Assortative Mating on Education: A Genetic Assessment*

Nicola Barban, *University of Oxford*

Elisabetta De Cao, *University of Oxford*

Sonia Oreffice, *University of Surrey and IZA*

Climent Quintana-Domeque, *University of Oxford and IZA*

November 28, 2016

Abstract

We combine an instrumental variables approach –exploiting *genetic variation* in polygenic scores and controlling for population stratification– with a stochastic linear bi-dimensional matching model –which allows us to *test the validity of the exclusion restriction*. Individuals match on an additive index of overall marital attractiveness which has two components: an observable component (education) and a non-observable component (anything else) to the econometrician. The model allows us to investigate whether *exogenous shocks* to the observable component satisfy the exclusion restriction, and if so, use them to identify the degree of assortative mating on the observable component. Using data from the Health Retirement Study, we find that polygenic scores satisfy the exclusion restriction and are relevant instruments, and cannot reject that the (estimated) degree of assortative mating on education is the same using OLS and IV. Our study illustrates how to combine quasi-experimental variation in spouses' characteristics with parsimonious matching models to investigate assortative mating.

JEL Classification Codes: D1, J1, J12.
Keywords: Matching Models, College, Exclusion Restriction, Instrumental Variables, Polygenic Scores, Schooling, HRS.

*See acknowledgments at the end of the paper. Any errors contained in the paper are our own.

1 Introduction

Assortative mating. Standard models of assortative mating tend to predict perfect assortativeness, while the data show imperfect assortativeness. Deviations from perfect assortative mating can be accounted for by introducing *randomness* into the matching process, measurement *error*, *simultaneity* bias or extending the relevant *dimensions* in which matching takes place, to include physical attributes or personality traits (Chiappori et al., 2012; Choo and Siow, 2006; Dupuy and Galichon, 2014; Galichon and Salanié, 2015; Siow, 2015). In practice, we may expect different sources of biases at the same time, so that whether assortative mating is under or over-estimated is, in the end, an empirical question.

Even if our interest is the statistical measurement of assortative mating and quasi-experimental variation in the attributes of the two sides of the market is available, a stylized matching model may give us leverage in the empirical investigation of assortative mating. Specifically, a parsimonious model can provide the necessary structure to investigate the validity of exclusion restrictions, which cannot be tested in standard instrumental variables settings without the additional structure provided by a model. Chiappori and Salanié (2016) emphasize that, if one is willing to provide answers to questions on the causes and consequences of educational homogamy, a theoretical framework is necessary given the *two-sided* nature of the marriage market.

This paper and its main findings. We present a parsimonious stochastic linear bi-dimensional matching model, where individuals match on an index of marital attractiveness, so that the matching is de facto one-dimensional as in Chiappori et al. (2012) and Hitsch et al. (2010), but attractiveness depends on two attributes: education and a non-education component, which is a linear combination of attributes other than education which can be correlated with education, for instance personality or physical attractiveness. Given the likely correlation between education and the non-education component, the chief issue for the econometrician is how to estimate assortative mating on education without observing the non-education component. If a valid instrumental variable were available, one could

investigate assortative mating on education by using the part of education uncorrelated with the non-education component.

We propose a potentially valid instrument for education, a polygenic score for education –i.e., a single quantitative measure of genetic predisposition based on genetic variants present in the entire genome (see Plomin et al., 2009)– but allow the score to be *correlated* with both the education and the non-education components, so that the exclusion restriction can be violated. Under our bi-dimensional linear model, the exclusion restriction will be violated when two conditions hold at the same time: (1) the polygenic score for education is correlated with the non-education component, and (2) the non-education component is relevant for overall marital attractiveness. Crucially, the exclusion restriction is testable under the (linear) structure provided by our parsimonious model.

We construct a polygenic score to predict educational attainment of married men and women using data from the Health and Retirement Study (HRS), building upon the recent findings from a large scale genome-wide association study (GWAS) of educational attainment (Okbay et al., 2016). Rather than focusing on a limited number of genetic variants, the polygenic scores (PSs) use the entire information in the DNA (or a large proportion of it) to construct a measure of genetic predisposition to higher educational attainment (Conley et al., 2015; Ward et al., 2014; Domingue et al., 2014; Plomin et al., 2009).

While genes (and polygenic scores) are considered to be randomly assigned at conception (Mendelian randomization), at least after accounting for population stratification, this is a necessary but *not* a sufficient condition to use them as valid instrumental variables. Polygenic scores for education may affect spousal education above and beyond their effects on own education, that is, polygenic scores for education may affect the own-non education component, and hence violate the exclusion restriction. In general, the exclusion restriction is *not* testable, however, this can be examined in our context after imposing the structure provided by our linear bi-dimensional stochastic matching model. To the best of our knowledge, this is the first study to test the validity of the exclusion restriction of polygenic scores

for education in a marital matching model.

Our empirical findings show that the validity of the exclusion restriction cannot be rejected for either the male or female polygenic scores of education. We then proceed to use the polygenic scores for education as instrumental variables for education. Specifically, we use *spousal* PSs as instrumental variables for *spousal* education, controlling for *own* PS, to reassess assortative mating in the marriage market. Both polygenic scores of education are relevant instruments. The ratios of the estimated OLS-IV coefficients on the husband's year of education and the husband's college degree indicator are 0.98 and 0.90, respectively. The ratios of the estimated OLS-IV coefficients on the wife's year of education and the wife's college degree indicator are both 0.70. Interestingly, we cannot reject the equality of OLS and IV returns to own education in the marriage market in terms of spousal's education. This finding suggests that previous studies based on OLS analyses are indeed informative about assortative mating on education. Incidentally, a similar conclusion has been reached by labor economists regarding the estimation of the returns to education in the labor market (Angrist and Krueger, 1991).[1]

Taken at face value our results are consistent with two potential explanations: first, the non-education component (e.g., physical attractiveness) is irrelevant for marital attractiveness, at least among the sample under analysis, that is, individuals who are on average 70 years old, who got married on average 40 years ago and have been married to each other ever since, and who are still alive (healthier); second, the non-education component (e.g., physical attractiveness) is uncorrelated with education, at least among our highly selected group of couples. Whatever the reason is, the bottom line is that for a group of individuals for whom education is (perhaps) the most important attribute on the marriage market (older generations), assortative mating on education can be assessed with standard regression techniques.

Traditionally, the purpose of marriage was division of labor and child rearing rather than

[1]If anything, returns to education in the labor market seem to be underestimated when using OLS rather than IV (Card, 2001).

an hedonic marriage and affinity on personality (Becker, 1973; Lundberg, 2012). While it seems plausible to assume that in traditional marriages education was the main attribute driving marital surplus, the sexual revolution transformed the marriage market and its institution toward hedonic marriages (Stevenson and Wolfers, 2007). Thus, we would expect to find different returns in the marriage market using OLS and genetic-IV on a sample of recently first-married couples. In other words, our empirical findings are only *internally* valid, but the parsimonious approach combining genes with a structural model described in this paper is *externally* valid, and hopefully will stimulate future research.

Our paper shows how new available data (such as genetic data) can be used in combination of stylized matching models to make progress on the measurement of assortative mating. The actual quantification of assortative mating on education in the marriage market is an important step forward in economic analysis and public policy, if only because assortative mating may have direct implications for the transmission of socioeconomic status and inequality across generations (Currie, 2011; Fernandez and Rogerson, 2001).

Contributions and related literature. Our approach provides a novel identification strategy to investigate the degree of assortative mating in the marriage market, while complementing recent research on genetic assortative mating. Assortative mating has been studied in economics since the seminal work by Becker (1973). In particular, many social scientists have documented a strong and increasing educational homogamy (e.g., Bruze, 2011; Chiappori et al., 2009; Greenwood et al., 2014; Schwartz and Mare, 2005). The very recent work by Larsen et al. (2015) claims that using the variation in male educational attainment induced by the WWII G.I. Bill may provide the most transparent identification strategy to date. While theirs is a clever identification strategy, it only applies to *one side* of the marriage market (men), and only exploits *cohort* variation. Earlier work had studied the impact of male scarcity on marital assortative mating using the large shock that WWI caused to the number of French men (Abramitzky et al., 2011), used quarter of birth as a (weak) instrument for female education, or data on twins to assess assortative mating and

how education is productive in marriage (Lefgren and McIntyre, 2006; Huang et al., 2009). More generally an IV approach to instrument for market conditions, such as sex ratios, had been used by Angrist (2002) and Charles and Luoh (2010), for instance.

There is also a literature on genetic assortativeness. Using data from the HRS, Domingue et al. (2014) find that spouses are more genetically similar than two people chosen at random. Guo et al. (2014) also find a positive similarity in genomic assortment in married couples by using the HRS and the Framingham Heart study. Conley et al. (2016), however, show that the increased level of assortative mating in education observed across birth cohorts from 1920 to 1955 does not correspond to an increase in similarity at the genotypic level. These articles use genetic information from large scale GWASs that are also the core of our analysis. While these studies are instrumental for our analysis, our work departs from them, if only because our focus is assortative mating on education, and not spousal resemblance at the genotypic level.[2]

Our research also broadly speaks to the increasing "genoeconomics" literature that studies the genetic determinants of socioeconomic outcomes (Beauchamp et al., 2011; Benjamin et al., 2007; Conley et al., 2014a). While a few studies in economics have used genome-wide polygenic score as an instrumental variable (see also von Hinke Kessler Scholder et al., 2016; Böckerman et al., 2016), we are the first to examine the validity of the exclusion restriction of a genome-wide polygenic score. By combining the IV-genetic approach with a parsimonious linear bi-dimensional stochastic matching model, we can test the validity of the exclusion restriction, directly tackling the issue of pleiotropy, which in the context of genome-wide scores leads to concerns about the number of potential pathways through which the score could influence the outcome. Hence, our study breaks new ground by complementing and expanding the economic literature using genes (or genetic markers) as instrumental variables (e.g., Cawley et al., 2011; Fletcher and Lehrer, 2011; Norton and Han, 2008; von Hinke Kessler Scholder et al., 2011, 2013, 2014, 2016).

[2]On the genetic similarity of spouses see also Zou et al. (2015).

Structure of the paper. The rest of the paper is organized as follows. Section 2 presents two stylized models of assortative mating –a one-dimensional deterministic linear matching model, and a bi-dimensional stochastic linear matching model– and defines the corresponding OLS and IV estimands. Section 3 defines the polygenic scores and the genetic IV. Section 4 describes the data and the construction of the polygenic scores. Section 5 contains our empirical analysis. Finally, Section 6 concludes the paper.

2 Measuring Assortative Mating in Theory

2.1 A one-dimensional deterministic linear matching model

Consider two populations (men and women) of equal size, normalized to one. Agents differ in their overall marital attractiveness: x is the female marital attractiveness index, and y is the male marital attractiveness index. Without loss of generality, assume that

$$x \sim U[a, b] \tag{1}$$

$$y \sim U[0, 1] \tag{2}$$

where $b > 1$ and $a > 0$. Positive assortative mating (PAM) on marital attractiveness implies that[3]

$$\frac{x - a}{b - a} = y \tag{3}$$

[3]Let the marital surplus be defined as $h(x, y)$, and assume that h is twice differentiable. In a transferable utility setting, super-modularity of the surplus function (i.e., $h_{xy} > 0$) implies PAM. In a non-transferable utility setting, if h is increasing in both x and y, then we have PAM.

Hence, we have the matching function

$$y = \beta_0 + \beta_1 x \tag{4}$$

where $\beta_0 = -\frac{a}{b-a}$ and $\beta_1 = \frac{1}{b-a}$. Thus, this model predicts *perfect* PAM on marital attractiveness

$$corr(y,x) = \frac{cov(y,x)}{\sqrt{var(y)}\sqrt{var(x)}} = \frac{cov(\beta_0 + \beta_1 x, x)}{\sqrt{var(\beta_0 + \beta_1 x)}\sqrt{var(x)}} = \frac{\beta_1 var(x)}{\beta_1 var(x)} = 1 \tag{5}$$

If overall marital attractiveness in the marriage market is fully captured by educational attainment, then this simple model predicts perfect PAM on education. Deviations from perfect PAM on education can be accounted for by introducing *randomness* into the matching process, or extending the relevant *dimensions* in which matching takes place (Chiappori et al., 2012, forthcoming; Choo and Siow, 2006; Dupuy and Galichon, 2014; Galichon and Salanié, 2015; Siow, 2015).

2.2 A bi-dimensional stochastic linear matching model

Stochastic matching functions. Suppose that the "stochastic" matching functions are given by

$$y = \alpha + \beta x + v_y \tag{6}$$

and

$$x = \alpha^{'} + \beta^{'} y + v_x \tag{7}$$

where y is the male overall marital attractiveness index, x is the female overall marital attractiveness index, both unobserved by the econometrician, and v_y and v_x are random

components, which can be thought of as being reduced form representations of search frictions.

Attractiveness indices. Let the overall marital attractiveness index be linear (Chiappori et al., 2012; Hitsch et al., 2010) and bi-dimensional, where the two attributes at stake are: E (education) and $\widetilde{E}$ (non-education), with $\widetilde{E}$ being unobservable to the econometrician, and homogeneously assessed by each side of the market.[4] More generally, $\widetilde{E}$ can be thought of as being a linear combination of attributes other than education and that are homogenously assessed in the population, for instance, physical attractiveness. Hence, the female and male attractiveness indices can be written as

$$y = \pi_0 + \pi_1 E_y + \pi_2 \widetilde{E}_y + u_y \tag{8}$$

and

$$x = \delta_0 + \delta_1 E_x + \delta_2 \widetilde{E}_x + u_x \tag{9}$$

where E_y is male education, $\widetilde{E}_y$ is the male non-education component and u_y is a random component (uncorrelated with both E_y and $\widetilde{E}_y$). E_x, $\widetilde{E}_x$ and u_x are similarly defined for women.[5] Crucially, the two components E and $\widetilde{E}$ are allowed to be correlated.

Education equations I. Substituting (8) and (9) into (6) we obtain the following equation for the male educational attainment

$$E_y = \frac{1}{\pi_1}(\alpha + \beta\delta_0 - \pi_0) + \frac{1}{\pi_1}\beta\delta_1 E_x + \frac{1}{\pi_1}\beta\delta_2 \widetilde{E}_x - \frac{1}{\pi_1}\pi_2 \widetilde{E}_y + \frac{1}{\pi_1}(\beta u_x + v_y - u_y) \tag{10}$$

Similarly, substituting (8) and (9) into (7) we obtain the following equation for the female

[4] The issue of heterogeneous preferences in the marriage market has recently attracted attention. Chiappori et al. (forthcoming) study bidimensional matching on education and smoking in the marriage market, allowing for heterogeneous preferences in the population regarding the desirability of spousal smoking.

[5] $\widetilde{E}_x = \sum_j^M w_j \widetilde{E}_{x,j}$ where w_j is the weight of each non-educational attribute. Similarly for $\widetilde{E}_y$.

educational attainment

$$E_x = \frac{1}{\delta_1}\left(\alpha^{'} + \beta^{'}\pi_0 - \delta_0\right) + \frac{1}{\delta_1}\beta^{'}\pi_1 E_y + \frac{1}{\delta_1}\beta^{'}\pi_2\widetilde{E}_y - \frac{1}{\delta_1}\delta_2\widetilde{E}_x + \frac{1}{\delta_1}\left(\beta^{'}u_y + v_x - u_x\right) \quad (11)$$

The parameters measuring assortative mating on education are

$$\rho = \frac{\beta\delta_1}{\pi_1} \quad (12)$$

and

$$\rho^{'} = \frac{\beta^{'}\pi_1}{\delta_1} \quad (13)$$

We can define our measure of assortative mating on education as $\sqrt{\rho\rho^{'}}$, which is the geometric mean of the two parameters. Incidentally, $\sqrt{\rho\rho^{'}} = \sqrt{\beta\beta^{'}}$, which is the geometric mean of the two parameters measuring assortative mating on the unobservable index of overall marital attractiveness. Hence, by measuring assortative mating on education (which is observable) we can measure assortative mating on the index of attractiveness (which is unobservable to us). However, given that we do not observe $\widetilde{E}_x$ and $\widetilde{E}_y$, we cannot recover ρ or $\rho^{'}$ using OLS, since OLS will suffer from omitted variables bias. In order to be able to recover ρ or $\rho^{'}$, we need additional information.

Auxiliary equations: Genes and educational attainment. Let z be a measure of genetic predisposition to higher educational attainment: z is expected to be positively correlated with E. Hence, we have the following auxiliary equations for male and female education

$$E_y = \theta_0 + \theta_1 z_y + \epsilon_y \quad (14)$$

and

$$E_x = \mu_0 + \mu_1 z_x + \epsilon_x \tag{15}$$

In addition, based on recent studies, z is allowed to be correlated with $\widetilde{E}$. For example, it has been shown that the polygenic score for education is associated with attention deficit hyperactivity disorder (de Zeeuw et al., 2014). Hence, we have the following auxiliary equations for the male and female non-education components

$$\widetilde{E}_y = \lambda_0 + \lambda_1 z_y + \eta_y \tag{16}$$

and

$$\widetilde{E}_x = \kappa_0 + \kappa_1 z_x + \eta_x \tag{17}$$

Education equations II. Substituting (16) and (17) into (10), we obtain a new equation for male education

$$E_y = \alpha^* + \beta^* E_x + \gamma^* z_x + \delta^* z_y + u^* \tag{18}$$

where

$$\alpha^* = \frac{\alpha + \beta\delta_0 + \beta\delta_2\kappa_0 - \pi_0 - \pi_2\lambda_0}{\pi_1} \tag{19}$$

$$\beta^* = \frac{\beta\delta_1}{\pi_1} \tag{20}$$

$$\gamma^* = \frac{\beta\delta_2\kappa_1}{\pi_1} \tag{21}$$

$$\delta^* = -\frac{\pi_2\lambda_1}{\pi_1} \tag{22}$$

$$u^* = \frac{\beta u_x + v_y - u_y + \beta\delta_2\eta_x - \pi_2\eta_y}{\pi_1} \tag{23}$$

Similarly, substituting (16) and (17) into (11), we obtain a new equation for female education

$$E_x = \alpha^{'*} + \beta^{'*}E_y + \gamma^{'*}z_x + \delta^{'*}z_y + u^{'*} \tag{24}$$

where

$$\alpha^{'*} = \frac{\alpha^{'} + \beta^{'}\pi_0 + \beta^{'}\pi_2\lambda_0 - \delta_0 - \delta_2\kappa_0}{\delta_1} \tag{25}$$

$$\beta^{'*} = \frac{\beta^{'}\pi_1}{\delta_1} \tag{26}$$

$$\gamma^{'*} = -\frac{\delta_2\kappa_1}{\delta_1} \tag{27}$$

$$\delta^{'*} = \frac{\beta^{'}\pi_2\lambda_1}{\delta_1} \tag{28}$$

$$u^{'*} = \frac{\beta^{'} u_y + v_x - u_x + \beta' \pi_2 \eta_y - \delta_2 \eta_x}{\delta_1} \tag{29}$$

IV estimands. Can our measures of genetic predisposition (polygenic scores) of education z_x and z_y be used as valid instrumental variables? In general, neither z_x nor z_y can be used as instrumental variables for E_x and E_y, respectively, since z_x appears in (18) as an explanatory variable (z_x affects E_y above and beyond its effect through E_x), and similarly z_y appears in (24) as an explanatory variable too (z_y affects E_x above and beyond its effect through E_y). Hence, as long as $\gamma^* \neq 0$, z_x does not satisfy the exclusion restriction and cannot be used as an instrumental variable for E_x. Similarly, as long as $\delta^{'*} \neq 0$, z_y does not satisfy the exclusion restriction and cannot be used as an instrumental variable for E_y. In general, the exclusion restriction is *not* falsifiable.[6] However, under our structural assumption, namely our bi-dimensional stochastic linear matching model, the exclusion restriction can be tested. In particular, one can run equation (18) by OLS and test the following hypothesis:

$$\begin{aligned} H_0 : \gamma^* = 0 \quad &\leftrightarrow \quad z_x \text{ satisfies the exclusion restriction} \\ H_1 : \gamma^* \neq 0 \quad &\leftrightarrow \quad z_x \text{ does not satisfy the exclusion restriction} \end{aligned} \tag{33}$$

If we cannot reject H_0, we can use z_x as an instrumental variable for E_x to recover β^*.

[6] Consider the simple linear model:

$$y = a + bx + cw + e \tag{30}$$

where w is unobserved. If we run the OLS regression

$$y = a + bx + dz + u \tag{31}$$

where $u = cw + e$, then

$$d_{OLS} = d + c\frac{cov(z, w)}{var(z)} \tag{32}$$

Thus, in general, without any additional structure, the exclusion restriction is not testable, since this OLS regression suffers from omitted variable bias. In other words, d_{OLS} will be different than zero even though $d = 0$ as long as $c \neq 0$ and $cov(z, w) \neq 0$.

Similarly, one can run equation (24) by OLS and test the following hypothesis:

$$\begin{aligned} H_0 : \delta'^* = 0 \quad &\leftrightarrow \quad z_y \text{ satisfies the exclusion restriction} \\ H_1 : \delta'^* \neq 0 \quad &\leftrightarrow \quad z_y \text{ does not satisfy the exclusion restriction} \end{aligned} \tag{34}$$

If we cannot reject H_0, we can use z_y as an instrumental variable for E_y to recover β'^*.

Note that in both cases the exclusion restriction can be satisfied because either the polygenic score of education E is uncorrelated with the non-education component $\widetilde{E}$ ($\kappa_1 = 0$ for z_x and $\lambda_1 = 0$ for z_y), or attractiveness only depends on education E ($\delta_2 = 0$ (female non-education is irrelevant) and $\pi_2 = 0$ (male non-education is irrelevant)), or both. In practice, we expect both $\kappa_1 \neq 0$ and $\lambda_1 \neq 0$, since the polygenic score for education is associated with attention deficit hyperactivity disorder (de Zeeuw et al., 2014), amongst others, which may be related to any (or many) of the attributes embedded into the non-education component. These issues are covered in much more detail in Section 3 (pp. 19-20) when discussing the exclusion restriction assumption. Our next step is thus to build a potentially valid genetic IV.

3 Building a Potentially Valid Genetic IV

3.1 Polygenic Scores

Recent advances in molecular genetics have made it possible and relatively inexpensive to measure millions of genetic variants in a single study. The most common type of genetic variation among people is called single nucleotide polymorphism (SNP). SNPs are genetic markers that have two variants called alleles. Since individuals inherit two copies for each SNP, one from each parent, there are three possible outcomes: 0, 1 or 2 copies of a specific allele. SNPs occur normally throughout a person's DNA. Each SNP represents a difference in a single DNA building block, called a nucleotide. For example, a SNP may indicate that,

in a certain stretch of DNA, a nucleotide cytosine is replaced with the nucleotide thymine among some individuals.

SNPs are usually indicated by their position in the DNA, their possible nucleotides and by an identification number. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome. A large part of current genetic research aims to identify the function of these genetic variants and their relationship to different diseases. GWASs have been used to identify SNPs associated to particular diseases or traits. A drawback of GWAS is that, given the polygenic nature of human diseases and traits, most variants identified confer relatively small increments in risk, and explain only a small proportion of heritability. A common solution is to use the results from a GWAS and compile a polygenic score (PS) for a phenotype aggregating thousands of SNPs across the genome and weighting them by the strength of their association.

There are two main reasons to use a PS to describe the genetic susceptibility to a trait in social sciences (Belsky and Israel, 2014; Schmitz and Conley, forthcoming). First, complex health outcomes or behaviors are usually highly polygenic, i.e., reflect the influence or aggregate effect of many different genes (Visscher et al., 2008). PSs assume that individuals fall somewhere on a continuum of genetic predisposition resulting from small contributions from many genetic variants. Second, a single genetic variant has too small of an effect in explaining complex phenotypes, i.e., no single gene produces a symptom or trait at a detectable level, unless the sample size is extremely high.

A PS for individual i can be calculated as the sum of the allele counts a_{ij} (0, 1 or 2) for each SNP $j = 1, \dots M$, multiplied by a weight w_j:

$$PS_i = \sum_{j=1}^{M} w_j a_{ij}$$

A standard choice of weights is to use the association coefficients derived from a GWAS. A common practice is to include SNPs based on their association strength (p-value). For

instance, it is possible to include in the PS only SNPs that reach genome-wide significance (5×10^{-8}) or those that reach a less stringent level of association. The most inclusive criterion is to include all the SNPs associations from a GWAS, weighting their effect using their effect size. Since SNPs are not independent in the genome but their occurrence varies according to a block structure called *linkage disequilibrium* (LD), PS are often calculated using only SNPs that are independent to each other.[7] These independent SNPs are then used to calculate the score, avoiding possible bias due to oversampling DNA regions highly genotyped. The range of possible values that a PS can take depends on the number of SNPs included, tending to a normal distribution if the number of independent SNPs included in the score is sufficiently high. For comparability purposes, we standardize a score by subtracting its mean and dividing it by its standard deviation.

Using PSs rather than single genetic markers has several advantages. First, they are "hypothesis-free" measures that do *not* require knowledge about the biological processes involved. This is particularly important when the phenotype of interest is complex, i.e., influenced by a large number of genes, or when its biological mechanisms are not yet fully understood (Belsky and Israel, 2014). Second, using a score, rather than single genes, is a possible *solution* to overcome the low predictive power of single genes, especially for behavioral traits. For example, the top genome-wide significants SNP from the most recent GWAS on educational attainment (Okbay et al., 2016) explains around 0.01% of the variation in years of schooling. A linear polygenic score from all measured SNPs explains 3.2% of the same variable. Third, complete genome-wide association results are *publicly* available. PSs can be calculated from consortia data for a range of phenotypes. The results published by these consortia are based on a meta-analysis of a large number of cohort studies. The predictive power of a polygenic score is inflated if the samples are not independent, i.e., the same sample was used in the original calculation of association results. For this reason, it is common to use genetic association results from independent studies or to rerun the

[7] To select independent SNPs we use a procedure called *clumping* that prevents that SNPs are highly correlated (in linkage disequilibrium).

association results excluding the cohort to which the score is applied, which is exactly how we proceed.

Of course, there are also drawbacks in using PSs rather than single genetic markers. First, the issue of pleiotropy (discussed in the next subsection), which in the context of genome-wide scores leads to concerns about the number of potential pathways through which the score could influence the outcome. Note, however, that our bi-dimensional model allows us to tackle the issue of pleiotropy. Second, the issue of population stratification, which refers to the situation in which there is a systematic relationship between the allele frequency and the outcome of interest in different subgroups of the population, and that we discuss in the next subsection. Finally, PSs are measured with measurement error (e.g., an $R^2 < 0.05$ for the education PS compared to heritability estimates of ~ 0.4; Branigan et al. (2013)).

3.2 Genetic IV

There is a growing literature both methodological and applied on the use of genetic data as instrumental variables. The motivation for using a genetic instrumental variable (IV) is the fact that individuals' genotypes are randomly allocated at conception, such a quasi-experimental design is called *Mendelian randomization* (Smith and Ebrahim, 2003).[8] However, randomization, while necessary is *not* a sufficient condition to use genetic data as valid instrumental variables.

There is a vast literature in statistics and epidemiology that focuses on methodological aspects related to genetic IV (e.g., Burgess et al., 2015; Davies et al., 2015; Didelez and Sheehan, 2007; Glymour et al., 2012; Kang et al., forthcoming; Lawlor et al., 2008; Sheehan et al., 2008; Smith and Ebrahim, 2003). More recently, von Hinke Kessler Scholder et al. (2016) carefully examine the conditions needed for genetic variants to be used as valid instrumental variables with the aim of disseminating these conditions in the economics and

[8]See also von Hinke Kessler Scholder et al. (2011); Cawley et al. (2011); Taylor et al. (2014) for a discussion about potential problems when exploiting Mendelian randomization as a genetic IV.

social sciences literature. As discussed before, in our study we consider one PS that contains *all* the information coming from the genetic markers of interest, instead of using one or few genetic variants, and testing each allele separately.[9] Hence, we improve on the existing literature.[10]

A valid instrument must satisfy the following assumptions:

A1. *Independence Assumption*

A2. *1st Stage or Relevance Assumption*

A3. *Exclusion Restriction Assumption*

A4. *Monotonicity Assumption*

The **Independence Assumption (A1)** requires that the polygenic score is as good as randomly assigned. Even if genotypes are randomly assigned at conception (Mendelian Randomization), the existence of *Population Stratification* can violate this assumption. Population stratification refers to the situation in which there is a systematic relationship between the allele frequency and the outcome of interest in different subgroups of the population.[11] Genetic similarity is often correlated with geographical proximity, because human genetic diversity is the result of the history of population migration, ethnic admixture and

[9]Recent articles by von Hinke Kessler Scholder et al. (2016) and Böckerman et al. (2016) use polygenic scores for body mass index as IV.

[10] The existing literature in economics has studied: the effect of obesity or body fat mass on labor market outcomes (Norton and Han, 2008), on medical costs (Cawley and Meyerhoefer, 2012), or on educational attainment (von Hinke Kessler Scholder et al., 2012); the impact of poor health on academic performance (Ding et al., 2009; Fletcher and Lehrer, 2011); the effect of cigarette smoking on BMI (Wehby et al., 2012); the effect of alcohol exposure in utero on child academic achievement (von Hinke Kessler Scholder et al., 2014); the effects of cigarette quitting during pregnancy on different health behaviors (Wehby et al., 2013); the effect of child/adolescent height on different health and human capital outcomes (von Hinke Kessler Scholder et al., 2013).

[11]Population stratification can lead to false positive associations, if variation in phenotype is due to cultural differences among subpopulations rather than biological differences (Tian et al., 2008). Human genetic diversity is the result of large-scale population movements, admixture, natural selection and genetic drift (Botigue et al., 2013). Population stratification is strongly correlated with the geographical distribution of individuals, since the number of common ancestors decreases exponentially with geographic distance. In European rural population, an individual's DNA can be used to infer their geographic origin with surprising accuracy, often within a few hundred kilometres (Novembre et al., 2008).

residential segregation. This may affect the marriage market since potential partners living in the same geographical area are more likely to share common ancestry.[12]

It is possible to control for the non-random distribution of genes across populations and account for differences in genetic structures within populations in three ways. First, genome-wide analysis should be based on ethnic *homogeneous* populations, for example restricting the analysis to individuals of European ancestry or controlling for geographical origin. Second, only *unrelated* individuals should be included in the analysis to avoid family structure or cryptic relatedness.[13] Last, population structure can be approximated by running a *principal components analysis* (PCA) on the entire genotype and using the principal components as control variables in the analysis (Price et al., 2006). PCA is the most common method used to control for population stratification in a GWAS. In our analysis, we focus on a uniform group of individuals (White and Non-Hispanic), and control for region of birth and for the first five genetic principal components in all our regressions.

The **1st Stage or Relevance Assumption (A2)** requires that the spousal polygenic score for education affects spousal education. While the use of one or few genetic variants can be weakly associated with education (weak instrument problem), our polygenic score is relevant and has been shown to robustly affect education (Rietveld et al., 2013; Okbay et al., 2016). Moreover, the score predicts education differences between siblings (Rietveld et al., 2014).

The **Exclusion Restriction Assumption (A3)** requires that the spousal polygenic score for education affects own education only through spousal education. In common genetic IV studies that investigate the effect of one individual's treatment on the *same* individual's outcome, by using a genetic variant of his as instrument, the exclusion restriction can be violated mainly in four situations (von Hinke Kessler Scholder et al., 2016): (i) when

[12]Genetic population stratification has a strong bearing in genetic spouse similarities as a consequence of ethnic homogamy and geographic proximity. Friends and spouses are more genotypically similar than randomly matched individuals even in ethnically homogeneous samples (Christakis and Fowler, 2014; Domingue et al., 2014). Moreover, individuals who are genetically similar are more likely to have been reared in a similar environment (urban versus non-urban setting), Conley et al. (2014b).

[13]Kinship in the sample that is not known to the investigator.

parents' behavior or preferences are affected by the genotype; (ii) when the mechanisms, through which genetic variants affect the exposure variable, imply changes in behaviors or preferences that affect directly the outcome; (iii) when the genetic instrument is correlated with other genetic variants that affect the outcome (*Linkage Disequilibrium*);[14] (iv) when disruptive influences of the risk factor on the outcome are limited by foetal or post-natal development processes (*Canalization*), which violates **A3** because it results in an indirect effect of the genotype on the outcome. In our matching context, the exclusion restriction will be violated if two conditions hold at the same time:

Condition 1 The polygenic score of education z is predictive of the non-education component $\widetilde{E}$. Two main reasons can account for this. First, it could be that the polygenic score for educational attainment leads to differential development across a variety of different endophenotypes. For example, the polygenic score for education is associated with attention deficit hyperactivity disorder (de Zeeuw et al., 2014). Second, the gene-environment correlation, which can be related to the four situations discussed above.

Condition 2 The non-education attribute is relevant for overall attractiveness. This depends on what is deemed to be attractive for each side of the marriage market, and may be different by gender, ethnicity, period of time, and geography. For example, among prime-age cohorts of white Americans, Chiappori et al. (2012) find that attractiveness is multidimensional, depending on body mass index and socioeconomic status for both men and women. However, matching patterns have changed in the US over time (Stevenson and Wolfers, 2007). Whether education (fully) summarizes attractiveness for one, both or none of the sides of the marriage market is thus an empirical question.

Hence, z_x will violate the exclusion restriction if and only if $\kappa_1 \neq 0$ (Condition 1) and $\delta_2 \neq 0$

[14] A similar situation occurs when one genetic variant has multiple functions (*Pleitropy*). In this case the exclusion restriction is violated if the pleiotropic effect directly influences the outcomes.

(Condition 2). By the same token, z_y will violate the exclusion restriction if and only if $\lambda_1 \neq 0$ and $\pi_2 \neq 0$. While the existing evidence suggests that Condition 1 is satisfied, whether Condition 2 holds or is not known ex ante, and will depend on the particular context and period of time being analyzed.

Finally, the **Monotonicity Assumption (A4)** requires that the spousal polygenic score affects spousal education for every "spouse" in the same direction.[15] However, with homogeneous causal effects, as in our stylized matching model, the monotonicity assumption is irrelevant.

4 Data Description

4.1 Health and Retirement Study

The data used in this paper come from the Health and Retirement Study (HRS), a national panel survey representative of Americans over the age of 50 and their spouses, interviewed every two years since 1992.[16] The survey contains detailed socio-demographic information. It consists of six cohorts: initial HRS cohort, born between 1931 and 1941 (first interviewed in 1992); the Study of Assets and Health Dynamics Among the Oldest Old (AHEAD) cohort, born before 1924 (first interviewed in 1993); Children of Depression (CODA) cohort, born between 1924 and 1930 (first interviewed in 1998); War Baby (WB) cohort, born between 1942 and 1947 (first interviewed in 1998); Early Baby Boomer (EBB) cohort, born between 1948 and 1953 (first interviewed in 2004) and Mid Baby Boomer (MBB) cohort, born between 1954 and 1959 (first interviewed in 2010).

Between 2006 and 2008, the HRS genotyped 12,507 respondents who provided DNA samples and signed consent. DNA samples were genotyped using the Illumina Human Omni-2.5 Quad BeadChip, with coverage of approximately 2.5 million single nucleotide

[15] Chaisemartin (2015) shows that IV estimates a causal effect under a weaker condition than monotonicity.
[16] For the non-genetic data, we used the RAND HRS Data files, Version N.

polymorphisms (SNPs). Current genetic data available for research also include imputation of approximately 21 million DNA variants from the 1000Genomes Project.[17] Following recommendations of the genotyping center, we removed individuals with a genotyping rate <95% and SNPs with minor allele frequency (MAF) less than 1%, with p-value less than 1×10^{-4} on the test for Hardy-Weinberg equilibrium, and with missing call rate greater than 5%. The resulting genetic sample includes 12,205 individuals and information for 8,391,857 genetic variants.

The survey interviews the respondents of eligible birth years at the time of their first interview, as well as their married spouses or partners, regardless of age. It includes any individual interviewed at least once. For our study we are interested in couples rather than in the longitudinal structure of the data, we therefore build a cross-section. The original sample (RAND HRS Data) contains 37,319 individuals: We focus on individuals for which the genetic data are available after the quality control described above, 12,205 in total, excluding 25,114 respondents from the original survey. We also restrict the sample to only White respondents, excluding Black and Hispanic respondents (2,157 and 770 individuals, respectively). We consider only heterosexual couples at their first marriage. In particular, we exclude never married partners, people that are divorced or widowed at the time of the first interview, and people that have been already married or widowed more than once when entering the survey. We also drop respondents whose spouse has never been interviewed, couples where the spousal age gap is ten years or more, couples in which at least one of the two spouses has zero years of education, and couples in which one of the two spouses was born outside the US or born in the US but with missing census division of origin.[18] This yields a working sample of 1,441 couples (2,882 individuals).

The main variables used in our empirical analysis are education and the polygenic scores

[17]For details on quality control of the HRS genetic data, please see here. Data are available for research via the database of Genotypes and Phenotypes.

[18]Census Divisions are groupings of states and the District of Columbia that are subdivisions of the four census regions (Northeast, Midwest, South, and West). There are nine Census divisions: New England, Mid Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, Pacific.

for education. Education is defined in two ways: the number of completed years of schooling, and an indicator equal to 1 if the individual has a college degree (or above), and 0 otherwise. We generated a polygenic score based on the most recent GWAS results on educational attainment available (Okbay et al., 2016). The same polygenic score is used in the analysis of years of education and college attainment, since the genetic correlation between the two measures is very high, with the point estimate suggesting a perfect genetic correlation.

Since the HRS was part of the educational attainment consortium, we obtained the list of association results calculated excluding the HRS from the meta-analysis from the Social Science Genetic Association Consortium.[19] Using these summary statistics, we constructed linear polygenic scores weighted for their effect sizes in the meta-analysis. We constructed the scores using the softwares PLINK and PRSice (Purcell et al., 2007; Euesden et al., 2015).[20] We generated 9 different scores, restricting the number of SNPs based on their association p-values in the GWAS results. We started from a score that considered the complete set of available SNPs (p-value<1) and then calculated the scores on a subset of SNPs with the following p-value thresholds: $5\times10^{-1}; 5\times10^{-2}; 5\times10^{-3}; 5\times10^{-4}; 5\times10^{-5}; 5\times10^{-6}; 5\times10^{-7}; 5\times10^{-8}$. All the scores are clumped using the genotypic data as a reference panel for *Linkage Disequilibrium* structure.

To ensure that the population stratification does not violate the **Independence Assumption (A1)**, we focus our analysis on a *homogeneous* subpopulation, White non-Hispanic, and we control for place of birth (Census division), year of birth, an indicator variable if the place of birth differs between spouses, and the first five genetic principal components for each individual using genome-wide principal components that function as ancestry markers (Price et al., 2006).[21]

These population controls allow to analyze genotypic variants that are not driven by

[19] Complete genetic association results on educational attainment are available here, see acknowledgments for data conditions.

[20] Genetic data are based on best call genotypes imputed to 1000 Genome.

[21] The results from a principal components analysis (PCA) on the entire genotype are available from the HRS genetic data.

specific ethnicity. Moreover, our polygenic scores are based on genome-wide association results on individuals of European ancestry and control for population structure. Once we control for population structure and individual's genes for education, it is safe to assume that spousal genes for education are as good as randomly assigned. Finally, given that both the PS for education and the principal components are generated regressors, the standard errors in our regression analysis are bootstrapped. IV estimates are calculated using 2SLS.

After assessing the relevance of the 9 polygenic scores,[22] and testing whether they satisfy the exclusion restriction based on our linear bi-dimensional stochastic matching model, our results (see Online Appendix) indicate that more inclusive scores (based on a less stringent p-value) are more relevant (A2). However, by including a larger number of SNPs the exclusion restriction assumption (A3) is more likely to be violated. Interestingly, the only PSs satisfying both the relevance condition (A2) and the exclusion restriction (A3) for both husbands and wives appears to be the PSs based on SNPs with a p-value$<5 \times 10^{-4}$, and these are the ones used in our analysis.

5 Results

5.1 Descriptive statistics

Table 1 provides the basic descriptive statistics for our sample of husbands and wives. These individuals were born between 1910 and 1961. On average, husbands –with 13.6 years of education– are more educated than their wives –with 13.4 years of education; 35% of husbands have a college degree while 24% of wives do.

[Table 1 about here]

Table 2 shows the correlation matrix for years of education and PSs: the correlation between husband's and wife's years of education is 0.568 (p-value$<$0.001), while that for

[22] p-value<1, p-value$<5\times10^{-1}$, p-value$<5\times10^{-2}$, p-value$<5\times10^{-3}$, p-value$<5\times10^{-4}$, p-value$<5\times10^{-5}$, p-value$<5\times10^{-6}$, p-value$<5\times10^{-7}$, p-value$<5\times10^{-8}$

their PSs for education is 0.066 (p-value$<$0.001). If anything, this indicates that there is both positive assortative mating on education and on PSs for education, but that the former is much stronger than the latter.

[Table 2 about here]

5.2 OLS versus IV estimates

Table 3 contains the first results regarding assortative mating on education. The first three columns display OLS estimates of regressions of wife's years of education on the husband's years of education. Column (2) adds the polygenic score (PS) for wife's education, and column (3) controls for both spouses' PSs. Column (1) shows that the point estimate of the coefficient on husband's education is 0.451. Once the wife's genetic score is accounted for, the point estimate decreases from 0.451 to 0.444, column (2). The coefficient does not change when adding both PSs, column (3). Note that column (3) displays the estimation of the regression equation (24) in our stochastic linear bi-dimensional matching model, and hence it allows us to test whether z_y satisfies the exclusion restriction. According to the estimates in this column, we cannot reject $\delta^{'*} = 0$: the point estimate is 0.004 (SE = 0.054). A similar qualitative picture emerges in the last three columns, (4), (5) and (6), where we replace years of education with a college degree (or above) indicator.

[Table 3 about here]

In Table 4 we run the same analysis as in Table 3 but now the husband's education is the dependent variable and the wife's education is the main explanatory variable. Overall the magnitude of the coefficients is larger than in Table 3, and we find evidence that z_x satisfies the exclusion restriction.

[Table 4 about here]

The findings in Tables 3 and 4 suggest that the PSs for education satisfy the exclusion restriction.

Table 5 begins with our instrumental variable analysis. The table contains two blocks of regressions, columns (1)-(3) for years of education, and columns (4)-(6) for college degree (or above). In column (1) we analyze whether the husband's PS satisfies the *instrument relevance* condition: the F-statistic for the husband's PS being irrelevant is 42.36, beyond the "rule of thumb" of 10 (Staiger and Stock, 1997; Stock and Yogo, 2005). Column (2) shows the reduced-form: interestingly, the role of the husband's PS is more than half that of the own PS. Finally, column (3) assesses assortative mating on years of education using 2SLS: the point estimate of the coefficient on husband's years of education is 0.451 (SE = 0.126), very similar to 0.444 (SE = 0.022), the point estimate obtained using OLS in Table 3. Looking at columns (4)-(6), we find similar results: the instrument appears to be relevant for college degree (44.25), the role of the husband's PS is more than half that of the own PS, and the 2SLS point estimate is 0.474 (SE = 0.136), which is similar to 0.427 (SE = 0.024) the OLS point estimate in Table 3. Both of the Hausman tests at the bottom of columns (3) and (6) *cannot* reject that the OLS and IV estimands are the same.

[Table 5 about here]

While Table 5 contains the IV (2SLS) analysis corresponding to the OLS analysis of Table 3, Table 6 displays the IV (2SLS) analysis corresponding to Table 4. The IV point estimate of the coefficient on wife's years of education is 0.925 (SE = 0.232) versus the OLS point estimate 0.643 (SE = 0.028), while the point estimate of the coefficient on wife's college is 0.759 (SE = 0.269) versus the OLS point estimate of 0.532 (SE = 0.026). The F-statistics for the wife's education PS are respectively 31.21 and 21.04. Both of the Hausman tests at the bottom of columns (3) and (6) *cannot* reject that the OLS and IV estimands are the same.

[Table 6 about here]

5.3 Discussion

Taken at face value our results are consistent with two potential explanations: first, the non-education component (e.g., physical attractiveness) is irrelevant for marital attractiveness, at least among the sample under analysis, that is, individuals who are on average 70 years old, who got married on average 40 years ago and have been married to each other ever since, and who are still alive (healthier); second, the non-education component (e.g., physical attractiveness) is uncorrelated with education, at least among our highly selected group of couples. Whatever the reason is, the bottom line is that for a group of individuals for whom education is (perhaps) the most important attribute on the marriage market (older generations), assortative mating on education can be assessed with standard regression techniques.

Traditionally, the purpose of marriage was division of labor and child rearing rather than an hedonic marriage and affinity on personality (Becker, 1973; Lundberg, 2012). While it seems plausible to assume that in traditional marriages education was the main attribute driving marital surplus, the sexual revolution transformed the marriage market and its institution toward hedonic marriages (Stevenson and Wolfers, 2007). Thus, we would expect to find different returns in the marriage market using OLS and genetic-IV on a sample of recently first-married couples. In other words, our empirical findings are only *internally* valid, but the parsimonious approach combining genes with a structural model described in this paper is *externally* valid, and hopefully will stimulate future research.

6 Conclusions

Our study illustrates how to combine quasi-experimental variation in the spouses' characteristics with parsimonious matching models so that the exclusion restriction can be investigated and, when satisfied, assortative mating can be assessed. In particular, this is the first paper to present an IV strategy to estimate assortative mating on education using

spousal genetic markers. While there is a burgeoning literature using genetic measures to analyze own behavior (von Hinke Kessler Scholder et al., 2011), such as the returns to health in schooling in the labor market, this often suffers from unconvincing exclusion restrictions, especially when neurotransmitters genes are used (Cawley et al., 2011). Using a stochastic linear bi-dimensional matching model, we are able to investigate whether the educational polygenic score satisfies the exclusion restriction for both men and women in a matching context. The validity of the exclusion restriction cannot be rejected for either the female or the male polygenic scores of education. Interestingly, we cannot reject that the (estimated) degree of assortative mating on education is the same using OLS and IV.

While our evidence is consistent with a positive causal effect of own's education on spousal's education, future research should try to pin down the exact mechanism behind the positive assortative mating in education: Will more educated wives become more likely to encounter potential husbands that are more educated? And, or will more educated wives become more attractive to more educated husbands, holding constant the likelihood of meeting an educated spouse? In other words, is it search or preferences? These are interesting questions that we leave for future research.

References

Abramitzky, R., A. Delavande, and L. Vasconcelos (2011): "Marrying up: the role of sex ratio in assortative matching," *American Economic Journal: Applied Economics*, 124–157.

Angrist, J. (2002): "How Do Sex Ratios Affect Marriage and Labor Markets? Evidence from America's Second Generation," *Quarterly Journal of Economics*, 117, 997–1038.

Angrist, J. D. and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106, 979–1014.

Beauchamp, J. P., D. Cesarini, M. Johannesson, M. J. H. M. van der Loos, P. D. Koellinger, P. J. F. Groenen, J. H. Fowler, J. N. Rosenquist, A. R. Thurik, and N. A. Christakis (2011): "Molecular genetics and economics," *Journal of Economic Perspectives*, 25, 57–82.

Becker, G. (1973): "A Theory of Marriage: Part I," *Journal of Political Economy*, 81, 813–846.

Belsky, D. W. and S. Israel (2014): "Integrating Genetics and Social Science: Genetic Risk Scores," *Biodemography and Social Biology*, 60, 137–155.

Benjamin, D. J., C. F. Chabris, E. L. Glaeser, V. Gudnason, T. B. Harris, D. I. Laibson, L. J. Launer, and S. Purcell (2007): "Genoeconomics," in *Biosocial Surveys*, ed. by M. Weinstein, J. W. Vaupel, K. W. Wachter, et al., Washington D.C.: National Academies Press, chap. 15, 304–335.

Böckerman, P., J. Cawley, J. Viinikainen, T. Lehtimäki, S. Rovio, I. Seppälä, J. Pehkonen, and O. Raitakari (2016): "The Effect of Weight on Labor Market Outcomes: An Application of Genetic Instrumental Variables," *National Bureau of Economic Research, Working paper n. 22200.*

Botigue, L. R., B. M. Henn, S. Gravel, B. K. Maples, C. R. Gignoux, E. Corona, G. Atzmon, E. Burns, H. Ostrer, C. Flores, J. Bertranpetit, D. Comas, and C. D. Bustamante (2013): "Gene flow from North Africa contributes to differential human genetic diversity in southern Europe," *Proceedings of the National Academy of Sciences*, 110, 11791–11796.

Branigan, A. R., K. J. McCallum, and J. Freese (2013): "Variation in the heritability of educational attainment: An international meta-analysis," *Social forces*, 92, 109–140.

Bruze, G. (2011): "Marriage Choices of Movie Stars: Does Spouse's Education Matter?" *Journal of Human Capital*, 5, 1–28.

Burgess, S., N. J. Timpson, S. Ebrahim, and G. D. Smith (2015): "Mendelian randomization: where are we now and where are we going?" *International Journal of Epidemiology*, 44, 379–388.

Card, D. (2001): "Estimating the Returns to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160.

Cawley, J., E. Han, and E. Norton (2011): "The Validity of Genes Related to Neurotransmitters as Instrumental Variables," *Health Economics*, 20, 884–888.

Cawley, J. and C. Meyerhoefer (2012): "The medical care costs of obesity: an instrumental variables approach," *Journal of Health Economics*, 31, 219–230.

Chaisemartin, C. D. (2015): "Tolerating Defiance? Identification of treatment effects without monotonicity," *University of Warwick, mimeo.*

Charles, K. K. and M. C. Luoh (2010): "Male Incarceration, the Marriage Market, and Female Outcomes," *Review of Economics and Statistics*, 92, 614–627.

Chiappori, P. A., M. Iyigun, and Y. Weiss (2009): "Investment in Schooling and the Marriage Market," *American Economic Review*, 99, 1689–1713.

Chiappori, P. A., S. Oreffice, and C. Quintana-Domeque (2012): "Fatter Attraction: Anthropometric and Socioeconomic Matching on the Marriage Market," *Journal of Political Economy*, 120, 659–695.

Chiappori, P.-A., S. Oreffice, and C. Quintana-Domeque (forthcoming): "Bidimensional Matching with Heterogeneous Preferences: Education and Smoking in the Marriage Market," *Journal of the European Economic Association*.

Chiappori, P. A. and B. Salanié (2016): "The Econometrics of Matching," *Journal of Economic Literature*, 54, 832–861.

Choo, E. and A. Siow (2006): "Who Marries Whom and Why," *Journal of Political Economy*, 114, 175–201.

Christakis, N. A. and J. H. Fowler (2014): "Friendship and natural selection," *Proceedings of the National Academy of Sciences*, 111, 10796–10801.

Conley, D., B. W. Domingue, D. Cesarini, C. Dawes, C. A. Rietveld, and J. D. Boardman (2015): "Is the effect of parental education on offspring biased or moderated by genotype?" *Sociological Science*, 2, 82–105.

Conley, D., J. Fletcher, and C. Dawes (2014a): "The emergence of socio-genomics," *Contemporary Sociology: A Journal of Reviews*, 43, 458–467.

Conley, D., T. Laidley, D. W. Belsky, J. M. Fletcher, J. D. Boardman, and B. W. Domingue (2016): "Assortative mating and differential fertility by phenotype and genotype across the 20th century," *Proceedings of the National Academy of Sciences*, 201523592.

Conley, D., M. L. Siegal, B. W. Domingue, K. M. Harris, M. B. McQueen, and J. D. Boardman (2014b): "Testing the key assumption of heritability estimates based on genome-wide genetic relatedness," *Journal of Human Genetics*, 59, 342–345.

Currie, J. (2011): "Inequality at Birth: Some Causes and Consequences," *American Economic Review*, 101, 1–22.

Davies, N. M., S. von Hinke Kessler Scholder, H. Farbmacher, S. Burgess, F. Windmeijer, and G. D. Smith (2015): "The many weak instruments problem and Mendelian randomization," *Statistics in Medicine*, 34, 454–468.

de Zeeuw, E. L., C. E. van Beijsterveldt, T. J. Glasner, M. Bartels, and E. A. E. G. E. D. J. J. H. S. S. G. A. C. C. A. R. M. M. G.-B. J. J. H. E. J. de Geusand Dorret I. Boomsma (2014): "Polygenic scores associated with educational attainment in adults predict educational achievement and ADHD symptoms in children," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 165, 510–520.

Didelez, V. and N. Sheehan (2007): "Mendelian randomization as an instrumental variable approach to causal inference," *Statistical Methods in Medical Research*, 16, 309–330.

Ding, W., S. F. Lehrer, J. N. Rosenquist, and J. Audrain-McGovern (2009): "The impact of poor health on academic performance: New evidence using genetic markers," *Journal of Health Economics*, 28, 578–597.

Domingue, B., J. Fletcher, D. Conley, and J. Boardman (2014): "Genetic and Educational Assortative Mating among US Adults," *Proceedings of the National Academy of Sciences*, 111, 7996–8000.

Dupuy, A. and A. Galichon (2014): "Personality Traits and the Marriage Market," *Journal of Political Economy*, 122, 1271–1319.

Euesden, J., C. M. Lewis, and P. F. O'Reilly (2015): "PRSice: Polygenic Risk Score software," *Bioinformatics*, 1466–1468.

Fernandez, R. and R. Rogerson (2001): "Sorting And Long-Run Inequality," *Quarterly Journal of Economics*, 116, 1305–1341.

Fletcher, J. M. and S. F. Lehrer (2011): "Genetic lotteries within families," *Journal of Health Economics*, 30, 647–659.

Galichon, A. and B. Salanié (2015): "Cupid's Invisible Hand: Social Surplus and Identification in Matching Models," *New York University, mimeo*.

Glymour, M. M., E. J. T. Tchetgen, and J. M. Robins (2012): "Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions," *American Journal of Epidemiology*, 175, 332–339.

Greenwood, J., N. Guner, G. Kocharkov, and C. Santos (2014): "Marry Your Like: Assortative Mating and Income Inequality," *American Economic Review, Papers & Proceedings*, 104, 348–353.

Guo, G., L. Wang, H. Liu, and T. Randall (2014): "Genomic Assortative Mating in Marriages in the United States," *PLOS ONE*, 9, e112322.

Hitsch, G., A. Hortaçsu, and D. Ariely (2010): "Matching and Sorting in Online Dating," *American Economic Review*, 100, 130–163.

Huang, C., H. Li, P. W. Liu, and J. Zhang (2009): "Why Does Spousal Education Matter for Earnings? Assortative Mating and Cross-Productivity," *Journal of Labor Economics*, 27, 633–652.

Kang, H., A. Zhang, T. T. Cai, and D. S. Small (forthcoming): "Instrumental variables estimation with some invalid instruments and its application to mendelian randomization," *Journal of the American Statistical Association*.

Larsen, M., T. McCarthy, J. Moulton, M. Page, and A. Patel (2015): "War and Marriage: Assortative Mating and the World War II G.I. Bill," *Demography*, 52, 1431–1461.

Lawlor, D. A., R. M. Harbord, J. A. C. Sterne, N. Timpson, and G. D. Smith (2008): "Mendelian randomization: using genes as instruments for making causal inferences in epidemiology," *Statistics in Medicine*, 27, 1133–1163.

Lefgren, L. and F. McIntyre (2006): "The relationship between women's education and marriage outcomes," *Journal of Labor Economics*, 24, 787–830.

Lundberg, S. (2012): "Personality and Marital Surplus," *IZA Journal of Labor Economics*, 1.

Norton, E. C. and E. Han (2008): "Genetic information, obesity, and labor market outcomes," *Health Economics*, 17, 1089–1104.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante (2008): "Genes mirror geography within Europe," *Nature*, 456, 98–101.

Okbay, A., J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers, C. A. Rietveld, P. Turley, G.-B. Chen, V. Emilsson, S. F. W. Meddens, et al. (2016): "Genome-wide association study identifies 74 loci associated with educational attainment," *Nature*, 533, 539–542.

Plomin, R., C. M. A. Haworth, and O. S. P. Davis (2009): "Common disorders are quantitative traits," *Nature Reviews Genetics*.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006): "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, 38, 904–909.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. De Bakker, M. J. Daly, et al. (2007): "PLINK: a tool set for whole-genome association and population-based linkage analyses," *American Journal of Human Genetics*, 81, 559–575.

Rietveld, C. A., D. Conley, N. Eriksson, T. Esko, S. E. Medland, A. A. Vinkhuyzen, J. Yang, J. D. Boardman, C. F. Chabris, C. T. Dawes, et al. (2014): "Replicability and robustness of genome-wide-association studies for behavioral traits," *Psychological science*, 25, 1975–1986.

Rietveld, C. A., S. E. Medland, J. Derringer, J. Yang, T. Esko, N. W. Martin, H.-J. Westra, K. Shakhbazov, A. Abdellaoui, A. Agrawal, et al. (2013): "GWAS of 126,559 individuals identifies genetic variants associated with educational attainment," *Science*, 340, 1467–1471.

Schmitz, L. and D. Conley (forthcoming): "Modeling Gene-Environment Interactions With Quasi-Natural Experiments," *Journal of Personality*.

Schwartz, C. and R. Mare (2005): "Trends in Educational Assortative Marriage from 1940 to 2003," *Demography*, 42, 621–646.

Sheehan, N. A., V. Didelez, P. R. Burton, and M. D. Tobin (2008): "Mendelian randomisation and causal inference in observational epidemiology," *PLOS Medicine*, 5, e177.

Siow, A. (2015): "Testing Becker's Theory of Positive Assortative Matching," *Journal of Labor Economics*, 33, 409–441.

Smith, G. D. and S. Ebrahim (2003): "'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?" *International Journal of Epidemiology*, 32, 1–22.

Staiger, D. and J. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stevenson, B. and J. Wolfers (2007): "Marriage and Divorce: Changes and their Driving Forces," *Journal of Economic Perspectives*, 21, 27–52.

Stock, J. and M. Yogo (2005): "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models*, ed. by D. W. Andrews, New York: Cambridge University Press, 80–108.

Taylor, A. E., N. M. Davies, J. J. Ware, T. VanderWeele, G. D. Smith, and M. R. Munafò (2014): "Mendelian randomization in health research: Using appropriate genetic variants and avoiding biased estimates," *Economics & Human Biology*, 13, 99–106.

Tian, C., P. K. Gregersen, and M. F. Seldin (2008): "Accounting for ancestry: population substructure and genome-wide association studies," *Human Molecular Genetics*, 17, R143–R150.

Visscher, P. M., W. G. Hill, and N. R. Wray (2008): "Heritability in the genomics era–concepts and misconceptions," *Nature Reviews Genetics*, 9, 255–266.

von Hinke Kessler Scholder, S., G. D. Smith, D. Lawlor, C. Propper, and F. Windmeijer (2011): "Mendelian randomization: the use of genes in instrumental variable analyses," *Health Economics*, 20, 893–896.

——— (2013): "Child height, health and human capital: evidence using genetic markers," *European Economic Review*, 57, 1–22.

——— (2016): "Genetic Markers as Instrumental Variables," *Journal of Health Economics*, 45, 131–148.

von Hinke Kessler Scholder, S., G. D. Smith, D. A. Lawlor, C. Propper, and F. Windmeijer (2012): "The effect of fat mass on educational attainment: examining

the sensitivity to different identification strategies," *Economics & Human Biology*, 10, 405–418.

von Hinke Kessler Scholder, S., G. L. Wehby, S. Lewis, and L. Zuccolo (2014): "Alcohol exposure in utero and child academic achievement," *Economic Journal*, 124, 634–667.

Ward, M. E., G. McMahon, B. St Pourcain, D. M. Evans, C. A. Rietveld, D. J. Benjamin, P. D. Koellinger, D. Cesarini, G. D. Smith, N. J. Timpson, et al. (2014): "Genetic variation associated with differential educational attainment in adults has anticipated associations with school performance in children," *PLOS ONE*, 9, e100248.

Wehby, G. L., J. C. Murray, A. Wilcox, and R. T. Lie (2012): "Smoking and body weight: evidence using genetic instruments," *Economics & Human Biology*, 10, 113–126.

Wehby, G. L., A. Wilcox, and R. T. Lie (2013): "The impact of cigarette quitting during pregnancy on other prenatal health behaviors," *Review of Economics of the Household*, 11, 211–233.

Zou, J. Y., D. S. Park, E. G. Burchard, D. G. Torgerson, M. Pino-Yanes, Y. S. Song, S. Sankararaman, E. Halperin, and N. Zaitlen (2015): "Genetic and socioeconomic study of mate choice in Latinos reveals novel assortment patterns," *Proc Natl Acad Sci US A*.

Table 1. Summary statistics

	N	Mean	SD	Min	Max
Husband's Year of Birth	1,441	1937	8.87	1910	1957
Husband's Years of Education	1,438	13.64	2.68	2	17
Husband's College	1,441	0.35	0.48	0	1
Husband's Education Polygenic Score (p-value$< 1 \times 10^{-4}$)	1,435	0.21	0.97	-2.94	3.04
Wife's Year of Birth	1,441	1939	8.85	1911	1961
Wife's Years of Education	1,438	13.35	2.20	3	17
Wife's College	1,441	0.24	0.43	0	1
Wife's Education Polygenic Score (p-value$< 1 \times 10^{-4}$)	1,429	0.20	0.96	-2.81	3.32

Source: Data are from the HRS (Rand, Version N).

Note: White non-Hispanic couples in their first marriage, with at most 10 years of age difference and born in the US.

Both spouses have been interviewed at least once and provided DNA sample.

Table 2. Correlation matrix for Years of Education and Polygenic Scores

	Husband's Years of Education	Wife's Years of Education	Husband's Education Polygenic Score	Wife's Education Polygenic Score
Husband's Years of Education	1.000	–	–	–
	(0.000)			
Wife's Years of Education	0.5676	1.000	–	–
	(0.000)	(0.000)		
Husband's Education Polygenic Score	0.1772	0.1041	1.000	–
	(0.000)	(0.000)	(0.000)	
Wife's Education Polygenic Score	0.1178	0.1450	0.0660	1.000
	(0.000)	(0.000)	(0.000)	(0.000)

Note: p-values in parentheses. The score is based on SNPs with p-value$<5 \times 10^{-4}$.

Table 3. Regressions of Wife's Education on Husband's Education controlling for Polygenic Scores

Measure of education	*Years of Education*			*College*		
	Wife's Education			Wife's Education		
	(1)	(2)	(3)	(4)	(5)	(6)
Husband's Education	0.451***	0.444***	0.444***	0.435***	0.428***	0.427***
	(0.022)	(0.022)	(0.022)	(0.024)	(0.024)	(0.024)
Wife's Education Polygenic Score		0.190***	0.190***		0.036***	0.036***
		(0.051)	(0.051)		(0.010)	(0.010)
Husband's Education Polygenic Score			0.004			0.004
			(0.054)			(0.010)
Observations	1,417	1,417	1,417	1,423	1,423	1,423
R-squared	0.34	0.34	0.34	0.27	0.27	0.27

Note: All regressions include wife's year of birth, wife's place of birth dummy variables,
an indicator if the place of birth differs between spouses, and their respective genetic principal components.
Bootstrapped standard errors in parentheses.
*** p <0.01, ** p <0.05, * p <0.1

Table 4. Regressions of Husband's Education on Wife's Education controlling for Polygenic Scores

Measure of education	*Years of Education*			*College*		
	Husband's Education			Husband's Education		
	(1)	(2)	(3)	(4)	(5)	(6)
Wife's Education	0.663***	0.648***	0.643***	0.547***	0.535***	0.532***
	(0.028)	(0.028)	(0.028)	(0.026)	(0.026)	(0.026)
Husband's Education Polygenic Score		0.328***	0.325***		0.060***	0.059***
		(0.063)	(0.063)		(0.010)	(0.010)
Wife's Education Polygenic Score			0.085			0.012
			(0.064)			(0.012)
Observations	1,417	1,417	1,417	1,423	1,423	1,423
R-squared	0.34	0.36	0.36	0.27	0.28	0.28

Note: All regressions include husband's year of birth, husband's place of birth dummy variables, an indicator if the place of birth differs between spouses, and their respective genetic principal components. Bootstrapped standard errors in parentheses.

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 5. Using the Husband's Polygenic Score as an Instrumental Variable for Husband's Education

Measure of education	*Years of Education*			*College*		
	Husband's Education	Wife's Education		Husband's Education	Wife's Education	
	FS	RF	2SLS	FS	RF	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
Husband's Education			0.451***			0.474***
			(0.126)			(0.136)
Wife's Education Polygenic Score	0.284***	0.316***	0.187***	0.039***	0.053***	0.034***
	(0.072)	(0.055)	(0.061)	(0.013)	(0.011)	(0.012)
Husband's Education Polygenic Score	0.456***	0.206***		0.078***	0.037***	
	(0.070)	(0.061)		(0.012)	(0.012)	
F-test instrument relevance	42.36	–	–	44.25	–	–
Hausman test p-value	–	–	[0.948]	–	–	[0.726]
Observations	1,417	1,417	1,417	1,423	1,423	1,423

Note: Control variables are described in Tables 3 and 4.

Bootstrapped standard errors in parentheses. Bootstrap Hausman test is reported.

*** p <0.01, ** p <0.05, * p <0.1

Table 6. Using the Wife's Polygenic Score as an Instrumental Variable for Wife's Education

Measure of education	*Years of Education*			*College*		
	Wife's Education	Husband's Education		Wife's Education	Husband's Education	
	FS	RF	2SLS	FS	RF	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
Wife's Education			0.925***			0.759***
			(0.232)			(0.269)
Husband's Education Polygenic Score	0.206***	0.457***	0.267**	0.035***	0.078***	0.051***
	(0.062)	(0.070)	(0.080)	(0.012)	(0.012)	(0.015)
Wife's Education Polygenic Score	0.302***	0.280***		0.053***	0.040***	
	(0.054)	(0.071)		(0.012)	(0.013)	
F-test instrument relevance	31.21	–	–	21.04	–	–
Hausman test p-value	–	–	[0.245]	–	–	[0.357]
Observations	1,417	1,417	1,417	1,423	1,423	1,423

Note: Control variables are described in Tables 3 and 4.

Bootstrapped standard errors in parentheses. Bootstrap Hausman test is reported.

*** p <0.01, ** p <0.05, * p <0.1

Online Appendix

Table A1. Assessing the validity of different polygenic scores for education based on different p-values

Panel A. Testing the exclusion restriction: coefficient and statistical significance

	$p<1$	$p<5\times10^{-1}$	$p<5\times10^{-2}$	$p<5\times10^{-3}$	$p<5\times10^{-4}$	$p<5\times10^{-5}$	$p<5\times10^{-6}$	$p<5\times10^{-7}$	$p<5\times10^{-8}$
Husband's Years of Education	0.102*	0.1	0.089*	0.03	0.004	-0.052	-0.118**	-0.113**	-0.076
Husband's College	0.038***	0.036***	0.028**	0.012	0.004	-0.002	-0.011	-0.01	-0.005
Wife's Years of Education	0.175**	0.156**	0.126*	0.136**	0.085	-0.021	-0.081	-0.086	-0.036
Wife's College	0.028*	0.025*	0.022*	0.029**	0.012	-0.001	-0.008	-0.009	-0.006

Panel B. Testing instrument relevance: First stage F-test

	$p<1$	$p<5\times10^{-1}$	$p<5\times10^{-2}$	$p<5\times10^{-3}$	$p<5\times10^{-4}$	$p<5\times10^{-5}$	$p<5\times10^{-6}$	$p<5\times10^{-7}$	$p<5\times10^{-8}$
Husband's Years of Education	57.83***	59.24***	57.02***	63.78***	31.21***	14.05***	4.80**	4.75**	2.17
Husband's College	46.60***	47.42***	33.69***	43.56***	21.04***	6.50**	2.39	3.00*	0.46
Wife's Years of Education	54.01***	54.14***	72.81***	48.07***	42.36***	37.62***	27.24***	16.66***	9.63***
Wife's College	54.95***	58.82***	76.36***	56.91***	44.25***	38.19***	28.30***	19.20***	11.12***

Acknowledgments

The Health and Retirement Study is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work: http://dx.doi.org/10.5281/zenodo.22558 This article builds on Okbay et al. (2016) GWAS meta-analysis on educational attainment. The data used for the construction of the genetic risk score were accessed under section 4 of the Data Sharing Agreement of the Social Science Genetic Association Consortium (www.thessgac.org). We gratefully acknowledge Aysu Okbay for providing genome-wide association results excluding HRS from the meta-analysis. Barban acknowledges funding from the European Research Council via an ERC Consolidator Grant SOCIOGENOME (615603 awarded to M. Mills, see www.sociogenome.com).