Italian twins in the labour market

Sonia Brescianini (Istituto Superiore di Sanità)
Lorenzo Cappellari (Università Cattolica di Milano and LISER)
Daniele Checchi (Università degli Studi di Milano)

We use administrative data on educational attainments and life-time earnings to study their correlations among Italian twins. Using the ACE decomposition, we find that heritability in education accounts to almost half of the variance, especially for younger birth cohorts. With respect to labour market outcomes, we find that only for the oldest cohorts there is a greater share of inequality that can be attributed to idiosyncratic factors compared to education, and symmetrically a lower share due to genetics, while the impact of shared environment remains stable among the youngest cohorts. We suggest that increased employment flexibility may be responsible for the decline in the environmental component. Using a larger sample of pseudo-twins (individuals sharing birth date, birth place and family name) we confirm previous results, providing evidence that heritability also drives labour market attachment and prosocial behaviour.

Motivation

> The difference of natural talents in different men is, in reality, much less than we are aware of; and the very different genius which appears to distinguish men of different professions, when grown up to maturity, is not upon many occasions so much the cause, as the effect of the division of labour. The difference between the most dissimilar characters, between a philosopher and a common street porter, for example, seems to arise not so much from nature, as from habit, custom, and education. When they came into the world, and for the first six or eight years of their existence, they were, perhaps, very much alike, and neither their parents nor playfellows could perceive any remarkable difference. About that age, or soon after, they come to be employed in very different occupations. The difference of talents comes then to be taken notice of, and widens by degrees, till at last the vanity of the philosopher is willing to acknowledge scarce any resemblance. [Smith, 1937, pp. 15–16]

Nature vs nurture: to what extent is the inequality attributable to differences in genetic make-up, and to what extent is it attributable to differences in life experience? What proportion of the variance in achievement is due to genetic factors, and what proportion to environmental factors? What indeed is the heritability of education, of occupation, of earnings?

Italy: a country with high intergenerational persistence in education and incomes, coupled with high (and rising) inequality in incomes.

Both could be accounted for by a large role played by heritability of unobservable abilities.

In the absence of adequate controls, inequality of opportunities would be largely underestimated.

When unobservable traits play a significant role, policies aiming to equality should be adequately targeted.

Twin data helps to shed light on the relevance of heritability.

Like other siblings, twins share the family of origin and also experience a common background outside the family, through schools and youth neighbourhoods, with any correlation of their outcomes reflecting those shared influences.

Unlike regular siblings, twins also share the date of birth, which potentially reinforces their exposure to shared influences both within and outside the family, making the correlations of their outcomes even stronger.

Furthermore, twins share the same in utero experiences.

When information on twin zygosity is available, it can be leveraged to decompose the between-twins correlation into components due to genetics (or pre-birth influences) and the shared environment (post-birth influences), an approach grounded in behavioural genetics and known as the ACE model (A = Additive genetic factors, C = Common/shared environmental factors, E = Unique environmental factors).

The study of kinship correlations posits that outcomes (or phenotypes) are the result of three <u>independently distributed</u> (i.e. $cov(A, C) = cov(A, E) = 0$) and <u>linearly additive</u> (i.e. $\frac{\partial^2 Y}{\partial A \partial C} = 0$) factors:

$$Y_i = A_i + C_{f(i)} + E_i$$

where i is the individual and $f(i)$ denotes her family, Y is the outcome, A is a genetic effect (genotype), C is a common environmental effect shared by family members, and E is an idiosyncratic effect unique to person i.

"The genotype is properly construed as the expected phenotype of persons with a given genetic constitution, the expectation being taken over the distribution of available environments…Bear in mind that *y* [the phenotype] is observed while *x* [the genotype], *u* [shared environment] and *v* [residual] are hypothetical constructs" (Goldberger Economica 1979).

If the individual has a kin j, its phenotype is similarly defined

$$Y_j = A_j + C_{f(j)} + E_j$$

Under previous assumptions

$$cov(Y_i Y_j) = cov(A_i A_j) + cov(C_i C_j)$$

or expressing in terms of correlations

$$\frac{cov(Y_i Y_j)}{var(Y)} = \rho_{Y_{ij}} = \frac{cov(A_i A_j)}{var(A)} \cdot \frac{Var(A)}{Var(Y)} + \frac{cov(C_i C_j)}{var(C)} \cdot \frac{Var(C)}{Var(Y)}$$

$$= \rho_{A_{ij}} \cdot \underbrace{\boldsymbol{\frac{Var(A)}{Var(Y)}}}_{heritability} + \rho_{C_{ij}} \cdot \boldsymbol{\frac{Var(C)}{Var(Y)}}$$

With one observable ($\rho_{Y_{ij}}$) and 4 unknowns, further assumptions are required.

TABLE 3
CORE FORMULAS OF THE KINSHIP CORRELATION MODELS*

Identical twin together	$A + B$
Identical twin apart	A
Sibling together	$CA + B$
Sibling apart	CA
Parent together	$DA + EB$
Parent apart	DA
Adoptive sibling	B
Adoptive parent	EB
Spouse	F

* A = Genetic component of variance
B = Common-environment component of variance
C = Genotypic correlation of siblings
D = Genotypic correlation of parent & child
E = Common-environment correlation of parent & child
F = Phenotypic correlation of spouses

The ACE model is popular in behavioural genetics and exploits the difference between identical and fraternal twins to identify heritability. The outcome variance is:

$$var(Y) = \sigma_A^2 + \sigma_C^2 + \sigma_E^2;$$

the identical (monozygotic) twins covariance is

$$cov(Y_i Y_j)^{MZ} = \mathbf{1} \cdot \sigma_A^2 + \mathbf{1} \cdot \sigma_C^2 \ ;$$

and the fraternal (dizygotic) twins covariance is

$$cov(Y_i Y_j)^{DZ} = \frac{\mathbf{1}}{\mathbf{2}} \cdot \sigma_A^2 + \mathbf{1} \cdot \sigma_C^2$$

Then with 3 moments (Var and Covs) and 3 unknowns, the model is exactly identified

$$2\left[Cov\left(Y_i Y_j\right)^{MZ} - Cov\left(Y_i Y_j\right)^{DZ}\right] = \sigma_A^2$$

$$2Cov\left(Y_i Y_j\right)^{DZ} - Cov\left(Y_i Y_j\right)^{MZ} = \sigma_C^2 \ \text{ and } \ Var(Y) - Cov\left(Y_i Y_j\right)^{MZ} = \sigma_E^2$$

“At one extreme, suppose that the population is composed entirely of clones who face diverse environments. Then the variance of [A] is zero, implying that heritability is zero. At the other extreme, suppose that the population is composed of genetically diverse persons who share the same environment. Then the variance of [C] is zero, implying that heritability is one.” (Manski 2011, p.88) $\rightarrow$ measuring heritability is a variance decomposition exercise without entailment on social policy.

The literature suggests that heritability accounts for 20-30% of the variance in education.

When considering labour market oucomes, this share of variance typically increases to 30-40%.

Preferences (i.e. risk aversion) also exhibit heritability.

But what are the policy implications ?

> [T]he existence of powerful genetic factors makes the extension of [income and welfare] . . . policies beyond a certain point unwise and ultimately counterproductive. The results of going against nature are a reduction in motivation to work, and a revolt against the eroding of differentials. [*The Times*, 26 May 1977]

But "if earnings capacity is diminished by genetically poor eyesight, don't provide eyeglasses ?" Goldberger 1979: suppose that genetic share A is high - a large proportion of the variance in earnings is indeed attributable to genetic variance - and that B is small - only a small proportion of earnings variance is attributable to common environment variance. What then would follow?

Proposition ➀ <u>Intergenerational mobility is low</u>: son's earnings closely resemble father's earnings. **Not true**: even if earnings heritability were 100 per cent, the earnings of sons and fathers would correlate about 0.5, of grandsons and grandfathers would correlate about 0.25.

Proposition ② Environmental improvements could not produce much change in an individual's earnings. **Not true**: the proposition rests on a confusion between variance decompositions and norms of reaction, that is between correlations and regressions.

Proposition ③ The genetic factor, an important determinant of an individual's earning capacity, cannot be changed. **Not true**: the proposition rests on a confusion between "genome" (the genetic constitution or set of genes) and genotype. The former, we may say, is inviolate. But the latter is a market valuation of the abilities associated with that genome.

Proposition ⑥ Attempts to equalize earnings, however desirable they may be on equity grounds, must lead to efficiency. **An untenable dichotomy**. On the one hand, we have policies that merely equalize opportunity [common environment]: they reduce the common environment variance, they involve no net social costs. On the other hand, we have policies that go beyond equalizing opportunity: they try to equalize outcomes, they compensate for poor genes, they involve net social costs. Surely the environmental-genetic dichotomy does not correspond to a high-cost-low-cost dichotomy. An allocation of earnings variance into environmental and genetic components tells us nothing about tradeoffs

“Research on heritability is fundamentally uninformative for policy analysis, but make a cautious argument that research using genes as covariates is potentially informative (p.83) … What has made research on heritability particularly controversial has been the inclination of some researchers to interpret the magnitude of heritability estimates as indicators of the potential responsiveness of individual achievement to social policy. In particular, large estimates of heritability have been interpreted as implying small potential policy effectiveness.” (Manski 2011, pg.87)

When analysing the effect of treatments, the linear independence of genetic and environmental components is untenable.

The present paper uses administrative data on twins to decompose variances in income and education variance into three components:

- ⇨ genetic
- ⇨ environment
- ⇨ idiosyncratic

Merging privacy protected data from two public administrations, while preserving individual anonymity, is one of the main achievements of the paper. It required more than three years of red tape, and imposed significant data limitation in order to prevent reversed identification.

Still it is the first paper on this topic in Italy.

Potential objections:

① are twins representative of the entire population ? → external validity of our results

② in the absence of legal obligation, are public registries of twins (voluntary participation) representative of the twin population ? potential self-selection into medical studies

③ when studying the variance of incomes, we restrict to couples of twins with pension contributions associated to positive dependent employment spells → potential self-selection into dependent employment → internal validity in the population of twins

④ the ACE model is based on strong assumptions: gene-environment independence, additivity, homogeneity of environments assumed for monozygotic (MZ) and dizygotic (DZ) twins, and the absence of assortative mating between parents. Richer data are required to assess their validity.

The data

Our dataset originates from the merging of two files: the file of twins enrolled in the Italian Twin Register (ITR), managed by the Istituto Superiore di Sanità-ISS and the one on administrative data on payroll taxes from the Istituto Nazionale per la Previdenza Sociale-INPS.

The ITR was established in 2001. It is a population-based voluntary registry of twins. Originally recruited on ‘possible twin pairs’ identified using the demographic information summarized in the fiscal code, late replaced by applying a population-based recruitment strategy in several municipalities.

Currently, a total of 29.000 twins are enrolled in the ITR (11.500 monozygotic and 16.700 dizygotic) resident throughout the country and belonging to a wide age range (from 0 to 95 years, mean 36.8 years).

INPS collects payroll taxes from all workers (private and public employees, self-employed, contract workers), and therefore covers (almost) all source of earnings in the Italian population. The administrative data regarding work careers (in terms of employment and unemployment/layout spell, parental and illnesses leaves) are contained in contributory archives (*estratti conto*) where the present data are obtained.

In order to merge the two archives we extracted from the universe of the Italian population a subsample of individuals sharing family name, born in the same day in the same municipality ("pseudo-twins") → they overestimate the true-twin population, because of homonymy → drop all twinnings with more than two individuals. We started with a population of pseudo-twins of 344.226 individuals from INPS, against a potential number of 480.000 twins estimated from Census data over the same time period (assuming a share of 1% twins among newborn over the same period).

Selection rule:
① subjects born between 1964 and 1996, observed in 2022, aged between 26 and 58,
② both twins recording at least on event in the dependent employment archive.

Merge:
344.226 pseudo-twins were matched with 13.600 twins in the ISS registry, ending with a working sample of 9.722 twins who experienced at least one job spell in the sample period.

birth year	age in 2022	new born (ISTAT)	estimated twins	pseudo twins (INPS)	real twins (ISS)	matched INPS-ISS
1964	58	1 016 120	20 528	11 830	150	105
1965	57	990 458	20 009	11 976	188	134
1966	56	979 940	19 797	12 178	200	145
1967	55	948 772	19 167	12 384	172	120
1968	54	930 172	18 791	12 030	166	121
1969	53	932 466	18 838	12 670	173	141
1970	52	901 472	18 212	11 956	192	150
1971	51	906 182	18 307	12 864	216	165
1972	50	888 203	17 943	12 490	200	144
1973	49	874 546	17 668	13 020	160	123
1974	48	868 882	17 553	13 524	214	165
1975	47	827 852	16 724	13 356	178	143
1976	46	781 638	15 791	12 486	184	149
1977	45	741 103	14 972	11 950	142	106
1978	44	709 043	14 324	11 256	164	116
1979	43	670 221	13 540	10 398	153	111
1980	42	640 401	12 937	10 182	168	142
1981	41	623 103	12 588	10 200	152	115
1982	40	619 097	12 507	9 808	200	142
1983	39	601 928	12 160	9 622	1592	1193
1984	38	587 871	11 876	9 104	2314	1750
1985	37	577 345	11 664	9 280	1710	1272
1986	36	555 445	11 221	8 916	578	441
1987	35	551 539	11 142	8 786	266	202
1988	34	569 698	11 509	9 208	322	246
1989	33	560 688	11 327	9 098	338	250
1990	32	569 255	11 500	8 996	314	232
1991	31	562 787	11 369	8 834	342	258
1992	30	567 841	11 472	8 540	424	279
1993	29	549 484	11 101	7 782	546	328
1994	28	533 050	10 769	7 264	622	361
1995	27	525 609	10 618	6 378	456	227
1996	26	528 103	10 669	5 838	404	146
Total		**23 690 314**	**478 592**	**344204**	**13 600**	**9 722**

Source: **New born**: nati vivi (legittimi+naturali) from Istat Serie Storiche - Tavola 2.5.1 - Nati vivi e nati morti per filiazione e sesso - Anni 1926-2014

Estimated twins: assuming 1% of deliveries, computed as $=2 \times \frac{0.01}{0.99}$

the degree of matching an inverted U-shape pattern: for older persons born in the 60's the match covers around 70% of registered twins, possibly because a fraction of them has already retired. When we consider individuals in their forties, the coverage rate rises above 80%, then declining in the youngest cohorts to 60%.

There are good reasons to partition the sample into three, in order to minimize the effect of self-sorting into the registry.

The matched file is exposed to the risk of overestimating the correlation in labour market experience among Italian twins (given our selection rules).

Conversely the available information on pseudo-twins (which possibly includes false twins) may lead to an attenuation biased estimate of the actual correlation because we are unable to identify identical twins.

Two step analysis: proceed initially with the matched sample, and then expand the analysis with the extended sample.

Imposed data limitations allows for three outcome variables:

⇨ years of education (from IRT or from year of first job)

⇨ the quintile in the distribution of the decennial averages (less than 30, 31-40, 41-50 and above 50) of the yearly income

⇨ the quartile in the distribution of decennial averages of the time fraction spent into employment.

The matched INPS-ISS file is still representative of the original ISS file:

- ✓ female are over represented
- ✓ the fraction of identical twins is around 45% in the twin registry, higher than the biological expectation of one third
- ✓ twins in Central regions are more likely to enroll.

Descriptive statistics for ISS and matched INPS-ISS files

	match INPS-ISS			*actual twins in ISS registry*		
demographics	mean	sd	obs	mean	sd	obs
female	0.57	0.50	9 722	0.57	0.49	13 600
age	38.81	7.47	9 722	38.27	7.74	13 600
years of education	12.54	2.73	7 501	12.81	2.75	10 068
age of leaving family	22.67	4.80	7 006	22.58	4.78	9 403
maximal educational attainment	% fraction		obs	% fraction		obs
primary	0.73		55	0.92		93
lower secondary	12.00		900	13.32		1 341
upper secondary (decleared after 26)	16.38		1 229	16.10		1 621
upper secondary (decleared before 25)	55.30		4 148	54.82		5 519
college degree (decleared before 25)	1.39		104	1.48		149
college degree (decleared after 26)	14.20		1 065	13.36		1 345
total	**100.00**		**7 501**	**100.00**		**10 068**
zygosity	% fraction		obs	% fraction		obs
identical twins (monozygote)	46.60		4 530	45.63		6 206
fraternal twins (dizygote) same sex	29.64		2 882	29.96		4 074
fraternal twins (dizygote) different sex	23.76		2 310	24.41		3 320
total	**100.00**		**9 722**	**100.00**		**13 600**

The matched file is partially different from the population of pseudo twins:

- ✓ younger population
- ✓ women and the less educated are overrepresented
- ✓ (pseudo) fraternal twins in the couple of different sex are more frequent in INPS data

Descriptive statistics for INPS and matched INPS-ISS files

	match INPS-ISS				*pseudo twins (INPS)*			
demographics	mean	sd	obs		mean	sd	obs	
female	0.57	0.50	9 722		0.457	0.498	344 204	
age	38.81	7.47	9 722		43.695	9.147	344 204	
years of education	12.54	2.73	7 501		15.862	4.510	343 267	
age of leaving family	22.67	4.80	7 006					
age of first job					22.14	5.28	343 267	
zygosity	% fraction		obs					
identical twins (monozygote)	46.60		4530		63.30		217 878	
fraternal twins (dizygote) same sex	29.64		2882					
fraternal twins (dizygote) different sex	23.76		2310		36.70		126 326	
total	**100.00**		**9722**		**100.00**		**344 204**	
income position (%)	age<30	age 31-40	age 41-50	age>50	age<30	age 31-40	age 41-50	age>50
first quintile	22.40	16.12	13.13	13.40	21.58	19.17	20.56	20.00
second quintile	17.15	19.66	14.15	13.29	18.53	21.09	18.65	20.56
third quintile	20.29	22.40	19.89	15.65	21.87	21.65	21.86	20.45
fourth quintile	17.61	19.98	21.45	22.75	19.14	18.93	18.57	18.92
fifth quintile	22.55	21.85	31.38	34.91	18.89	19.17	20.35	20.06
total	**9 323**	**8 064**	**2 247**	**888**	**324 829**	**289 170**	**199 372**	**78 203**
worked time position (%)	age<30	age 31-40	age 41-50	age>50	age<30	age 31-40	age 41-50	age>50
first quartile	32.08	24.69	20.65	18.92	25.15	24.78	24.54	24.36
second quartile	26.05	30.85	24.34	22.07	26.62	27.76	26.72	27.77
third quartile	21.97	44.46	55.01	28.94	24.76	47.46	48.74	23.71
fourth quartile	19.90			30.07	23.46			24.16
total	**9 323**	**8 064**	**2 247**	**888**	**324 829**	**289 170**	**199 372**	**78 203**

Results from matched ISS-INPS file

We have partitioned the data into three groups:
the ***young*** (born after 1985– 33% of the sample),
the ***adult*** (born between 1983 and 1985 – 41% of the sample)
the ***old*** (born before 1983 –26% of the sample).

These three groups have presumably experienced different economic and institutional settings when entering the labour market: the oldest group entered the labour market before the wave of reforms aimed at increasing employment flexibility hit the Italian labour market (between the late 1990s and the early 2000s). The intermediate group has entered the labour market right after some major reforms had entered into effect (Pacchetto Treu in 1997 and Biagi Law in 2003). Finally, the youngest group entered the labour market after labour flexibility had been fully implemented and amid the financial crisis of 2008.

Rank correlations (Spearman) between twins

	years education	permanent income quintile	permanent workdays quartile
identical twins - all sample	**0.71**	**0.52**	**0.46**
fraternal twins - all sample	**0.43**	**0.27**	**0.26**
Identical twins – male	0.72	0.56	0.49
Fraternal twins – same sex: male	0.44	0.31	0.32
Identical twins – female	0.70	0.48	0.42
Fraternal twins – same sex: female	0.53	0.30	0.26
Fraternal twins – different sex	0.34	0.21	0.23
Identical twins - all sample – born after 1983	0.78	0.68	0.73
Fraternal twins - all sample – born after 1983	0.48	0.37	0.47
Identical twins - all sample – born 1983-1985	0.50	0.52	0.51
Fraternal twins - all sample – born 1983-1985	0.25	0.21	0.37
Identical twins - all sample – born before 1985	0.46	0.45	0.40
Fraternal twins - all sample – born before 1985	0.24	0.25	0.29

Note: bootstrapped 50 replications – all significant at 0.01

Outcome correlation among identical twins is higher than among fraternal twins, especially when they are male. Correspondingly, the correlation is the lowest among different sex fraternal twins.

Correlation halves when passing from the youngest to the oldest group. While this is expected when looking at labour market experience, due to the idiosyncratic components, it is more surprising when looking at years of education.

For each cohort we have estimated the so-called the ACE model that is popular in behavioural genetics. The ACE posits that outcomes (or phenotypes) are the result of three orthogonal and linearly additive factors:

$$Y_i = A_i + C_{f(i)} + E_i$$

where i is the individual and $f(i)$ denotes her family, Y is the outcome, A is an additive genetic effect, C is a common environmental effect shared by family members, and E is an idiosyncratic effect unique to person i.

Each component $x = (A, C, E)$ is drawn from a zero-mean distribution with variance σ_x^2.

This basic formulation of the ACE model rests on several assumptions:

① Orthogonality of the three factors rules out the possibility of gene-environment correlation, i.e. individuals or families do not sort into environments on the basis of their genes.

② The linear specification excludes the possibility of gene-environment interactions, a circumstance in which the environment mediates genetic expressions.

③ spouses are not sorted on genes, implying that DZ twins share on average half of their genes, while genetic assortative mating would imply a larger sharing for DZ's.

④ absence of dominance of the gene variants someone receives from one parent on the variants received from the other parent.

⑤ the extent of environmental sharing is the same for MZ and DZ twins.

Under this set of assumptions, the model provides sufficient information for the identification of the three variance components σ_A^2, σ_C^2 and σ_E^2.

Namely, the outcome variance is:

$$var(Y) = \sigma_A^2 + \sigma_C^2 + \sigma_E^2;$$

the MZ twins covariance is

$$cov(YY')^{MZ} = \sigma_A^2 + \sigma_C^2 \ ;$$

and the DZ twins covariance is

$$cov(YY')^{DZ} = 0.5\sigma_A^2 + \sigma_C^2 \ ;$$

These are three equations in three unknown parameters, which are therefore identified. The model parameters can be used to compute the **degree of heritability**, that is the share of cross-sectional dispersion accounted for by the **genetic component** $\frac{\sigma_A^2}{\sigma_A^2+\sigma_C^2+\sigma_E^2}$.

To estimate the model we assume normality of the factors and use a Mixed-Model approach (even though semi-parametric models yield similar results).

This is essentially a two-equations SURE (one for each twins) in which the moment conditions implied by the model are imposed on the variance-covariance matrix of the errors.

Whenever the estimated shared environment is negligible either statistically or substantively, we turned to a restricted version of the ACE model, the AE model, that constrains the shared environmental component to be equal to zero.

The estimating equations always include gender as controls.

(A) Years of education								
	(1) age < 37 born after 1985		(2) age 37-39 born 1983-1985		(3) age > 39 born before 1983		(4) entire sample	
A genetics	43.68	***	62.88	***	50.6	***	46.88	***
C environment	35.77	***	6.66	***	22.71	***	26.55	***
E idiosyncratic	20.54	***	30.45	***	26.69	***	26.56	***
observations	1200		5524		3344		10068	
(B) life-cycle earnings (quintile average over 4 decades)								
	(1) age < 37 born after 1985		(2) age 37-39 born 1983-1985		(3) age > 39 born before 1983		(4) entire sample	
A genetics	48.44	***	49.95	***	23.46	***	49.20	***
C environment	0.79				27.72	***	3.59	
E idiosyncratic	50.77	***	50.05	***	48.8	***	47.19	***
observations	2970		4215		2537		9722	
(C) life-cycle employment (quartile average over 4 periods)								
	(1) age < 37 born after 1985		(2) age 37-39 born 1983-1985		(3) age > 39 born before 1983		(4) entire sample	
A genetics	43.85	***	38.29	***	21.22	**	39.50	***
C environment	2.21		6.73		18.64	**	6.86	*
E idiosyncratic	53.94	***	54.97	***	60.13	***	53.62	***
observations	2970		4215		2537		9722	

Note: controls include gender – statistical significance: *** p<0.01, ** p<0.05, * p<0.1

Main findings:
① the estimated heritability in educational attainment fluctuates over cohorts (between 40 and 50%).

② for the oldest cohort the environmental component is relevant (20%) along all dimensions (education and labour market) while for younger cohorts the idiosyncratic factor dominates (around 50%)

③ for younger cohorts the shared environment virtually does not contribute to account for labour market inequality

But labour market outcomes are not strictly comparable across cohorts, due to different experience.

ACE decomposition (percent) of labour market outcomes, by decades

earnings quintile								
	age < 37		age 37-39		age > 39			
	1st decade (age 20-30)	2nd decade (age 31-40)	1st decade (age 20-30)	2nd decade (age 31-40)	1st decade (age 20-30)	2nd decade (age 31-40)	3rd decade (age 41-50)	4th decade (age 51-60)
A genetics	42.39 ***	42.82 ***	37.38 ***	44.25 ***	20.97 ***	25.89 ***	12.15	24.57 *
C environment	5.08		8.30		21.16 ***	21.65 ***	31.86 ***	21.05 *
E idiosyncratic	52.53 ***	57.18 ***	54.32 ***	55.75 ***	57.86 ***	52.45 ***	55.98 ***	54.37 ***
observations	2942	1349	4031	4214	2350	2501	2247	888
workdays quartiles								
	age < 37		age 37-39		age > 39			
	1st decade (age 20-30)	2nd decade (age 31-40)	1st decade (age 20-30)	2nd decade (age 31-40)	1st decade (age 20-30)	2nd decade (age 31-40)	3rd decade (age 41-50)	4th decade (age 51-60)
A genetics	41.95 ***	10.47	40.23 ***	21.28 ***	20.83 *	32.19 ***	24.43 ***	26.95 ***
C environment	4.02	16.05	7.34	5.79	18.56 **	2.69	3.63	
E idiosyncratic	54.03 ***	73.48 ***	52.43 ***	72.92 ***	60.60 ***	65.12 ***	71.93 ***	73.04 ***
observations	2942	1349	4031	4214	2350	2501	2247	888

Note: controls include gender – statistical significance: *** $p<0.01$, ** $p<0.05$, * $p<0.1$

The oldest cohort remains characterized by a greater impact of shared environment and a lower impact of the genetic component compared to younger cohorts even if we limit the observation to outcomes measured soon after labour market entry (i.e. age 20-30), which are available for all cohorts. Puzzle: <u>why old and young cohorts are structurally so different ?</u>

Suggested interpretations:

⇨ young cohorts in our data have entered the labour market after the introduction of a set of reforms aimed at increasing flexibility (the OECD EPL index declined from 4.75 over 6.00 in 1997 to 2.00 in 2003) → lower entry wages and reduced career paths

⇨ greater relevance of genetic vs shared environmental determinants of inequality suggests that labor flexibility has amplified the impact of inherent abilities and traits, partly determined by genetics, on labor market success and career progression.

⇨ A cohort born at the turn of the year 1982 entered tertiary education or labour market in 1999. The year 1999 is also the start of the Bologna process, separating 3-year BA courses from an additional 2-year MA courses → this reform has made college access less dependent on parental background (reduced share of environmental component)

No detectable geographical heterogeneity in our sample.

Robustness

To address the robustness of our results to the presence of assortative mating in genes, we estimate a version of the ACE model in which we impose a genetic sharing larger than 50% for fraternal twins → the genetic component expands reducing (by construction) the contribution of the shared environment.

ACE decomposition with varying degrees of hypothesized genetic correlation among DZ twins ($corrA_{DZ}$)
Years of education

	(1)		(2)		(3)		(4)	
	$corrA_{DZ}$ =0.50		$corrA_{DZ}$=0.55		$corrA_{DZ}$ =0.60		$corrA_{DZ}$ =0.65	
A genetics	46.88	***	52.09	***	58.6	***	66.95	***
C environment	26.55	***	21.34	***	14.82	***	6.48	
E idiosyncratic	26.56	***	26.56	***	26.56	***	26.56	***
observations	10068		10068		10068		10068	

Note: controls include gender – statistical significance: *** p<0.01, ** p<0.05, * p<0.1

Some predetermined characteristics that we observe like year and region of birth may well interact with the three ACE factors, and that are likely to reflect the impact of environmental (rather than genetic) influences. To the extent that these observable characteristics are not orthogonal to ACE factors, we would expect variance decompositions to change when we residualise outcomes on the observables.

Results for education go in the expected direction $\rightarrow$ controlling for year and region of birth reduces the proportion of variance that is attributed to shared environment.

Moving to labour market outcomes due to lack of convergence we report estimates from the restricted AE specification. The lack of convergence of the model with shared environment suggests that year and region of birth represent a relevant source of shared environmental influences in the process that determines labour market outcomes.

ACE and AE decompositions controlling for year and region of birth

(A) Years of education						
	(1) controlling for gender		(2) controlling for gender and birth year		(3) controlling for gender, birth year and region of birth	
A genetics	46.88	***	50.53	***	50.56	***
C environment	26.55	***	20.80	***	20.58	***
E idiosyncratic	26.56	***	28.67	***	28.84	***
observations	10068		10068		10068	
(B) Life-cycle earnings (quintile average over 4 decades)						
	(1) controlling for gender		(2) controlling for gender and birth year		(3) controlling for gender, birth year and region of birth	
A genetics	53.22	***	49.61	***	46.16	***
E idiosyncratic	46.77	***	50.38	***	53.83	***
observations	9722		9722		9722	
(C) Life-cycle employment (quartile average over 4 periods)						
	(1) controlling for gender		(2) controlling for gender and birth year		(3) controlling for gender, birth year and region of birth	
A genetics	47.33	***	44.02	***	41.85	***
E idiosyncratic	52.67	***	55.97	***	58.14	**
observations	9722		9722		9722	***

Evidence from pseudo-twins

Using pseudo-twins data identified by matching the condensed surname, date and place of birth available in the tax code may underestimate twin correlations in outcomes due to potential homonymy, but allow the expansion of outcomes that is not available in the matched INPS-ISS twin sample.

Pseudo-twins can be studied in their

⇨ *labour market attachment* through unemployment or absenteeism (respectively proxied by events of unemployment subsidy or illnesses spells),

⇨ *fertility decisions* (proxied by use of parental leaves),

⇨ *prosocial attitudes* (captured by event of blood donations, since Italian workers are entitled to one day off in such event)

⇨ *religious vocations* (proxied by contribution in the clergy pension fund).

Additional information available for pseudo-twins: frequency of specific events

	Obs	Mean	Std. Dev.	Max
unemployment/layoff events				
age<30	324 829	0.102	0.191	1
age 31-40	289 170	0.155	0.265	1
age 41-50	199 372	0.163	0.277	1
age>50	78 203	0.180	0.313	1
absence for illness events				
age<30	324 829	0.041	0.103	1
age 31-40	289 170	0.071	0.157	1
age 41-50	199 372	0.087	0.187	1
age>50	78 203	0.098	0.230	1
parental leave events				
age<30	324 829	0.019	0.074	1
age 31-40	289 170	0.056	0.151	1
age 41-50	199 372	0.035	0.141	1
age>50	78 203	0.030	0.151	1
pro-social (absence for blood donation)				
age<30	324 829	0.004	0.037	1
age 31-40	289 170	0.013	0.086	1
age 41-50	199 372	0.019	0.118	1
age>50	78 203	0.016	0.114	1
religious vocation (contributing to clergy pension fund)				
age<30	324 829	0.000	0.021	1
age 31-40	289 170	0.001	0.028	1
age 41-50	199 372	0.001	0.031	1
age>50	78 203	0.001	0.030	1

Larger sample sizes, consistent with life-cycle dynamics.

Pseudo twins correlations

	years education	permanent income decile	permanent workdays quartile	unemploy ment/lay-off events	absence for illness events	parental leave events	pro-social (blood donations)	religious vocation (clergy)
Same gender	0.545	0.425	0.363	0.274	0.242	0.271	0.191	0.086
Mixed gender	0.300	0.158	0.134	0.113	0.113	-0.051	0.050	-0.006

Note: bootstrapped 50 replications – all significant at 0.01

Because mixed gender twins are DZ, the correlations estimated on mixed gender pseudo twins are directly comparable to those of mixed gender DZ twins:

⇨ years of education for mixed gender pairs are similar(0.30 vs 0.34 estimated on mixed gender DZ actual twins)

⇨ (permanent) earnings and working time obtain lower correlation (0.15 and 0.13 vs 0.21 and 0.23 respectively) for mixed gender DZ actual twins.

⇨ for additional outcomes the same-gender correlations are relatively sizeable and always larger than mixed-gender ones.

In the absence of zygosity information in the pseudo-twins sample, we rely on the gender composition of the pair to derive the ACE decomposition of the covariances.

To do this we first notice that on average one third of twin births is made by MZ twins (who are all same-gender pairs), one third is made by same-gender DZ twins and the remaining third is given by mixed-gender DZ twins. Therefore, a same-gender (SG) pair of pseudo twins will be composed by MZ twins and DZ twins in equal proportions, such as the ACE covariance decomposition for this pair is given by:

$$cov(YY')^{SG} = 0.75\sigma_A^2 + \sigma_C^2 \ .$$

Mixed gender (MG) pairs will be entirely composed of DZ twins

$$cov(YY')^{MG} = cov(YY')^{DZ} = 0.5\sigma_A^2 + \sigma_C^2$$

ACE decomposition (percent) of lifecycle outcomes – pseudo twins

	Years of education		Life-cycle earnings		Life-cycle employment	
A genetics	52.57	***	34.45	***	34.21	***
C environment			5.05	***		
E idiosyncratic	47.42	***	60.48	***	65.78	***
observations	344 204		344 204		345 136	

Note: controls include gender – statistical significance: *** p<0.01, ** p<0.05, * p<0.1

Heritability accounts for a little more than half of the variance in years of education, while the rest is due to purely idiosyncratic variation, against 30% of educational dispersion in the actual twins data → pseudo-twins approach tends to overestimate idiosyncratic variation and to underestimate the inequality that comes from shared factors.

ACE decomposition (percent) of lifecycle outcomes by birth cohort– pseudo twins

(A) Life-cycle earnings

	1 born in 1982 or before		2 born after 1982	
A genetics	33.46	***	34.96	***
C environment	4.21	**	2.43	**
E idiosyncratic	62.31	***	62.59	***
observations	226 558		117 646	

(B) Life-cycle employment

	1 born in 1982 or before		2 born after 1982	
A genetics	29.15	***	32.25	***
E idiosyncratic	70.84	***	67.74	***
observations	226 558		117 646	

Note: controls include gender – statistical significance: *** p<0.01, ** p<0.05, * p<0.1

For permanent earnings the pattern is analogous to the one found in the twins data, that is shared family environment tends to lose relevance among younger cohorts exposed to more flexible labour markets.

We use an AE specification or additional dimensions to lack of convergence of the ACE specification. We find that the genetic component exhibits a declining weight as long as we move towards responsible individual choices:

AE decomposition (percent) of additional outcomes – pseudo twins

	Unemployment		Absence		Parental leave		Blood donations		Religious vocation	
A genetics	30.11	***	24.34	***	12.25	***	19.87	***	7.73	***
E idiosyncratic	69.88	***	75.65	***	87.74	***	80.12	***	92.26	***

Note: controls include gender – statistical significance: *** p<0.01, ** p<0.05, * p<0.1 –
Number of observations 344 204 - For blood donations and religious vocation we impose a unit correlation of genetic factors among same sex pseudo twins

Conclusions

Genetics component matters, contributing to a large fraction of inequality in education and labour market outcomes (between 30 to 50%)

The environment matters for education, but less and less for labour market outcomes in young generations.

Labour market flexibility emphasizes idiosyncratic dimensions, as well as individual (unobservable) component $\rightarrow$ reduced inequality of opportunities?

Genetics plays a reduced role (not negligible, though) in social attitudes

Appendix

Table A1 – Descriptive statistics for ISS

	(a) identical twins (ISS)			(b) fraternal twins (ISS) – same sex			(c) fraternal twins (ISS) – other sex		
Demographics	mean	sd	obs	mean	sd	obs	mean	sd	obs
Female	0.60	0.48	6206	0.57	0.49	4074	050	0.50	3220
Age	39.15	7.84	6206	37.39	7.73	4074	37.68	7.35	3220
Years of education	12.96	2.80	4574	12.73	2.70	2774	12.59	2.63	2373
Age of leaving family	22.99	5.12	4921	22.22	4.54	2593	22.13	4.31	2236
Income quintile	3.03	1.29	4530	2.97	1.32	2882	2.97	1.31	2310
Work duration quartile	2.23	0.87	4530	2.18	0.86	2882	2.19	0.85	2310

	(a) age < 37 born after 1985			(b) age 37-39 born 1983-1985			(c) age > 39 born before 1983		
Demographics	mean	sd	obs	mean	sd	obs	mean	sd	obs
Female	0.52	0.49	4612	0.57	0.49	5616	0.63	0.48	3372
Age	30.68	3.28	4612	37.97	0.76	5616	49.11	5.40	3372
Years of education	12.87	2.47	1200	12.25	1.97	5524	13.71	3.56	3344
Age of leaving family	21.27	3.01	1128	19.90	1.15	5217	27.61	5.12	3058
Income quintile	2.71	1.37	2970	3.01	1.27	4215	3.32	1.20	2537
Work duration quartile	2.04	0.93	2970	2.22	0.83	4215	2.37	0.75	2537

Table A2 – AE decomposition (percent) of lifecycle outcomes

	(a) Age < 37 born after 1985		(b) Age 37-39 born 1983-1985		(c) Age > 39 born before 1983	
(1) Years of schooling						
A genetics	79.83	***	69.88	***	73.86	***
E idiosyncratic	20.17	***	30.12	***	26.14	***
(2) Life-cycle earnings (average over 4 periods)						
A genetics	49.37	***	49.95	***	53.63	***
E idiosyncratic	50.63	***	50.05	***	46.37	***
(3) Life-cycle employment (average over 4 periods)						
A genetics	46.44	***	45.98	***	42.10	**
E idiosyncratic	53.56	***	54.02	***	57.90	***

Note: controls include gender – statistical significance: *** p<0.01, ** p<0.05, * p<0.1

Table A3 – AE decomposition (percent) of labour market outcomes, by decades

(1) Earnings quintile	(a) Age < 37 born after 1985		(b) Age 37-39 born 1983-1985		(c) Age > 39 born before 1983			
	1st decade (age 20-30)	2nd decade (age 31-40)	1st decade (age 20-30)	2nd decade (age 31-40)	1st decade (age 20-30)	2nd decade (age 31-40)	3rd decade (age 41-50)	4th decade (age 51-60)
A genetics	48.30 ***	42.82 ***	46.81 ***	44.25 ***	44.50 ***	49.70 ***	47.26	47.63 *
E idiosyncratic	51.70 ***	57.18 ***	53.19 ***	55.75 ***	55.50 ***	50.30 ***	52.74 ***	52.37 ***

(2) Workdays quartiles	(a) Age < 37 born after 1985		(b) Age 37-39 born 1983-1985		(c) Age > 39 born before 1983			
	1st decade (age 20-30)	2nd decade (age 31-40)	1st decade (age 20-30)	2nd decade (age 31-40)	1st decade (age 20-30)	2nd decade (age 31-40)	3rd decade (age 41-50)	4th decade (age 51-60)
A genetics	46.66 ***	29.61	48.43 ***	28.18 ***	41.61 *	35.25 ***	28.60 ***	26.96 ***
E idiosyncratic	53.34 ***	70.39 ***	51.57 ***	71.82 ***	58.39 ***	64.75 ***	71.40 ***	73.04 ***

Note: controls include gender – statistical significance: *** p<0.01, ** p<0.05, * p<0.1