfactuals constructed with some other methods. Whether the difference between the new method and its alternatives is empirically relevant will be investigated in the next section.

## 5. Empirical analysis

### 5.1. Data

For the empirical analysis, decennial census data of the United States are used from four census waves between 1980 and 2010. The census wave-specific contingency tables are presented in Table 2. Details on the construction of the data used are presented in Appendix B in the online appendix (see: https://doi.org/10.1017/dem.2021.1).

Our sample covers those heterosexual young couples where the men are aged 30–34 years.[18] We do not distinguish between officially married couples and couples in a consensual union.[19] Hereafter, by "marriage" we mean both types of union.

In addition to couples, our data also cover single individuals from the same age group. Data on single people are used by the CS method, but not by the IPF algorithm and the new method.

Our variable on the highest level of education can take five values, i.e., "less than primary completed", "primary completed", "secondary completed", "some college", and "university completed". For the main analysis, we merge the lowest two categories and also the next two categories. So, we work with the following three categories: "less than high school", "high school completed", and "university completed".[20]

### 5.2. Stylized facts

Three stylized facts are documented by Tables 3 and 4. First, the studied period was characterized by an *educational expansion*. Second, the *educational gender gap has closed and then it has reversed*: in 2010, women in relationships with young men

[18] Another sample covers those heterosexual couples where the women are aged 30–34 years. We use this sample in one of our robustness checks. The choice of the age group guarantees to observe individuals almost only with finalized educational attainment (or with a BA if still in study assigning them already to the highest education category in this paper). It also ensures not to have overlapping observations: no individual is observed twice in our decennial census data. In addition, it helps to keep the age-related variation in marital preferences low in each generation under study.

[19] The marital status variable in the census data has one common category: "married/in union". By using another dataset with distinct categories for married and cohabiting couples, Eika et al. (2019) find that classifying the cohabiting couples as married does not materially change the estimates of the marital sorting parameters.

[20] Choo and Siow (2006) use the same three categories as we do. While the education variable in Eika et al. (2019) can take four categories: "high school dropouts", "high school graduates", "some college", and "college graduates".

**Table 2.** The contingency tables for the US from four census waves

(a) Couples 1980

| | Wife/female partner | | | |
|---|---|---|---|---|
| | *L* | *M* | *H* | Total |
| Husb./male p. | | | | |
| *L* | 468,151 | 429,125 | 16,417 | 913,693 |
| *M* | 436,459 | 2,786,413 | 296,450 | 3,519,322 |
| *H* | 33,235 | 901,832 | 921,656 | 1,856,723 |
| Total | 937,845 | 4,117,370 | 1,234,523 | 6,289,738 |

(b) Singles 1980

| | *L* | *M* | *H* | Total |
|---|---|---|---|---|
| Female | 470,715 | 1,398,382 | 506,246 | 2,375,343 |
| Male | 400,071 | 1,223,829 | 664,115 | 2,288,015 |

(c) Couples 1990

| | Wife/female partner | | | |
|---|---|---|---|---|
| | *L* | *M* | *H* | Total |
| Husb./male p. | | | | |
| *L* | 367,228 | 430,219 | 19,076 | 816,523 |
| *M* | 373,347 | 3,726,106 | 532,355 | 4,631,808 |
| *H* | 19,244 | 750,676 | 1,053,370 | 1,823,290 |
| Total | 759,819 | 4,907,001 | 1,604,801 | 7,271,621 |

(d) Singles 1990

| | *L* | *M* | *H* | Total |
|---|---|---|---|---|
| Female | 512,808 | 2,390,532 | 807,477 | 3,710,817 |
| Male | 621,937 | 2,558,696 | 874,855 | 4,055,488 |

(e) Couples 2000

| | Wife/female partner | | | |
|---|---|---|---|---|
| | *L* | *M* | *H* | Total |
| Husb./male p. | | | | |
| *L* | 358,332 | 347,489 | 24,252 | 730,073 |
| *M* | 286,655 | 2,917,020 | 613,873 | 3,817,548 |
| *H* | 16,308 | 547,492 | 1,261,129 | 1,824,929 |
| Total | 661,295 | 3,812,001 | 1,899,254 | 6,372,550 |

(*Continued*)

**Table 2.** (*Continued.*)

(f) Singles 2000

| | *L* | *M* | *H* | Total |
|---|---|---|---|---|
| Female | 427,016 | 2,291,427 | 945,353 | 3,663,796 |
| Male | 574,146 | 2,540,118 | 993,819 | 4,108,083 |

(g) Couples 2010

| | Wife/female partner | | | |
|---|---|---|---|---|
| | *L* | *M* | *H* | Total |
| Husb./male p. | | | | |
| *L* | 314,032 | 295,373 | 28,823 | 638,228 |
| *M* | 191,865 | 2,310,584 | 733,754 | 3,236,203 |
| *H* | 11,594 | 425,542 | 1,411,180 | 1,848,316 |
| Total | 517,491 | 3,031,499 | 2,173,757 | 5,722,747 |

(h) Singles 2010

| | *L* | *M* | *H* | Total |
|---|---|---|---|---|
| Female | 497,157 | 2,589,723 | 1,262,563 | 4,349,443 |
| Male | 728,746 | 3,035,923 | 1,132,050 | 4,896,719 |

*Source*: Data are from the international version of Integrated Public Use Microdata Series (IPUMS) from the Minnesota Population Center.

*Notes*: Our sample covers heterosexual couples where the men are aged 30–34 years, and single people from the same age group. The variable on the highest level of education can take three different values, where *L* stands for "low level of education" corresponding to not having completed high school; *M* denotes "medium level of education" corresponding to having a high school degree; and *H* stands for "high level of education" corresponding to holding a university diploma.

**Table 3.** Educational distribution of married/in union young men and that of their spouses in the US between 1980 and 2010 (in %)

| | Share of married/in union individuals who attained | | | | | |
|---|---|---|---|---|---|---|
| | No high school | | High school | | Diploma | |
| Year | Women | Men | Women | Men | Women | Men |
| 1980 | 15 | 15 | 65 | 56 | 20 | 30 |
| 1990 | 10 | 11 | 67 | 64 | 22 | 25 |
| 2000 | 10 | 11 | 60 | 60 | 30 | 29 |
| 2010 | 9 | 11 | 53 | 57 | 38 | 32 |

*Source*: Authors' calculations using data in Table 2.

**Table 4.** The proportion of educationally homogamous couples in the US between 1980 and 2010 (in %)

| | Share of homogamous couples who attained | | | |
|---|---|---|---|---|
| | No high school | High school | Diploma | All types |
| Year | (1) | (2) | (3) | (4) = (1) + (2) + (3) |
| 1980 | 7 | 44 | 15 | 66 |
| 1990 | 5 | 51 | 14 | 71 |
| 2000 | 6 | 46 | 20 | 71 |
| 2010 | 5 | 40 | 25 | 71 |

*Source*: Authors' calculations using data in Table 2.

were more educated than their spouses on average, although it was just the opposite 30 years earlier. Third, the *proportion of educationally homogamous couples has increased.*

The first two stylized facts clearly show that the structural availability of marriageable men and women with a given qualification level has changed. Whether these changes can fully explain the observed increase in the prevalence of homogamy is undoubtedly an important empirical question.[21] An equally relevant and exciting question is whether the answer to the previous question depends on the choice of the method applied for the analysis. This is addressed in the next section.

## *5.3. The empirical application of the new method*

Let us apply the methodology introduced in section 4 to the data described in subsection 5.1. It involves the following three steps. First, we construct the contingency tables under the counterfactuals. This is the step where we apply the new method introduced in subsection 4.2.2. Second, we calculate the share of educationally homogamous couples. For that, we simply take the sum of the diagonal elements in each of the observed contingency tables and divide it by the corresponding total number of couples. We obtain $f(A_{1980}, P_{1980})$, $f(A_{1990}, P_{1990})$, $f(A_{2000}, P_{2000})$, and $f(A_{2010}, P_{2010})$. Similarly, we calculate the share of educationally homogamous couples in each of the six counterfactual tables. We obtain $f(A_{1980}, P_{1990})$, $f(A_{1990}, P_{1980})$, $f(A_{1990}, P_{2000})$, $f(A_{2000}, P_{1990})$, $f(A_{2000}, P_{2010})$, $f(A_{2010}, P_{2000})$. Finally, we apply the decomposition scheme in equation (11) three times (or, equation (12)).

### *5.3.1. Results of the main analysis*

Figure 1 presents the outcome of the main decomposition. It reports the extent to which certain drivers contributed to the changes in the share of educationally homogamous couples in the US between 1980 and 2010. Figure 1b shows the results for the investigated three decades separately, while these outcomes are aggregated in Figure 1a.

[21] This question is studied for various countries and periods by several papers. See, for instance, Naszodi and Mendonca (2019b), Hu and Qian (2016) and references therein.

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

The results presented in Figure 1 cover not only the decompositions performed with the new method, but also those obtained with the CS solution, and the IPF algorithm.[22] Regarding some other methods, we cannot present decompositions with them. This is because the transformed contingency table has a negative cell when it is obtained with any of the statistical approaches, where the covariance, or the Pearson's correlation, or the regression coefficient is kept fixed.[23]

Figure 1a suggests that the components are robust to the choice of the method when the decomposition is applied to the period between 1980 and 2010.[24] Interestingly, it is not the case with all the decade-specific components (see Figure 1b). *For the period between 1980 and 1990, the sign of the decomposed effect* of varying marital preferences across the groups of young American adults from different cohorts *depends on whether the new method is applied or any of its alternatives.*[25] Whereas *for the period between 2000 and 2010, it is the magnitude of the same effect that is sensitive to the choice of the method* (see the dark bars of Figure 1b).

The analysis in this section had a limited scope in documenting differences between the empirical findings of the methods. The next subsection checks the robustness of some findings obtained with the new method. While in subsection 5.4, we use some survey evidence and the decade-specific components to support our claim that the new method is better than its two alternatives.

### 5.3.2. Robustness checks

In this subsection, we investigate whether our decompositions obtained with the new method are robust to some choices; specifically, the choice on the number of educational categories (3 in the main specification, 4 in its two alternatives), and the definition of young couples (men are aged 30–34 years vs. women are aged 30–34 years).

We present not only the decade-specific results (see Figure 2b), but also the aggregate results (see Figure 2a). However, we interpret only the decade-specific results since we use them for method selection. Lack of their robustness would undermine the concept of selection. Fortunately, this is not the case: Figure 2b shows that *the sign of the decade-specific components is sensitive neither to the definition of young couples, nor to the educational categories considered*. Moreover, the magnitude of the marital preference-component in the last decade is also robust (see the dark bars in Figure 2b belonging to the period 2000–2010).

## 5.4. A supplementary analysis

When different methods deliver contrasting results, it calls for checking how well these results fit other evidence. To study whether the CS solution, the IPF

[22] For the empirical implementation of the IPF algorithm, we use the "mipfp" package in R developed by Barthélemy and Suesse (2018) (see https://github.com/jojo-/mipfp). For the CS method, we solve the system of quadratic equations numerically.

[23] We repeat our point made in Footnote 12 that a negative frequency is consistent with anyone or both of the following cases: (i) the transformation method is not adequate, (ii) the elements of conditions of the counterfactual are not possible to happen jointly.

[24] All three methods suggest that we can attribute an *economically significant increase* (>2 percentage points) in the share of homogamous couples to the varying marital preferences between 1980 and 2010 (see the dark bars of Figure 1a).

[25] This sensitivity cannot be due to sample variation and sampling variation. First, the new method and the IPF are applied to the very same set of data. Second, we work with census data that are large and representative, by construction, to the population.

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

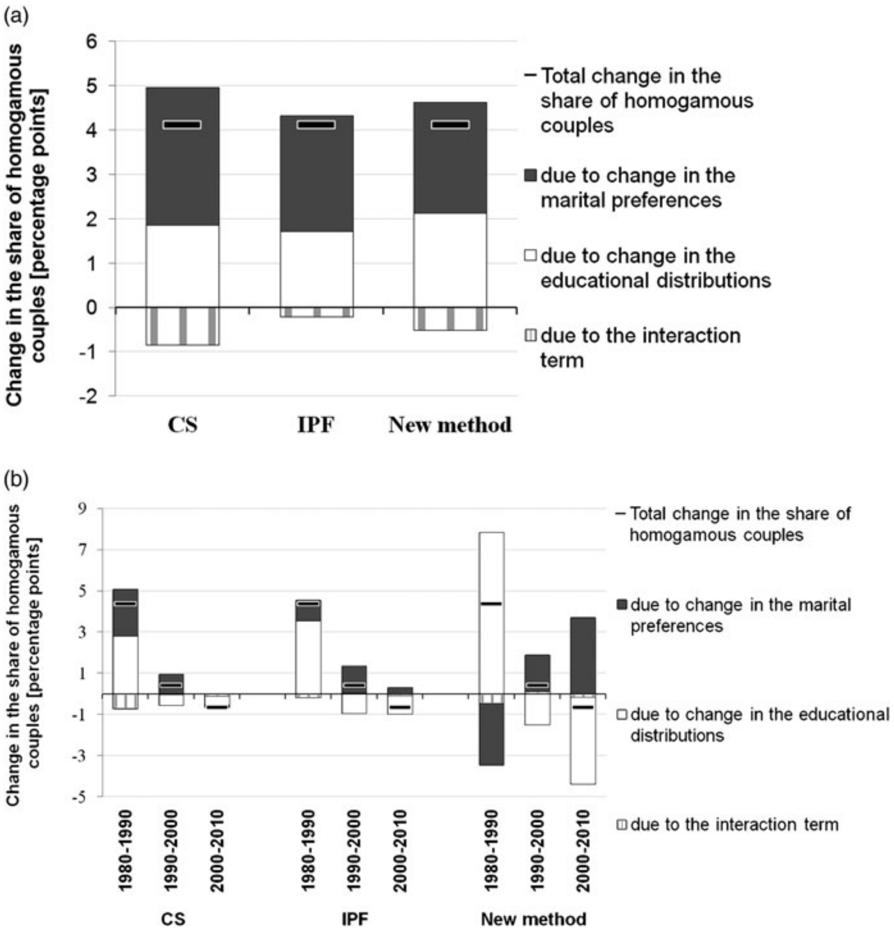

**Figure 1.** The long-horizon and short-horizon decompositions of changing prevalence of marital homogamy in the US with counterfactuals constructed by three different methods.
*Source*: Authors' calculations using data in Table 2.
*Notes*: The decompositions are performed with the decomposition scheme in equation (11) for each of the three decades (1980–1990, 1990–2000, and 2000–2010), and with the three methods (Choo–Siow method, IPF algorithm, and the new method). The results are presented in 1b. The corresponding aggregate components obtained with the decomposition scheme in equation (12) are presented in 1a.

algorithm or the new method provides us with a more realistic view, we use survey evidence from the Pew Research Center on the views of different generations about spousal education.

The Changing American Family survey conducted in 2010 informs us about two shares: the share of men who say it is very important for a woman to be well-educated in order to be a good wife/partner; and the share of women who say it is very important for a man to be well-educated in order to be a good husband/partner.[26]

[26] The survey data can be downloaded from https://www.pewsocialtrends.org/dataset/changing-american-family/.

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

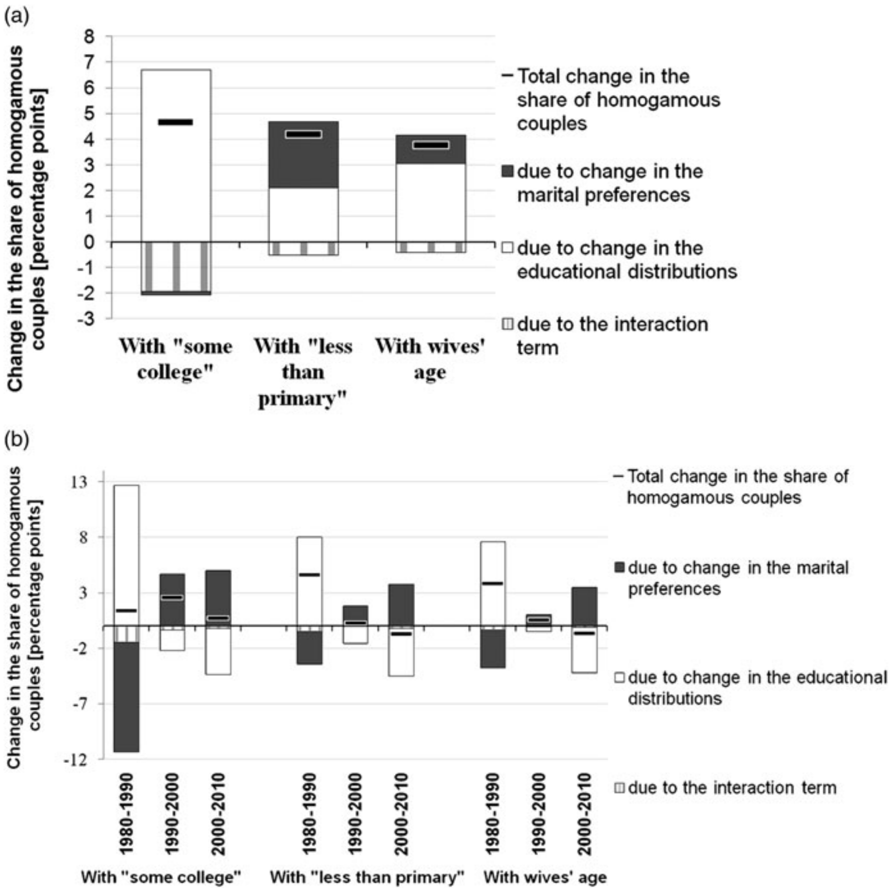

**Figure 2.** The long-horizon and short-horizon decompositions of changing prevalence of marital homogamy in the US with counterfactuals constructed by the new method.

*Source*: Authors' calculations using data in Table 2.

*Notes*: In the first case, we introduced the educational category "some college" by splitting the middle education category of the main analysis. In the second case, we introduced the education category "less than primary completed" by splitting the lowest category of the main analysis. In the third case, we defined young couples with the age of the wives/female partners by restricting it between 30 and 34 years. The decade-specific components in 2b are summed in order to obtain the results of the long-horizon decompositions reported in 2a. The decomposition scheme used is the same as in the main analysis.

This survey is not designed to identify directly those respondents who have preferences for homogamy. Therefore, we cannot learn from it what share of the male and female respondents prefer to mate with others who have the same education level as they do.

Obviously, the observed shares and the shares that cannot be identified from the survey capture the prevalence of different types of preferences in society. However, as it is pointed out by Hitsch *et al.* (2010) and others, "both types of preferences can lead to empirically observed assortative mating patterns [Becker (1973), Browning *et al.* (2008), and Kalmijn (1998)] and are thus indistinguishable using data on marriages only." On the one hand, the point they make questions whether the effects identified in the previous section are really the effects of changing preferences for

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press

partners with the same education level or, rather, they are the effects of changing preferences for well-educated partners. On the other hand, their point motivates us to use the variation in the observed shares across generations as a proxy for the effect of changing marital preferences irrespective of the exact type of these preferences.

When analyzing the survey data, our primary focus is on the responses of the early baby boomers (who were in the age group of 30–34 in the census year 1980) and the late boomers (who were in the same age group 10 years later in 1990) since the conflicting findings presented in subsection 5.3 were related to the revealed preferences of these two generations.

Figure 3 shows that in 2010 spousal education was viewed to be very important by 35% of the women respondents and 34% of the men respondents among the late boomers. These shares are lower than the corresponding shares in the generation of early boomers (around 39% and 45%, respectively). The detected variation across generations suggests that the changing composition of the studied age group with respect to its members' marital/mating preferences has a negative effect on the share of educationally homogamous young couples in the US between 1980 and 1990. The latter finding is in line with the result obtained by the new method, but not with those of the CS method and the IPF algorithm (see the dark bars in Figure 1b for the period 1980–1990).

Moreover, Figure 3 shows that there is a remarkable difference not only between the early boomers and the late boomers but also between the early generation-X (who were in the age group of 30–34 in 2000) and the late generation-X (who were in the same age group in 2010) regarding their preferences. In 2010, spousal education was viewed to be very important by about 41% of the female and 32% of the male respondents in the early generation-X. These shares are higher in the late generation-X since those are close to 46% and 45%, respectively. This survey evidence suggests that the share of educationally homogamous young couples would have increased massively between 2000 and 2010 if the educational composition of the early generation-X and that of the late generation-X had been the same. This result is again more in line with the finding obtained by the new method than with the finding obtained with its alternatives (see the dark bars of Figure 1b for the period 2000–2010).

All in all, the reviewed survey evidence yield support for the new method.[27]

## 6. Conclusion

Counterfactual analysis is in the focus of several research papers in the assortative mating literature. The typical question addressed is how the marriage patterns would have changed in the absence of change in the education levels of men and women.

In this paper, we proposed a new method that can provide an answer to the typical question. We compared it with some conventional methods for constructing counterfactuals, such as the well-known IPF algorithm and the method relying on the Choo–Siow model. Our empirical analysis performed on US census data illustrated the following point. Some answers to the new method are different from those provided with the IPF algorithm, and the CS method. It shows that the choice of method can be crucial.

Motivated by the detected lack of robustness to the method, we proposed an empirical method selection criterion. The supplementary analysis in this paper checked whether the results of the new method or that of its alternatives are closer to some survey evidence on Americans' marital/mating preferences. This analysis supports the new method.

[27] In our Appendix C, we discuss some potential limitations of this supplementary analysis (see: https://doi.org/10.1017/dem.2021.1).

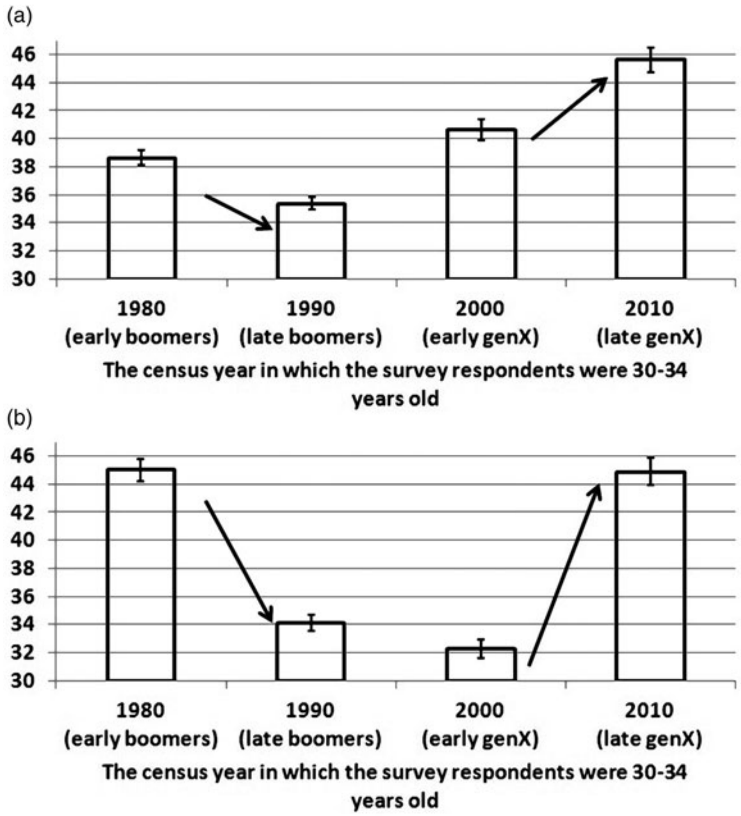

**Figure 3.** Generation-specific views from the opposite sex on the importance of spousal education in the US in 2010.
*Source*: Authors' calculations based on the answers to the survey questions number 23 and number 24 in the Changing American Family survey conducted by the Pew Research Center in 2010.
*Notes*: Answering the corresponding survey questions was refused by 3 women (aged 35, 55, and 87 in 2010) and 1 men (aged 57 in 2010), while the questions were answered by 289 women and 237 men in the age groups studied. Out of the 289 women 84 were in the age group 60–64 (representing early boomers), 92 were in the age group 50–54 (representing late boomers), 60 were in the age group 40–44 (representing early generation-X) and 53 were in the age group 30–34 (representing late generation-X) in 2010. Out of the 237 men respondents 56 were in the age group 60–64, 75 were in the age group 50–54, 61 were in the age group 40–44, 45 were in the age group 30–34 in the same year. The 95% symmetric confidence intervals are obtained with the approximation proposed by Agresti and Coull (1998).

It remains partly for future research to test systematically the assumptions behind the new method and to apply the test to other methods as well. It is also on our research agenda to recover a micro foundation of the Liu–Lu measure as an aggregate matching function. We have already started to address some points of this research agenda. In a follow-up paper, we investigate whether the empirical findings obtained with the new method are sensitive to the assumption which rules out the possibility of remaining single.[28]

[28] In Naszodi and Mendonca (2019a), we use dating data in addition to aggregate census data to identify the effect of changing preferences on marriage matching patterns and to estimate the share of singles refrained from the marriage market.

**Supplementary material.** The online appendix of this article can be found at https://doi.org/10.1017/dem.2021.1. The new method (implemented in Excel, Visual Basic, and R) can be downloaded from http://dx.doi.org/10.17632/x2ry7bcm95.1.

**Acknowledgements.** The authors thank Pierre-André Chiappori, Alfred Galichon, and Erik Plug for the helpful discussions and gratefully acknowledge comments from Liliana Cuccu, Andrea Barta, Sven Langedijk, Péter Mihályi, Eszter Naszodi, Tamás K. Papp, Sylke Schnepf, András Simonovits, Kornél Steiger, Iván Szelényi, Márta Ujvári, Stefano Verzillo, and two anonymous referees.

**Author contribution.** Anna Naszodi formed the concept of the research, wrote the first draft of the manuscript, and revised the paper before re-submission. She reviewed the literature and positioned the paper relative to some related papers on the topic; collected data from IPUMS and the Pew Research Center; implemented the new method developed in the paper in excel and Visual Basic; performed the final empirical analysis. Francisco Mendonca made some first-round data analysis using census data. He identified thereby the discrepancy of performing the decomposition for Portugal by using only the endpoints of the sample and recognized that the solution of the new method is unique. He implemented the new method in R. Also, he contributed by proof reading the manuscript and formatting its tables. In addition, he contributed indirectly by working on the companion paper entitled "Changing educational homogamy: Shifting preferences or evolving educational distribution?".

**Disclaimer.** The views expressed in this paper are those of the authors and do not necessarily reflect the official views of the European Commission.

## References

Agresti, Alan and Brent A. Coull (1998) Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician* 52(2), 119–126.

Altham, Patricia (1970) The measurement of association of rows and columns for an r×s contingency table. *Journal of the Royal Statistical Society Series* B 32 (1), 63–73.

Altham, Patricia, and Joseph Ferrie (2007) Comparing contingency tables tools for analyzing data from two groups cross-classified by two characteristics. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 40, 3–16.

Barthélemy, Johan and Thomas Suesse (2018) Mipfp: an R package for multidimensional array fitting and simulating multivariate Bernoulli distributions. *Journal of Statistical Software, Code Snippets* 86 (2), 1–20.

Becker, Gary S. (1973) A theory of marriage: part I. *Journal of Political Economy* 81(4), 813–846.

Biewen, Martin (2012) Additive decompositions with interaction effects. Discussion Papers No. 6730. IZA.

Breen, Richard and Leire Salazar (2005) Has increased women's educational attainment led to greater earnings inequality in the UK? Working paper No. 216. Estudio.

Breen, Richard and Leire Salazar (2011) Educational assortative mating and earnings inequality in the United States. *American Journal of Sociology* 117(3), 808–843.

Browning, Martin and Pierre-André Chiappori, and Yoram Weiss (2008) The economics of the family. http://www.tau.ac.il/~weiss/fam_econ/.

Chen, Liang, Eugene Choo, Alfred Galichon and Simon Weber (2019) Matching function equilibria: Existence, uniqueness and estimation. 13 May 2019. https://ssrn.com/abstract=3387335.

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

Choo, Eugene and Aloysius Siow (2006) Who marries whom and why. *Journal of Political Economy* 114(1), 175–201.

Dagsvik, John (2000) Aggregation in matching markets. *International Economic Review* 41(1), 27–58.

DiNardo, John, Nicole Fortin and Thomas Lemieux (1996) Labor market institutions and the distribution of wages, 1973–1992: a semiparametric approach. *Econometrica* 64(5), 1001–1044.

Dupuy, Arnaud and Alfred Galichon (2014) Personality traits and the marriage market. *Journal of Political Economy* 122(6), 1271–1319.

Eika, Lasse and Magne Mogstad and Basit Zafar (2019) Educational assortative mating and household income inequality. *Journal of Political Economy* 127(6), 2795–2835.

Fernandez, Raquel (2001) Education, segregation and marital sorting: Theory and an application to UK data (Working Paper No. 8377). National Bureau of Economic Research. Retrieved from DOI 10.3386/w8377 .

Fernandez, Raquel, Nezih Guner and John Knowles (2005) Love and money: a theoretical and empirical analysis of household sorting and inequality. *The Quarterly Journal of Economics* 120(1), 273–344.

Fernandez, Raquel and Richard Rogerson (2001) Sorting and long-run inequality. *Quarterly Journal of Economics* 116, 1305–1341.

Fienberg, Stephen E. (1970) An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics* 41(3), 907–917.

Greenwood, Jeremy, Nezih Guner and Georgi Kocharkov and Cezar Santos (2014) Marry your like: assortative mating and income inequality. *American Economic Review* 104(5), 348–353.

Greenwood, Jeremy, Nezih Guner, Georgi Kocharkov and Cezar Santos (2015) Corrigendum to Marry Your Like: Assortative Mating and Income Inequality. Retrieved from http://www.jeremygreenwood.net/papers/ggksPandPcorrigendum.pdf.

Hitsch, Günter J., Ali Hortaçsu and Dan Ariely (2010) What makes You click? – Mate preferences in online dating. *Quantitative Marketing and Economics* 8(4), 393–427.

Hu, Anning and Zhenchao Qian (2016) Does higher education expansion promote educational homogamy? Evidence from married couples of the post-80s generation in Shanghai, China. *Social Science Research* 60, 148–162.

Kalmijn, Matthijs (1998) Intermarriage and homogamy: causes, patterns, trends. *Annual Review of Sociology* 24, 395–421.

Kremer, Michael (1997) How much does sorting increase inequality? *The Quarterly Journal of Economics* 112(1), 115–139.

Liu, Haoming and Jingfeng Lu (2006) Measuring the degree of assortative mating. *Economics Letters* 92(3), 317–322.

Menzel, Konrad (2015) Large matching markets as two-sided demand systems. *Econometrica* 83(3), 897–941.

Naszodi, Anna and Francisco Mendonca (2019a) Changing educational homogamy: Shifting preferences or evolving educational distribution? Unpublished manuscript, under review.

Naszodi, Anna and Francisco Mendonca (2019b) Like marries like. JRC Science for Policy Briefs Series JRC115102, March. European Commission, Joint Research Centre. URL: https://ec.europa.eu/jrc/sites/jrcsh/files/fairness_pb2019_assortative_mating_jrc115102.pdf.

Sinkhorn, Richard (1964) A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics* 35(2), 876–879.

Sinkhorn, Richard (1967) Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly* 74(4), 402–405.

Stephan, Frederick F. and W. Edwards Deming (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics* 11(4), 427–444.

**Cite this article:** Naszodi A, Mendonca F (2023). A new method for identifying the role of marital preferences at shaping marriage patterns. *Journal of Demographic Economics* **89**, 1–27. https://doi.org/10.1017/dem.2021.1

https://doi.org/10.1017/dem.2021.1 Published online by Cambridge University Press