Barban, Nicola; De Cao, Elisabetta; Oreffice, Sonia; Quintana-Domeque, Climent

Working Paper

Assortative Mating on Education: A Genetic Assessment

IZA Discussion Papers, No. 12563

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Barban, Nicola; De Cao, Elisabetta; Oreffice, Sonia; Quintana-Domeque, Climent (2019) : Assortative Mating on Education: A Genetic Assessment, IZA Discussion Papers, No. 12563, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/207389

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12563

Assortative Mating on Education: A Genetic Assessment

Nicola Barban
Elisabetta De Cao
Sonia Oreffice
Climent Quintana-Domeque

AUGUST 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12563

Assortative Mating on Education: A Genetic Assessment

Nicola Barban
University of Essex

Elisabetta De Cao
London School of Economics

Sonia Oreffice
University of Exeter and IZA

Climent Quintana-Domeque
University of Exeter and IZA

AUGUST 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Assortative Mating on Education: A Genetic Assessment

We investigate assortative mating on education using a sample of couples from the Health and Retirement Study. We estimate a reduced-form linear matching function, which links wife's education to husband's education and both wife's and husband's unobservable characteristics. Using OLS we find that an additional year in husband's education is associated with an average increase in wife's education of 0.4 years. To deal with omitted variable bias due to unobservable characteristics, we use a measure of genetic propensity (polygenic score) for husband's education as an instrumental variable. Assuming that our instrument is valid, our 2SLS estimate suggests that an additional year in husband's education increases wife's education by about 0.5 years. Since greater genetic propensity for educational attainment has been linked to a range of personality and cognitive skills, we allow for the possibility that the exclusion restriction is violated using the plausible exogenous approach by Conley et al. (2012). 'True' assortativeness on education cannot be ruled out, as long as one standard deviation increase in husband's genetic propensity for education directly increases wife's education by less than 0.2 years.

JEL Classification: C36, D1, J1, J12

Keywords: education, genetic scores, instrumental variables, plausibly exogenous, HRS

Corresponding author:
Climent Quintana-Domeque
University of Exeter
Rennes Drive
Exeter, EX4 4PU
United Kingdom
E-mail: C.Quintana-Domeque@exeter.ac.uk

1 Introduction

Assortative mating on education —that individuals with similar education match with one another more frequently than would be expected under a random mating pattern or that partners' educational attainments are positively correlated— has been studied in economics since the seminal work by Becker (1973). 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), while Eika et al. (forthcoming) have shown that assortative mating has been declining over time among college graduates, whereas the low-educated have been increasingly sorting into internally homogeneous marriages.

Given that education is correlated with many other characteristics (e.g., cognitive ability, parental background), assessing whether individuals 'truly' match on education —rather than on other correlated characteristic(s)— and quantifying the extent of 'real' educational assortativeness are relevant empirical challenges. For one thing, assortative mating can play an important role in the transmission of advantage and disadvantage across generations (e.g., Fernandez and Rogerson, 2001). In this paper, we estimate a reduced-form linear matching function —which links education of the wife to education of the husband and the unobservable characteristics of both spouses— using genetic data.

We use a polygenic score for husband's educational attainment as an instrumental variable for husband's education. The polygenic score —a single quantitative measure of genetic predisposition based on genetic variants present in the entire genome (see Plomin et al., 2009)— is constructed to predict educational attainment of married individuals using data from the Health and Retirement Study (HRS), building upon the recent findings from a large scale GWAS of educational attainment (Lee et al., 2018), and following recent work in economics (Barth et al., 2018; Papageorge and Thom, 2016).[1]

[1]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; Domingue et al., 2014; Plomin et al., 2009; Ward et al., 2014).

Our OLS estimates of the matching function show that an additional year in husband's education is associated with an average increase of 0.4 years in wife's education. We also find that a one standard deviation increase in husband's educational attainment polygenic score (EA PGS) increases husband's education by about 0.6 years and wife's education by about half of this magnitude, 0.3. Our IV (2SLS) estimates of the matching function, using husband's EA PGS as the (excluded) instrument for wife's education, suggest that an additional year in husband's education increases wife's education by about 0.5 years.

While the educational attainment polygenic score (EA PGS) is a relevant instrument for education and is 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 the EA PGS as a valid instrumental variable.[2] Polygenic scores for spousal education may affect own education above and beyond their effects on spousal education, that is, polygenic scores for spousal education may affect the unobservable characteristics in the matching function, and hence violate the exclusion restriction.[3]

Greater polygenic propensity for educational attainment has been linked to higher cognitive aptitude, self-control, and interpersonal skills in childhood (Belsky et al., 2016; Rabinowitz et al., 2019), and more recently, to larger brains (Elliott et al., 2018),[4] but also to lower scores on the ADHD (attention deficit hyperactivity disorder) index (de Zeeuw et al., 2014). Since intelligence and personality, amongst other attributes, are relevant in the marriage market (Dupuy and Galichon, 2014; Lundberg, 2012), it is important to assess the consequences of departing from the exclusion restriction.

To allow for the possibility that the exogeneity condition is violated, we relax the exclusion restriction following the approach in Conley et al. (2012) whose implementation is

[2]As discussed by Beauchamp et al. (2011), the use of genetic markers as instrumental variables was anticipated by Davey Smith and Ebrahim (2003) and pioneered in economic research by Ding et al. (2009), who used genes as instrumental variables for health in studying the impact of health on academic outcomes.

[3]Many researchers argue that it is very unlikely (if not impossible) that any markers satisfying the exclusion restriction will ever be found in many economic applications (Conley, 2009; Cawley et al., 2011).

[4]Brain size is positively related with cognitive scores (Elliott et al., 2018).

carefully discussed by Clarke and Matta (2018).[5] In particular, we allow for the husband's EA PGS to have a direct effect on wife's education. Assuming a positive direct effect smaller than 0.2 –i.e., a one standard deviation increase in husband's EA PGS directly increases wife's education by less than 0.2 years– we cannot rule out 'true' positive assortativeness on education. In other words, it seems safe to conclude that people actually match on education.

There is an extensive literature on educational assortative mating which tries to assess the 'real' assortativeness, via adjusting for observable characteristics (Oreffice and Quintana-Domeque, 2010; Chiappori et al., 2016), using within-siblings or within-twins variation (Huang et al., 2009; Giuntella et al., 2019) or instrumental variables (Lefgren and McIntyre, 2006).

Larsen et al. (2015) claim that using the variation in male educational attainment induced by the WWII G.I. Bill may provide the most transparent identification strategy to date: their findings suggest that the additional education received by returning veterans caused them to "sort" into wives with significantly higher levels of education. While theirs is an interesting identification strategy, it 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 (Lefgren and McIntyre, 2006), or data on twins to assess assortative mating and how education is productive in marriage (Huang et al., 2009) or, more recently, on siblings to assess how husband's education affects wife's education (Giuntella et al., 2019).[6]

Our work is also related to studies on genetic assortativeness.[7] These articles use genetic

[5] `plausexog` in STATA: https://ideas.repec.org/c/boc/bocode/s457832.html

[6] 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.

[7] 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.

information from large scale GWASs that are also the core of our analysis. While they 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.[8] Moreover, we find that genetic assortativeness on education polygenic scores is much smaller than assortativeness on education, and that it essentially disappears after controlling for education and population stratification, consistent with recent work by Barth et al. (2018).

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., 2014). 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., 2019), we are the first to rely on molecular data to exploit potential exogenous variation in educational attainment, allowing for the possibility that our instrument violates the exogeneity condition using the approach by Conley et al. (2012), in a marriage market application.

As we shall see, our IV results are valid for a range of small violations 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.[9] Hence, our work complements and expands 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).

The rest of the paper is organized as follows. Section 2 presents the reduced-form linear matching function and how to identify the coefficient of interest. Section 3 defines the polygenic score for education and how to handle potential deviations from IV assumptions. Section 4 describes the data sources, the construction of the polygenic score and presents

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

[9] Recent work by van Kippersluis and Rietveld (2018a) and van Kippersluis and Rietveld (2018b) expands the plausible exogenous approach in Conley et al. (2012) in a world with *heterogeneous* first-stage effects but with *constant* reduced-form effects.

some descriptive statistics. Section 5 contains the OLS estimates of the matching function, first-stage and reduced-form equations, IV estimates of the matching function, and bounds based on Conley et al. (2012)'s approach. Section 6 concludes the paper.

2 A Reduced-Form Matching Function

While several studies have used (and estimated) linear matching functions linking spouses incomes, occupations and/or human capital measures (Lam and Schoeni, 1993, 1994; Ermisch et al., 2006; Oreffice and Quintana-Domeque, 2010), these were based on implicit or explicit statistical (linear) decomposition exercises rather than derived from equilibrium matching models.

Recently, Giuntella et al. (2019) derive a linear matching function from a linear model where assortative mating takes place on a mate desirability index. In addition, they assume that the desirability index is the sum of two components —an observable characteristic (e.g., education) and an unobservable characteristic (e.g., family background)— which are jointly normally distributed, and that individuals prefer to marry those with a high desirability index. These assumptions allow them to obtain the following reduced-form linear matching function:[10]

$$x_i = a_0 + b_0 y_j + c_0 v_j + d_0 u_i, \tag{1}$$

which links education of the wife, x_i, to education of the husband, y_j, and the unobservable characteristics of both spouses, v_j and u_i, and where the parameter b_0 captures the degree of assortative mating on education.

Since v_j and u_i are both unobservable to the econometrician, they get subsumed in the

[10]The online appendix borrows the derivation of the reduced-form linear matching function from Giuntella et al. (2019).

structural error term ϵ_{ij}:

$$x_i = a_0 + b_0 y_j + \epsilon_{ij}, \tag{2}$$

where $\epsilon_{ij} = c_0 v_j + d_0 u_i$. The unconditional OLS estimand of b_0 is given by

$$b_0^{OLS} = \frac{Cov(x_i, y_j)}{Var(y_j)} = b_0 + \frac{Cov(\epsilon_{ij}, y_j)}{Var(y_j)} = b_0 + c_0 \frac{Cov(v_j, y_j)}{Var(y_j)} + d_0 \frac{Cov(u_i, y_j)}{Var(y_j)} \neq b_0. \tag{3}$$

Thus, the unconditional OLS estimand differs from b_0. Indeed, to identify b_0 additional information is required. This can take three main forms: (1) using proxies for v_j and u_i, (2) adding structure on v_j and u_i, and (3) finding (at least) one instrumental variable z_j for y_j.

(1) Using proxies for v_j and u_i. While v_j and u_i are not observable, one might use proxies V_j and U_i instead, such that:

$$v_j = V_j + e_j, \tag{4}$$

$$u_i = U_i + f_i, \tag{5}$$

with $E[e_j] = E[f_i] = E[V_j e_j] = E[U_i f_i] = E[e_j f_i] = 0$. The regression equation becomes

$$x_i = a_0 + b_0 y_j + c_0 V_j + d_0 U_i + \eta_{ij}, \tag{6}$$

with $\eta_{ij} = c_0 e_j + d_0 f_i$. Of course, this approach will be as good as the proxies V_j and U_i are for the unobservable characteristics v_j and u_i. This approach is implicitly used in any empirical analysis using OLS regressions with control variables to assess assortative mating (e.g., Oreffice and Quintana-Domeque, 2010; Chiappori et al., 2016).

(2) Adding structure on v_j and u_i. Another possibility is to use information on pairs of same-sex siblings for husbands (i.e., brothers) and wives (i.e., sisters). Then, the matching equation can be written as

$$x_{is} = a_0 + b_0 y_{js'} + c_0 v_{js'} + d_0 u_{is}, \tag{7}$$

where i denotes the woman within the group of sisters s and j denotes the man within the group of brothers s'. Indeed, Giuntella et al. (2019) assume that the unobservable attribute for the woman i and the man j can be decomposed as follows:

$$v_{js'} = \theta_{s'} + \epsilon_{js'}, \tag{8}$$

$$u_{is} = \phi_s + \varepsilon_{is}, \tag{9}$$

with $E[\theta_{s'}\epsilon_{js'}] = E[\phi_s\varepsilon_{is}] = 0$, $E[y_{js'}\epsilon_{js'}] = E[y_{js'}\varepsilon_{is}] = 0$ and $E[\epsilon_{js'}] = E[\varepsilon_{is}] = 0$. This assumption means that two brothers j and j' in the group s' (resp. two sisters i and i' in the group s) who have the same level of $y = y_j = y_{j'}$ (resp. $x = x_i = x_{i'}$) are (on average) perfect substitutes on the marriage market since they share the same $\theta_{s'}$ (resp. ϕ_s). Under this assumption, the matching equation becomes

$$x_{is} = a_0 + b_0 y_{js'} + c_0 \theta_{s'} + d_0 \phi_s + e_{isjs'}, \tag{10}$$

where $\theta_{s'}$ (resp. ϕ_s) is a vector of same-sex sibling fixed effects (FE) for the husbands (resp. wives) and $e_{isjs'} = c_0\epsilon_{js'} + d_0\varepsilon_{is}$. The conditional on same-sex sibling FE OLS estimand of b_0 is given by

$$b_0^{c,OLS} = \frac{Cov(x_{is}, \tilde{y}_{js'})}{Var(\tilde{y}_{js'})} = b_0, \tag{11}$$

where $\tilde{y}_{js'}$ is the residual of an OLS regression of $y_{js'}$ on husbands' ($\theta_{s'}$) and wives' (ϕ_s) same-sex siblings fixed effects, that is:

$$y_{js'} = \tau_0 + \tau_1\theta_{s'} + \tau_2\phi_s + \tilde{y}_{js'}. \tag{12}$$

Even if the identifying assumption holds in *theory*, one of the key limitations of this approach, already acknowledged by Giuntella et al. (2019), is that including both male and female same-sex sibling fixed effects at the same time is very demanding in *practice*. Thus, the researcher ends up including either husband's or wife's sibling fixed effects, but not both at the same time.

(3) Finding (at least) one instrumental variable z_j for y_j. Suppose that we have information on a potential valid instrument z_j for y_j. In particular, suppose that z_j is a measure of genetic predisposition to higher educational attainment (Lee et al., 2018). Then, the IV estimand of b_0 is given by

$$b_0^{IV} = \frac{Cov(x_i, z_j)}{Cov(y_j, z_j)} = b_0 + \frac{Cov(\epsilon_{ij}, z_j)}{Cov(y_j, z_j)} = b_0 + c_0\frac{Cov(v_j, z_j)}{Cov(y_j, z_j)} + d_0\frac{Cov(u_i, z_j)}{Cov(y_j, z_j)}. \tag{13}$$

The two well-known conditions for instrument validity are the following:

IV1: Relevance. The instrument z_j must be correlated with the endogenous variable y_j:

$$Cov(y_j, z_j) \neq 0$$

IV2: Exogeneity. The instrument z_j must be uncorrelated with the structural error term ϵ_{ij}:

$$Cov(\epsilon_{ij}, z_j) = c_0\frac{Cov(v_j, z_j)}{Cov(y_j, z_j)} + d_0\frac{Cov(u_i, z_j)}{Cov(y_j, z_j)} = 0$$

If these two conditions hold, the IV estimand of b_0 allows us to recover b_0:

$$b_0^{IV} = \frac{Cov(x_i, z_j)}{Cov(y_j, z_j)} = b_0. \tag{14}$$

This is the approach we follow to identify b_0 in this paper. In the next section we discuss the construction of a potentially valid genetic IV, its potential violation of the exclusion restriction and how to relax this last one.

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. Genome-wide association studies (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 identified variants 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 (PGS) 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 PGS to describe the genetic susceptibility to a trait in social sciences (Belsky and Israel, 2014; Schmitz and Conley, 2017). 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). PGSs 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 large.

A PGS 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, that is:

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

where weights w_j are transformations of GWAS coefficients. A polygenic score is therefore a linear combination of the effects of multiple SNPs on the trait of interest. SNPs are not independent in the genome but their occurrence varies according to a block structure called *linkage disequilibrium* (LD). Using unadjusted GWAS coefficients as weights could reduce the accuracy of PGSs and yield to imprecise estimates (Barth et al., 2018). Different methods have been proposed to account for linkage disequilibrium in the construction of polygenic scores.

In this paper, we use a Bayesian method called LDpred (Vilhjálmsson et al., 2015) to derive the weights w_j. Weights are based on the association results from the GWAS on educational attainment by Lee et al. (2018), where HRS has been removed from the analysis,

to avoid overfitting.[11] LDpred assumes a point-normal mixture prior for the distribution of effect sizes and takes into account the correlation structure of SNPs by estimating the LD patterns from a reference sample of unrelated individuals. The weight for each variant is set to be equal to the mean of the posterior distribution (approximated via MCMC simulation) after accounting for LD. LDpred requires an assumption about the fraction of SNPs which are truly associated with the outcome. A common choice for polygenic traits, followed in this study, is to assume that all the SNPs are associated with the outcome of interest (Barcellos et al., 2018; Barth et al., 2018). The scale of PSs depends on the number of SNPs included in the score. For comparability purposes, we standardize a score by subtracting its mean and dividing it by its standard deviation.

Using PGSs 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 significant SNP from the most recent GWAS on educational attainment (Lee et al., 2018) explains around 0.01% of the variation in years of schooling. A linear polygenic score from all measured SNPs explains 10.6% of the same variable in the HRS sample. 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

[11] An alternative method to estimate weights for a polygenic score consists of selecting independent SNPs with a statistical procedure called *pruning*. The selected independent SNPs are then used to calculate the score, avoiding possible bias due to oversampling DNA regions that are highly genotyped. The range of possible values that a PGS 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. Simulation studies have shown that LDpred leads to more precise estimates for polygenic scores in case of highly polygenic traits (Vilhjálmsson et al., 2015)

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

3.2 Genetic IV

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., 2016; Lawlor et al., 2008; Sheehan et al., 2008; Davey Smith and Ebrahim, 2003; Windmeijer et al., 2018). von Hinke Kessler Scholder et al. (2016) and van Kippersluis and Rietveld (2018b) carefully examine the conditions needed for genetic variants to be used as valid instrumental variables.[12]

The reduced-form linear matching function clarifies the necessary requirements, relevance and exogeneity, for a valid instrument.[13] The relevance assumption (**IV1**) requires that the spousal polygenic score for education, z_j, is linearly related with spousal education, y_j. 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; Lee et al., 2018). Moreover, the score predicts education differences between siblings (Rietveld et al., 2014). The exogeneity

[12]von Hinke Kessler Scholder et al. (2016) and Böckerman et al. (2019) use polygenic scores for body mass index as IV. Previous studies based on candidate genes have investigated: 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). More recently, Cawley et al. (2019) investigate whether an individual's BMI is affected by the polygenic risk score for BMI of their full sibling when controlling for the individual's own polygenic risk score for BMI. They do not find evidence for such an effect.

[13]The usual 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* (Davey Smith and Ebrahim, 2003). However, randomization is *not* a sufficient condition to use genetic data as valid instrumental variables, as recently emphasized by Davies et al. (2018) and van Kippersluis and Rietveld (2018b).

assumption (**IV2**) requires that the spousal polygenic score, z_j, is uncorrelated with the structural error term, which as we have seen is a sum of the unobservable attributes for both the wife, u_i, and the husband, v_j.[14]

3.3 Assessing and deviating from IV assumptions

Instrument relevance (**IV1**) can be assessed by means of F-tests with well-known rules of thumb (Staiger and Stock, 1997; Stock and Yogo, 2005). The exogeneity assumption (**IV2**) has two components: independence and exclusion (Angrist and Pischke, 2014). As recently emphasized by van Kippersluis and Rietveld (2018a) and van Kippersluis and Rietveld (2018b) in the context of using polygenic scores as instrumental variables, independence is naturally satisfied when genetic variants are used as IV due to mendelian randomization (Davey Smith and Ebrahim, 2003). However, the exclusion restriction is more difficult to assess. Consider the following equation:

$$x_i = a_0 + b_0 y_j + \gamma z_j + \epsilon_{ij}. \tag{15}$$

The exclusion restriction is satisfied when $\gamma = 0$, however, this cannot be directly assessed, since estimation of γ via OLS will generate biased and inconsistent coefficient estimates.

To allow for the possibility that the exogeneity condition is violated, $\gamma \neq 0$, and that, the husband's polygenic score for education affects wife's education above and beyond husband's education, we follow Conley et al. (2012) and implement "plausibly exogenous" estimation as carefully explained by Clarke and Matta (2018). The standard exogeneity assumption (**IV2**) requires γ to be zero in equation (15). However, when invoking "plausible exogeneity"

[14] 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 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*); (iv) when disruptive influences of the risk factor on the outcome are limited by foetal or post-natal development processes (*Canalization*).

we replace this assumption with the assumption that γ is close to, but not necessarily equal to, 0.[15] In our context, γ can be different from zero because z_j is correlated with u_i, v_j, or both.

This deviation from the exogenity assumption can take different forms: either the support of γ can be assumed or distributional assumptions about γ can be made. We follow the "union of confidence interval" (UCI) approach, which consists in finding bounds for the IV when the exclusion restriction is violated by choosing a range of values for γ.

4 Data Description

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 2012, the HRS genotyped about 20,000 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] van Kippersluis and Rietveld (2018a) and van Kippersluis and Rietveld (2018b) suggests finding an estimate of the direct effect γ based on the reduced-form effect of the instrument for a sample with a zero first-stage. Their approach allows to exploit Mendelian randomization which is pleiotropy-robust. The approach consists in using the estimate of γ (if we reject that $\gamma = 0$) as an input for the plausibly exogenous approach. Their 'beyond plausible exogenous' approach relies on two assumptions: (1) the coexistence of heterogeneous first-stage effects with homogeneous direct effects across the zero-first-stage group and the full sample, and (2) the selection into the zero-first-stage subgroup should not be driven by the husband's EA PGS and wife's education.

[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 15,445 individuals and information for 8,391,857 genetic variants.

The survey interviews the respondents of eligible birth years repeatedly, as well as their married spouses or partners regardless of age. Since we are interested in couples rather than in the longitudinal structure of the data, we build a cross-section. We include any individual interviewed at least once. 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, 15,445 in total, excluding 21,874 respondents from the original survey. We restrict the sample to only White respondents, also excluding Hispanics. We consider only heterosexual couples at their first marriage. In particular, we exclude never married partners, people who are divorced or widowed at the time of the first interview, and people who 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 those in which at least one of the two spouses was born outside the US or born in the US but with missing census division of origin.[18] We also exclude individuals born before 1920 who might have been exposed to the Spanish Flu and born after 1959 which is the end of the last HRS cohort (Mid Baby Boomer). This yields a working sample of 1,562 couples (3,124 individuals).

[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.

The main variables used in our empirical analysis are education (measured as the number of completed years of schooling) and the polygenic scores for education (EA PGSs). As previously discussed, we generated a polygenic score based on the most recent GWAS results on educational attainment available (Lee et al., 2018). 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 software LDpred and PLINK (Vilhjálmsson et al., 2015).[20]

Table 1 provides the basic descriptive statistics for our sample of husbands and wives. These individuals were born between 1920 and 1959. On average, husbands –with 13.6 years of education– are more educated than their wives –with 13.3 years of education. If we compare our sample of husbands and wives with its non-genotyped counterpart (N=2,468), i.e., where at least one of the spouses has not been genotyped, we find that our sample is on average about three years younger, and more educated (0.51 years more for wives, 0.79 years more for husbands). This is consistent with Barth et al. (2018). See Tables A1 and A2 in the online appendix.

[Table 1 about here]

In Table 2, panel A, we present some correlates of years of education and educational attainment polygenic scores (EA PGSs). The correlates indicate that there is a strong positive linear relationship between spousal years of education (0.57, p-value$<$0.0000), consistent with positive assortative mating in education. We also find some evidence, albeit weaker, of assortative mating on the EA PGS (0.13, p-value$<$0.0000). The correlation in

[19]Because of data sharing agreements, results are calculated from association results that exclude also 23andMe from the meta-analysis. 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. LD structure is estimated from the HRS genotypic data (only individuals with European ancestry) using a LD window of $M/3000$, where M is the number of included SNPs. The prior used to construct the score assumes that there is a probability $p = 1$ that a SNP has a non-zero association.

EA PGSs is less than one quarter of the correlation in years of education. Moreover, the correlation between adjusted EA PGSs, obtained after regressing the spousal EA PGS on spousal years of education, and the 10 principal components of the spousal genetic data to account for population stratification (Beauchamp et al., 2011)[21], is much smaller (0.05, p-value=.0496).

In panel B, we present some tabulations to explore assortative mating in both years of education and EA PGSs, both above and below the median. The tabulations for education in the first row indicate that the probability of a husband being low educated if the wife is low educated is 82%, while the probability of a husband being low educated if the wife is high educated is 18%. The second row reveals that the probability of a husband being highly educated if the wife is highly educated is 63% while the probability of him being high educated if the wife is low educated is 37%. This again reveals the presence of positive assortative mating on education ($\chi^2(1) = 332.1$, p-value=0.000). Indeed, the main diagonal (low-low, high-high) contains 73% of couples, while under random educational matching we would expect 50% of couples in the main diagonal.

Finally, in Panel C, the tabulations for (unadjusted) EA PGSs reveal some degree of assortative mating ($\chi^2(1) = 10.2$, p-value=0.001), but much lower than that for education. Indeed, the probability of a husband having a low (high) EA PGS if the wife has a low (high) EA PGS is 54%, while the probability of a husband having a high (low) EA PGS if the wife has a low (high) EA PGS is 46%. The main diagonal (low-low, high-high) contains 54% of couples, while under random score matching we would expect 50% of couples in the

[21]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. Genetic similarity is often correlated with geographical proximity. 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. 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 (Beauchamp et al., 2011; Price et al., 2006). PCA is the most common method used to control for population stratification in a GWAS. In our application, the first ten genetic principal components for each spouse using genome-wide principal components function as ancestry markers.

main diagonal. Focusing on the adjusted EA PGSs, the tabulation reveals that the main diagonal contains 51% of couples, and that we cannot reject that spousal EA PGSs are independent ($\chi^2(1) = 0.4328$, p-value$= 0.511$). Our findings are consistent with Barth et al. (2018), who fail to reject the null hypothesis of random sorting in EA PGSs.

[Table 2 about here]

Interestingly, we find similar correlations and contingency tables for education among the non-genotyped respondents. The correlation is 0.61 (p-value<0.001) and the entries in the contingency table are also very similar or essentially the same for the conditional probability of husband's low education. See Table A3 in the online appendix.

5 OLS and IV estimates

5.1 OLS estimates of the matching equation

We first present our OLS estimates of equation (2) in Table 1, in column 1, and additional versions of it with control variables, in columns 2-5. Column 1 shows that an additional year in husband's education is positively associated with an average increase of 0.46 in the number of years of wife's education, and that 32% of the variation in wife's education is explained by husband's education. This positive correlation is consistent with previous research (e.g., see Table 7 in Chiappori et al., 2018). In column 2 we add demographic controls (i.e., year of birth of each spouse, and 8 indicators for each spouse region (Census division) of birth). The estimated coefficient remains very similar, 0.45, and the explanatory power increases from 32% to 34%.

[Table 3 about here]

From column 3 to column 5 we use genetic information. In column 3 we add the wife's EA PGS for education, without any other control variables, and obtain a similar coefficient for

husband's education, 0.43 (instead of 0.46), and a higher explanatory power, 35% (instead of 32%). A one standard deviation increase in the wife's EA PGS is associated with an increase in the wife's educational attainment of 0.38 years. In column 4 we account for population stratification adding the wife's and husband's ten first principal components of the principal component analysis to genotypic data: the coefficient on husband's education is estimated at 0.42, and the one on the wife's EA PGS at 0.35. Finally, in column 5, we look at the relationship between wife's and husband's education netting out the influence of the wife's EA PGS, population stratification and demographic controls. Our results indicate that an additional year in husband's education is associated with an average increase of 0.41 in the number of years of wife's education.[22] The OLS point estimates range from 0.46 (95% CI: 0.42,0.50), in column 1, to 0.41 (95% CI: 0.37,0.45), in column 5.[23]

5.2 OLS estimates of first-stage and reduced-form equations

Table 4 contains the estimates of the first-stage regression equation, where husband's education is regressed against husband's EA PGS (our plausible exogenous instrument). In the first three columns, we do not adjust for the wife's EA PGS. Column 1, which does not include any controls, shows that a one standard deviation increase in the husband's PS is associated with an average increase in husband's education of 0.67 years. When accounting for population stratification (column 2), the magnitude decreases to 0.62, and after adding demographic characteristics to the population stratification controls (column 3), our point estimate is 0.64.

[Table 4 about here]

When focusing on columns 4-6, which adjust for the wife's EA PGS, we find that a one standard deviation increase in the husband's EA PGS is associated with an average

[22]Following the procedure in Oster (2019), we find that the estimated assortativeness coefficient is zero if the relative importance of unobservable controls u_i and v_j in column 5 is about 72% of that of the observed controls.

[23]Similar qualitative results are obtained when using a binary indicator for college education (Table A4).

increase in husband's education of 0.61 years (column 4). Assuming that polygenic scores are as good as randomly assigned, at least after accounting for population stratification, column 5 shows a substantial causal effect of polygenic scores on educational attainment: a one standard deviation increase in husband's EA PGS (resp. wife's EA PGS) increases average husband's education by 0.58 (resp. 0.39) years. Finally, column 6 shows similar magnitudes, 0.60 (resp. 0.40) netting out the influence of demographic characteristics. The row on instrument relevance reports Kleibergen-Paap rk Wald F statistics above 90, well above and beyond the critical values in Stock and Yogo (2005), so that we conclude that instrument relevance is reasonably satisfied.[24]

In Table 5 we turn to the reduced form estimates, where wife's education is regressed against husband's EA PGS (our plausible exogenous instrument). As in Table 4, we present two set of estimates, without adjusting (columns 1-3) and adjusting for the wife's EA PGS (columns 4-6). Column 1, which does not include any controls, shows that a one standard deviation increase in the husband's EA PGS is associated with an average increase in wife's education of 0.38 years. After adding population stratification controls (column 2), our point estimate decreases to 0.32, and after including demographic characteristics in addition to the controls for population stratification (column 3), our point estimate becomes 0.34.

[Table 5 about here]

We then shift our attention to reduced-form estimates adjusted for the wife's EA PGS (columns 4-6). Column 4 shows that a one standard deviation increase in husband's EA PGS (resp. wife's EA PGS) is associated with an average increase in wife's education of 0.31 (resp. 0.56) years. In column 5 we assume conditional random assignment of polygenic scores, after accounting for population stratification, and find that a one standard deviation increase in husband's (resp. wife's) EA PGS increases wife's education by 0.28 (resp. 0.51) years. Similar effects are found after netting out the influence of demographic characteristics,

[24]Similar qualitative results are obtained when using a binary indicator for college education (Table A5).

0.29 (resp. 0.52).[25]

5.3 IV estimates of the matching equation

If our instruments are *valid* (i.e., relevant and exogenous), the causal effect of husband's education on wife's education is given by the reduced-form coefficient on the husband's EA PGS inflated (divided) by the first-stage coefficient on the husband's EA PGS. Looking at the regression conditional on the wife's EA PGS estimates in Tables 4 and 5, columns 4-6, these ratios are 0.509 (column 4 divided by column 1), 0.473 (column 5 divided by column 2) and 0.488 (column 6 divided by column 3), which are well-known to be numerically equivalent to the IV point estimates using 2SLS displayed in Table 6. An additional year in husband's education increases average wife's educational attainment by about 0.5 years: in words, and in our matching context, this means that if we take two men, who are observationally equivalent in the marriage market, and we increase the educational attainment of one of them by one more year of education, the one with higher education will be expected to have a wife with about half a year more of education.[26]

[Table 6 about here]

5.4 Plausible exogenous IV estimates

As discussed in the introduction, polygenic scores for spousal education may affect own education above and beyond their effects on spousal education, and hence violate the exclusion restriction. In this subsection, we relax the exclusion restriction following the "union of confidence interval" (UCI) approach developed by Conley et al. (2012). The UCI approach consists in finding bounds for the IV when the exclusion restriction is violated by choosing a range of values for γ.

[25]Similar qualitative results are obtained when using a binary indicator for college education (Table A6).
[26]Similar qualitative results are obtained when using a binary indicator for college education (Table A7).

Table 7 compares the 95% CI intervals/bounds for OLS, IV and UCI estimates, for models that account for the wife's EA PGS. Both OLS and IV estimates generate a similar 95% lower bound (0.334 for IV, column 2; 0.366 for OLS, column 3). By construction, since we choose a range of values for γ between 0 and 0.2, the 95% upper bounds in the IV and the UCI cases are exactly the same. The bite of the UCI approach comes from $\gamma = 0.2$. In that case, we can see that the 95% lower bound ranges from -0.021 (column 2) to 0.044 (column 1).

[Table 7 about here]

Our interpretation of Table 7 is that we need to have a sufficiently large and positive γ, $\gamma \geq 0.2$, to rule out 'true' positive assortative mating on education. In words, a one standard deviation in husband's EA PGS must directly increase wife's education in 0.2 or more years to nullify the observed educational assortativeness.

Figure 1 below displays the UCI 95% range of lower bounds for column 3 in Table 7. For completeness, we also present Figure 2, which displays the UCI 95% bounds analysis for column 3 in Table 7 when $\gamma \in [-0.2, 0.2]$.

[Figure 1 about here]

[Figure 2 about here]

All in all, we interpret our findings as consistent with people actually matching on education. Indeed, the results are very similar using OLS and IV, and the bounds analysis suggests that, for mild violations of the exclusion restriction, $\gamma < 0.2$, our IV findings are able to reveal 'true' positive assortative matching on education.

5.5 Discussion

The reduced-form linear matching function we estimate is based on Giuntella et al. (2019), but both the identification strategy and the data used in our analysis differ from theirs.

First, as previously discussed, we use an IV strategy while Giuntella et al. (2019) use a siblings-FE strategy. Both strategies have their pros and cons. While a sibling-FE identification strategy can be certainly more transparent than an IV, it can only circumvent omitted variable bias so long as the unobservable characteristics (of both spouses) are constant within pairs of siblings. Of course, the advantage of an IV strategy, namely being able to deal with omitted variable bias due to any type of unobservable characteristic(s), comes at a key cost: the reliability of the untestable exclusion restriction. In this paper, we have relaxed the exclusion restriction using the Conley et al. (2012) approach.

Second, while we use a sample from a nationally representative survey (HRS), Giuntella et al. (2019) use a sample from the administrative birth records from Florida. Our results refer to an HRS sample of individuals who are on average 70 years old, still alive, who got married on average 40 years ago and have been married to each other ever since. Instead, Giuntella et al. (2019) findings are for a sample of parents born in the state of Florida whose children were born in Florida and whose brothers (or sisters) were born in Florida and whose children were born in Florida too.

Given all these differences, both in methods and in samples, it is reassuring that both studies find evidence on 'true' assortativeness on education.

6 Conclusions

This is the first paper to present a genetic-IV strategy to estimate the causal effect of education in the marriage market. Our IV (2SLS) estimates of the matching function suggest that an additional year in husband's education increases wife's education by about 0.5 years. Even if the husband's educational attainment polygenic score (EA PGS) has a direct effect on wife's education over and above husband's education, we cannot rule out a positive causal effect of husband's on wife's education, so long as one standard deviation increase in husband's EA PGS directly increases wife's education by less than 0.2 years.

Future research should try to pin down the exact mechanism behind the causal effect of husband's education on wife's education: Do more educated husbands become more likely to encounter potential wives that are more educated? And/or do more educated husbands become more attractive to more educated wives, holding constant the likelihood of meeting an educated spouse? In other words, is the causal effect of husband's education on wife's education mainly due to search or preferences? (Bruze, 2011).

References

Abramitzky, R., A. Delavande, and L. Vasconcelos (2011): "Marrying up: the role of sex ratio in assortative matching," *American Economic Journal: Applied Economics*, 3, 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 J. S. Pischke (2014): *Mastering'metrics: The path from cause to effect*, Princeton University Press.

Barcellos, S. H., L. S. Carvalho, and P. Turley (2018): "Education can reduce health differences related to genetic risk of obesity," *Proceedings of the National Academy of Sciences*, 115, E9765–E9772.

Barth, D., N. W. Papageorge, and K. Thom (2018): "Genetic Endowments and Wealth Inequality," *Journal of Political Economy*.

Böckerman, P., J. Cawley, J. Viinikainen, T. Lehtimäki, S. Rovio, I. Seppälä, J. Pehkonen, and O. Raitakari (2019): "The effect of weight on labor market outcomes: An application of genetic instrumental variables," *Health Economics*, 28, 65–77.

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.

Belsky, D. W., T. E. Moffitt, D. L. Corcoran, B. Domingue, H. Harrington, S. Hogan, R. Houts, S. Ramrakha, K. Sugden, B. S. Williams, R. Poulton, and A. Caspi (2016): "The Genetics of Success: How Single- Nucleotide Polymorphisms Associated With Educational Attainment Relate to Life-Course Development," *Psychological Science*, 27, 957–972.

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.

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.

Cawley, J., E. Han, J. Kim, and E. C. Norton (2019): "Testing for family influences on obesity: The role of genetic nurture," *Health Economics*, 28, 937–952.

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.

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 (2016): "Black–White Marital Matching: Race, Anthropometrics, and Socioeconomics," *Journal of Demographic Economics*, 82, 399–421.

Chiappori, P.-A., S. Oreffice, and C. Quintana-Domeque (2018): "Bidimensional Matching with Heterogeneous Preferences: Education and Smoking in the Marriage Market," *Journal of the European Economic Association*, 16, 161–198.

Clarke, D. and B. Matta (2018): "Practical considerations for questionable IVs," *Stata Journal*, 18, 663–691(29).

Conley, D. (2009): "The Promise and Challenges of Incorporating Genetic Data into Longitudinal Social Science Surveys and Research," *Biodemography and Social Biology*, 55, 238–251.

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 (2014): "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*, 113, 6647–6652.

Conley, T. G., C. B. Hansen, and P. E. Rossi (2012): "Plausibly Exogenous," *The Review of Economics and Statistics*, 94, 260–272.

Davey Smith, G. and S. Ebrahim (2003): "'Mendelian randomization': can genetic

epidemiology contribute to understanding environmental determinants of disease?" *International Journal of Epidemiology*, 32, 1–22.

Davies, N. M., M. V. Holmes, and G. Davey Smith (2018): "Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians," *BMJ*, 362.

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, E. A. Ehliand, G. E. Daviesand, J. J. Hudziakand, SSGAC, C. A. Rietveldand, M. M. Groen-Blokhuisand, J. J. Hottengaand, E. J. de Geusand, and D. 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.

Eika, L., M. Mogstad, and B. Zafar (forthcoming): "Educational Assortative Mating and Household Income Inequality," *Journal of Political Economy*.

Elliott, M. L., D. W. Belsky, K. Anderson, D. L. Corcoran, T. Ge, A. Knodt, J. A. Prinz, K. Sugden, B. Williams, D. Ireland, R. Poulton, A. Caspi, A. Holmes, T. Moffitt, and A. R. Hariri (2018): "A Polygenic Score for Higher Educational Attainment is Associated with Larger Brains," *Cerebral Cortex*, 29, 3496–3504.

Ermisch, J., M. Francesconi, and T. Siedler (2006): "Intergenerational mobility and marital sorting," *American Economic Review*, 116, 659–679.

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.

Giuntella, O., G. La Mattina, and C. Quintana-Domeque (2019): "Assortative mating on human capital: at birth and in adulthood?" *mimeo, University of Exeter*.

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.

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 (2016): "Instrumental Variables Estimation With Some Invalid Instruments and its Application to Mendelian Randomization," *Journal of the American Statistical Association*, 111, 132–144.

Lam, D. and R. F. Schoeni (1993): "Effects of family background on earnings and returns to schooling: evidence from Brazil," *Journal of Political Economy*, 101, 710–40.

——— (1994): "Family ties and labor market in the United States and Brazil," *Journal of Human Resources*, 29, 1235–58.

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.

Lee, J. J., R. Wedow, A. Okbay, E. Kong, O. Maghzian, M. Zacher, M. Johannesson, P. Koellinger, P. Turley, P. Visscher, et al. (2018): "Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment," *Nat Genet, in press*, 1112–1121.

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.

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.

Oreffice, S. and C. Quintana-Domeque (2010): "Anthropometry and socioeconomics among couples: Evidence in the United States," *Economics and Human Biology*, 8, 373–384.

Oster, E. (2019): "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 37, 187–204.

Papageorge, N. W. and K. Thom (2016): "Genes, Education, and Labor Market Outcomes: Evidence from the Health and Retirement Study," *IZA, DP n. 10200*.

Plomin, R., C. M. A. Haworth, and O. S. P. Davis (2009): "Common disorders are quantitative traits," *Nature Reviews Genetics*, 10, 872–878.

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.

Rabinowitz, J. A., S. I. Kuo, W. Felder, R. J. Musci, A. Bettencourt, K. Benke, D. Y. Sisto, E. Smail, G. Uhl, B. S. Maher, A. Kouzis, and N. S. Ialongo (2019): "Associations between an educational attainment polygenic score with educational attainment in an African American sample," *Genes, Brain and Behavior*, 18, e12558.

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 (2017): "Modeling Gene-Environment Interactions With Quasi-Natural Experiments," *Journal of Personality*, 85, 10–21.

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.

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

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.

van Kippersluis, H. and C. A. Rietveld (2018a): "Beyond plausibly exogenous," *The Econometrics Journal*, 21, 316–331.

——— (2018b): "Pleiotropy-robust Mendelian randomization," *International Journal of Epidemiology*, 47, 1279–1288.

Vilhjálmsson, B. J., J. Yang, H. K. Finucane, A. Gusev, S. Lindström, S. Ripke, G. Genovese, P.-R. Loh, G. Bhatia, R. Do, et al. (2015): "Modeling linkage disequilibrium increases accuracy of polygenic risk scores," *The American Journal of Human Genetics*, 97, 576–592.

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.

Windmeijer, F., H. Farbmacher, N. Davies, and G. D. Smith (2018): "On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments," *Journal of the American Statistical Association*, 0, 1–12.

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," *Proceedings of the National Academy of Sciences*, 112, 13621–13626.

Table 1: Descriptive statistics

	Mean	SD	Min	Max
Husband's Year of Birth	1937.89	9.04	1920	1959
Husband's Years of Education	13.62	2.69	2	17
Husband's EA PGS	0.19	1.07	–3.20	3.83
Wife's Year of Birth	1940.12	8.92	1920	1959
Wife's Years of Education	13.32	2.19	3	17
Wife's EA PGS	0.15	1.05	–3.19	3.16

Note: The number of observations is 1,562. The descriptive statistics are based on white non-Hispanic couples in their first marriage, with at most 10 years of age difference and born in the US. Individuals born between 1920 and 1959. Both spouses have been interviewed at least once and provided DNA sample.

Table 2: Correlations and contingency tables for education and educational attainment polygenic scores

Panel A. Correlations	Husband's Education	Wife's Education	Husband's PS	Wife's PS
Husband's Education	1			
	[0.000]			
Wife's Education	0.5647	1		
	[0.000]	[0.000]		
Husband's EA PGS	0.2648	0.1867	1	
	[0.000]	[0.000]	[0.000]	
Wife's EA PGS	0.1986	0.2881	0.1319	1
	[0.000]	[0.000]	[0.000]	[0.000]

Correlation between Adjusted Wife's EA PGS and Adjusted Husband's EA PGS	0.0496
	[0.0498]

Panel B. Contingency table for education, conditional probability of husband's education

	Wife's Low Education	Wife's High Education
Husband's Low Education	82.17	17.83
Husband's High Education	36.96	63.04
Pearson test	$\chi^2(1) = 332.1$	
	[0.000]	

Panel C. Contingency table for EA PGSs, conditional probability of husband's EA PGSs

	Unadjusted		*Adjusted*	
	Wife's Low EA PGS	Wife's High EA PGS	Wife's Low EA PGS	Wife's High EA PGS
Husband's Low EA PGS	54.03	45.97	50.83	49.17
Husband's High EA PGS	45.97	54.03	49.17	50.83
Pearson test	$\chi^2(1) = 10.2$		$\chi^2(1) = 0.4328$	
	[0.001]		[0.511]	

Note: In Panel A adjusted husband's (wife's) EA PGS is the residual from a regression of the husband's (wife's) EA PGS on husband's (wife's) years of education and 10 principal components of the husband's (wife's) genetic data. In Panels B and C: low is defined as below the median and high is defined as above the median; Each cell reports the conditional probability of husband's education (EA PGS) given his wife's education (EA PGS). The row probabilities sum to 100. Adjusted conditional probabilities are based on the residual EA PGSs. p-values are reported in brackets.

Table 3: OLS estimates of the Matching Function

	Dependent variable: Wife's Education				
	(1)	(2)	(3)	(4)	(5)
Husband's Education	0.461***	0.448***	0.432***	0.422***	0.407***
	(0.020)	(0.021)	(0.020)	(0.021)	(0.021)
Wife's EA PGS			0.383***	0.350***	0.365***
			(0.043)	(0.044)	(0.044)
PCAs	No	No	No	Yes	Yes
Demographics	No	Yes	No	No	Yes
Observations	1,562	1,562	1,562	1,562	1,562
R-squared	0.319	0.337	0.351	0.368	0.387

Note: Robust standard errors in parentheses. PCAs: 10 first principal components of the husband's (wife's) genetic data. Demographics: year of birth of the wife, year of birth of the husband, 8 indicators of the wife's region (Census division) of birth and 8 indicators of the husband's region (Census division) of birth.
*** p<0.01, ** p<0.05, * p<0.1

Table 4: OLS estimates of the First Stage

	Dependent variable: Husband's Education					
	(1)	(2)	(3)	(4)	(5)	(6)
Husband's EA PGS	0.667***	0.620***	0.636***	0.611***	0.583***	0.602***
	(0.060)	(0.062)	(0.061)	(0.060)	(0.061)	(0.060)
Wife's EA PGS				0.426***	0.388***	0.398***
				(0.060)	(0.062)	(0.061)
PCAs	No	Yes	Yes	No	Yes	Yes
Demographics	No	No	Yes	No	No	Yes
Observations	1,562	1,562	1,562	1,562	1,562	1,562
R-squared	0.070	0.106	0.144	0.097	0.128	0.166
Instrument relevance	122.816	100.857	109.241	105.226	91.265	99.920

Note: Instrument relevance: Kleibergen-Paap rk Wald F statistic. See footer in Table 3.

Table 5: OLS estimates of the Reduced Form

	Dependent variable: Wife's Education					
	(1)	(2)	(3)	(4)	(5)	(6)
Husband's EA PGS	0.384***	0.324***	0.338***	0.311***	0.276***	0.294***
	(0.050)	(0.050)	(0.049)	(0.048)	(0.049)	(0.048)
Wife's EA PGS				0.561***	0.511***	0.523***
				(0.050)	(0.051)	(0.050)
PCAs	No	Yes	Yes	No	Yes	Yes
Demographics	No	No	Yes	No	No	Yes
Observations	1,562	1,562	1,562	1,562	1,562	1,562
R-squared	0.035	0.083	0.127	0.106	0.139	0.185

Note: See footer in Table 3.

Table 6: IV (2SLS) estimates of the Matching Function

	Dependent variable: Wife's Education					
	(1)	(2)	(3)	(4)	(5)	(6)
Husband's Education	0.576***	0.522***	0.531***	0.509***	0.473***	0.488***
	(0.063)	(0.068)	(0.066)	(0.067)	(0.071)	(0.069)
Wife's EA PGS				0.344***	0.327***	0.329***
				(0.054)	(0.053)	(0.053)
PCAs	No	Yes	Yes	No	Yes	Yes
Demographics	No	No	Yes	No	No	Yes
Observations	1,562	1,562	1,562	1,562	1,562	1,562

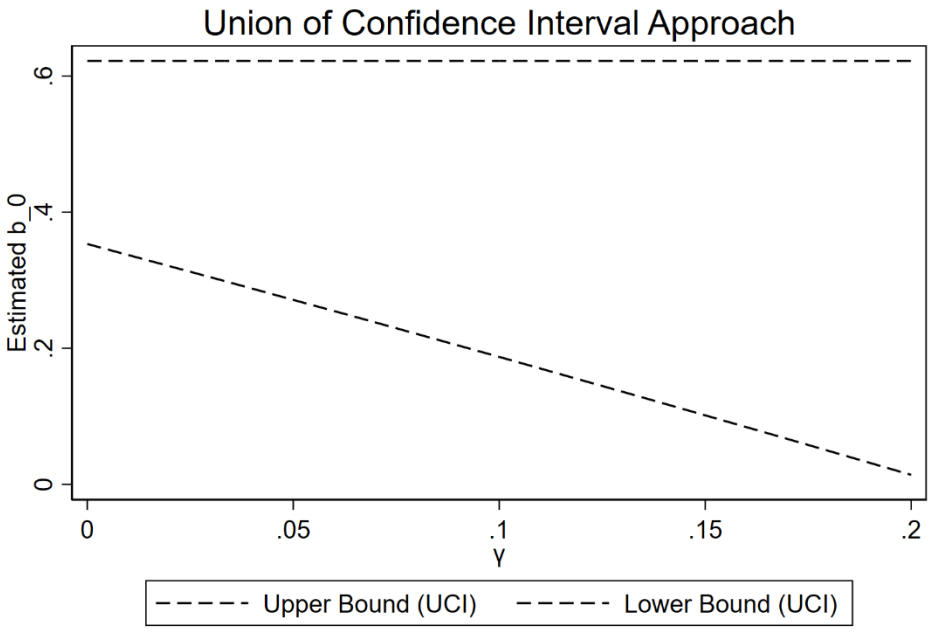
Note: (Excluded) instrumental variable: Husband's EA PGS. See footer in Table 3.

Table 7: Comparison of assortativeness estimates – OLS, IV (2SLS) and IV (UCI) bounds

	(1)	(2)	(3)
OLS 95% CI	[0.391,0.472]	[0.382,0.463]	[0.366,0.449]
IV (2SLS) 95% CI	[0.378, 0.640]	[0.334,0.611]	[0.353,0.622]
IV (UCI) 95% bounds with $\lambda \in [0, 0.2]$	[0.044,0.640]	[–0.021,0.611]	[0.014,0.622]
PCAs	No	Yes	Yes
Demographics	No	No	Yes

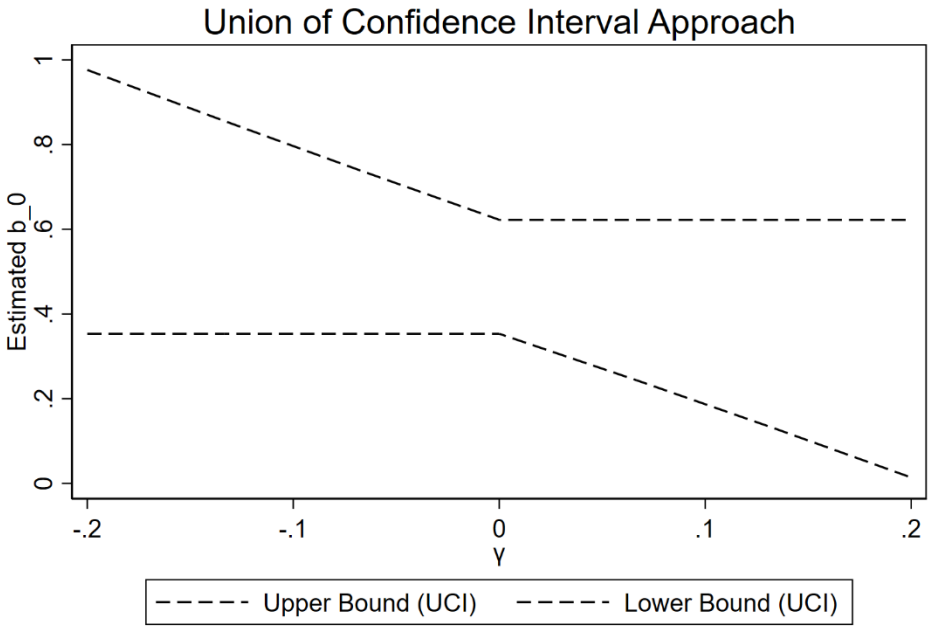
Note: All the estimated CI/bounds are based on regressions that adjust for the Wife's EA PGS. IV bounds are based on the approach developed by Conley et al. (2012) and estimated using the `plausexog` STATA command as described by Clarke and Matta (2018): IV (UCI: Union of Confidence Interval) approach consists in finding bounds for the IV when the exclusion restriction is violated ($\lambda \neq 0$) by choosing a range of values for λ, in our case, with a minimum of 0 and a maximum of 0.2.

Figure 1:



Note: This graph was generated using plausexog (Clarke and Matta, 2018).

Figure 2:



Note: This graph was generated using plausexog (Clarke and Matta, 2018).

Online Appendix (Not for publication)

Derivation of the matching function

The following derivation is borrowed from Giuntella et al. (2019). They consider two populations (men and women) of equal size, normalized to 1. Individuals match on mate desirability: I_i and J_j represent the mate desirability index of female i and male j, respectively.

A1: Additive separability. The mate desirability index is additively separable in two components:

$$I_i = \gamma x_i + u_i \tag{1}$$

$$J_j = \delta y_j + v_j \tag{2}$$

where x_i (resp. y_j) is an attribute observable to the social scientist (e.g., education) and u_i (resp. v_j) is an attribute not observable to the social scientist (e.g., family background).

A2: Normality. Observable and unobservable attributes are jointly normally distributed:

$$\begin{pmatrix} x_i \\ u_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x \\ \mu_u \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_{x,u} \\ \sigma_{x,u} & \sigma_u^2 \end{pmatrix} \right]$$

and

$$\begin{pmatrix} y_j \\ v_j \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_y \\ \mu_v \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{y,v} \\ \sigma_{y,v} & \sigma_v^2 \end{pmatrix} \right].$$

Hence, the mate desirability indices are normally distributed, $J_j \sim N(\mu_J, \sigma_J^2)$ and $I_i \sim N(\mu_I, \sigma_I^2)$.

A3: Marital surplus function. The marital surplus –the difference in the utility generated by a couple formed by individuals i and j with desirability indices I_i and J_j and the utility levels of i and j when single– is produced according to the function $h(I_i, J_j)$. This function is twice continuously differentiable, strictly increasing in both I and J (i.e., $h_I > 0$ and $h_J > 0$) and strictly super-modular (i.e., $h_{IJ} > 0$).

A4: There are no search frictions. The environment is frictionless: each woman (resp. man) is assumed to have free access to the pool of all potential male (resp. female) spouses, with perfect knowledge of the characteristics of each other.

Given that there are no search frictions (**A4**), matching arises due to preferences, and the assumption on the marital surplus (**A3**) function guarantees PAM on the desirability index:[1]"high desirable men marry high desirable women". Men with a high value of J will marry women with a high value of I, i.e., the fraction of men above J will marry the fraction of women above I. Moreover, given the normality assumption (**A2**), one can write:

$$1 - \Phi\left(\frac{J_j - \mu_J}{\sigma_J}\right) = 1 - \Phi\left(\frac{I_i - \mu_I}{\sigma_I}\right), \tag{3}$$

where Φ is the standard normal cumulative distribution function. Hence, they have:

$$I_i = \left\{\mu_I - \mu_J\left(\frac{\sigma_I}{\sigma_J}\right)\right\} + \left(\frac{\sigma_I}{\sigma_J}\right) J_j. \tag{4}$$

[1]The assumption that marital surplus is increasing in both arguments guarantees PAM in a non-transferable utility context. The assumption that the cross-derivative is strictly positive gurantees PAM in a transferable utility context.

In compact notation, they write:

$$I_i = \alpha + \beta J_j. \tag{5}$$

Finally, given additive separability (**A1**), they can substitute (1) and (2) into (5), and express x_i (female observable characteristic) in terms of y_j (male observable characteristic) and the unobservable characteristics of both, u_i and v_j:

$$x_i = a_0 + b_0 y_j + c_0 v_j + d_0 u_i. \tag{6}$$

This is the matching function, linking the observable characteristic of the wife, x_i, to the observable characteristic of the husband, y_j, and the unobservable characteristics of both spouses, v_j and u_i. The parameter b_0 captures the degree of assortative mating with respect to the observable characteristic.

References

Giuntella, O., G. La Mattina, and C. Quintana-Domeque (2019): "Assortative mating on human capital: at birth and in adulthood?" *mimeo, University of Exeter.*

Table A1: Descriptive statistics in the non-genotyped sample

		Mean	SD	Min	Max
Husband's Year of Birth	2,468	1934.88	9.91	1920	1959
Husband's Years of Education	2,464	12.83	2.89	1	17
Wife's Year of Birth	2,468	1937.24	9.73	1920	1959
Wife's Years of Education	2,461	12.81	2.28	1	17

Note: The descriptive statistics are based on white non-Hispanic couples in their first marriage, with at most 10 years of age difference and born in the US. Individuals born between 1920 and 1959. Couples in which at least one of the spouses has not been genotyped.

Table A2: Average differences between genotyped and non-genotyped samples

	Difference (Genotyped – Non-Genotyped)	Standardized difference
Husband's Year of Birth		
	3.01***	0.311
	(0.304)	
Husband's Years of Education		
	0.792***	0.279
	(0.090)	
Wife's Year of Birth		
	2.87***	0.302
	(0.299)	
Wife's Years of Education		
	0.513***	0.227
	(0.072)	

Note: In the first column, each row displays the coefficient (and robust standard error) of an OLS regression of the variable in each row on a constant and an indicator of genotyped sample. The second column displays the standardized difference, where the variable in each row has been standardized to have mean 0 and standard deviation 1.
*** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1

Table A3: Correlations and contingency tables for education in the non-genotyped sample

Panel A. Correlation	Husband's Education
Wife's Education	0.6140
	[0.000]

Panel B. Contingency table for education, conditional probability of husband's education

	Wife's Low Education	Wife's High Education
Husband's Low Education	82.11	17.89
Husband's High Education	33.55	66.45
Pearson test	$\chi^2(1) = 599.138$	
	[0.000]	

Note: N=2,548. In Panel B: low is defined as below the median and high is defined as above the median; Each cell reports the conditional probability of husband's education given his wife's education. The row probabilities sum to 100. p-values are reported in brackets.

Table A4: OLS estimates of the Matching Function

	Dependent variable: Wife's College				
	(1)	(2)	(3)	(4)	(5)
Husband's College	0.435***	0.428***	0.408***	0.401***	0.392***
	(0.023)	(0.023)	(0.023)	(0.024)	(0.024)
Wife's EA PGS			0.071***	0.064***	0.068***
			(0.009)	(0.009)	(0.009)
PCAs	No	No	No	Yes	Yes
Demographics	No	Yes	No	No	Yes
Observations	1,562	1,562	1,562	1,562	1,562
R-squared	0.239	0.262	0.269	0.291	0.313

Note: Robust standard errors in parentheses. PCAs: 10 first principal components of the husband's (wife's) genetic data. Demographics: year of birth of the wife, year of birth of the husband, 8 indicators of the wife's region (Census division) of birth and 8 indicators of the husband's region (Census division) of birth.
*** p<0.01, ** p<0.05, * p<0.1

Table A5: OLS estimates of the First Stage

	Dependent variable: Husband's College					
	(1)	(2)	(3)	(4)	(5)	(6)
Husband's EA PGS	0.111***	0.103***	0.104***	0.103***	0.098***	0.099***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Wife's EA PGS				0.064***	0.059***	0.062***
				(0.011)	(0.011)	(0.011)
PCAs	No	Yes	Yes	No	Yes	Yes
Demographics	No	No	Yes	No	No	Yes
Observations	1,562	1,562	1,562	1,562	1,562	1,562
R-squared	0.062	0.094	0.109	0.081	0.110	0.126
Instrument relevance	109.58	88.367	90.178	93.144	79.340	81.877

Note: Instrument relevance: Kleibergen-Paap rk Wald F statistic. See footer in Table A4.

Table A6: OLS estimates of the Reduced Form

	Dependent variable: Wife's College					
	(1)	(2)	(3)	(4)	(5)	(6)
Husband's EA PGS	0.069***	0.056***	0.059***	0.056***	0.047***	0.051***
	(0.010)	(0.010)	(0.010)	(0.009)	(0.010)	(0.009)
Wife's EA PGS				0.095***	0.086***	0.091***
				(0.010)	(0.010)	(0.010)
PCAs	No	Yes	Yes	No	Yes	Yes
Demographics	No	No	Yes	No	No	Yes
Observations	1,562	1,562	1,562	1,562	1,562	1,562
R-squared	0.030	0.072	0.104	0.084	0.115	0.150

Note: See footer in Table A4.

Table A7: IV (2SLS) estimates of the Matching Function

	Dependent variable: Wife's College					
	(1)	(2)	(3)	(4)	(5)	(6)
Husband's College	0.617***	0.539***	0.563***	0.547***	0.486***	0.516***
	(0.078)	(0.083)	(0.083)	(0.083)	(0.087)	(0.087)
Wife's EA PGS				0.060***	0.058***	0.059***
				(0.011)	(0.011)	(0.011)
PCAs	No	Yes	Yes	No	Yes	Yes
Demographics	No	No	Yes	No	No	Yes
Observations	1,562	1,562	1,562	1,562	1,562	1,562

Note: (Excluded) instrumental variable: Husband's EA PGS. See footer in Table A4.