FINNISH CENTRE OF
EXCELLENCE IN TAX
SYSTEMS RESEARCH

FIT Working Paper 26

Nurfatima Jandarova and Aldo Rustichini

Individual Characteristics and Earnings

Tampere University

UNIVERSITY OF HELSINKI

VATT INSTITUTE FOR ECONOMIC RESEARCH

CENTRES OF EXCELLENCE IN RESEARCH

INDIVIDUAL CHARACTERISTICS AND EARNINGS

NURFATIMA JANDAROVA AND ALDO RUSTICHINI

Abstract. We study how observed individual characteristics affect earnings of individuals. The characteristics we study are individual personality traits (including cognitive ability) and family background. We make use of data providing information on the individual characteristics rather than estimating them as latent variables.

Their contribution may be indirect (facilitating the acquisition of education) or direct (perhaps affecting productivity). We estimate the fraction of these two contributions through regression analysis and structural model, and find that the contribution of both pathways is significant.

These characteristics may be in part determined endogenously. To estimate the proportion due to original individual characteristics we use measures provided by Polygenic Scores for education years and fluid intelligence. The marginal effects of these scores is significant and high. The indirect contribution (operating though acquisition of college) is around one third of the total effect.

(Nurfatima Jandarova) Centre of Excellence in Tax Systems Research, Tampere University, Tampere, Finland

Email address: nurfatima.jandarova@gmail.com

(Aldo Rustichini) Department of Economics, University of Minnesota, 1925 4th Street South 4-101, Hanson Hall, Hanson Hall, Minneapolis, MN, 55455

Email address: aldo.rustichini@gmail.com

JEL: J24, I21

Keywords: education, cognitive skills, noncognitive skills, family background, polygenic score

Date: October 4, 2024.

1. Introduction

We study how observed individual characteristics affect earnings of individuals. The characteristics we study are cognitive ability, bit also personality traits likely to affect earnings and family background. An important feature of our study is the use of the data (UK Household Longitudinal Study (UKHLS), also known as Understanding Society) which provides measurements of these characteristics. In particular, it contains several tests on cognitive abilities and Big 5 personality scores. We can use this information to construct cognitive and non-cognitive scores, respectively, rather than estimating them as latent variables. The dataset also offers detailed information on wages and earnings. [1] The use of individual data is essential: the adoption of aggregate data (as for example in the Lynn and Vanhanen (2002)) is subject to natural and justified criticisms (as discussed in Ervik (2003)). In particular, the IQ data may be unreliable and the cross country comparison difficult. In addition, and more substantially, a simple bivariate correlation does not allow any conclusion on the causality direction.

There are two ways in which these traits may affect earnings: one through the educational attainment, the other an effect on earnings independent of education. An important contribution of the paper is to disentangle the two components.

The analysis examines first simple descriptive statistics on the relation between earnings, education and individual characteristics. We then estimate the way in which these individual characteristics affect the probability of achieving higher education (college degree). In our study we do not assume specific functional forms but estimate the best within a rich family of models.

The history of the analysis of individual characteristics and earnings is wide. What we add to this tradition is a dataset that includes precise information on cognitive skills, personality, family background, earnings and genotype of individuals. We use the individual genotypes, recent genome-wide association studies and methodologies to construct polygenic scores for educational attainment. Our study continues a tradition of investigation (going back to Griliches and Mason (1972); Griliches (1976) [2]; see also Lindqvist and Vestman (2011) [3] A closer analysis of the data

[1] An analysis of the education acquisition in post-war UK, relying on the same data, is reported in Ichino et al. (2022). The emphasis is on the evaluation of real and counter-factual higher education policies for given distribution of characteristics in the population.

[2] These studies used the AFQt for a smaple of army veterans and the National Longitudinal Survey of Young Men respectively

[3] The authors conclude that

> We find strong evidence that men who fare poorly in the labor market in the sense of unemployment or low annual earnings lack non-cognitive rather than cognitive ability. However, cognitive ability is a stronger predictor of wages for skilled workers and of earnings above the median.

seem to add a new perspective to earlier conclusions of studies investigating a similar question (see eg Ashenfelter et al. (2000) [4]) In this analysis no information on personality traits and family background was available, so no direct comparison among these factors was possible. In Hanushek et al. (2015) the *PIAAC* data set is used to estimate in 23 countries and they find that on the average of the countries, a one SD increase in numeracy skills is associated with an 18 percent wage increase among prime-age workers. There is substantial heterogeneity across countries from a maximum of 28 per cent (USA) to a minimum of 14 (Sweden). The UK is near the top with 22.5 per cent. The important contribution Heckman et al. (2006), which like ours attempts at explaining a large fraction of labor market and (in their case) behavioral outcomes through the use of a low-dimensional vector of cognitive and non-cognitive skills. One of distinctive features of the paper is showing that the estimated contribution of latent variables representing these skills may be very different from that obtained though their noisy measurements (following methods in Hansen et al. (2004)). A further step, that we propose here, is to use direct although potentially incomplete evidence on the genetic evidence available on these latent factors. This additional data may help in improving their estimate (as we discuss later in section 6). The use of the genetic evidence extends analysis of the issue of returns that was developed earlier on the basis of identical twins (see for example Isacsson (1999); Ashenfelter and Krueger (1994); Miller et al. (1995); Ashenfelter and Zimmerman (1997); Bonjour et al. (2003)). The limits of identical twins methodology are well known (a word of caution was already in Bound and Solon (1999)), and thus the genetic data make the conclusions more robust.

Organization of the paper. The paper is organized as follows. In section 2 we present the data used in the analysis below. Notations and definitions are presented in section 3. Descriptive and linear regression analysis is reported in section 4; non linear estimations and structural models are estimated in section 5. The analysis using genetic data is developed in section 6. Section 7 concludes.

[4] The conclusions of these earlier investigations were more skeptical:

> The results of all these studies are surprisingly consistent: they indicate that the return to schooling is not caused by an omitted correlation between ability and schooling. Moreover, we find no evidence that the return to schooling differs significantly by family background or by the measured ability of the student.

2. Data

We first review the descriptive evidence in the data concerning the relationship between cognitive ability and earnings. We begin by describing the data set used in this paper and key variables of interest.

2.1. **Data.**

In our analysis we use the UK Household Longitudinal Study (UKHLS), also known as Understanding Society. This is the largest household panel study in the UK, covering about 40,000 individuals in each wave since 2009. The participants were sampled from the UK population in 2009 and are followed every year. Starting from wave 2, the follow-up sample also includes former British Household Panel Survey (BHPS) [5] respondents.

The survey encompasses a wide range of topics, including education, employment and cognitive abilities: we review here those that are relevant for our investigation.

2.1.1. *Education.* The survey contains a variable describing the highest qualification reached by the individual. The variable has six categories: degree, other higher degree, A-level or equivalent, GCSE or equivalent, other and no qualification. This variable is updated in every wave, taking into account newly acquired qualifications, if applicable. We convert this categorical variable to a binary degree indicator D_i that takes value of 1 whenever individual i reports having a degree as highest qualification in any wave.

2.1.2. *Wages.* In each wave the respondents are asked about their employment status, jobs and earnings. We use monthly labour earnings and usual hours worked in a month to construct hourly wages. We then deflate the hourly wages using the CPI excluding rent, maintenance repairs and water charges, an index recommended by the UKHLS (see Fisher et al. (2019)).

2.1.3. *Cognitive score.* In wave 3 the participants were administered a set of five cognitive tests: word recall (immediate and delayed), serial 7 subtraction, number series, verbal fluency and numeric ability. The UKHLS then summarizes the results into counts of correct answers to each test. There are 40,889 individuals with non-missing test results out of the full sample of 49,692 respondents in wave 3. We then estimate the cognitive score using the maximum likelihood confirmatory factor

[5] The BHPS is a predecessor of the UKHLS. The BHPS ran from 1991 to 2008 covering about 10,000 individuals. In the final wave of the BHPS, the respondents were asked if they wished to continue as part of the UKHLS; about 80% chose to continue.

analysis, adapting the model of Johnson and J.Bouchard (2005). For more detailed information, see section H.1.

2.2. **Big 5 score.** In wave 3, adult respondents were also given a short 15-item Big 5 personality test. Each test is a separate statement (for example, ”I see myself as someone who does a thorough job”) that the respondents can disagree (1) or agree (7) with. The score for each domain - agreeableness, conscientiousness, extraversion, neuroticism and openness - is a rounded average of answers to the corresponding three sub-domain questions. There are 40,544 individuals with non-missing Big 5 personality test results (81.59% of full sample). Furthermore, 38,034 individuals have both non-missing Big 5 and cognitive test results. We use the principal component analysis to combine the five domain scores into single Big 5 score. For more detailed information, see section H.2.

2.3. **Family score.** All adult respondents are asked basic questions about their parents. In particular, we use highest educational qualification and employment status at the time the respondents were 14 years old. First, we convert the highest educational qualification into years of schooling by assigning average years of education among individuals of same gender and birth cohort with the corresponding qualification. Second, we convert the categorical variable with parent's employment status - working, not working, deceased, absent - into four indicator variables, separately for each parent. We then combine the years of education and parental status indicators into family advantage score using the principal component analysis. We again use the first component, which captures 23% of the data variation, and assigns positive weights to education and working status and negative weights to having deceased or absent parents.

Our working sample consists of wave 3 respondents with non-missing cognitive and Big 5 scores. We also restrict our sample to those born between 1950 and 1989 with non-missing degree indicator. Furthermore, we select only those who have been observed in the survey at least once between ages 25 and 65. This filter helps us remove individuals who have not yet completed their education phase or those who have only been observed past retirement. The final sample consists of 26,564 individuals and 234,757 person-wave observations.

We generate two working datasets: panel and cross-sectional. The panel dataset contains up to 12 observations for each person in our working sample. The observations correspond to waves of the UKHLS survey (2009-22).

Using observations from all waves of the UKHLS, we also create predicted lifetime earnings. We first estimate wage age profiles using fixed effects estimator and allowing gender and college degree to change the slope of the age profiles. In our estimations we use the restriction dictated by the economic theory that wage age profile is flat towards the end of the career (Heckman et al. (1998); Lagakos et al. (2018)). For more detailed information, see section H.3. Using the estimated profiles, we create predicted earnings over all ages and calculate discounted present value of predicted lifetime earnings for each individual in the analysis sample.

In the cross-sectional analysis we use two measures of predicted wages: predicted wage at age 45 and discounted present value of predicted wages over the lifecycle.

3. Notation and Definitions

We summarize below the notation and variables used to model acquisition of education and determination of earnings. Θ is the set of intelligence values; X the set of family background; Y the set of non-cognitive factors affecting education acquisition. The product space of these individual characteristics is $Z \equiv \Theta \times X \times Y$.

Individuals can acquire human capital: we denote $H \equiv \{nc, c\}$ set of human capital values, where nc stand for non-college, and c college. Human capital can be acquired by provision of effort: we let $E \subseteq \mathbb{R}_+$ a compact set of efforts levels; $\pi(e, \theta)$ probability of achieving a college education with effort e.

A population consists of set of individuals of different ages, with $A \equiv \{1, \ldots, L\} \subseteq \mathbb{N}$ is the set of productive ages, in units of one year; $\mathbb{N}$ is the set of biological ages. Individuals discount the future: $\delta \in (0, 1)$ is the discount factor, or, in different interpretation, the probability of dying in one year.

The next variables describe the population. The distribution of characteristics is denoted by the vector $(\xi(z) : z \in Z)$, $\sum_z \xi(z) = 1 - \delta$, where $\xi(z) \geq 0$ is fraction of children of type Z, entering the population every year; $\Delta(H \times Z \times A)$ probability distributions of the population, generic element μ. We denote $w(h, \theta, a)$ wage of type (h, θ, a). Finally $C(e; z)$ is effort cost at effort e for type z. Note that the (x, y) pair only enters into the effort cost in education acquisition.

3.1. Wages and equilibrium.

At steady state, the sequence of wages for different ages is given for the individual. We assume the utility in a year of a person is his or her wages for the current year. The only way the agent can affect his own future stream of incomes is though the education, for which an effort cost is paid in the period before graduation age.

We define the future discounted sum of wage incomes, starting at the first earning year, for a given μ^*:

$$W(h, \theta, \delta) \equiv \sum_{i=gr}^{ret} \Bigg(\delta^{i-1} w^*(h, \theta, i) \tag{1}$$

Note that the μ^* is constant over time.

The agent solves the optimal choice of education at a time before graduation:

$$\max_{e \in E} \left(\pi(e) \left(W(c, \theta, \delta) - W(nc, \theta, \delta) \right) - C(e; z) \right). \tag{2}$$

We will consider cost of effort functions of the form:

$$C(e; z) = \frac{c(e)}{\Gamma(z)} \tag{3}$$

This problem (2) gives an optimal policy $e^*(z)$ for any type z as function of the wage schedule, w^*, which appears implicitly in the W function.

4. Regression Data Analysis

In this section, we discuss the natural regression analysis of the key variables of interest. These observations will motivate the theoretical framework we present in section 5.

First, we explore the relationship between cognitive score, Big 5 and family advantage scores on probability of obtaining a college degree. Figure (1) plots average share of individuals with college degree in each quintile of individual characteristics. As expected, people with higher cognitive score and more favorable family background are more likely to have a college degree. The figure also suggests that there may be some complementarity between the two variables: improvement in family background appears to have slightly larger effect on college shares when cognitive ability is also high.

The regression estimates presented in Table 1 support similar conclusions. Higher values of cognitive, Big 5 and family advantage scores are all associated with higher chances of getting a college degree. A one standard deviation (sd) higher cognitive ability score increases the probability of getting a college degree by about 14-16 percentage points (pp), while 1 sd higher family advantage

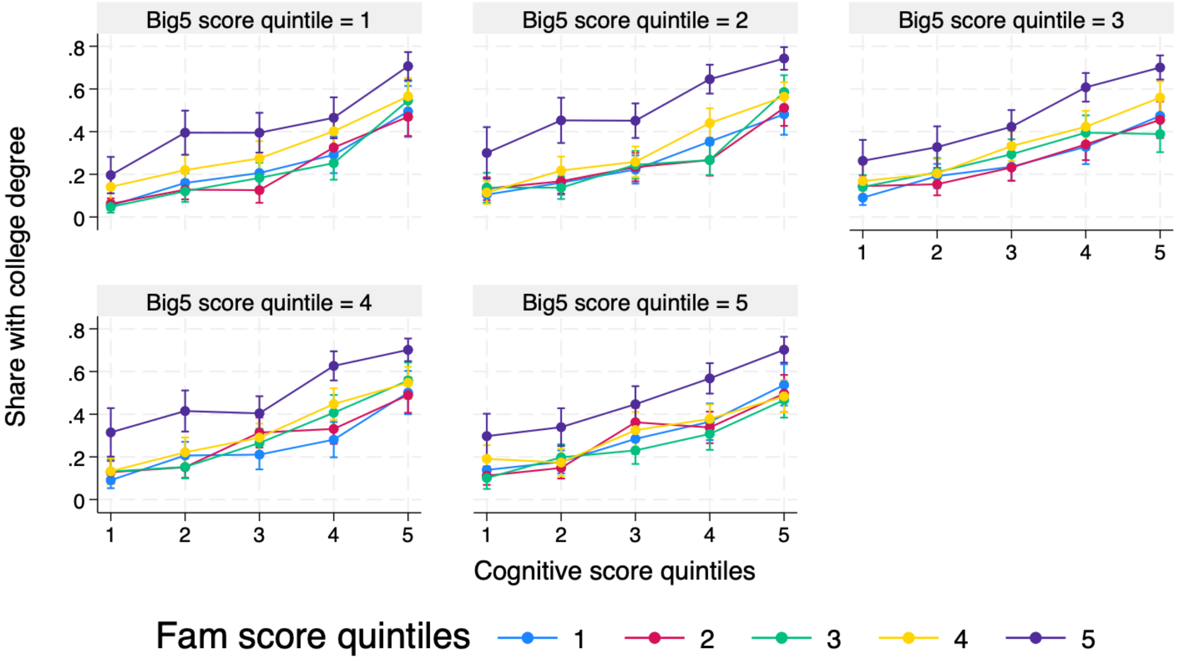


Figure 1. College share by quintiles of individual characteristics

Notes: the figure plots average share college degree in the working sample by quintiles of individual characteristics scores. The sample includes UKHLS wave 3 observations born between 1950 and 1989, with non-missing college, cognitive and Big 5 scores. Observations were weighted using the survey response weights. Error bars correspond to 95% confidence intervals.

score increases it by about 5-10 pp. The results also suggest mild relationship between personality traits and educational attainment (up to 2 pp per 1 sd higher Big 5 score). The estimates display positive complementarity between cognitive ability and family background, although the magnitude is mild. The table also reports the evolution of these associations across cohorts. While the correlation with intelligence remained nearly flat, the effect of favorable family background became stronger over time.

Second, we are interested in the relationship between wages, college indicator and individual characteristics. Figure 2 plots average log predicted wages at age 45 by college indicator and cognitive score quintiles. As expected, both college degree and higher cognitive ability are associated with higher wages. The difference in log wages between top and bottom quintiles (cognitive score premium) appears to be similar between older workers with and without college degree. However, in the younger cohort the cognitive score premium is stronger among workers with college degree.

	Born in 1950-64		Born in 1960-79		Born in 1980-94	
	(1) OLS	(2) Logit ME	(3) OLS	(4) Logit ME	(5) OLS	(6) Logit ME
Cog score	0.137***	0.152***	0.151***	0.172***	0.139***	0.160***
	(0.005)	(0.006)	(0.005)	(0.007)	(0.008)	(0.011)
Fam score	0.049***	0.053***	0.076***	0.087***	0.083***	0.096***
	(0.005)	(0.008)	(0.006)	(0.008)	(0.008)	(0.011)
Big 5 score	0.006	0.015*	0.010	0.016*	0.016*	0.022*
	(0.005)	(0.006)	(0.006)	(0.006)	(0.008)	(0.009)
Cog score × Fam score	0.028***		0.036***		0.035***	
	(0.004)		(0.004)		(0.006)	
Cog score × Big 5 score	-0.004		-0.005		0.001	
	(0.004)		(0.005)		(0.006)	
Obs.	9,539	9,539	10,586	10,586	5,440	5,440

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1. Probability of college degree by individual characteristics

Notes: the table reports regression results with college indicator as the dependent variable separately in each birth cohort. Logit ME are estimates of the marginal effects after logit regression computed to be comparable to the OLS estimates. All individual characteristics scores are standardized to mean 0 and standard deviation 1. The sample includes UKHLS wave 3 observations born between 1950 and 1989, with non-missing college, cognitive and Big 5 scores. The regressions were weighted using the survey response weights. Standard errors clustered at the survey sampling unit are reported in parentheses.

Table 2 reports estimation results of log predicted wages (OLS) or level of predicted wages (Poisson) at age 45 on college indicator and individual characteristics. Both sets of coefficients can be interpreted as wage semi-elasticities. For example, workers with a college degree have about 40% higher wages compared to workers without a degree. Similarly, a 1 sd higher cognitive ability score is associated with almost 15% increase in real hourly wages of 1950-64 birth cohort. The cognitive score premium almost halves in 1980-89 birth cohort, consistent with the visual evidence above.

4.1. **Panel data analysis.** The analysis reported in the two tables Figure 2 and Table 2 is based on cross-sectional variation in the predicted wages at age 45. In addition to that we can use the panel dimension of the UKHLS to explore if age profiles differ with intelligence. For this part of the analysis we use observed wages in the data instead of the predicted series. The estimated profiles are presented in Figure 3. Note that within transformation of the panel data removes level differences between college and non-college workers. The estimates confirm that age profiles of college-educated workers are steeper. Among college-educated workers, the slope of wage-age

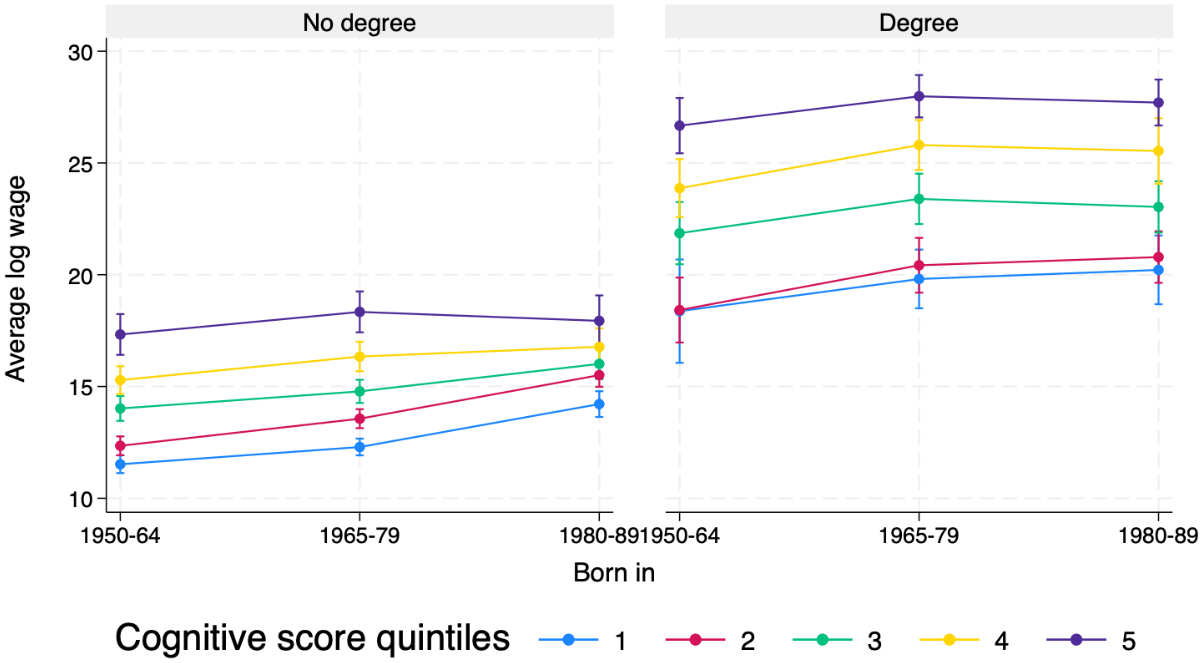


Figure 2. Average wage by college and cognitive score quintiles

Notes: The figure plots average predicted log real hourly wages at age 45 across birth cohorts by college indicator and cognitive score quintiles. The sample includes UKHLS wave 3 observations born between 1950 and 1989, with non-missing college indicator, cognitive and Big 5 scores. Observations were weighted using the survey response weights. Error bars correspond to 95% confidence intervals.

profile seems increase slightly with cognitive ability score. However, the magnitude of an additional effect from higher cognitive score is mild at best.

The above estimates suggest that while college degree and higher cognitive ability each contribute to higher wages, their interaction has negligible effect on either the level or the slope lifetime wage profiles.

Finally, we estimate a simple SEM system of wages and college attainment as functions of individual characteristics to get an approximate estimate of their indirect and direct effects. The results are reported in Table 3.

5. Non-linear estimation

5.1. **Model of effort.** In this section we proceed beyond descriptive statistics. We formulate and test a model of the probability of acquiring college, and estimate the relevant parameters.

We begin by examining the the restrictions imposed on the probability of getting a college degree that can be derived from the assumption that such probability is induced by an optimal

	Born in 1950-64		Born in 1960-79		Born in 1980-89	
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	Poisson	OLS	Poisson	OLS	Poisson
Male	0.220***	0.244***	0.189***	0.208***	0.114***	0.147***
	(0.018)	(0.014)	(0.015)	(0.012)	(0.021)	(0.015)
College	0.403***	0.434***	0.436***	0.434***	0.374***	0.381***
	(0.026)	(0.021)	(0.019)	(0.015)	(0.021)	(0.019)
Cog score	0.125***	0.151***	0.121***	0.140***	0.069***	0.079***
	(0.012)	(0.009)	(0.010)	(0.008)	(0.014)	(0.011)
Fam score	0.027*	0.029**	0.022*	0.035***	0.034**	0.030**
	(0.011)	(0.009)	(0.010)	(0.008)	(0.012)	(0.010)
Big 5 score	0.005	0.003	-0.003	0.019*	0.018	0.022*
	(0.012)	(0.008)	(0.012)	(0.009)	(0.011)	(0.009)
College × Cog score	0.013	-0.005	0.020	-0.007	0.030	0.027
	(0.025)	(0.020)	(0.019)	(0.014)	(0.022)	(0.017)
College × Fam score	-0.007	-0.007	0.035*	0.027	-0.001	0.002
	(0.025)	(0.018)	(0.017)	(0.015)	(0.018)	(0.017)
College × Big 5 score	-0.022	-0.016	-0.013	-0.025	-0.002	-0.007
	(0.027)	(0.017)	(0.018)	(0.014)	(0.019)	(0.016)
Constant	2.364***	2.492***	2.461***	2.584***	2.621***	2.687***
	(0.014)	(0.011)	(0.014)	(0.011)	(0.016)	(0.013)
Obs.	6,933	6,896	8,676	8,582	4,394	4,377

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2. Determinants of wage at 45: college and individual characteristics

Notes: the table reports regression results with predicted wages at age 45 as the dependent variable. Column 1 reports results from OLS regression of log predicted wages at age 45 and column 2 reports results from Possion regression of level of predicted wages at age 45. The individual characteristics scores are standardized to mean 0 and standard deviation 1. The sample includes UKHLS wave 3 observations born between 1950 and 1989, with non-missing college, cognitive and Big 5 scores. The regressions were weighted using the survey response weights. Standard errors clustered at the survey sampling unit are reported in parentheses.

choice of effort. The conclusion we report is that the restrictions are only the natural qualitative ones; specifically, that of the function being increasing in the incentive of obtaining the degree, upper-semicontinuous in the variable expressing the net benefit.

5.2. **Formulation of the problem.** We consider the space of efforts $\mathbb{R}_+$. A probability of obtaining the college degree is a function of the effort. We define the set of such functions:

Definition 5.1. Π *is the set of functions* $\pi : \mathbb{R}_+ \to [0,1]$ *that are strictly increasing, concave,* $\pi(0) = 0$, $\lim_{e\to+\infty} \pi(e) = 1$, *continuous at* 0.

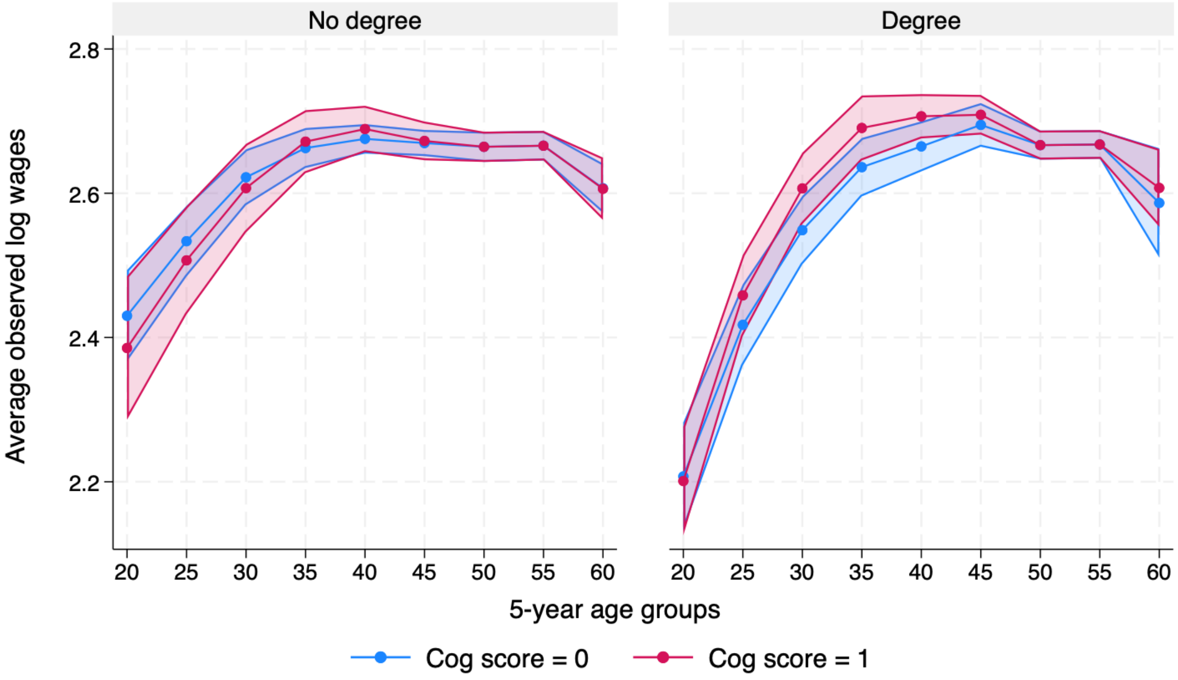


FIGURE 3. Wage age profiles, college and cognitive ability

Notes: the figure plots estimated age profiles of log real hourly wages by college and cognitive score. The estimates were obtained using fixed effects regression weighted using cross-sectional response weight from wave 3 and with standard errors clustered at an individual level. The sample includes UKHLS wave 3 observations born between 1950 and 1989, with non-missing college, cognitive and Big 5 scores. Each individual has up to twelve wage observations. Shaded areas correspond to 95% confidence intervals.

Taking into account the form (3) of the effort cost, the optimal effort choice problem can be formulated as follows:

Definition 5.2. *For a given discounted wage difference* $\Delta(W, \delta)$ *and a cost reduction value* $\Gamma(z)$*, the optimal effort problem solves:*

$$\max \left\{ \pi(x)\Delta(W, \delta) - \frac{c(x)}{\Gamma(z)} : x \geq 0 \right\} \Bigg(\tag{4}$$

We denote by the generic term A the quantity which is really relevant in effort choice:

$$A \equiv \Gamma(x)\Delta W(\theta, \delta). \tag{5}$$

By relabeling the effort in terms of the units of cost we can formulate the problem of the choice of the optimal effort for any non-negative real number A as:

$$H(A; \pi) \equiv \ \text{argmax} \ \{\pi(x)A - x : x \in \mathbb{R}_+\} \tag{6}$$

	Born in 1950-64		Born in 1960-79		Born in 1980-89	
	Poisson	Logit	Poisson	Logit	Poisson	Logit
	Pred. wage 45	College	Pred. wage 45	College	Pred. wage 45	College
Male	0.244***	0.168**	0.207***	-0.027	0.147***	-0.245**
	(0.016)	(0.056)	(0.013)	(0.051)	(0.018)	(0.084)
College	0.435***		0.448***		0.398***	
	(0.020)		(0.015)		(0.019)	
Cog score	0.154***	0.896***	0.144***	0.793***	0.098***	0.704***
	(0.009)	(0.040)	(0.007)	(0.034)	(0.009)	(0.049)
Fam score		0.314***		0.401***		0.424***
		(0.047)		(0.038)		(0.049)
Big 5 score		0.090*		0.072*		0.095*
		(0.035)		(0.029)		(0.040)
Cog score × Fam score		0.007		0.064		0.081
		(0.046)		(0.040)		(0.046)
Cog score × Big 5 score		-0.113**		-0.084**		-0.051
		(0.037)		(0.032)		(0.044)
Marginal effect of cognitive score on predicted wages at age 45						
Indirect effect	0.064		0.069		0.056	
	(0.004)		(0.003)		(0.004)	
Total effect	0.218		0.214		0.154	
	(0.009)		(0.007)		(0.010)	
Obs.	9,496		10,488		5,419	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 3. SEM of wages, college and cognitive ability

Notes: the table reports SEM regression results with predicted wages at age 45 and college as the dependent variables. The individual characteristics scores are standardized to mean 0 and standard deviation 1. The sample includes UKHLS wave 3 observations born between 1950 and 1989, with non-missing college, cognitive and Big 5 scores. The regressions were weighted using the survey response weights. Standard errors clustered at the survey sampling unit are reported in parentheses.

It is convenient to present the problem optimization problem in a form, which is closer to the classical formulation of the conjugate function in convex analysis. Thus we define:

$$h(\alpha; \pi) \equiv \ \text{argmax}\ \{\pi(x) - \alpha x : x \in \mathbb{R}_+\}. \tag{7}$$

The maximization problems in (6) at A has the same solution as the problem (7) for $\alpha = \frac{1}{A}$.

5.3. **Estimation problem.** Our data describe a frequency of college degree for a given vector z of individual characteristics, an observed vector of wage paths $W(h, \theta, a)$.

Our unknowns are the function giving the probability of college degree for a given effort, represented by the function π, and the discount factor δ. We assume the function Γ to be linear, with unknown coefficients. Our problem is to characterize the set of functions $P : \mathbb{R}_+ \to [0,1]$ of the variable A such that, for some $\pi \in \Pi$:

$$P(A) = \pi(H(A;\pi)). \tag{8}$$

An equivalent problem, and the one we consider below, is that of finding a function Q such that $Q(\alpha) = \pi(h(\alpha;\pi))$. In general H and h are multivalued closed valued functions.

Recalling the definition (5.1) of probability of college for given effort, we define:

Definition 5.3. *The set of endogenous probabilities is the set $\mathcal{Q}$ of multivalued functions $Q : \mathbb{R}_+ \to [0,1]$ that are decreasing, closed valued, with $\lim_{\alpha\to 0} Q(\alpha) = 1$, $Q(\overline{\alpha}) = 0$ for some $\overline{\alpha} > 0$.*

The function Q icorresponds to the observed data (a probability of college associated with the combination of incentives –the college premium – and the cost of effort associated with a profile of family background and personality). A simple example will illustrate a small difficulty in determining whether a function $\pi \in \Pi$ exists that induces a given function Q. Let $\Lambda(\alpha) \equiv (1 + e^{-\alpha})^{-1}$ be the logit, and define:

$$Q(\alpha) = \min\{2(1 - \Lambda(\alpha)) + C, 1\}) \tag{9}$$

For any $C \geq 0$ the function in (9) is decreasing, $Q(0) = 1$, $\lim_{\alpha\to+\infty} Q(\alpha) = 0$. But with $C = 0$, at α close to 0 the derivative $\frac{Q'}{\alpha}$ is approximately $\frac{1}{\alpha}$, so it is not integrable. So at $C = 0$ there is no π that induces the Q globally.

5.4. **Differentiable probability.** We deal first with the simpler case in which the multivalued function Q is a continuously differentiable function. Below we consider $0 < \underline{\alpha} < \overline{\alpha} < +\infty$.

Proposition 5.4. *For any function $Q \in \mathcal{Q}$ which is continuously differentiable strictly decreasing in the interval $[\underline{\alpha}, \overline{\alpha}]$, with $Q(\underline{\alpha}) = 0$, $Q(\overline{\alpha}) = 1$, there exists a continuously differentiable function $\pi \in \Pi$ such that for all $\alpha \in \mathbb{R}_+$, $Q(\alpha) = \pi(h(\alpha;\pi))$.*

Proof. Note that $Q(\alpha) = 0$ for $\alpha > \overline{\alpha}$, so we may take the boundary condition

$$h(\overline{\alpha}) = 0. \tag{10}$$

Consider the ordinary differential equation

$$\frac{dh}{d\alpha} = \frac{Q'}{\alpha}, \alpha > 0. \tag{11}$$

We now define the function π as the solution of

$$\pi(h(\alpha)) = Q(\alpha) \tag{12}$$

The function h satisfies the equation (11), which is the first order necessary and sufficient conditions for the problem (7), namely $\pi'(h(\alpha)) = \alpha$. Thus our claim follows. □

A similar result can be proved for the more general case described in definition (5.3).

Proposition 5.5. *For any function $Q \in \mathcal{Q}$, defined in the interval $[\underline{\alpha}, \overline{\alpha}]$, that satisfies: $Q(\underline{\alpha}) = 0$, $Q(\overline{\alpha}) = 1$, there exists a continuous concave function $\pi \in \Pi$ such that for all $\alpha \in \mathbb{R}_+$, $Q(\alpha) = \pi(h(\alpha; \pi))$.*

Proof. Construct a sequence of nested partitions of the interval $[\underline{\alpha}, \overline{\alpha}]$ not including the (at most countable) points at which Q is not single valued with norm tending to zero. Call n the index of the partitions. Set $h^n(\overline{\alpha}) = 0$ and for every element in the partition call

$$\Delta h_k^n \equiv \frac{Q(\alpha_k^n) - Q(\alpha_{k-1}^n)}{\alpha_k^n} \tag{13}$$

and define Q^n and h^n at each point in the partition as the sum of the partial increments. To each such Q^n corresponds a function $\pi^n \in \Pi$ which is piece-wise linear concave, uniformly Lipschitz (in α and n). By Ascoli-Arzelà's theorem there is a limit function π which is the claimed function. □

5.5. **Coefficient Estimations.** In this section we present the estimation results for different functional forms for the dependence of the probability of college on effort. The comparison between estimated coefficients and marginal effects obtained in different models will provide an estimate of the robustness of the results to different functional specifications. We consider the following models:

(1) linear: $P(A) = A$;
(2) logit: $P(A) = \frac{\exp(A)}{1+\exp(A)}$;
(3) cutoff power: $P(A) = \min\{\max\{A, 0\}, 1\}^{\upsilon}, \upsilon \in \mathbb{R}_+$;

(4) power of logit: $P(A) = \left(\frac{\exp(A)}{1+\exp(A)}\right)^{\upsilon}, \forall \upsilon \in \mathbb{R}_{+}$.

The family of power functions in (3) is rich, in the sense that it allows a set of functions of different curvatures. The reason for the introduction of the power of logit model (number 4 in the list) is that (3) poses a restriction on the predicted relationship between independent variables and college outcome that is not observed in the data: once the parameters in the function Γ are given, when the value of A is negative the fraction of college should be zero, independently of the value of A. This is not what we observe, hence we enrich the set of models considered allowing (4).

An important component of A is the discounted wage difference over the working lifetime, denoted $\Delta W(\theta, \delta)$, which we call *college premium*. To compute this variable we fit the following linear regression:

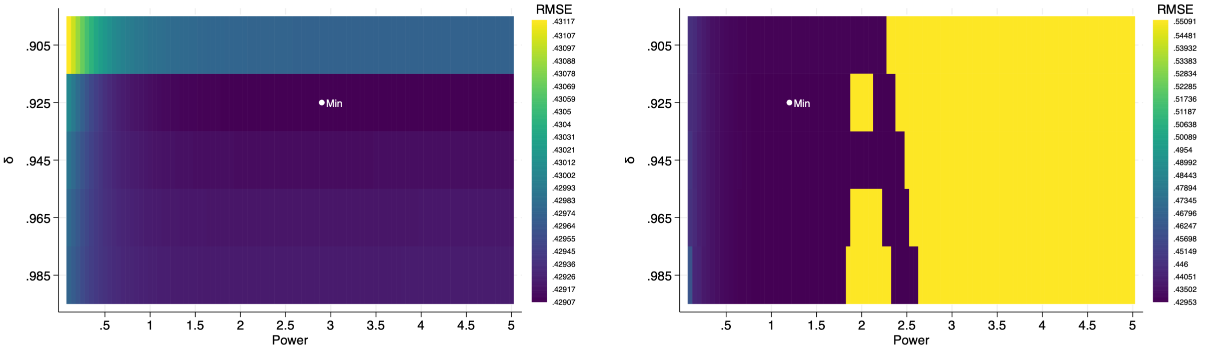
$$W_{i,\delta} = \varphi_0 + \varphi_1 Z_i + \varphi_2 \cdot \mathbb{1}\{H_i = c\} + \varphi_3 \cdot Z_i \cdot \mathbb{1}\{H_i = c\} + u_i \tag{14}$$

where $W_{i,\delta}$ is the discounted present value of annualised predicted wages[6] (in thousands of pounds) given the discount factor δ and Z_i is the vector of cognitive, Big 5 and family advantage scores. Then, we calculate $\Delta W(Z, \delta) = \hat{\varphi}_2 + \hat{\varphi}_3 Z_i$ as average college difference in annualised predicted wages given individual characteristics in Z.

This allows us to create variable equivalent to A in the dataset and fit the models (1)-(4) to the observed college indicator. We do the estimation of models (3) and (4) in two steps, borrowing the idea from threshold regression. First, we fix the values of the discount factor δ and power υ. In particular, we choose a coarse grid for $\delta = (0.905, 0.925, 0.945, 0.965, 0.985)$ motivated by prior estimates of discount factor in the literature. For power, however, we choose finer and wider grid ranging from 0 to 5, since we do not have a strong prior. In the second step, we choose the pair (δ, υ) that minimises the root mean squared error ($RMSE$) of the residuals. Figure (4) reports the heat-map for the $RMSE$ in the power of logit model (4a) and cutoff power (4b) estimations. The optimal pair (δ, υ) is indicated by the white dot. The minimum value of $RMSE$ is achieved at a discount factor of 0.905, the same for both models[7]. With that discount, the min-$RMSE$ power is 2.95 for the power of logit model ($RMSE = 0.4291$), and 1.2 for the cutoff power model ($RMSE = 0.4292$).

[6] Annualised wages = hourly wages × 40 hours × 52 weeks

[7] We, therefore, use same value when fitting models (1) and (2).



(A) Power of logit model

(B) Power model

FIGURE 4. Heat-maps of $RMSE$ for the power of logit and power model. Panels as indicated in the caption.

Notes: the figure plots the $RMSE$ of residuals from first-step estimations of cutoff power (right) and power of logit models (left) across tentative values of the discount factor δ and power υ. Higher values of $RMSE$ are coloured lighted, and lower values - darker. The optimal pair $(\hat{\delta}, \hat{\upsilon})$ is indicated by a white dot in each plot. The estimation sample is based on UKHLS wave 3 born between 1950 and 1989 with non-missing college, cognitive and Big 5 scores.

5.6. **Marginal Effects.** The next table (4) presents the estimated marginal effects for different estimation models. In all cases we use the optimal pair of estimated functions and discount for OLS, logit and the two classes of models. The first column reports results for the linear model, the second the standard logit, the third for the power function, and the fourth the model with the power of logit. The college premium is computed at the discount factor value that minimizes the $RMSE$, 0.905. All independent variables, including the college premium, are standardized, so the sizes of the marginal effects are comparable.

The marginal effects are substantially stable across different models. The marginal effect of the standardized cognitive score is the largest, and is approximately twice as large as that of the family background, and two to three times as large as that of the Big 5 score and of the standardized lifetime college premium.

Table (4) reports the marginal effects of variables with the standard deviation as unit. To evaluate the effect of college premium we may prefer to measure the college premium variable as lifetime discounted difference of the two values (college and non-college) in *thousands of pounds*. This variable has a mean of 36.9, and standard deviation of 9.8. From the two last models in table (4) we see that the additional £1,000 in college premium raises the probability of college by about half of a percentage points (the marginal value is 0.00429 by direct estimation).

In summary, the results of the model estimation are largely in agreement with those obtained with more elementary means, as already reported in figure (1) and in table (1).

	(1)	(2)	(3)	(4)
	OLS	Logit	Cutoff power	Logit power
Cog score	0.122***	0.096***	0.113***	0.109***
	(0.005)	(0.007)	(0.005)	(0.006)
Fam score	0.062***	0.053***	0.059***	0.058***
	(0.005)	(0.006)	(0.005)	(0.005)
Big 5 score	0.024***	0.033***	0.028***	0.028***
	(0.003)	(0.004)	(0.003)	(0.004)
College premium, std	0.029***	0.060***	0.041***	0.042***
	(0.006)	(0.009)	(0.007)	(0.008)
Obs.	25,588	25,588	25,588	25,588
RMSE	0.43	0.43	0.43	0.43
College premium mean	37.85	37.85	37.85	37.85
College premium sd	10.10	10.10	10.10	10.10
Delta	0.905	0.905	0.905	0.905
Power			1.20	2.95

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4. Probability of College: Cognitive skills, personality and family background. Marginal effects

Notes: the table reports marginal effects estimated with different models. The sample includes UKHLS wave 3 observations born between 1950 and 1989, with non-missing college, cognitive and Big 5 score. The individual characteristics scores and college premium are standardized to mean 0 and standard deviation 1. Conventional standard errors are reported in parentheses.

6. PGS Analysis

A common criticism of an analysis that takes cognitive skills as an exogenous characteristic of an individual notes that cognitive ability at age 18 or later is not in fact an exogenous variable, but is the outcome of a rich network of family background, social interactions, and school education (for the direct causal effect of education on intelligence, see Ritchie and Tucker-Drob (2018); Hegelund et al. (2020); Deary and Johnson (2010); Hansen et al. (2004); Ashenfelter and Krueger (1994) use identical twins to estimate effect of schooling.). A related but different criticism is the one presented Heckman et al. (2006), who point out that the estimates obtained using as independent variables latent factors may be very different from those obtained using noisy measurements (as are our cognitive scores). A difficulty inherent in the program they propose, however, is the weakness of the identification. To illustrate this in our case, consider a simple model in which wages are determined by cognitive and non-cognitive latent factors, for which noisy measurements are available. It is easy to see that different parameter vectors of the contribution of latent factors to wages and to noise in the measurements may produce the same observable data.

To address these potential criticisms and establish a proper evaluation of the effect size of cognitive ability on earnings, we no longer consider the measurement provided by the variable θ. We rely instead on the information available for a subset of the original sample of the genetic component of the individual's attitude to acquire education and cognitive skills.

To this end, we compute Polygenic Score (PGS, sometimes called Polygenic Index) for variable of interest. We focus on years of education (EY) as the main trait for PGS as it is one of the most well-studied phenotypes in the genome-wide association studies. A PGS (or PGI) is a single score, specific to each individual, that gives a measure of the attitude or predisposition of that individual for a given trait. The computation of a PGS relies on a Genome-Wide Association Study ($GWAS$). A $GWAS$ identifies common genetic variants that contribute to specific traits (see Bush and Moore (2012); Zeng et al. (2015); Pearson and Manolio (2008); Uffelmann et al. (2021)). The variants considered in the GWA study are Single Nucleotide Polymorphisms (SNP)'s, that is, variations at a single position in a DNA sequence among individuals. A GWA study spans the entire genome (hence the Genome-Wide qualification) and involve a very large number of individuals[8]. Each PGS, for a specific phenotype, is computed as a weighted sum of a person's string of SNP's. The weights are obtained from a $GWAS$, which estimates a coefficient measuring how much any variant is associated with the trait of interest.

We relied on the genotype information available for a subset of the original sample. More detailed information on the data used and the procedure followed is provided in sections I.1 and I.2 of the Appendix.

A first simple descriptive comparison of the effects of the individual characteristics on the probability of obtaining college is provided in figure (5): A steeper curve in the figure suggests a larger role of that factor in explaining college variable. Thus, observed cognitive score appears to be the most important factor explaining college variable in the data. Recall from Table 1 that a 1 sd higher observed cognitive score is associated with about 15 pp higher college probability. Family advantage score curve is as steep as the cognitive score curve in the upper half of the data. The EY PGS curve is slightly flatter than that of cognitive score and, finally, the Big 5 score curve is

[8] For example, the latest GWAS of educational attainment (EY) uses genotypes from 3 million individuals (Okbay, 2022). GWAS for fluid intelligence score uses genotypes of more than 250 thousand individuals in the UK Biobank (Savage et al., 2018).

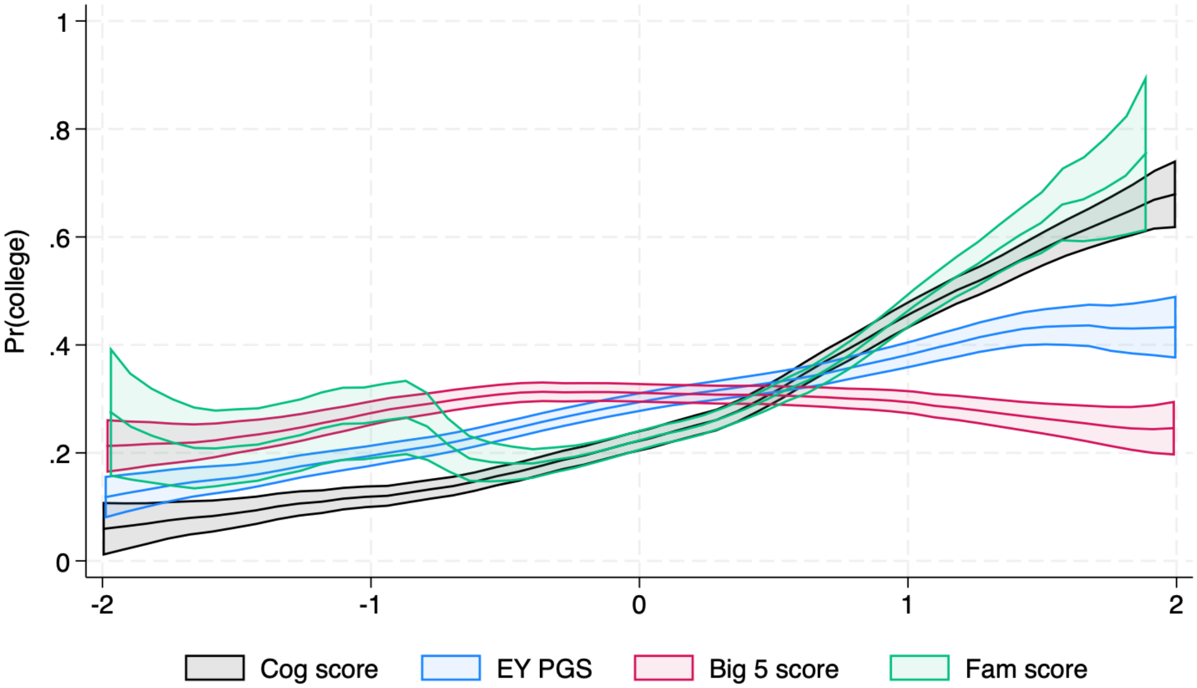


Figure 5. Individual characteristics and Pr(College)

Notes: the figure plots average college shares given individual scores (observed IQ score, EY PGS, family advantage score and Big 5 score) in the METADAC subsample.

almost completely flat except a hump in the middle. In the figure all variables are standardized to have mean zero and standard deviation one. It should be kept in mind that two thirds of the data are in the interval between -1 and 1.

Table (5) corresponds to table (1), with the measurement of cognitive ability (the variable θ) replaced by the $EYPGS$. The marginal effects of the $EYPGS$ are reduced by approximately one third of those for the cognitive score, but the significance level of the estimated coefficients are similar. In both cases they are substantially larger than the corresponding values for personality traits ($Big5score$), while the marginal effects of EY PGS and $Famscore$ are now comparable in magnitude.

Table (6) below reports the result of the SEM for college indicator (logit regression) and log-wages at 45. The marginal effect on log-wages of the $EYPGS$ is substantially lower than the one for college, but remains significant statistically and economically. A 1 sd higher EY PGS is associated with up to 4 % higher wages. The point estimate for the younger cohort is slightly lower. However, the sample size for the younger generation is approximately one third contributing

	Born in 1950-64		Born in 1960-79		Born in 1980-89	
	(1) OLS	(2) Logit ME	(3) OLS	(4) Logit ME	(5) OLS	(6) Logit ME
EY PGS	0.088*** (0.010)	0.085*** (0.011)	0.105*** (0.012)	0.110*** (0.014)	0.087*** (0.020)	0.097*** (0.023)
Fam score	0.066*** (0.011)	0.077*** (0.013)	0.101*** (0.012)	0.131*** (0.017)	0.104*** (0.021)	0.129*** (0.026)
Big 5 score	0.031** (0.010)	0.031** (0.010)	0.000 (0.013)	0.001 (0.013)	0.007 (0.020)	0.007 (0.021)
EY PGS × Fam score	0.032** (0.012)		0.032* (0.013)		0.012 (0.021)	
EY PGS × Big 5 score	0.013 (0.010)		-0.004 (0.012)		0.004 (0.018)	
Obs.	1,863	1,863	1,472	1,472	537	537

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5. College, individual characteristics and PGS
Notes: the table reports marginal effects of individual characteristics and EY PGS on college probability. Columns 1, 3 and 5 report simple regression coefficients from linear probability model. Columns 2, 4, and 6 report marginal effects after logit estimations. The estimations are run separately by 15-year birth cohort groups.

to larger standard errors. So, the coefficients of EY PGS between the two youngest birth cohorts are statistically indistinguishable.

In addition to the direct effect conditional on education, EY PGS also has an indirect effect through college acquisition. The decomposition is reported in the lower part of Table 6. For example, a 1 sd higher EY PGS in the middle cohort is associated with a 0.5 log points increase in college odds ratio, while college degree is associated with 40% higher predicted wages. This translates to 4% higher predicted wages for a 1 sd higher EY PGS via college degree channel. It is clear from the table that the association between EY PGS and predicted wages is largely mediated through college acquisition. About half of total association between EY PGS and predicted wages can be attributed to college degree.

The table (7) below reports the results of the non-linear estimates. It corresponds to the earlier table (4), with the cognitive score variable replaced by the $EYPGS$. The coefficient of EY PGS is now lower than that cognitive score in Table (4). At the same time, the coefficients of the rest of the variables go up.

	Born in 1950-64		Born in 1960-79		Born in 1980-89	
	Poisson Pred. wage 45	Logit College	Poisson Pred. wage 45	Logit College	Poisson Pred. wage 45	Logit College
Male	0.332***	0.203	0.267***	-0.042	0.210***	0.045
	(0.012)	(0.113)	(0.012)	(0.121)	(0.019)	(0.189)
College	0.441***		0.418***		0.408***	
	(0.013)		(0.012)		(0.019)	
EY PGS	0.037***	0.511***	0.040***	0.534***	0.020*	0.421***
	(0.006)	(0.066)	(0.006)	(0.068)	(0.010)	(0.101)
Fam score		0.460***		0.638***		0.559***
		(0.084)		(0.090)		(0.119)
Big 5 score		0.183**		0.003		0.033
		(0.062)		(0.065)		(0.093)
EY PGS × Fam score		0.072		0.040		-0.054
		(0.087)		(0.095)		(0.128)
EY PGS × Big 5 score		0.033		-0.024		0.018
		(0.064)		(0.068)		(0.086)
Marginal effect of EY PGS on predicted wages at age 45						
Indirect effect	0.040		0.044		0.036	
	(0.004)		(0.005)		(0.008)	
Total effect	0.077		0.084		0.055	
	(0.008)		(0.007)		(0.012)	
Obs.	2,309		1,986		703	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6. SEM of wages and college in METADAC subsample

Notes: the table reports SEM regression results with predicted log wages at age 45 and college indicator as the dependent variables. EY PGS, family advantage and Big 5 scores are standardised to mean 0 and standard deviation 1. The regressions are run in METADAC subsample born in 1950-89 with non-missing college, cognitive and Big 5 scores. Standard errors are reported in parentheses.

7. Conclusions

This study proposed to estimate the effect of individual characteristics, including but not limited to cognitive skills, as well as family background, to the earnings of individuals. We relied on a large dataset which included detailed information on the dependent and independent variables. We provided a quantitative estimate of the effect of cognitive ability on earnings, by examining two pathways: along one, cognitive ability directly affects earnings, perhaps through changes in productivity. Along the other, congitive skills affect the probability of acquiring higher education.

	LPM	Logit	Cutoff power	Logit power
PGS EY	0.048***	0.033*	0.043**	0.037*
	(0.014)	(0.016)	(0.014)	(0.015)
Fam score	0.078***	0.131***	0.110***	0.125***
	(0.011)	(0.013)	(0.012)	(0.013)
Big 5 score	0.071***	0.074***	0.075***	0.073***
	(0.017)	(0.020)	(0.018)	(0.019)
College premium	0.075***	0.079**	0.077**	0.076**
	(0.022)	(0.027)	(0.024)	(0.026)
Obs.	3,872	3,872	3,872	3,872
Delta	0.905	0.905	0.905	0.905
Power			1.20	2.95
RMSE	0.4351	0.4327	0.4340	0.4333
College premium mean	45.78	45.78	45.78	45.78
College premium sd	10.59	10.59	10.59	10.59

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7. Non-linear estimations with EY PGS

Notes: the table reports marginal effects estimated with the different models as indicated. EY PGS, family advantage and Big 5 scores are standardised to mean 0 and standard deviation 1. The regressions are run in METADAC subsample born in 1950-89 with non-missing college, cognitive and Big 5 scores. Standard errors are reported in parentheses.

The three set of factors (cognitive skills, personality and family background) contribute to earnings, but the size of the marginal effect of cognitive skills is larger than each of the other two (see table 3). A similar conclusion holds for the comparative estimate of the marginal effect on the probability of college acquisition (see table 4).

The data set included information on the genotype of a good sized subset of the original sample. Using this additional information we could provide an estimate of the effect of these original characteristics on education acquisition and earnings (see table 6 in the main text for the Polygenic score on education years, and table 11 in the appendix for that of fluid intelligence). The estimate relying on the Polygenic score on education years cannot separate between the relative contribution of cognitive skills and other traits, but they do show that the joint contribution of this characteristics acquired at birth are substantial. The results using the *PGS* for cognitive ability show that the contribution of factors underlying cognitive skills is important.

References

Ashenfelter, O. and A. Krueger (1994): "Estimates of the economic return to schooling from a new sample of twins," *The American economic review*, 1157–1173.

Ashenfelter, O., C. Rouse, et al. (2000): "Schooling, intelligence, and income in America," *Meritocracy and economic inequality*, 89.

Ashenfelter, O. and D. J. Zimmerman (1997): "Estimates of the returns to schooling from sibling data: Fathers, sons, and brothers," *Review of Economics and Statistics*, 79, 1–9.

Bonjour, D., L. F. Cherkas, J. E. Haskel, D. D. Hawkes, and T. D. Spector (2003): "Returns to education: Evidence from UK twins," *American Economic Review*, 93, 1799–1812.

Bound, J. and G. Solon (1999): "Double trouble: on the value of twins-based estimation of the return to schooling," *Economics of Education Review*, 18, 169–182.

Bush, W. S. and J. H. Moore (2012): "Chapter 11: Genome-wide association studies," *PLoS computational biology*, 8, e1002822.

Card, D. and T. Lemieux (2001): "Can falling supply explain the rising return to college for younger men? A cohort-based analysis," *The quarterly journal of economics*, 116, 705–746.

Deary, I. J. and W. Johnson (2010): "Intelligence and education: causal perceptions drive analytic processes and therefore conclusions," *International journal of epidemiology*, 39, 1362–1369.

Demange, P. A., M. Malanchini, T. T. Mallard, and et al. (2021): "Investigating the Genetic Architecture of Noncognitive Skills Using GWAS-by-subtraction," *Nature Genetics*, 53, 35–44.

Ervik, A. O. (2003): "IQ and the Wealth of Nations." *The Economic Journal*, 113, F406–F408.

Fisher, P., L. Fumagalli, N. Buck, and S. Avram (2019): "Understanding Society and its income data," Working Paper 2019 - 08, University of Essex.

Griliches, Z. (1976): "Wages of very young men," *Journal of Political Economy*, 84, S69–S85.

Griliches, Z. and W. M. Mason (1972): "Education, income, and ability," *Journal of political Economy*, 80, S74–S103.

Hansen, K. T., J. J. Heckman, and K. J. Mullen (2004): "The effect of schooling and ability on achievement test scores," *Journal of econometrics*, 121, 39–98.

Hanushek, E. A., G. Schwerdt, S. Wiederhold, and L. Woessmann (2015): "Returns to skills around the world: Evidence from PIAAC," *European Economic Review*, 73, 103–130.

HECKMAN, J. J., L. LOCHNER, AND C. TABER (1998): "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents," *Review of Economic Dynamics*, 1, 1–58.

HECKMAN, J. J., J. STIXRUD, AND S. URZUA (2006): "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior," *Journal of Labor economics*, 24, 411–482.

HEGELUND, E. R., M. GRØNKJÆR, M. OSLER, J. DAMMEYER, T. FLENSBORG-MADSEN, AND E. L. MORTENSEN (2020): "The influence of educational attainment on intelligence," *Intelligence*, 78, 101419.

ICHINO, A., A. RUSTICHINI, AND G. ZANELLA (2022): "College education, intelligence, and disadvantage: policy lessons from the UK in 1960-2004," *CEPR Discussion Paper No. DP17284*.

ISACSSON, G. (1999): "Estimates of the return to schooling in Sweden from a large sample of twins," *Labour Economics*, 6, 471–489.

JOHNSON, W. AND T. J.BOUCHARD (2005): "The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized," *Intelligence*, 33, 393–416.

LAGAKOS, D., B. MOLL, T. PORZIO, N. QIAN, AND T. SCHOELLMAN (2018): "Life Cycle Wage Growth across Countries," *Journal of Political Economy*, 126, 797–849.

LINDQVIST, E. AND R. VESTMAN (2011): "The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment," *American Economic Journal: Applied Economics*, 3, 101–128.

LYNN, R. AND T. VANHANEN (2002): *IQ and the wealth of nations*, Bloomsbury Publishing USA.

MILLER, P., C. MULVEY, AND N. MARTIN (1995): "What do twins studies reveal about the economic returns to education? A comparison of Australian and US findings," *The American Economic Review*, 85, 586–599.

NI, G., J. ZENG, J. A. REVEZ, AND ET AL. (2021): "A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts," *Biological Psychiatry*, 90, 611–620.

OKBAY, A. E. A. (2022): "Polygenic Prediction of Educational Attainment within and between Families from Genome-Wide Association Analyses in 3 Million Individuals," *Nature Genetics*, 54, 437–449.

OTTAVIANO, G. I. AND G. PERI (2012): "Rethinking the effect of immigration on wages," *Journal of the European economic association*, 10, 152–197.

Pearson, T. A. and T. A. Manolio (2008): "How to interpret a genome-wide association study," *Jama*, 299, 1335–1344.

Privé, F., J. Arbel, and B. J. Vilhjálmsson (2021): "LDpred2: Better, Faster, Stronger," *Bioinformatics*, 36, 5424–5431.

Ritchie, S. J. and E. M. Tucker-Drob (2018): "How much does education improve intelligence? A meta-analysis," *Psychological science*, 29, 1358–1369.

Savage, J. E., P. R. Jansen, S. Stringer, and et al. (2018): "Genome-Wide Association Meta-Analysis in 269,867 Individuals Identifies New Genetic and Functional Links to Intelligence," *Nature Genetics*, 50, 912–919.

Uffelmann, E., Q. Q. Huang, N. S. Munung, J. De Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma (2021): "Genome-wide association studies," *Nature Reviews Methods Primers*, 1, 59.

Vilhjálmsson, B. J. e. a. (2015): "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores," *The American Journal of Human Genetics*, 97, 576–592.

Zeng, P., Y. Zhao, C. Qian, L. Zhang, R. Zhang, J. Gou, J. Liu, L. Liu, and F. Chen (2015): "Statistical analysis for genome-wide association study," *Journal of biomedical research*, 29, 285.

Appendix A. Description of the Economy

Appendix B. Production Function

We allow for separate aggregation of labor inputs of different intelligence but same human capital (see equations 1 and 2 in (Ottaviano and Peri (2012)); see also (Card and Lemieux (2001))). For a given allocation of the population into labor types

$$\mu \equiv (\mu(h,\theta,i) : h \in H, \theta \in \Theta, i \in \mathbb{N})$$

given a vector of positive coefficients $(b(h) : h \in H)$ and

$$(a(h,\tau,j) : h \in H, \tau \in \Theta, j \in A)$$

the total production equal to

$$F(\mu; b, a, \rho, \sigma) \equiv \sum_{h\in H} \bigg(b(h)C(h)^{\rho} \bigg)^{\frac{1}{\rho}} \tag{15}$$

where

$$C(h) \equiv \Bigg(\sum_{\tau\in\Theta, j\in A} \bigg(a(h,\tau,j)\mu(h,\tau,j)^{\sigma} \Bigg)^{\frac{1}{\sigma}} \Bigg(\tag{16}$$

so that

Proposition B.1. *The wage for type* (h,θ,i) *is:*

$$w(h,\theta,i) = F(\mu)^{1-\rho}C(h)^{\rho-\sigma}b(h)a(h,\theta,i)\mu(h,\theta,i)^{\sigma-1} \tag{17}$$

Proof. The equation (17) follows from:

$$\begin{aligned} w(h,\theta,i) = \frac{1}{\rho} & \sum_{h\in H} \bigg(b(h)C(h)^{\rho} \bigg)^{\frac{1-\rho}{\rho}} \\ & \times b(h)\rho C(h)^{\rho-1} \\ & \times \frac{1}{\sigma} \Bigg(\sum_{\tau\in\Theta, j\in A} \bigg(a(h,\tau,j)\mu(h,\tau,j)^{\sigma} \Bigg)^{\frac{1-\sigma}{\sigma}} \Bigg(\\ & \times \sigma a(h,\theta,i)\mu(h,\theta,i)^{\sigma-1} \end{aligned}$$

□

When $\rho = \sigma$ we have the standard CES with coefficient for the type (h, θ, i) equal to $b(h)a(h, \theta, i)$. In this case the wage equation is:

$$w(h, \theta, i) = F(\mu)^{1-\rho} b(h) a(h, \theta, i) \mu(h, \theta, i)^{\rho-1} \tag{18}$$

The wage equation (17) gives the equilibrium wage as a continuous function of μ^*.

Appendix C. Population Process

The dynamical system for μ is simple.

The initial condition for $i \leq G$

$$\mu(NC, \theta, i) = (1-\delta)^i \xi(\theta), \mu(C, \theta, i) = 0. \tag{19}$$

For the G year:

$$\mu(C, \theta, G) = (1-\delta)^G \xi(\theta) \Pi(e^*(\theta), \theta), \tag{20}$$

$$\mu(NC, \theta, G) = (1-\delta)^G \xi(\theta) - \mu(C, \theta, G).$$

For any other age:

$$\mu(h, \theta, i+1) = (1-\delta)\mu(h, \theta, i). \tag{21}$$

Appendix D. Equilibrium Conditions

D.1. **Equilibrium.** This is the equilibrium at steady state in the dynamic model, or simply the equilibrium in the static model. The steady state is characterized by a probability distribution

Definition D.1. *An equilibrium is a vector*

$$(w^*(h, \theta, i), \mu^*(h, \theta, i), e^*(z)) : (h, \theta, i) \in H \times Z \times \mathbb{N})$$

such that:

(1) for every z, $e^*(z)$ is optimal for type z in the year when decides the education investment, given the vector of wages w^*;

(2) w^* is the equilibrium wage vector, equal to the marginal product of the corresponding type at μ^*;

(3) μ^* is the steady state distribution over characteristics in the population, given e^* and ξ.

D.2. **Parameters.** The set of parameters are:

(1) (a, b, ρ, σ) for the production function;

(2) ξ distribution of characteristics

Appendix E. Equilibrium algorithm

We assume:

Assumption E.1. *(1) The function Π is continuous and strictly increasing in both arguments;*

(2) The function $C(\cdot; z)$ is continuously differentiable and strictly convex for every z.

The equilibrium can be found as a fixed point on the policy function e.

Proposition E.2. *Under assumption (E.1) an equilibrium exists.*

Proof. The algorithm

(1) For a given policy $(e(\theta) : \theta \in \Theta))$ we derive the implied invariant distribution μ;

(2) From this distribution we get wages and can thus compute the optimal education choice policy for every types θ;

A fixed point of this iteration is an equilibrium. The two maps, one from the optimal policy to distribution over characteristics of the population (in section (C)) and the other from distribution to wages (equation 17) are continuous, and the set of population policies compact. □

Appendix F. Linear Coefficients model

Here we consider the model with coefficients in the production function linear in intelligence. We denote the logit:

$$\Lambda(x) \equiv \frac{e^x}{1+e^x} \tag{22}$$

We allocate the population to deciles, so θ ranges in a ten-elements set, and let $p = 0.1$. The human capital h is either 0 (no college) or 1 (college).

We assume the production function to be:

$$F(\mu) = \left(\sum_{\theta,h} \left(\alpha_h\theta + \beta h)\mu(\theta,1)^\rho\right)^{\frac{1}{\rho}} \right(\tag{23}$$

where $(\alpha_1, \alpha_0, \beta)$ are real valued parameters to be estimated, and that measure respectively: (α_1, α_0) the effect of intelligence on productivity (depending on h) and β the shift due to college education.

The supply of college graduates for each class of intelligence θ is a logit of the difference between $w(\theta, 1)$ and $w(\theta, 0)$, multiplied a factor $\Gamma(\theta, x)$ where x is the pair of family and big 5 variables, and the function Γ is linear.

We denote by $\nu_X(x|\theta)$ the conditional probability on the traits and family characteristics, for a given θ. An equilibrium is a vector of $\mu(\theta, 1) \equiv \zeta(\theta)$ that solve:

$$\forall\theta, \zeta(\theta) = pPr(h = 1|\theta, \mu) \tag{24}$$

$$= p\int_X \Lambda\left(\Gamma(\theta,x)(F(\mu)^{1-\rho}[(\alpha_1\theta+\beta)\zeta(\theta)^{\rho-1} - \alpha_0\theta(p-\zeta(\theta))^{\rho-1}])\right) d\nu_X(x|\theta)$$

Note that F can be written as a function of ζ.

For each fixed ρ, the parameters are

$$(\Gamma_\theta, \Gamma_{x_2}, \Gamma_{x_2}, \alpha_1, \alpha_0, \beta) \tag{25}$$

with $\Gamma_\theta = 1 - \Gamma_{x_1} - \Gamma_{x_2}$. The following is clear:

1. For each μ there is a unique solution to the 10 equations (24).
2. For each vector of parameters there is a unique μ that solves (24).

We minimize the distance between the observed and predicted frequencies.

Appendix G. Additional Information on Data And Analysis

G.1. **Coefficient Estimations.** In Table (8) we report the coefficients of the power model.

Table 8. Non-linear estimation results: Power Model

	(1) Pr(college)
α_θ	0.007^{***} (0.000)
α_{Fam}	0.003^{***} (0.000)
α_{Big5}	0.002^{***} (0.000)
Γ intercept	-0.073^{***} (0.003)
Intercept	0.183^{***} (0.023)
Obs.	31,571

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In Table (9) we report the coefficients of the logit and power model.

Table 9. Non-linear estimation: Exponential Power. The table reports the estimated coefficients.

	(1) Pr(college)
α_θ	0.003*** (0.000)
α_{Fam}	0.002*** (0.000)
α_{Big5}	0.001*** (0.000)
Γ intercept	-0.033*** (0.006)
Intercept	-2.357*** (0.425)
Obs.	31,571

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix H. PGS for Intelligence and EA

The results of the analysis in section (6) were obtained relying on the *PGS* for educational attainment. In this section we reconsider those results using instead the *PGS* for Intelligence. We document that nothing substantial changes.

Table (10) reports the equivalent of table (5), replacing the *PGSEY* with the *PGSIQ*:

	Born in 1950-64		Born in 1960-79		Born in 1980-94	
	(1) OLS	(2) Logit ME	(3) OLS	(4) Logit ME	(5) OLS	(6) Logit ME
IQ PGS	0.063***	0.063***	0.075***	0.082***	0.046*	0.049*
	(0.010)	(0.010)	(0.012)	(0.013)	(0.021)	(0.023)
η_{Fam}	0.077***	0.097***	0.111***	0.152***	0.110***	0.132***
	(0.011)	(0.013)	(0.013)	(0.017)	(0.021)	(0.026)
η_{Big5}	0.035***	0.034***	0.002	-0.002	0.013	0.012
	(0.010)	(0.010)	(0.013)	(0.013)	(0.020)	(0.021)
IQ PGS × η_{Fam}	0.007		0.010		0.009	
	(0.011)		(0.012)		(0.021)	
IQ PGS × η_{Big5}	0.004		0.024		0.018	
	(0.011)		(0.013)		(0.020)	
Obs.	1,863	1,863	1,472	1,472	537	537

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10. College, individual characteristics and PGS

Notes: the table reports marginal effects of individual characteristics and IQ PGS on college probability. Columns 1, 3 and 5 report simple regression coefficients from linear probability model. Columns 2, 4, and 6 report marginal effects after logit estimations. The estimations are run separately by 15-year birth cohort groups.

In table (11) we report the results of the *GSEM* presented in table (6), but with the *PGS* computed for the IQ rather than for education years.

	Born in 1950-64		Born in 1960-79		Born in 1980-89	
	Poisson Pred. wage 45	Logit College	Poisson Pred. wage 45	Logit College	Poisson Pred. wage 45	Logit College
Male	0.332***	0.174	0.267***	-0.024	0.207***	-0.001
	(0.012)	(0.112)	(0.012)	(0.119)	(0.019)	(0.186)
College	0.450***		0.427***		0.416***	
	(0.013)		(0.012)		(0.019)	
IQ PGS	0.022***	0.374***	0.030***	0.406***	-0.001	0.218*
	(0.006)	(0.060)	(0.006)	(0.066)	(0.010)	(0.098)
η_{Fam}		0.568***		0.738***		0.570***
		(0.087)		(0.094)		(0.118)
η_{Big5}		0.205***		-0.011		0.055
		(0.060)		(0.065)		(0.091)
IQ PGS $\times\eta_{\text{Fam}}$		-0.071		-0.110		-0.026
		(0.079)		(0.090)		(0.111)
IQ PGS $\times\eta_{\text{Big5}}$		-0.012		0.126		0.080
		(0.064)		(0.066)		(0.092)
Marginal effect of intelligence on predicted wages at age 45						
Indirect effect	0.028		0.032		0.020	
	(0.004)		(0.005)		(0.009)	
Total effect	0.050		0.061		0.019	
	(0.007)		(0.008)		(0.013)	
Obs.	2,309		1,986		703	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 11. SEM of wages and college in METADAC subsample

Notes: the table reports SEM regression results with predicted log wages at age 45 and college indicator as the dependent variables. The SEM estimations allow correlated error terms between the two equations. IQ PGS, Family advantage (η_{Fam}) and Big 5 ($\eta_{\text{Big 5}}$) scores are standardised to mean 0 and standard deviation 1. The regressions are run in METADAC subsample born in 1950-94 with non-missing college and individual scores. Standard errors are reported in parentheses.

In the table (12) below we do the same for table (7):

	LPM	Logit	Cutoff power	Logit power
PGS EY	0.048***	0.033*	0.043**	0.037*
	(0.014)	(0.016)	(0.014)	(0.015)
Fam score	0.078***	0.131***	0.110***	0.125***
	(0.011)	(0.013)	(0.012)	(0.013)
Big 5 score	0.071***	0.074***	0.075***	0.073***
	(0.017)	(0.020)	(0.018)	(0.019)
College premium	0.075***	0.079**	0.077**	0.076**
	(0.022)	(0.027)	(0.024)	(0.026)
Obs.	3,872	3,872	3,872	3,872
Delta	0.905	0.905	0.905	0.905
Power			1.20	2.95
RMSE	0.4351	0.4327	0.4340	0.4333
College premium mean	45.78	45.78	45.78	45.78
College premium sd	10.59	10.59	10.59	10.59

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 12. Non-linear estimations with IQ PGS

Notes:

Robustness of GSEM results between subsamples and earnings prediction methods

	Full cont.		Full binned		GTP binned		MDAC binned			
	Poisson	Logit	Poisson	Logit	Poisson	Logit	Poisson	Logit	Poisson	Logit
	Pred. wage 45	College	Pred. wage 45	College	Pred. wage 45	College	Pred. wage 45	College	Pred. wage 45	College
Male	0.184***	-0.014	0.209***	-0.014	0.262***	0.033	0.280***	0.042	0.282***	0.082
	(0.003)	(0.030)	(0.004)	(0.030)	(0.008)	(0.067)	(0.008)	(0.078)	(0.008)	(0.075)
College	0.424***		0.377***		0.366***		0.362***		0.434***	
	(0.004)		(0.004)		(0.008)		(0.008)		(0.008)	
θ	0.132***	0.757***	0.122***	0.757***	0.125***	0.943***	0.131***	0.940***		
	(0.002)	(0.017)	(0.002)	(0.017)	(0.005)	(0.045)	(0.005)	(0.051)		
EY PGS									0.034***	0.491***
									(0.004)	(0.043)
η_{Fam}		0.324***		0.324***		0.440***		0.419***		0.564***
		(0.018)		(0.018)		(0.055)		(0.060)		(0.055)
η_{Big5}		0.097***		0.097***		0.077		0.087		0.079*
		(0.015)		(0.015)		(0.039)		(0.045)		(0.040)
$\theta \times \eta_{\text{Fam}}$		0.052**		0.052**		0.013		0.029		
		(0.019)		(0.019)		(0.057)		(0.064)		
$\theta \times \eta_{\text{Big5}}$		-0.080***		-0.080***		-0.071		-0.084		
		(0.017)		(0.017)		(0.043)		(0.050)		
EY PGS $\times \eta_{\text{Fam}}$										0.022
										(0.057)
EY PGS $\times \eta_{\text{Big5}}$										0.017
										(0.040)
Marginal effect of intelligence on predicted wages at age 45										
Indirect effect	0.061		0.054		0.060		0.059		0.041	
	(0.001)		(0.001)		(0.003)		(0.003)		(0.003)	
Total effect	0.193		0.176		0.185		0.190		0.075	
	(0.002)		(0.002)		(0.005)		(0.005)		(0.005)	
Obs.	25,420		25,356		5,374		4,998		4,998	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 13. Robustness of GSEM results between subsamples

H.1. **Cognitive score.** In wave 3 (with data collected in the period 2011-13), the UKHLS administered cognitive ability tests to the participants. In addition to variables coding each answer given by each individual, the UKHLS has a set of derived variables counting the number of correct answers to each test. We construct the cognitive score using the counts of correct answers given by an individual in each test. Specifically we use the following variables:

(1) A score of *memory*, based on a word recall task: individuals are asked to recall as many as they can from a list of ten words. There were two parts, immediate (variable *cgwri_dv*: participants were asked to recall words immediately after presentation) and delayed (variable *cgwrd_dv*: participants were asked to recall words after the *serial 7 subtractions* test).

(2) A score *serial 7 subtractions* (variable *cgs7cs_dv* for number of correct subtractions, as opposed to the number of correct answers). Individuals were asked to subtract 7 from the previous number, starting from 100, for five times in sequence.

(3) A score *number series* (variables *cgns1sc6_dv* and *cgns2sc6_dv*). This task requires the respondent to look at a series of numbers. One number is missing from the series, and the respondent must provide the missing number in the series, after identifying the pattern in the series.

(4) A score of *verbal fluency* (variable *cgvfc_dv*). Individuals were asked to name as many unique animals as they could in one minute.

(5) A score of *numerical ability* (variable *cgna_dv*). The variable is coded on a 0 to 5 scale, given by the number of items answered correctly. The tasks consisted of solving simple numerical problems based on everyday day life examples (such as computing the correct change after a purchase).

The correlation between the six scores is positive for all pairs. We combine the scores into a single cognitive ability score using confirmatory factor analysis, following Johnson and J.Bouchard (2005). We first match the UKHLS cognitive tests to the tests used by Johnson and J.Bouchard (2005) based on the description of tasks. We also standardize the scores within each year of birth and gender cells to abstract from possible age-related differences in performances.

H.2. **Big 5 personality score.** In wave 3, the UKHLS had also administered short 15-item Big 5 personality test to adult respondents. In each question, the respondents are asked to respond to a statement with a number from 1 (does not apply to me at all) to 7 (applies to me perfectly). The

UKHLS provides a set of derived variables with average scores across questions in each sub-domain. Specifically,

(1) an *agreeableness* variable (*big5a_dv*) combines scores from statements about being rude (reverse coded), forgiving, and kind;
(2) a *conscientiousness* variable (*big5c_dv*) combines scores from statements about doing thorough job, being lazy (reverse coded) and being efficient;
(3) an *extraversion* variable (*big5e_dv*) combines scores from statements about being talkative, sociable and reserved (reverse coded);
(4) a *neuroticism* variable (*big5n_dv*) combines scores from statements about worrying a lot, being nervours and relaxed (reverse coded); and
(5) an *openness* variable (*big5o_dv*) combines scores from statements about being original, artistic and having active imagination.

All five domains have expected correlation signs: agreeableness, conscientiousness, extraversion and openness are positively correlated with each other and neuroticism score is negatively correlated with the rest. Even though all variables should have similar scales, we standardise them within each 5-year birth cohort and gender cells. This way, we ensure all variables have mean zero and standard deviation one, removing possible age and gender differences. Then, we run principal component analysis and use the first component (which captures 36% of variance in the data) as the Big 5 score.

H.3. **Predicted earnings.** We use the UKHLS dataset to construct also a panel of earnings across twelve waves for the individuals in the working sample. Thus, for each individual we observe at most twelve years of earnings history. These twelve observations cover different sections along the age profile of wages depending on birth cohorts.

We address the issue of the span covered by earnings data by estimating the wage profiles, and using predicted wages. In particular, we fit a fixed-effects regression in Equation (26).

$$\log w_{it} = \alpha + \sum_{a \in \mathcal{A}} \Bigg(\beta_a + \gamma_a \mathbf{X}_i + \delta_t + \mu_i + v_{it} \tag{26}$$

where $\mathcal{A} = [20, 65]$, δ_t are time FEs, β_a are age FEs and γ_a are age FEs interacted with individual characteristics in $\mathbf{X}_i$ (gender and degree indicator). To overcome collinearity between age and time we need to impose an additional restriction to the coefficients. Motivated by a similar application of economic theory in Lagakos et al. (2018), we force the age profiles to be flat between ages 51 and 60. We achieve this by excluding relevant age indicators from the regression equation.

We then calculate predicted wages net of year effects as follows

$$\hat{w}_{ia} = \exp\left(\hat{\alpha} + \hat{\beta}_a + \hat{\gamma}_a \mathbf{X}_i + \hat{\mu}_i\right) \Bigg(\qquad \forall a \in \mathcal{A}$$

where $\hat{\beta}_a = 0 \quad \forall a \in [51, 60]$.

We also compute the discounted present value of predicted wages

$$DPV(w)_i = \sum_{a \in \mathcal{A}} \Bigg(\frac{\hat{w}_{i,a}}{(1+r)^a}$$

Appendix I. Information on GWAS and PGS

I.1. **Details on polygenic scores.** The GWAS coefficients used to generate the polygenic scores were downloaded from Okbay (2022); Savage et al. (2018); Demange et al. (2021). Okbay (2022) estimate GWAS for years of education. The first study (Okbay (2022)) uses various datasets with a total sample of size of more than three million individuals. A sample of 269,867 individuals, part of the UK Biobank dataset, is used in Savage et al. (2018) to estimate GWAS of fluid intelligence score. Finally Demange et al. (2021) estimate GWAS of latent cognitive and noncognitive factors using previously published estimates for years of education and cognitive test scores. They estimate a structural equation model where cognitive latent factor affects both the education and cognitive test score GWAS, while noncognitive factor affects only education GWAS.

Given the GWAS estimates and genotype information in the METADAC we compute polygenic scores for each of these phenotypes using a simple linear scoring method.

$$PGS_i = \sum_k \Bigg(\beta_k g_{ik} \tag{27}$$

where β_k is the GWAS coefficient of SNP k and $g_{ik} \in \{0, 1, 2\}$ is the genotype of individual i at locus k. However, naive scoring using full set of matched SNPs produces biased polygenic score due to correlation of genotypes between loci. Such correlation is called linkage disequilibrium (LD) and results from the fact that variants are inherited in blocks.

The approach we adopt here uses all variants scaling them down according to linkage disequilibrium. The method was introduced by Vilhjálmsson (2015) as LDpred and later updated by Privé et al. (2021) as LDpred2. A detailed comparison of different methods of computation of the scores is in Ni et al. (2021).

I.2. **METADAC dataset.** As part of the general dataset, UKHLS collected genotyping information on 9,920 individuals. This information was collected in waves 2 and 3. Our working sample consists of individuals born in 1950-89 with non-missing college indicator, intelligence score, Big 5 personality score and parental background information. Furthermore, we keep individuals who have been observed at least once between ages 20 and 65 over 8 waves. We have genotype information for 5,579 individuals over a total sample size of 26,643. The genotyped individuals are of higher age, higher intelligence than the general sample.

The METADAC provides genotyped calls at 518,542 variants, with no inputed variants. The genotyping rate is high: 95% of all variants have fewer than 32 (18) missing observations in the full (working) sample. All variants are bi-allelic SNPs (that is, only two of the four possible nucleotides appear at that locus). 130,356 SNPs (or 25% of variants) are fixed in the full METADAC sample, i.e., all individuals in the sample have the same allele at a given locus. The number of fixed SNPs rises in the working METADAC sample to 145,107 SNPs (or 28% of variants)

I.3. **Predicted earnings profile.** Due to privacy issues, the individual identifiers in the METADAC and the UKHLS are different, so it is impossible to link the two datasets. Therefore, we cannot use predicted DPV of lifetime earnings from the main analysis sample and need to repeat the prediction process again in the METADAC. This is further complicated by the fact that the METADAC only provides 50-quantiles of earnings and 5-hour bins of hours worked instead of continuous variables. This means that we have adjust the prediction algorithm to account for this.

First, we generate separate variables for lower and upper bounds of earnings quantiles and hours bins. For simplicity, we use mid-points of hours bins in the rest of the analysis. That is, if in the original data a person has worked 30-34 hours a week, we assume that she worked exactly 32 hours. We then calculate lower and upper bounds of hourly wages by dividing respective thresholds of monthly earnings quantiles on mid-point estimates of hours worked in four weeks.

Recall that in the main analysis sample, predicted earnings are based on the wage-age profile estimation in Equation (26). Denote the lower bound of hourly wages corresponding to earnings quantile q at time t as $w_{qt}^{(1)}$ and the upper bound - $w_{qt}^{(2)}$. Then, the regression equation becomes

$$\Pr\left(\ln w_{it} \in \left[w_{qt}^{(1)}, w_{qt}^{(2)}\right]\right) = \Pr\left(v_{it} + \mu_i \in \left[w_{qt}^{(1)} - \omega(a,i,t), w_{qt}^{(2)} - \omega(a,i,t)\right]\right) \tag{28}$$

where $\omega(a,i,t) \equiv \alpha + \sum_{a \in \mathcal{A}} (\beta_a + \gamma_a X_i) + \delta_t$ is predicted level of wages given age a, individual characteristics X_i and time period t. We fix its value given the coefficients from Equation (26) in the main analysis sample: $\hat{\omega}(a,i,t) = \hat{\alpha} + \sum_{a \in \mathcal{A}} \left(\hat{\beta}_a + \hat{\gamma}_a X_i\right) + \hat{\delta}_t$. What is left then is to estimate values v_{it} and μ_i that would fit the dataset the best. We assume that $\mu_i \sim \mathcal{N}\left(m_\mu, s_\mu^2\right)$ and $v_{it} \sim \mathcal{N}\left(0, s_v^2\right)$, where values m_μ, s_μ^2, s_v^2 correspond to moments of $\hat{\mu}_i$ and $\hat{v}_{it}$ from Equation (26) in the subsample of genotyped individuals[9]. In particular, we estimate of $m_\mu = -0.09$, $s_\mu^2 = 0.35$ and $s_v^2 = 0.19$.

We estimate Equation (28) using the generalised structural equation model (GSEM) in Stata, which allows us to specify an interval regression with latent variable at the individual level. To avoid confusion with the individual fixed effect from Equation (26), we denote the predicted value of the latent variable by $\tilde{\mu}_i^{\text{RE}}$. Then, as in the main analysis sample, we compute predicted wages at all ages net of time trends:

$$\tilde{w}_{ia}^{\text{RE}} = \hat{\alpha} + \sum_{a \in \mathcal{A}} \left(\hat{\beta}_a + \hat{\gamma}_a X_i \right) + \tilde{\mu}_i^{\text{RE}}$$

Finally, we compute the discounted present values of predicted lifetime earnings:

$$DPV(w)_i^{\text{RE}} = 40 \cdot 52 \cdot \sum_{a \in \mathcal{A}} \left(\frac{\tilde{w}_{ia}^{\text{RE}}}{(1+r)^a} \right)$$

As a robustness check of the prediction algorithm, we assign the observations in the main analysis sample to the earnings quantiles and hours bins from the METADAC. For example, all observations with weekly hours worked between 30 and 34 are assigned to 30-34 bin. We then repeat the same steps described above and obtain $DPV(w)_i^{\text{RE}}$ in the main analysis sample. This allows us to directly compare the predicted earnings from fixed effects regression in Equation (26) and from the interval regression with latent variable in Equation (28).

Figure 6 compares the distribution of individual effects predicted from the FE estimation ($\hat{\mu}_i^{\text{FE}}$) and from interval regression with latent variable ($\tilde{\mu}_i^{\text{RE}}$). The left panel compares the two in the main analysis sample, and the right panel plots the distribution in the METADAC sample. It is clear that distributions are quite similar. Furthermore, Figure 7a shows that $\hat{\mu}_i^{\text{FE}}$ and $\tilde{\mu}_i^{\text{RE}}$ in the main analysis sample are highly positively correlated ($\rho = 0.91$). This means that the predicted earnings from either prediction algorithm should be similar to each other. Indeed, Figure 7b shows that the predicted earnings from the two prediction algorithms are highly correlated with each other ($\rho = 0.93$).

Finally, Table 13 confirms that there are no significant differences in the main estimation results between subsamples and prediction methods.

[9] The UKHLS data contains an indicator that is equal to 1 if an individual has been genotyped (part of METADAC) and 0 otherwise. Therefore, even though we cannot link between the two dataset, we can still compute sample statistics for genotyped individuals in the main analysis sample.

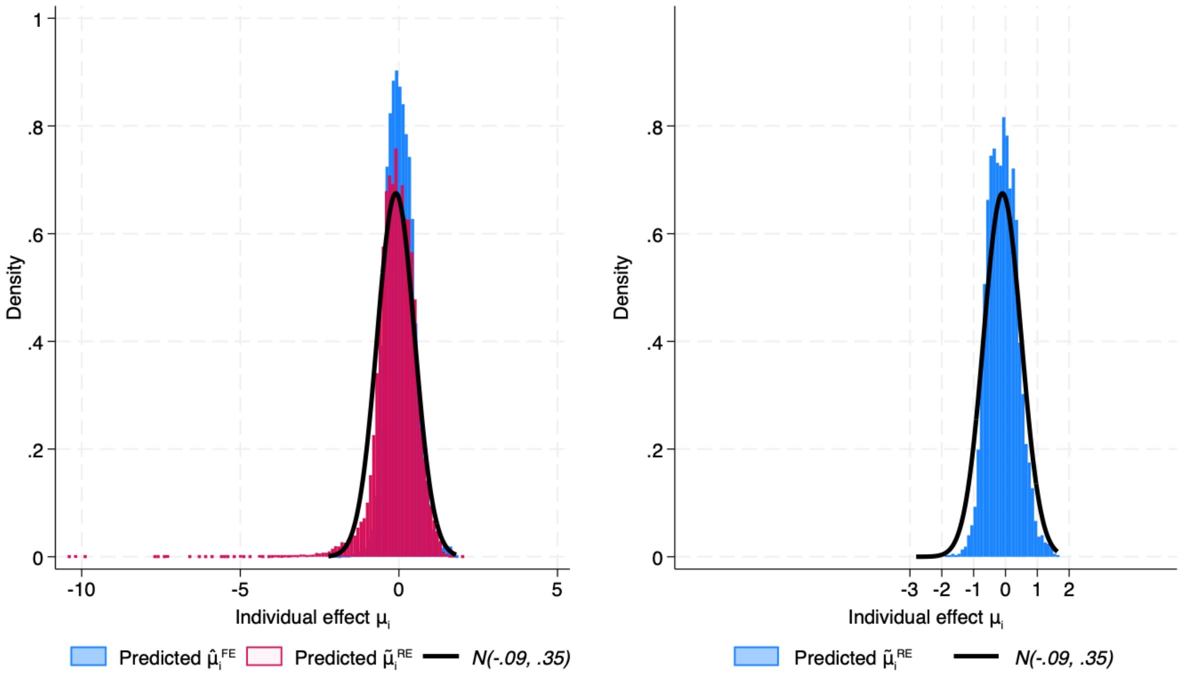
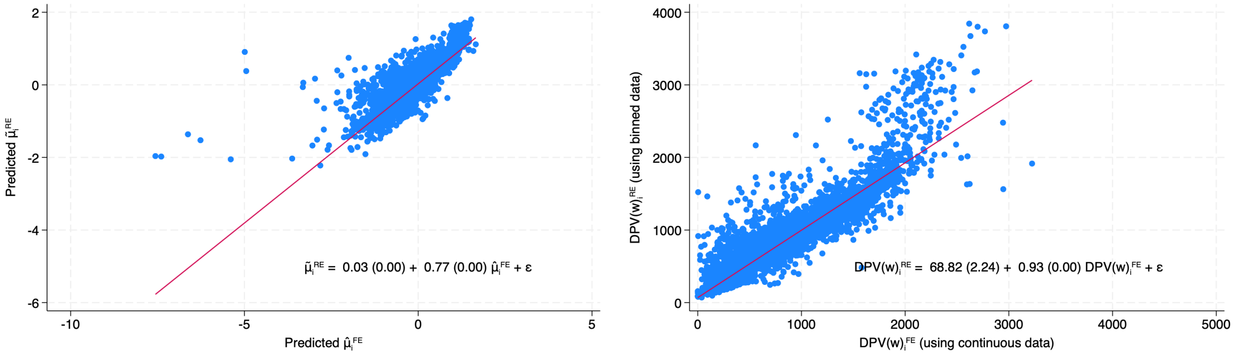


Figure 6. Distribution of predicted individual effects

Note: the figure plots the distribution of predicted individual effects μ_i. The left panel overlays the histograms of $\hat{\mu}_i^{\mathrm{FE}}$ and $\tilde{\mu}_i^{\mathrm{RE}}$ in the main analysis sample. The right panel plots the histogram of $\tilde{\mu}_i^{\mathrm{RE}}$ in the METADAC sample. Both panels show the density line of $\mathcal{N}(m_\mu, s_\mu^2)$.



(a) Predicted individual effects

(b) Predicted discounted present value of earnings

Figure 7. Comparison between prediction algorithms in the full sample

Note: the figure shows scatterplots of predicted individual effects μ_i (left panel) and discounted present values of lifetime earnings $DPV(w)_i$ (right panel) between prediction methods in the main analysis sample. To mimic the data limitations of the METADAC, we group the earnings and hours worked to same bins and run the interval regressions with latent variable to obtain predicted wages. The plots also show simple regression lines (red line) and regression equations with standard errors in parentheses.