ISSN 1471-0498

DEPARTMENT OF ECONOMICS
DISCUSSION PAPER SERIES

Assortative Mating on Education: A Genetic Assessment

Nicola Barban, Elisabetta De Cao, Sonia Oreffice,

and Climent Quintana-Domeque

Number 791
March 2016

Manor Road Building, Oxford OX1 3UQ

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, IAE-CSIC and IZA*

March 2016

Abstract

Social scientists have overwhelmingly documented a strong and increasing educational homogamy between spouses. When estimating sorting by education, the presence of measurement error in the education variables or random factors in the matching process may underestimate the actual degree of assortative mating, simultaneity bias may overestimate it, while omitting other individual characteristics relevant in the marriage market may under- or overestimate it. We address these issues using an instrumental variables approach based on exploiting *genetic variation* in polygenic scores and controlling for population stratification. Specifically, we instrument *spousal* education with his/her educational polygenic score while controlling for *own* educational polygenic score. If the exclusion restriction is *satisfied*, our findings indicate that (1) assortative mating is *underestimated* when using OLS, and that (2) male education is correlated with other matching-relevant socioeconomic characteristics, while female education is productive *per se* in the matching. If the exclusion restriction is *not* satisfied, our evidence is consistent with (2). This suggests that individual socioeconomic attractiveness in the marriage market is *multidimensional* for men, but can be summarized with education for women.

JEL Classification Codes: D1, J1, J12.
Keywords: Matching, Years of Education, College, Polygenic Scores, HRS.

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

1 Introduction

Assortative mating on education. Assortative mating in the marriage market 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). Chiappori and Salanié (forthcoming) 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. However, a more immediate concern is to make sure that the actual degree of assortative mating in the data is correctly measured, neither under- nor over-estimated.

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 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; Siow, 2015). Consider the analysis of assortative mating on education: The presence of measurement error in the education variable(s) or random factors in the matching process may underestimate the actual degree of assortative mating, simultaneity bias may overestimate it, while omitting other individual characteristics relevant in the marriage market may under- or overestimate it. 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.

This paper and its main findings. We address these estimation issues by reassessing assortative mating on education exploiting *genetic variation*, using information on a variety of genes known to be related to this socioeconomic attribute. We construct a genetic score designed 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 (Rietveld et al., 2013). 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 (Belsky et al., 2012; Conley et al., 2015; Ward et al., 2014). Specifically, we use *spousal* PSs as instrumental variables for *spousal* education, controlling for *own* PS, to reassess assortative mating in the marriage market.

Assuming that the exclusion restriction holds, we find that the actual degree of positive assortative mating (PAM) on education is typically *underestimated*. The ratios of the estimated OLS-IV coefficients on wife's year of education and on wife's college degree indicator are 0.81 and 0.73, respectively; those on husband's years of education and on husband's college degree indicator are 0.63 and 0.56, respectively. While we cannot reject the hypotheses that the OLS and IV coefficients on wife's year of education and college degree indicator are the same, we do reject the hypotheses that the OLS and IV coefficients on husband's year of education and college degree indicator are the same at the 5% level of statistical significance.

Regardless of the validity of the exclusion restriction, our evidence suggests an important gender difference in the role that education may play in shaping individual socioeconomic attractiveness in the marriage market and ultimately in the measured PAM. We show that this differential role of education is consistent with female education being "what men really value" in a potential spouse, whereas women may also value other male characteristics correlated with education, highlighting the importance of accounting for multiple dimensions in the marriage market (Chiappori et al., 2012, 2015; Dupuy and Galichon, 2014). At least for older cohorts (HRS data), women's socioeconomic attractiveness can be summarized by education while men's attractiveness appears to be multidimensional.

The main purpose of our paper is to show how an IV approach can be helpful in the analysis of matching patterns in (at least) two *different* ways: it can help us to *quantify*

the actual degree of assortativeness (on education), as long as the exclusion restriction is *satisfied*; and, *regardless* of the validity of the exclusion restriction, it may allow us to assess whether assortative mating in the marriage market takes place along additional characteristics correlated with education. Given that our instrument is "very" relevant, a small violation of the exclusion restriction is unlikely to seriously bias our IV estimates (Conley et al., 2012). Both the actual quantification of assortativeness and the understanding of the multidimensionality of matching in the marriage market are important steps 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 study and estimate the degree of assortative mating in the marriage market, while complementing recent research on genetic assortative mating. 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.

Using data from the HRS, Domingue et al. (2014) find that spouses are more genetically similar than two people chosen at random, and show that genetic assortative mating is one third of the magnitude of educational assortative mating. Guo et al. (2014) also find a positive similarity in genomic assortment in married couples by using the HRS and the

Framingham Heart study. Both 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.

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). We complement 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) in two main ways. First, we consider one polygenic score that contains *all* the information coming from the markers of interest instead of using only one or few genetic variants. Second, the fact that we focus on assortative mating using the *spousal* PS to instrument for her education helps us to *more credibly* satisfy the exclusion restriction, as illustrated below:

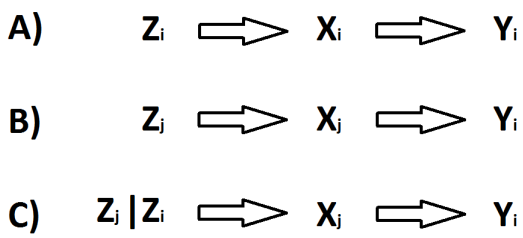

Figure 1: Diagrammatic representation of the exclusion restriction

Case A) in Figure 1 is the standard diagrammatic representation of IV: Z_i is the instrument for the endogenous explanatory variable X_i where Y_i is the outcome variable. Case B) is the present case: the instrument and the endogenous explanatory variables refer to individual j while the outcome variable refers to individual i. Case C) is a slight modification of B), since it controls for the corresponding variable Z for individual i. However, as we will see, even if the exclusion restriction is *not* satisfied, the IV approach combined with the *two-sided* nature of the marriage market proves useful in assessing whether matching in the marriage market is *multidimensional*.

Structure of the paper. The rest of the paper is organized as follows. Section 2 presents two stylized models of perfect and imperfect positive assortative mating on education. Section 3 discusses how to measure assortative mating in practice, defining the corresponding OLS and IV estimands. Section 4 defines the polygenic scores and the genetic IV. Section 5 describes the data and the construction of the polygenic scores. Section 6 contains our empirical analysis. Section 7 interprets our evidence. Finally, Section 8 concludes the paper.

2 Measuring Assortative Mating in Theory

2.1 A deterministic case

Consider two populations (men and women) of equal size, normalized to one. Agents differ in their educational attainment: x denotes the educational attainment for women, and y the educational attainment for men. Without loss of generality, assume that

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

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

where $b > 1$ and $a > 0$. Positive assortative mating (PAM) implies that

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

Hence, we have the matching function

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

where $\beta_0 = a$ and $\beta_1 = (b-a)$. Thus, this bare-bones matching model predicts *perfect* PAM

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

2.2 A stochastic case

Suppose that agents match on x and y, but we (the econometricians) observe x^* and y^*, such that

$$y^* = y + \epsilon \tag{6}$$

$$x^* = x + u \tag{7}$$

where $cov(y,\epsilon) = cov(x,u) = 0$. In that case, we can rewrite (4) as follows

$$y^* = \beta_0 + \beta_1 x^* - \beta_1 u + \epsilon \tag{8}$$

With measurement error, our bare-bones model predicts PAM, *not* perfect PAM

$$\begin{aligned} corr(y^*,x^*) = \frac{cov(y^*,x^*)}{\sqrt{var(y^*)}\sqrt{var(x^*)}} = \frac{cov(y,x)}{\sqrt{var(y)}\sqrt{var(x)}} \frac{\sqrt{var(y)}\sqrt{var(x)}}{\sqrt{var(y^*)}\sqrt{var(x^*)}} = \\ = corr(y,x)\sqrt{\frac{var(y)}{var(y)+var(\epsilon)}}\sqrt{\frac{var(x)}{var(x)+var(u)}} < 1 \end{aligned} \tag{9}$$

While this case is one with measurement error, one could similarly consider extensions in which the identification problem arises due to omitted relevant dimensions (omitted variable bias) or due to the simultaneity between y and x (simultaneity bias).[1] Of course, in practice,

[1] We can think that education of the husband (wife) is the quantity "supplied" of education by the male (female) side of the market, and education of the wife (husband) is the quantity "demanded" of education by the female (male) side of the market.

we may expect different sources of biases at the same time, so that whether assortative mating is under or over-estimated is an empirical question.

3 Measuring Assortative Mating in Practice

3.1 OLS

Suppose that the "stochastic" matching functions are given by

$$y = \alpha + \beta x + v_y \tag{10}$$

and

$$x = \alpha^{'} + \beta^{'} y + v_x \tag{11}$$

where v_y and v_x are random components.

As long as $var(x) = var(y)$, the population slopes of these matching functions are the same

$$\beta_{OLS} = \frac{cov(y,x)}{var(x)} = \frac{cov(y,x)}{var(y)} = \beta^{'}_{OLS} \tag{12}$$

Moreover,

$$\beta_{OLS} = \beta^{'}_{OLS} = corr(y,x) \tag{13}$$

However, β_{OLS} is not necessarily the parameter of interest. For example, x and y may be proxies for the underlying variables on which sorting actually takes place (e.g., education can be a proxy for ability). In the previous section, we showed that with measurement error, the

degree of positive assortative mating will be underestimated.[2] More fundamental problems are those due to omitted variables and simultaneity. Indeed, given the simultaneity implied by equations (10) and (11), the OLS estimands do not recover the parameters of interest

$$\beta_{OLS} = \beta + \left(\frac{\beta^{'}}{1 - \beta\beta'} \right) \frac{var(v_y)}{var(x)} \tag{14}$$

$$\beta^{'}_{OLS} = \beta^{'} + \left(\frac{\beta}{1 - \beta\beta'} \right) \frac{var(v_x)}{var(y)} \tag{15}$$

In that case, instrumental variables can help us to accurately measure the degree of assortativeness.[3]

3.2 IV

Suppose that we have a valid instrument z_x for x and a valid instrument z_y for y. If we instrument x in (10) with z_x, we obtain the following estimand

$$\beta_{IV} = \frac{cov(y, z_x)}{cov(x, z_x)} \tag{16}$$

By the same token, if we instrument y in (11) with z_y, we obtain the following estimand

$$\beta^{'}_{IV} = \frac{cov(x, z_y)}{cov(y, z_y)} \tag{17}$$

As long as $var(x) = var(y)$, (16) and (17) can be rewritten as

$$\beta_{IV} = \frac{corr(y, z_x)}{corr(x, z_x)} \tag{18}$$

[2]One possibility to tackle this problem is using repeated independent reports of the same variable as instrumental variables (Ashenfelter and Krueger, 1994).

[3]Since $\beta\beta^{'} < 1$, simultaneity bias will lead to an overestimation of β and $\beta^{'}$.

$$\beta'_{IV} = \frac{corr(x, z_y)}{corr(y, z_y)} \tag{19}$$

An important remark is in order. In general, these two estimands β_{IV} and β'_{IV} are different, since the instrumental variables z_x and z_y need not have the same correlations with x and y. The next section builds the instrumental variables z_x and z_y using genetic variation.

4 Building a Genetic IV

4.1 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.

A PS is based on genetic variants present in the entire genome and yield a single quantitative measure of genetic predisposition. 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.[4] 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

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

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 recent GWAS on educational attainment (Rietveld et al., 2013) explains around 0.02% of the variation in years of schooling. A linear polygenic score from all measured SNPs explains approximately 2-3% 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.[5] 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 association results excluding the cohort to which the score is applied, which is exactly how we proceed.

4.2 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

[5]For example, educational attainment (SSGAC; Rietveld et al. (2013)), body mass index (GIANT consortium; Voight et al. (2010)), cardiovascular disease (CHARGE consortium; Ganesh et al. (2009)), smoking behavior (TAG consortium; Tobacco et al. (2010)), psychiatric disorders (PGC consortium; Ripke et al. (2013)) and reproductive aging (REPROGEN consortium; Elks et al. (2010)).

quasi-experimental design is called *Mendelian randomization* (Smith and Ebrahim, 2003).[6] 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 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. Hence, we improve on the existing literature.[7]

4.3 Assumptions for a valid genetic IV

A valid instrument must satisfy the following assumptions:

A1. *Independence Assumption*

A2. *1st Stage or Relevance Assumption*

A3. *Exclusion Restriction Assumption*

A4. *Monotonicity Assumption*

[6] 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.

[7] 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).

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.[8] Genetic similarity is often correlated with geographical proximity, because human genetic diversity is the result of the history of population migration, ethnic admixture and residential segregation. This may affect the marriage market since potential partners living in the same geographical area are more likely to share common ancestry.[9]

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.[10] 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.

[8]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).

[9]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).

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

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 it has been shown to robustly affect education (Rietveld et al., 2013).

The **Exclusion Restriction Assumption (A3)** requires that the spousal polygenic score for education affects own education only through spousal education. Given that we are conditioning on the individual polygenic score and that our endogenous and instrumental variables belong to the spouse, assumption **A3** is more likely to hold than in common genetic IV studies considering the effect of one individual's treatment on the *same* individual's outcome, by using a genetic variant of his as instrument. In such studies, 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*);[11] (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.

Finally, the **Monotonicity Assumption (A4)** requires that the spousal polygenic score affects spousal education for every "spouse" in the same direction.[12] However, with homogeneous causal effects, the monotonicity assumption is irrelevant.

[11] 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.

[12] Chaisemartin (2015) shows that IV estimates a causal effect under a weaker condition than monotonicity.

5 Data Description

5.1 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.[13] 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 polymorphisms (SNPs). Current genetic data available for research also include imputation of approximately 21 million DNA variants from the 1000Genomes Project.[14] 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

[13] For the non-genetic data, we used the RAND HRS Data files, Version N.

[14] 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.

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, and couples 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.[15] This yields a working sample of 1,443 couples (2,886 individuals).

The main variables used in our empirical analysis are education and the polygenic scores for education. Education is defined in two ways: the number of completed years of schooling (from 0 to 17), and an indicator equal to 1 if the individual has a college degree (or above), and 0 otherwise. We generate two polygenic scores, one for years of education and one corresponding to holding a college degree (or above). Both these polygenic scores for years of education and college degree were calculated based on the most recent GWAS results available (Rietveld et al., 2013).

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.[16] Using these summary statistics, we constructed linear polygenic scores weighted for their effect sizes in the meta-analysis. We constructed all

[15]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.

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

scores using the softwares PLINK and PRSice (Purcell et al., 2007; Euesden et al., 2015).[17] For each phenotype, we considered the complete set of available SNPs. Results 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).[18]

These population controls allow to analyze genotypic variants that are not driven by 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 say that spousal genes for education are as good as randomly assigned. Finally, given that both the PSs for education and the principal components are generated regressors, the standard errors in our regression analysis are bootstrapped. IV estimates are calculated using 2SLS.

6 Results

6.1 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 13.3 years of education; 35% of husbands have a college degree while 24% of wives do.

[Table 1 about here]

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

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

Table 2 shows the correlation matrix for years of education and PSs: the correlation between husband's and wife's years of education is 0.560 ($p-value < 0.001$), while that for their PSs for education is 0.161 ($p-value < 0.001$). 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. In line with Domingue et al. (2014), the correlation between polygenic scores is about 30% of the correlation between years of education.

[Table 2 about here]

6.2 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 husband's years of education on the wife's years of education. Column (2) adds the polygenic score (PS) for husband's education, and column (3) controls for both spouses' PSs. Column (1) shows that the point estimate of the coefficient on wife's education is 0.650. Once the husband's genetic score is accounted for, the point estimate decreases from 0.650 to 0.634, column (2). The coefficient further decreases from 0.634 to 0.630 when adding both PSs, but the difference in point estimates is very small (column (3)). A similar qualitative picture emerges in the last three columns, (4), (5) and (6), where we replace years of education (0-17) 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 wife's education is the dependent variable and the husband's education is the main explanatory variable. Overall the magnitude of the coefficients is lower than in Table 3.

[Table 4 about here]

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 wife's PS satisfies the *instrument relevance* condition: the $F-statistic$ for the wife's PS being irrelevant is $F(1,1396)=36.93$, 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 wife'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 wife's years of education is 0.783 (SE = 0.188), which is larger than 0.634 (SE = 0.029), the one obtained using OLS in Table 3. Looking at columns (4)-(6), we find similar results: the instrument appears to be relevant for college degree ($F(1,1402)=25.27$), the role of the wife's PS is less than half that of the own PS, and the 2SLS point estimate is 0.726 (SE = 0.206), which is higher than 0.532 (SE = 0.027) the OLS point estimate in Table 3. Interestingly, 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. Table 6 shows the assortative mating on education with respect to the wife. The IV point estimate of the coefficient on husband's years of education is 0.692 (SE = 0.113) versus the OLS point estimate 0.438 (SE = 0.022), while the point estimate of the coefficient on husband's college is 0.761 (SE = 0.153) versus the OLS point estimate of 0.428 (SE = 0.024). Moreover, the IV point estimate of the coefficient on husband's years of education is 0.692 (SE = 0.113), smaller compared to the point estimate of the coefficient on wife's years of education as reported in Table 5, 0.783 (SE = 0.188). The IV point estimate of the coefficient on husband's college, instead, is 0.761 (SE = 0.153) higher than the point estimate of the coefficient on husband's years of education 0.692 (SE = 0.113). The $F-statistic$s for the husband's genetic years of education and college PS are respectively $F(1,1396)=42.70$ and

$F(1, 1402) = 35.31$. While the results for men are qualitatively similar to those for women, now, both Hausman tests at the bottom of columns (3) and (6) *do* reject that the OLS and IV estimands are the same at the 5% significance level.

[Table 6 about here]

Overall, our results consistently suggest that, if anything, the actual degree of positive assortative mating on education is *underestimated*.[19] The ratios of the estimated OLS-IV coefficients on wife's year of education and on wife's college degree indicator are 0.81 and 0.73, respectively; those on husband's years of education and on husband's college degree indicator are 0.63 and 0.56, respectively. While we cannot reject the hypotheses that the OLS and IV coefficients on wife's year of education and on college degree indicator are the same, we do reject the hypotheses that the OLS and IV coefficients on husband's year of education and on college degree indicator are the same at the 5% levels of statistical significance.

7 Interpreting the gender OLS-IV differences

Suppose that the stochastic matching functions are given by

$$y = \alpha + \beta x + \gamma w_x + v_y \tag{20}$$

and

$$x = \alpha^{'} + \beta^{'} y + \gamma^{'} w_y + v_x \tag{21}$$

[19] Quarters of birth are weak instrumental variables in our sample for both education variables (years of education and college degree): the F-statistics are between 0.02 and 0.44, much below 10. Results using quarter of birth dummies as instrumental variables are available upon request.

where $cov(x, w_x) \neq 0$, $cov(y, w_y) \neq 0$, $cov(z_x, w_x) \neq 0$ and $cov(z_y, w_y) \neq 0$, so we are allowing the exclusion restriction to be violated. The OLS estimands of β and $\beta^{'}$ are then

$$\beta_{OLS} = \frac{cov(y, x)}{var(x)} = \beta + \gamma \frac{cov(w_x, x)}{var(x)} \tag{22}$$

and

$$\beta^{'}_{OLS} = \frac{cov(y, x)}{var(y)} = \beta^{'} + \gamma^{'} \frac{cov(w_y, y)}{var(y)} \tag{23}$$

The IV estimands of β and $\beta^{'}$ are then

$$\beta_{IV} = \frac{cov(y, z_x)}{cov(x, z_x)} = \beta + \gamma \frac{cov(w_x, z_x)}{cov(x, z_x)} \tag{24}$$

and

$$\beta^{'}_{IV} = \frac{cov(x, z_y)}{cov(y, z_y)} = \beta^{'} + \gamma^{'} \frac{cov(w_y, z_y)}{cov(y, z_y)} \tag{25}$$

We cannot reject $\beta_{OLS} = \beta_{IV}$, hence

$$\gamma \frac{cov(w_x, x)}{var(x)} = \gamma \frac{cov(w_x, z_x)}{cov(x, z_x)} \tag{26}$$

Thus, if the exclusion restriction is violated ($cov(z_x, w_x) \neq 0$), it must be the case that we cannot reject $\gamma = 0$, which is consistent with female socioeconomic attractiveness in the marriage market being driven by education. However, we reject $\beta^{'}_{OLS} = \beta^{'}_{IV}$, hence

$$\gamma^{'} \frac{cov(w_y, y)}{var(y)} \neq \gamma^{'} \frac{cov(w_y, z_y)}{cov(y, z_y)} \tag{27}$$

so that we reject $\gamma^{'} = 0$, which is consistent with male socioeconomic attractiveness being a bundle of characteristics correlated with education.

Our empirical findings are consistent with $\gamma = 0$ and $\gamma' \neq 0$. Hence, while we cannot conclude that assortative mating is underestimated if the exclusion restriction is violated, we can still infer that female socioeconomic attractiveness in the marriage market is driven by education, but male socioeconomic attractiveness is driven by education and other correlated variables. In other words, our findings are in line with socioeconomic attractiveness in the marriage market being different by gender. Male socioeconomic attractiveness is the ability to generate *income*, which is correlated with education but also with other indicators such as prestige or occupational/social status. For women, instead, education is productive *per se* in marriage from a socioeconomic perspective, as it is the input in the household production function of high-quality household goods which are valued by men. Indeed, Bertrand et al. (2015) emphasize that gender identity norms limit further gender convergence in the labor market, and that women do much more household production than men even more so among high-educated couples. Moreover, Lefgren and McIntyre (2006) show that women's education has a cross-productivity effect in marriage, and Chiappori et al. (2012) measure female socioeconomic attractiveness using *education*, while they measure male socioeconomic attractiveness using the hourly *wage*.

8 Conclusions

This is the first paper to present an IV strategy to estimate assortative mating on education using spousal genetic markers. We instrument spousal education with his/her educational polygenic score while controlling for own educational polygenic score, and find that assortative mating on education is, if anything, underestimated. 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). In addition, given that our instrument is "very" relevant, a small

violation of the exclusion restriction is unlikely to seriously bias our IV estimates.

We show how an IV approach to study assortative mating can shed light not only on the quantification of the actual degree of assortativeness, but more generally, can help to assess whether matching takes place along more than one dimension. While the first result depends on the validity of the exclusion restriction, the second does not. Thus, we interpret the gender OLS-IV differences as indicating that socioeconomic attractiveness appears to be multidimensional for men but *not* for women, at least among older cohorts in the US.

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.

Ashenfelter, O. and A. Krueger (1994): "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review*, 84, 1157–1173.

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, R. Houts, G. G. Bennett, A. K. Biddle, J. A. Blumenthal, J. P. Evans, H. L. Harrington, K. Sugden, B. Williams, et al. (2012): "Polygenic risk, rapid childhood growth, and the development of obesity: evidence from a 4-decade longitudinal study," *Archives of Pediatrics & Adolescent Medicine*, 166, 515–521.

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.

Bertrand, M., J. Pan, and E. Kamenica (2015): "Gender Identity and Relative Income Within Households," *Quarterly Journal of Economics*, 130, 571–614.

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.

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, 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.

——— (2015): "Bidimensional Matching with Heterogeneous Preferences in the Marriage Market," *University of Oxford, mimeo.*

Chiappori, P. A. and B. Salanié (forthcoming): "The Econometrics of Matching," *Journal of Economic Literature.*

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., 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.

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

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.

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.

Elks, C. E., J. R. B. Perry, P. Sulem, D. I. Chasman, N. Franceschini, C. He, K. L. Lunetta, J. A. Visser, E. M. Byrne, D. L. Cousminer, et al. (2010): "Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies," *Nature Genetics*, 42, 1077–1085.

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.

Ganesh, S. K., N. A. Zakai, F. J. A. van Rooij, N. Soranzo, A. V. Smith, M. A. Nalls, M. H. Chen, A. Kottgen, N. L. Glazer, A. Dehghan, et al. (2009): "Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium," *Nature Genetics*, 41, 1191–1198.

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 (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.

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.

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., 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.

Ripke, S., C. O'Dushlaine, K. Chambert, J. L. Moran, A. K. Kähler, S. Akterin, S. E. Bergen, A. L. Collins, J. J. Crowley, M. Fromer, et al. (2013): "Genome-wide association analysis identifies 13 new risk loci for schizophrenia," *Nature Genetics*, 45, 1150–1159.

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.

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.

Tobacco, G. Consortium, et al. (2010): "Genome-wide meta-analyses identify multiple loci associated with smoking behavior," *Nature Genetics*, 42, 441–447.

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.

Voight, B. F., L. J. Scott, V. Steinthorsdottir, A. P. Morris, C. Dina, R. P. Welch, E. Zeggini, C. Huth, Y. S. Aulchenko, G. Thorleifsson, et al. (2010): "Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis," *Nature Genetics*, 42, 579–589.

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.

Table 1. Summary statistics

	N	Mean	SD	Min	Max
Husband's Year of Birth	1,443	1937	9	1910	1957
Husband's Years of Education	1,440	13.63	2.70	0	17
Husband's College	1,443	0.35	0.48	0	1
Husband's Years of Education Polygenic Score	1,434	0.24	0.93	-2.41	4.56
Husband's College Polygenic Score	1,437	-0.10	1.00	-3.16	4.25
Wife's Year of Birth	1,443	1939	9	1911	1961
Wife's Years of Education	1,440	13.34	2.22	0	17
Wife's College	1,443	0.24	0.42	0	1
Wife's Years of Education Polygenic Score	1,428	0.25	0.94	-2.81	5.18
Wife's College Polygenic Score	1,431	-0.09	0.99	-2.85	4.88

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	0.5600	1.000	–	–
	(0.000)	(0.000)		
Husband's Education Polygenic Score	0.1779	0.1482	1.000	–
	(0.000)	(0.000)	(0.000)	
Wife's Education Polygenic Score	0.1272	0.1716	0.1608	1.000
	(0.000)	(0.000)	(0.000)	(0.000)

Note: p-values in parentheses.

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

	Husband's Years of Education			Husband's College		
	(1)	(2)	(3)	(4)	(5)	(6)
Wife's Years of Education	0.650***	0.634***	0.630***			
	(0.029)	(0.029)	(0.030)			
Husband's Years of Education Polygenic Score		0.284***	0.280***			
		(0.066)	(0.066)			
Wife's Years of Education Polygenic Score			0.059			
			(0.070)			
Wife's College				0.547***	0.532***	0.529***
				(0.027)	(0.027)	(0.027)
Husband's College Polygenic Score					0.050***	0.050***
					(0.013)	(0.013)
Wife's College Polygenic Score						0.012
						(0.012)
Observations	1,419	1,419	1,419	1,425	1,425	1,425
R-squared	0.34	0.34	0.34	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 4. Regressions of Wife's Education on Husband's Education controlling for Polygenic Scores

	Wife's Years of Education			Wife's College		
	(1)	(2)	(3)	(4)	(5)	(6)
Husband's Years of Education	0.448***	0.438***	0.431***			
	(0.021)	(0.022)	(0.022)			
Wife's Years of Education Polygenic Score		0.264***	0.256***			
		(0.055)	(0.055)			
Husband's Years of Education Polygenic Score			0.133**			
			(0.054)			
Husband's College				0.435***	0.428***	0.419***
				(0.024)	(0.024)	(0.024)
Wife's College Polygenic Score					0.042***	0.042***
					(0.011)	(0.011)
Husband's College Polygenic Score						0.028***
						(0.011)
Observations	1,419	1,419	1,419	1,425	1,425	1,425
R-squared	0.33	0.34	0.34	0.27	0.27	0.28

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

	Years of Education			College		
	Wife	Husband		Wife	Husband	
	FS	RF	2SLS	FS	RF	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
Wife's Years of Education			0.783***			
			(0.188)			
Husband's Years of Education Polygenic Score	0.347***	0.499***	0.227**			
	(0.063)	(0.078)	(0.097)			
Wife's Years of Education Polygenic Score	0.385***	0.301***				
	(0.063)	(0.080)				
Wife's College						0.726***
						(0.206)
Husband's College Polygenic Score				0.062***	0.082***	0.037*
				(0.012)	(0.014)	(0.019)
Wife's College Polygenic Score				0.060***	0.044***	
				(0.012)	(0.013)	
F-test instrument relevance	36.93***	–	–	25.27***	–	–
Hausman test	–	–	0.64	–	–	0.79
Hausman test p-value	–	–	[0.423]	–	–	[0.374]
Observations	1,419	1,419	1,419	1,425	1,425	1,425

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 Husband's Polygenic Score as an Instrumental Variable for Husband's Education

	Years of Education			College		
	Husband	Wife		Husband	Wife	
	FS	RF	2SLS	FS	RF	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
Husband's Years of Education			0.692***			
			(0.113)			
Wife's Years of Education Polygenic Score	0.301***	0.386***	0.178**			
	(0.079)	(0.063)	(0.070)			
Husband's Years of Education Polygenic Score	0.508***	0.351***				
	(0.078)	(0.062)				
Husband's College						0.761***
						(0.153)
Wife's College Polygenic Score				0.042***	0.059***	0.027**
				(0.013)	(0.012)	(0.014)
Husband's College Polygenic Score				0.082***	0.063***	
				(0.014)	(0.012)	
F-test instrument relevance	42.70***	–	–	35.31***	–	–
Hausman test	–	–	4.04**	–	–	5.04**
Hausman test p-value	–	–	[0.044]	–	–	[0.023]
Observations	1,419	1,419	1,419	1,425	1,425	1,425

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

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 Rietveld et al. (2013) 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). Barban acknowledges funding from the European Research Council via an ERC Consolidator Grant SOCIOGENOME (615603 awarded to M. Mills, see www.sociogenome.com). Per the data sharing policy of the Social Science Genetic Association Consortium, we acknowledge the authors on the original GWAS meta-analysis on educational attainment (Rietveld et al., 2013): Abdel Abdellaoui, Arpana Agrawal, Eva Albrecht, Behrooz Z. Alizadeh, Jüri Allik, Najaf Amin, John R. Attia, Stefania Bandinelli, John Barnard, François Bastardot, Sebastian E. Baumeister, Jonathan Beauchamp, Daniel J. Benjamin, Kelly S. Benke, David A. Bennett, Klaus Berger, Lawrence F. Bielak, Laura J. Bierut, Jeffrey A. Boatman, Dorret I. Boomsma, Patricia A. Boyle, Ute Bültmann, Harry Campbell, David Cesarini, Christopher F. Chabris, Lynn Cherkas, Mina K. Chung, Dalton Conley, Francesco Cucca, George Davey-Smith, Gail Davies, Mariza de Andrade, Philip L. De Jager, Christiaan de Leeuw, Jan-Emmanuel De Neve, Ian J. Deary, George V. Dedoussis, Panos Deloukas, Jaime Derringer, Maria Dimitriou, Gudny Eiriksdottir, Niina Eklund, Martin F. Elderson, Johan G. Eriksson, Tõnu Esko, Daniel S. Evans, David M. Evans, Jessica D. Faul, Rudolf Fehrmann, Luigi Ferrucci, Krista Fischer, Lude Franke, Melissa E. Garcia, Christian Gieger, Håkon K. Gjessing, Patrick J.F. Groenen, Henrik Grönberg, Vilmundur Gudnason, Sara Hägg, Per Hall, Jennifer R. Harris, Juliette M. Harris, Tamara B. Harris, Nicholas D. Hastie, Caroline Hayward, Andrew C. Heath, Dena G. Hernandez, Wolgang Hoffmann, Adriaan Hofman, Albert Hofman, Rolf Holle, Elizabeth G. Holliday, Christina Holzapfel, Jouke-Jan Hottenga, William G. Iacono, Carla A. Ibrahim-Verbaas, Thomas Illig, Erik Ingelsson, Bo Jacobsson, Marjo-Riitta Järvelin, Min A. Jhun, Magnus Johannesson, Peter K. Joshi, Astanand Jugessur, Marika Kaakinen, Mika Kähönen, Stavroula Kanoni, Jaakkko Kaprio, Sharon L.R. Kardia, Juha Karjalainen, Robert M. Kirkpatrick, Philipp D. Koellinger, Ivana Kolcic, Matthew Kowgier, Kati Kristiansson, Robert F. Krueger, Zóltan Kutalik, Jari Lahti, David Laibson, Antti Latvala, Lenore J. Launer, Debbie A. Lawlor, Sang H. Lee, Terho Lethimäki, Jingmei Li, Paul Lichtenstein, Peter K. Lichtner, David C. Liewald, Peng Lin, Penelope A. Lind, Yongmei Liu, Kurt Lohman, Marisa Loitfelder, Pamela A. Madden, Patrick K.E. Magnusson, Tomi E. Mäkinen, Pedro Marques Vidal, Nicolas W. Martin, Nicholas G. Martin, Marco Masala, Matt McGue, George McMahon, Sarah E. Medland, Osorio Meirelles,

Andres Metspalu, Michelle N. Meyer, Andreas Mielck, Lili Milani, Michael B. Miller, Grant W. Montgomery, Sutapa Mukherjee, Ronny Myhre, Marja-Liisa Nuotio, Dale R. Nyholt, Christopher J. Oldmeadow, Ben A. Oostra, Lyle J. Palmer, Aarno Palotie, Brenda Penninx, Markus Perola, Katja E. Petrovic, Wouter J. Peyrot, Patricia A. Peyser, Ozren Polašek, Danielle Posthuma, Martin Preisig, Lydia Quaye, Katri Räikkönen, Olli T. Raitakari, Anu Realo, Eva Reinmaa, John P. Rice, C.A. Rietveld, Susan M. Ring, Samuli Ripatti, Fernando Rivadeneira, Thais S. Rizzi, Igor Rudan, Aldo Rustichini, Veikko Salomaa, Antti-Pekka Sarin, David Schlessinger, Helena Schmidt, Reinhold Schmidt, Rodney J. Scott, Konstantin Shakhbazov, Albert V. Smith, Jennifer A. Smith, Harold Snieder, Beate St Pourcain, John M. Starr, Jae Hoon Sul, Ida Surakka, Rauli Svento, Toshiko Tanaka, Antonio Terracciano, Alexander Teumer, A. Roy Thurik, Henning Tiemeier, Nicholas J. Timpson, André G. Uitterlinden, Matthijs J.H.M. van der Loos, Cornelia M. van Duijn, Frank J.A. van Rooij, David R. Van Wagoner, Erkki Vartiainen, Jorma Viikari, Peter M. Visscher, Veronique Vitart, Peter K. Vollenweider, Henry Völzke, Judith M. Vonk, Gérard Waeber, David R. Weir, Jürgen Wellmann, Harm-Jan Westra, H.-Erich Wichmann, Elisabeth Widen, Gonneke Willemsen, James F. Wilson, Alan F. Wright, Jian Yang, Lei Yu, Wei Zhao. The views presented in this article may not reflect the opinions of these collaborators.