Genetically Informed Estimation of Marginal Returns to Education

Pietro Biroli, Elisabetta De Cao, **Tomeu López-Nieto Veitch**, Rodrigo Pinto

5th Human Capital Workshop, Banca d'Italia

March 20, 2026

Outline

Background: Returns to Education

- Returns to education have **long been a central focus of economic research** given their impact on:
  - Productivity
  - Social mobility
  - Inequality
- It has been so for, at least, the last 60 years and is **still ongoing** (Becker, 1964; Mincer, 1974; Nybom, 2017; Westphal, Kamhöfer, and Schmitz, 2022).
- Why?
  - Difficult to account for **heterogeneity in skills and initial endowments**
  - How causal are the estimated returns?

The missing piece: Genetics

- Lack of genetic data has been a **key limitation** in the literature:
  - Core component of individuals' initial endowments and skills.
- **Genes and Education interplay** in shaping returns to education:
  - Genes **influence selection** into education
  - **Returns** to education may **vary depending on one's genetic makeup**
- Has this been studied before?
  - Genetic basis on the return to an additional year of compulsory schooling (Barcellos, Carvalho, and Turley, 2021)
  - But that reform (ROSLA) did not induce students to acquire a college degree
  - Yet **most wage returns to education stem from college**, with estimates of over a 50% college wage premium in the UK (Blundell, Green, and Jin (2016))

Research Question

Does higher education amplify or mitigate genetic inequalities in earnings?

What do we contribute?

(1) **Theoretical**

- Extend the MTE framework to account for **group-specific differences in resistance to treatment**
- **Disentangle selection from genuine heterogeneity** in between-group differences in treatment effects

(2) **Empirical**

- Incorporate **genetic data** to shed light on heterogeneity in skills and initial endowments
- Better **understand the heterogeneity and inequalities** in the returns to education

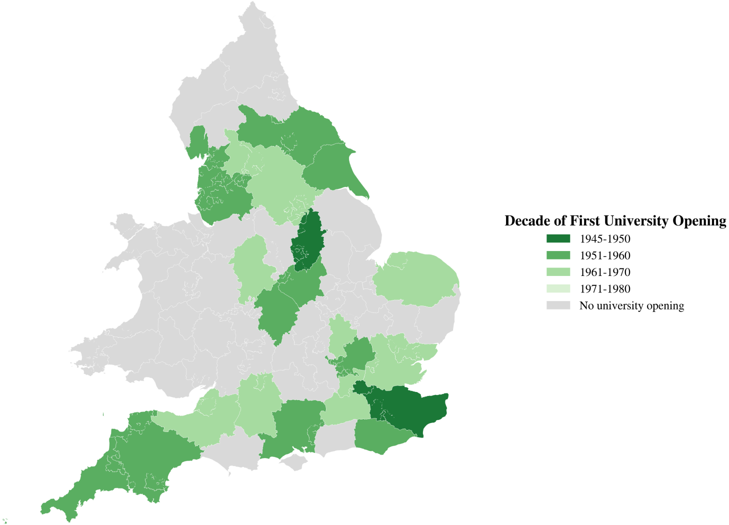
Outline

Outline

University Expansions and Distance to University

- **Distance to college** is a classic instrument for education
  - But it may still correlate to family background and local labour markets
  - e.g. Card (1995); Carneiro, Heckman, and Vytlacil (2011)
- **College expansions** provide more plausibly exogenous variation in access
  - e.g. Berlingieri, Gathmann, and Quinckhardt (2022); Ichino, Rustichini, and Zanella (2025)
- **In this study:** we exploit the mid-20th-century expansion of university access in the UK

UK University Expansion



Outline

Data

(1) **Genetic and Individual Socio-Economic Data** (UK Biobank):

- Very **large sample** born between 1945 and 1970
- **Genetic data**
- Educational attainment
- Occupational data
- Birth and current place coordinates (1km precision)

(2) **Higher Education Institutions Data** (source: Gov.uk):

- **Own digitisation**
- Location coordinates for each campus
- Type of institution: vocational, university...
- Dates of opening, closure and granted university status

Data (II)

(3) **Local Socioeconomic Context Data**:

(a) Unemployment:

- Historical Unemployment (source: Labour Gazette)
- Current Unemployment (source: NOMIS)

(b) Educational Level in the district of birth (source: 1951 Census)

(c) Share of high- and low- skilled workers in the district of birth (source: 1951 Census)

(d) Population and Mortality (source: GBHD)

Data (III): Caveats

- The UKB does not directly provide data on wages, we use the **imputed log-hourly wages** following Kweon et al. (2020); Barcellos, Carvalho, and Turley (2021).
  - Imputation is based primarily on 4-digit Standardised Occupation Codes (SOCs)
  - Prior work shows that this is a useful proxy for wages
- The UKB also lacks direct information on **parental education**. Like Barcellos, Carvalho, and Turley (2021), we proxy it with:
  - Educational level in the district of birth
  - Number of siblings

Outline

Reduction in Distance to University

(1) We measure the **reduction in distance** to the nearest university, $\Delta Dist$:

$$\Delta Dist = Dist\_prebirth - Dist\_postbirth$$

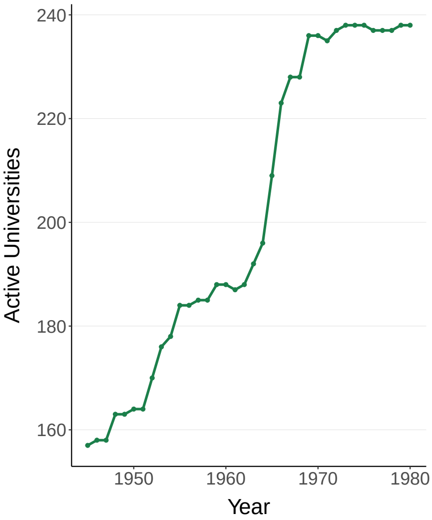
Motivation   Summary Statistics

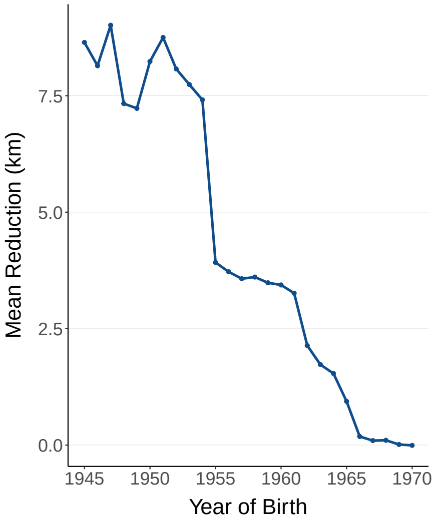
- *Dist_prebirth*: distance to nearest university, active at year of birth
- *Dist_postbirth*: distance to nearest university, active by age 15

(2) UK university expansion targeted disadvantaged areas (Robbins, 1963) $\Longrightarrow$ Control for:

- Number of active universities at birth year within 50km of birth-location
- Population Density and Growth in the district of birth
- Average Unemployment Rate in the district of birth (1945-1960)
- Urban/Rural Area Classification of the birth postcode
- Share of high- and low-Socioeconomic Status (SES) in the district of birth
- Infant Mortality Rate (IMR) in the district of birth
- Region and Cohort fixed effects

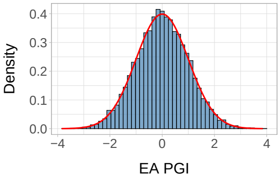
Access to University and Reduction in Distance





Measuring Genetic Predisposition to Education

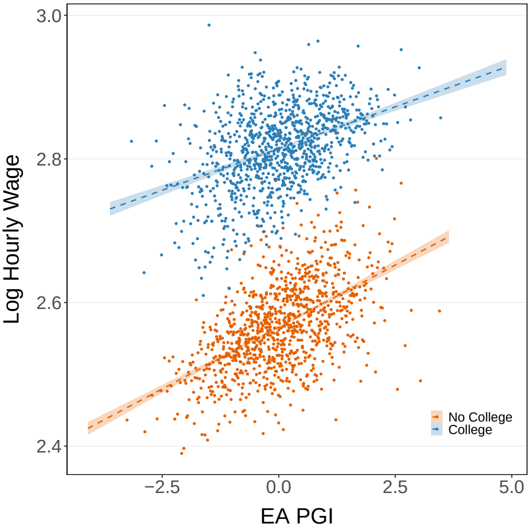
- We use the Educational Attainment Polygenic Index (EA PGI)
- PGIs are genetic scores that summarise the **combined effect of thousands of genetic variants** associated with a given trait (e.g. EA)
- A PGI is the **best linear predictor** of a trait **based on genetic variation**
    - The latest EA PGI (Okbay et al., 2022) explains about 12-16% of the variance in EA.
    - They are standardised scores and are approximately normally distributed



More on PGIs

Outline

Evidence of Gene-Environment Interplay?



Conventional Estimation of GxE

- **First Stages:**

$$E_i = \pi_0 + \pi_1 G_i + \pi_2 Z_i + \pi_3 (Z_i \times G_i) + \delta X_i + \mu_{r(i)} + \alpha_{c(i)} + v_{1,i}$$
$$(G_i \times E_i) = \theta_0 + \theta_1 G_i + \theta_2 Z_i + \theta_3 (Z_i \times G_i) + \eta X_i + \mu_{r(i)} + \alpha_{c(i)} + v_{2,i}$$

- **Second Stage:**

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 \widehat{E}_i + \beta_3 (\widehat{G_i \times E_i}) + \gamma X_i + \mu_{r(i)} + \alpha_{c(i)} + \varepsilon_i \tag{1}$$

- Where:
  - Y: Log-hourly imputed wage
  - E: College dummy
  - G: Educational Attainment PGI
  - Z: Reduction in Distance to university
  - X: Vector of controls including:
    - Socioeconomic measures (e.g. unemployment, educational level in the district)
    - First 20 genetic PCs to control for population stratification (Price et al., 2006)
    - E and G are interacted with all demeaned X's (Biroli et al., 2026)
  - $\mu_{r(i)}, \alpha_{c(i)}$: Region of birth and cohort fixed-effects
  - v_1, v_2, ε: idiosyncratic error terms

Second Stage Results

	Continuous PGI			Discrete EA PGI
	$G \times E$	G	No G	
	(1) Wage	(3) Wage	(5) Wage	(7) Wage
College	0.494***	0.583***	0.520***	0.678***
	(0.079)	(0.073)	(0.078)	(0.136)
College × EA PGI	-0.123***			
	(0.045)			
EA PGI	0.047**	-0.010		
	(0.017)	(0.008)		
College × EA PGI Medium				-0.108
				(0.137)
College × EA PGI High				-0.283*
				(0.139)
EA PGI Medium				0.024
				(0.039)
EA PGI High				0.090
				(0.050)
Num.Obs.	158223	158223	158223	158223
R2 Adj.	0.233	0.187	0.161	0.183
First Stage F-Statistic	46.58	92.48	94.71	31.34
Controls (X)	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
Genetic PCs	20	20	0	20

Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. UKB weights have been applied (van Alten et al., 2024). All specifications are estimated by 2SLS. College and its interactions are instrumented using Reduction in Distance and its interactions with PGI.

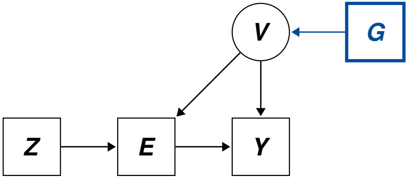
Potential Issue: Selection on Unobservables

- **Self-selection in returns to education** has been widely documented (e.g. Carneiro, Heckman, and Vytlacil (2011))
- If there is self-selection that depends on G **our estimate** ($\widehat{\beta_3^{\text{TSLS}}}$) will be biased (Hollenbach, Schmitz, and Westphal, 2026)
- **Solution:** Use the **Marginal Treatment Effect (MTE)** framework (Heckman and Vytlacil, 2005) to account for heterogeneity and potential self-selection in *G*

Outline

Outline

Roy Model with a Categorical Moderator



- ***E***: College Attainment dummy
- ***Z***: Reduction in Distance to University (IV)
- ***Y***: Imputed Log Hourly Wage
- ***V***: Unobserved confounder vector affecting both E and Y (e.g. skills, motivation)
- ***G***: EA PGI (tertiles)

IV Model with G Setup

- Modified Roy model (Roy, 1951):
  - **Choice Equation:** $E = f_E(Z, V), \quad E \in \{0, 1\}$
  - **Outcome Equation:** $Y = f_Y(E, V, \varepsilon_Y)$
  - **Confounding Vector:** $\boldsymbol{V} = f_V(G, \varepsilon_V)$
  - **Independence Condition:** $G, Z, \varepsilon_Y, \varepsilon_V$ are mutually independent

- Standard IV conditions:
  - Relevance
  - Exclusion
  - Exogeneity: $Z, G, \varepsilon_Y, \varepsilon_V$ are mutually independent
  - Monotonicity

Resistance to Treatment Ranks

- For any measurable $\psi : \mathbb{R}^d \to \mathbb{R}$ we have that $S \equiv \psi(\boldsymbol{V})$ is a random variable:

$$F_S(s) = Pr(S \leq s)$$

- For individuals in group $g \in \mathcal{G}$:

$$F_S^g(s) = Pr(S \leq s \,|\, G = g)$$

- Which allows to **rank individuals based on their unobservables**:

$$U \equiv F_S(S) \sim Unif(0,1)$$
$$U^g \equiv F_S^g(S) \,\big|\, G = g \sim \text{Unif}(0,1)$$

Treatment Rules

- Following Heckman and Vytlacil (2005), an individual attains college whenever:

$$E = \mathbb{1}\{P(Z) \geq U\}, \qquad E = \mathbf{1}\{P^g(Z) \geq U^g\} \text{ on } \{G = g\} \tag{2}$$

  - $\boldsymbol{P(Z)} \equiv P(E = 1 \mid Z)$ is the pooled propensity score
  - $\boldsymbol{P^g(Z)} \equiv P(E = 1 \mid Z, G = g)$ is the group-specific propensity score

- U, U^g **rank** individuals according to their **unobserved resistance to college**
  - $u \in U$: pooled quantile of resistance
  - $u^g \in U^g$: group-specific quantile of resistance

- Individuals at $P(Z) = U$ and $P^g(Z) = U^g$ on $\{G = g\}$ are indifferent

Marginal Treatment Effects

- The Marginal Treatment Effect (MTE) is the expected effect of treatment conditioned on unobserved resistance to treatment U. For the **pooled** population:

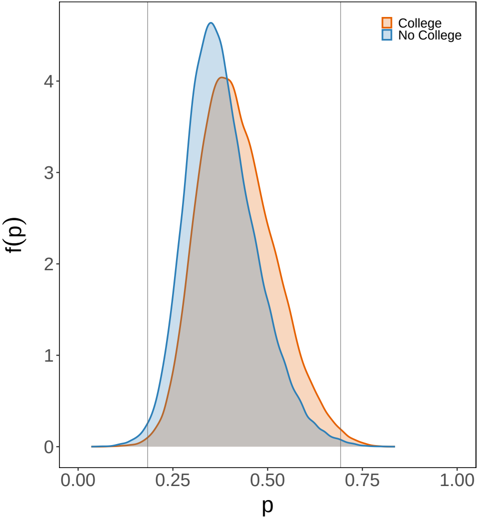
$$\text{MTE}(u) \equiv \mathbb{E}(Y_1 - Y_0 | U = u)$$

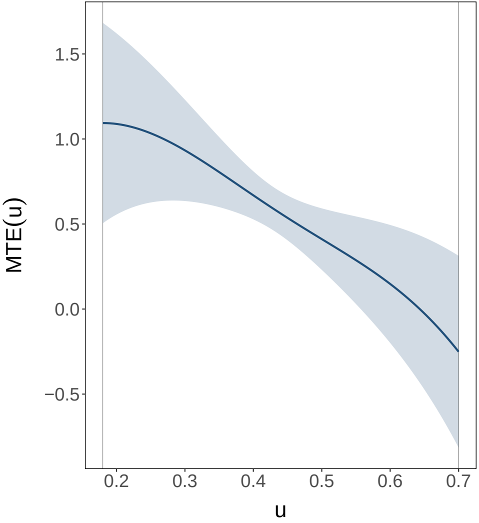
- Analogously, the **group-specific** MTE is:

$$\text{MTE}^g(u) \equiv \mathbb{E}(Y_1 - Y_0 | U^g = u^g, G = g)$$

Outline

Results (I): Pooled Propensity and MTE





Treatment Effects and MTE

- Following Carneiro, Heckman, and Vytlacil (2011) we can compute different treatment effects, τ, as weighted averages of the MTE:

$$\tau = \int_0^1 \text{MTE}(u)\, w^{\tau}(u) du \tag{3}$$

- For instance:

$$\text{ATE} = \mathbb{E}(Y_1 - Y_0) = \int_0^1 \text{MTE}(u)\, du$$

$$\text{ATT} = \mathbb{E}(Y_1 - Y_0 \mid E = 1) = \int_0^1 \text{MTE}(u)\, w^{ATT}(u)\, du$$

$$\text{ATU} = \mathbb{E}(Y_1 - Y_0 \mid E = 0) = \int_0^1 \text{MTE}(u)\, w^{ATU}(u)\, du$$

Results (II): Returns to college

	Support		Treatment Effects*		
	$u_{\min}$	$u_{\max}$	ATE	ATT	ATU
Pooled MTE					
Unconditional	0.18	0.70	0.545***	0.937***	0.261
			[0.40, 0.70]	[0.57, 1.30]	[-0.01, 0.53]
Conditional on G in X	0.05	0.90	0.530***	0.967***	0.235
			[0.36, 0.68]	[0.62, 1.36]	[-0.02, 0.54]

* Approximations to the parameters, based on the empirical common support of the propensity score. Bootstrap 95% confidence intervals (1000 replicates). Conditional on G defines the MTE model that includes the EA PGI (G) as an additional observed covariate in X.

Large and Heterogenous Returns to College

- College yields, on average, **large returns:** over 50%
- Yet the pooled MTE reveals **substantial heterogeneity:**
  - very large gains for those at low resistance to college
  - much smaller (possibly null) gains for those at high resistance
  - **Selection into gains**, consistent with e.g. Carneiro, Heckman, and Vytlacil (2011); Nybom (2017)

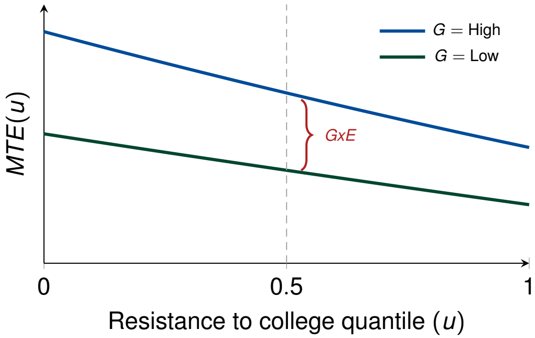
How does this heterogeneity differ by EA PGI group?

Outline

Outline

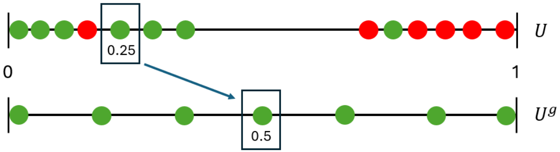
Estimating GxE via MTE

- We want to compare High PGI ($G = H$) individuals to Low PGI ($G = L$)
- Difference in MTE^{g}s gives the interaction effect (e.g. Hollenbach, Schmitz, and Westphal (2026))



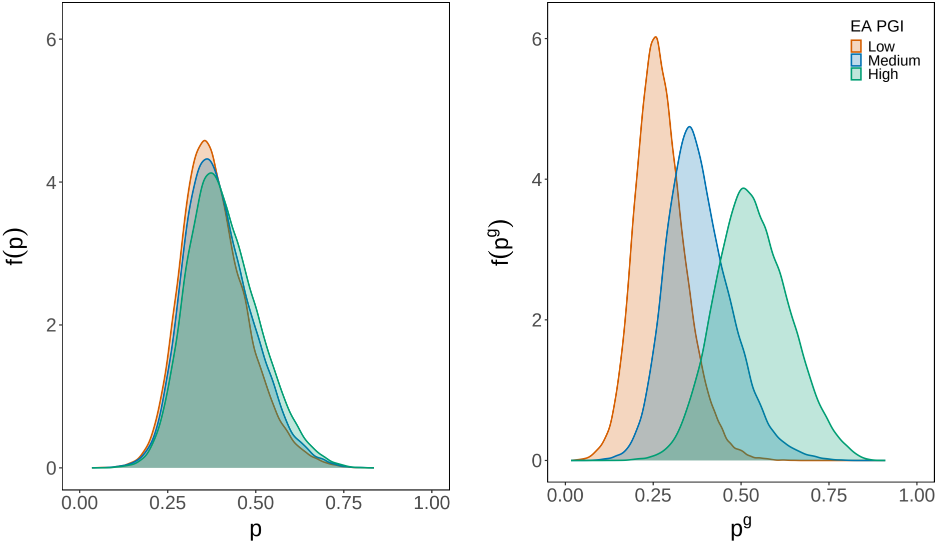
The Problem: Rank Comparability

- Is resistance universal? Most likely, $U \neq U^H \neq U^L$



- A fixed quantile might imply very different latent resistance across groups
- Differences in MTE may combine genuine heterogeneity and selection patterns

Results (III): PScore by EA PGI Tertile

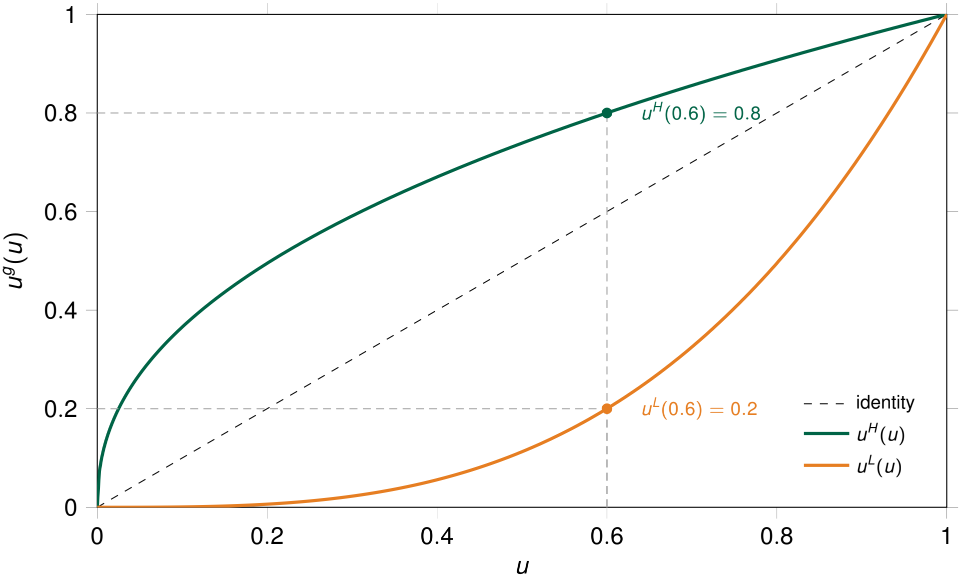


Relationship between Resistances (U, U^g)

- Key feature: We allow pooled and group-specific ranks **to differ**
  - Different rankings unless $F_S^g(S) \equiv F_S(S)$
- Relationship between U and U^g is governed by the warp $u^g : [0,1] \to [0,1]$:

$$u^g(u) = F_S^g(F_S^{-1}(u)) \tag{4}$$

- **Intuition:** A level u of resistance corresponds to a quantile of resistance U in the whole population, but to quantile U^g within group $G = g$
  - If $u^g < u$: Group $G = g$ is more resistant to college
  - If $u^g > u$: Group $G = g$ is less resistant to college


$u^H(0.6) = 0.8$
$u^L(0.6) = 0.2$
identity
$u^H(u)$
$u^L(u)$
$u^g(u)$
u
0
0.2
0.4
0.6
0.8
1

Properties of the warp

- By (2), the warp is identified via the **calibration identity**:

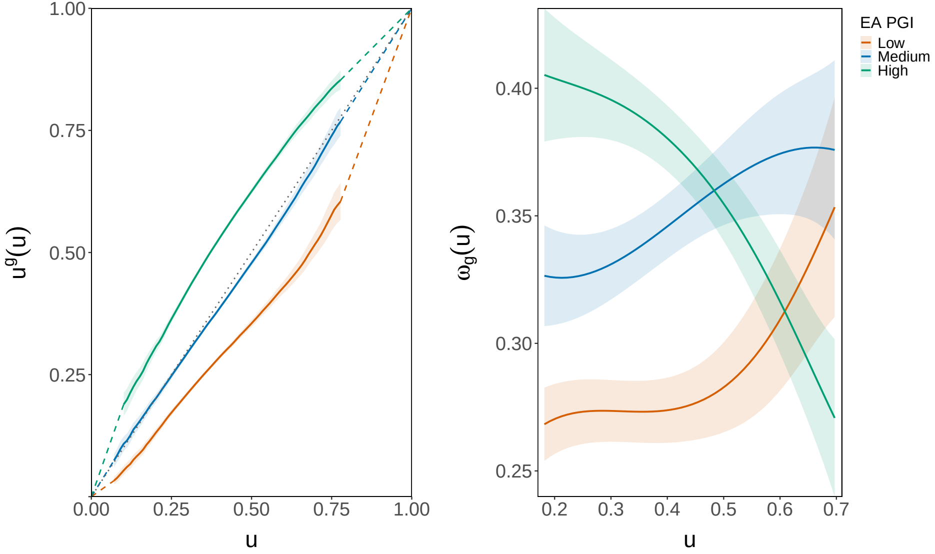
$$u^g(P(Z)) = P^g(U) \tag{5}$$

- The derivative yields the tilt weights:

$$\omega_g(u) = u^{g\prime}(u) = \frac{f_S^g(F_S^{-1}(u))}{f_S(F_S^{-1}(u))} \tag{6}$$

  - “how much group *g* contributes to the unconditional resistance at rank u”
  - $\sum_g \pi_g \omega_g(u) = 1, \quad \forall u \in U$

Results (IV): Estimated warp and tilt weights functions



Comparing group-specific MTEs

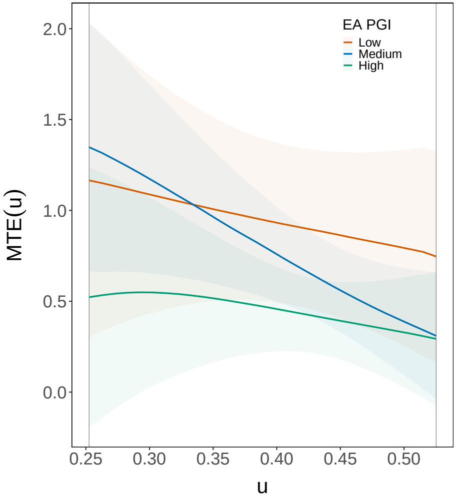
- Groups exhibit **clearly different latent resistances** to college
- To compare them, we express all MTEs on a **common pooled resistance scale**
- We do this using the **warp** $u^g(u)$:

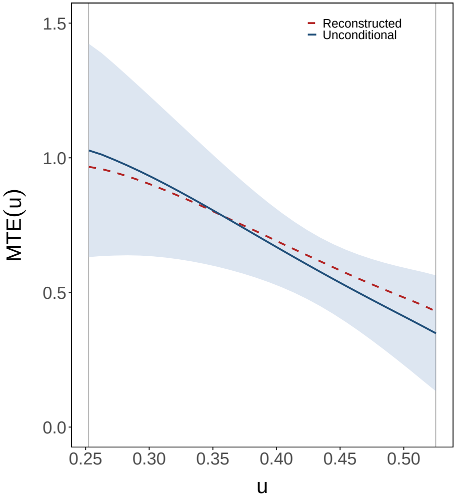
$$\widetilde{\text{MTE}}^{g}(u) = \text{MTE}^g\big(u^g(u)\big) \tag{7}$$

- The pooled MTE is then a **weighted, warped average** of group-specific MTEs:

$$\text{MTE}(u) = \sum_{g \in \mathcal{G}} \pi_g \, \text{MTE}^g\left(u^g(u)\right) \omega_g(u) \tag{8}$$

Results (V): Group-specific MTE





Results (VI): Returns to college by EA PGI

	Support				Treatment Effects*		
	$u^g_{\min}$	$u^g_{\max}$	$u_{\min}$	$u_{\max}$	ATE	ATT	ATU
Group-specifc MTE							
Low EA PGI	0.17	0.38	0.25	0.53	0.960***	1.060***	0.868***
					[0.51, 1.41]	[0.50, 1.62]	[0.38, 1.36]
Med EA PGI	0.25	0.51	0.25	0.53	0.816***	1.097***	0.543***
					[0.50, 1.14]	[0.59, 1.60]	[0.25, 0.83]
High EA PGI	0.37	0.65	0.25	0.53	0.444***	0.487***	0.393***
					[0.17, 0.72]	[0.04, 0.93]	[0.14, 0.64]

*Treatment effects are evaluated over the observed common support of the resistance distribution. In the absence of full support, the reported estimates correspond to integrals over the identified region and therefore approximate the population treatment parameters defined over the full support. Entries report point estimates with bootstrap 95% confidence intervals in brackets (1000 replications).

- Group-specific results suggest **larger returns to college for Low EA PGI** individuals
- This pattern is consistent with **college mitigating inequality in earnings**

What drives these between-group differences?

Outline

Disentangling Gene-Environment Interplay

- **Pathways of Gene-Environment Interplay:**

  (a) Variation in genetic makeup is linked to variation in educational attainment
  (rGE: gene–environment correlation)

  (b) Returns to education vary depending on one's genetic makeup
  ($G \times E$: gene–environment interaction)

- $\boxed{\textbf{Aim:}}$
  Disentangle *GxE* from *rGE* **to understand** how genetic predisposition *G* shapes the **heterogeneity in the returns to college.**

Decomposition of GxE: Example for the ATE (I)

- We are interested in the **difference in returns to college** between the high and low EA PGI groups ($G = H$ vs. $G = L$) in the full population, i.e.,

$$\Delta^{\text{ATE}} = \text{ATE}^{H} - \text{ATE}^{L}$$

- The group-specific ATE is:

$$\text{ATE}^{g} = \int_0^1 \text{MTE}^{g}(u)\,\omega_g(u)\,du, \quad g \in \{H, L\} \tag{9}$$

- Hence,

$$\Delta^{\text{ATE}} = \underbrace{\int_0^1 \text{MTE}^{H}(u)\,\omega_H^{\text{ATE}}(u)\,du}_{\text{ATE}^{H}} - \underbrace{\int_0^1 \text{MTE}^{L}(u)\,\omega_L^{\text{ATE}}(u)\,du}_{\text{ATE}^{L}}$$

Decomposition of GxE: Example for the ATE (II)

- The **tilt weights** $\omega_G(u) = \frac{f_S^G(F_S^{-1}(u))}{f_S(F_S^{-1}(u))}$ depend only on selection:
  - Variation in G linked to variation in $E \implies$ $\boxed{\text{rGE}}$
- $\text{MTE}^g(u) = \mathbb{E}[Y_1 - Y_0 | U^g = u, G = g]$ measures the **treatment effect at a given resistance** level
  - Different returns at a given resistance (linked to propensity) to college $\implies$ $\boxed{\text{GxE}}$

Decomposition of GxE: Example for the ATE (III)

- Suppose there are no differences in $\mathrm{MTE}^g(u)$ across G:

$$\mathrm{MTE}^H(u) = \mathrm{MTE}^L(u) = \overline{\mathrm{MTE}}(u)$$

- Then,

$$\Delta^{\mathrm{ATE}} = \int_0^1 \overline{\mathrm{MTE}}(u)\,(\omega_H(u) - \omega_L(u))\,du$$

- The difference in the ATE arise from the weights $\boxed{\text{rGE}}$

Decomposition of GxE: Example for the ATE (IV)

- Instead, suppose there are no distributional differences across G, then:

$$\omega_H = \omega_L = \overline{\omega}(u)$$

- Then,

$$\Delta^{\text{ATE}} = \int_0^1 \left(\text{MTE}^H(u) - \text{MTE}^L(u) \right) \overline{\omega}(u) du$$

- All differences in the ATE arise from the difference in the MTE $\boxed{G \times E}$

Decomposition of GxE: Example for the ATE (V)

- Put together:

$$\Delta^{\text{ATE}} = \underbrace{\int_0^1 \left(\text{MTE}^H(u) - \text{MTE}^L(u)\right) \overline{\omega}(u)\, du}_{\boxed{G \times E}} + \underbrace{\int_0^1 \overline{\text{MTE}}(u) \left(\omega_H(u) - \omega_L(u)\right) du}_{\boxed{\text{rGE}}} \tag{10}$$

- Where:
  - $\overline{\omega}(u) = \frac{1}{2}\left(\omega_H(u) + \omega_L(u)\right)$
  - $\overline{\text{MTE}}(u) = \frac{1}{2}\left(\text{MTE}_H(u) + \text{MTE}_L(u)\right)$

- This **approach is also applicable to other treatment effects** See generalisation

Results (VII): Decomposition of ATE

Comparison	G$\times$E (MTE)	rGE (Composition)	Total
Medium vs Low	-0.143	-0.006	-0.150
	[-0.59, 0.30]	[-0.02, 0.00]	[-0.60, 0.30]
High vs Low	-0.505	0.007	-0.499
	[-0.96, -0.05]	[-0.01, 0.02]	[-0.96, -0.04]
High vs Medium	-0.367	***0.018	-0.349
	[-0.70, -0.03]	[0.00, 0.04]	[-0.69, -0.01]

This table decomposes differences in average treatment effects (ATE) across EA PGI tertiles into a sorting component (rGE) and a return heterogeneity component (G$\times$E) following Proposition 3. Bootstrap 90% confidence intervals based on 1000 replications are reported in brackets.

Outline

Conclusions

- We **extended the MTE framework** to make group-specific returns comparable and decompose the differences into *rGE* and $G \times E$
- We find **large and heterogeneous returns to college**
- Returns are larger for **Low EA PGI** individuals, pointing to **negative** $G \times E$
- The decomposition suggests that this gap is driven mainly by $G \times E$, with only a **limited role for rGE**
- This pattern suggests that college may **mitigate genetic inequality in earnings**

Thank you for listening!

tomeu.lopeznieto@unibo.it
https://tomeulnv.github.io/

References I

Barcellos, Silvia H., Leandro S. Carvalho, and Patrick Turley. 2021. “The Effect of Education on the Relationship between Genetics, Early-Life Disadvantages, and Later-Life SES.” URL https://www.nber.org/papers/w28750.

Becker, Gary S. 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. New York: Columbia University Press.

Berlingieri, Francesco, Christina Gathmann, and Matthias Quinckhardt. 2022. “College Openings and Local Economic Development.” Tech. rep., IZA Discussion Paper No. 15364. Available at SSRN: https://ssrn.com/abstract=4135927 or http://dx.doi.org/10.2139/ssrn.4135927.

Biroli, Pietro, Titus Galama, Stephanie Von Hinke, Hans van Kippersluis, Cornelius A. Rietveld, and Kevin Thom. 2026. “The Economics and Econometrics of Gene–Environment Interplay.” *The Review of Economic Studies* 93:144–180. URL https://dx.doi.org/10.1093/restud/rdaf034.

Blundell, Richard, David Green, and Wenchao (Michelle) Jin. 2016. “Big historical increase in numbers did not reduce graduates’ relative wages.” Institute for Fiscal Studies (IFS) Press Release. URL https://ifs.org.uk/news/big-historical-increase-numbers-did-not-reduce-graduates-relative-wages. Published on August 18, 2016. Accessed March 23, 2025.

Card, David. 1995. “Using Geographic Variation in College Proximity to Estimate the Return to Schooling.” In *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, edited by L.N. Christofides, E.K. Grant, and R. Swidinsky. University of Toronto Press, 201–222.

Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil. 2011. “Estimating Marginal Returns to Education.” *American Economic Review* 101:2754–81. URL http://www.aeaweb.org/articles.php?doi=10.1257/aer.101.6.2754.

Heckman, James J. and Edward Vytlacil. 2005. “Structural Equations, Treatment Effects, and Econometric Policy Evaluation.” *Econometrica* 73 (3):669–738.

Hollenbach, Johannes, Hendrik Schmitz, and Matthias Westphal. 2026. “Gene-environment interactions with essential heterogeneity.” *The Economic Journal* :ueag010URL https://doi.org/10.1093/ej/ueag010.

Ichino, Andrea, Aldo Rustichini, and Giulio Zanella. 2025. “College, Cognitive Ability, and Socioeconomic Disadvantage: Policy Lessons from the UK in 1960–2004.” *The Review of Economic Studies* 0:1–45.

Kweon, Hyeokmoon, Casper Burik, Richard Karlsson Linnér, Ronald de Vlaming, Aysu Okbay, Daphne Martschenko, Kathryn Harden, Thomas A. DiPrete, and Philipp Koellinger. 2020. “Genetic Fortune: Winning or Losing Education, Income, and Health.” Tech. rep., Tinbergen Institute Discussion Paper 2020-053/V. URL https://ssrn.com/abstract=3682041. Available at SSRN: https://ssrn.com/abstract=3682041 or http://dx.doi.org/10.2139/ssrn.3682041.

Mincer, Jacob. 1974. *Schooling, Experience, and Earnings*. New York: National Bureau of Economic Research.

Nybom, Martin. 2017. “The Distribution of Lifetime Earnings Returns to College.” *Journal of Labor Economics* 35 (4):903–952.

Okbay, Aysu, You Wu, Ning Wang, and et al. 2022. “Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals.” *Nature Genetics* 54 (4):437–449.

References II

Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. 2006. "Principal components analysis corrects for stratification in genome-wide association studies." *Nature Genetics 2006 38:8* 38:904–909. URL https://www.nature.com/articles/ng1847.

Robbins, Lionel. 1963. *Higher Education: Report of the Committee Appointed by the Prime Minister Under the Chairmanship of Lord Robbins, 1961-63*. London: Her Majesty's Stationery Office (HMSO).

Roy, A. D. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3 (2):135–146.

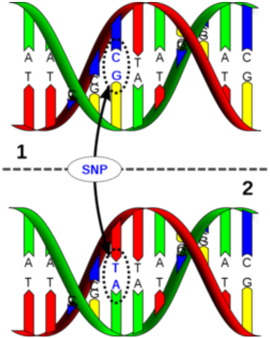
The 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. 2015. "A global reference for human genetic variation." *Nature* 526 (7571):68–74.

van Alten, S., B. W. Domingue, J. Faul, T. Galama, and A. T. Marees. 2024. "Reweighting UK Biobank corrects for pervasive selection bias due to volunteering." *International Journal of Epidemiology* 53 (3):dyae054. URL https://doi.org/10.1093/ije/dyae054.

Westphal, Matthias, Daniel A. Kamhöfer, and Hendrik Schmitz. 2022. "Marginal College Wage Premiums Under Selection Into Employment." *The Economic Journal* 132 (646):2231–2272. URL https://doi.org/10.1093/ej/ueac021.

Primer in Genetics

- Human genome: series of 3 billion letter pairs (A,G,T,C)
- **Genetic variants:** one-letter changes across individuals (single nucleotide polymorphisms, SNPs)
- About 10m **SNPs** (The 1000 Genomes Project Consortium et al., 2015)
- Genome-wide association studies **(GWAS) have identified genome-wide significant relationships between specific SNPs and phenotypes** of all sorts.
- We **use SNPs** identified from large, replicated GWAS **to create summarised genetic scores** to study gene-by-environment interplay

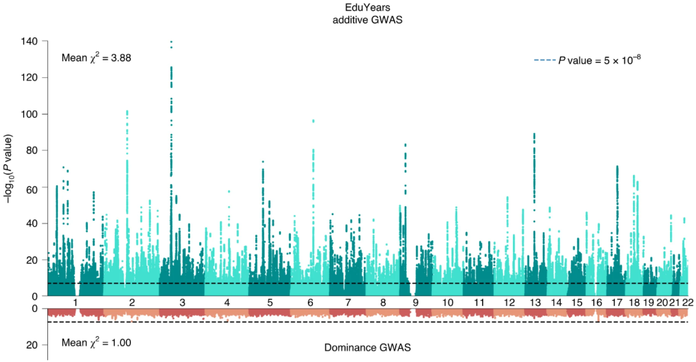


Polygenic Index (PGI)

Discovery stage:

GWAS (Okbay et al., 2022)

Back



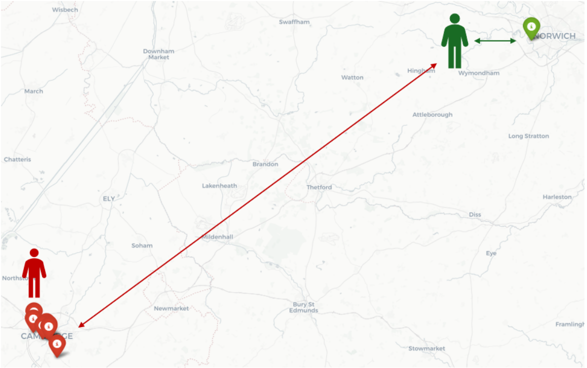
Prediction stage:

$$PGI_i = \sum_{j=1}^{J} \beta_j \cdot G_{ij}$$

where G_{ij} is the genotype for individual i at SNP j, and β_j is the OLS association between SNP j and the outcome.The PGI is normalized to have mean zero and standard deviation one.

Who benefitted from the University Expansion?

- Studies include **distance to university as a cost shifter** in pursuing higher education
- However, **birth location determines if and by how much the expansion actually shifted the costs**



Reduction in Distance: Summary Statistics

Back

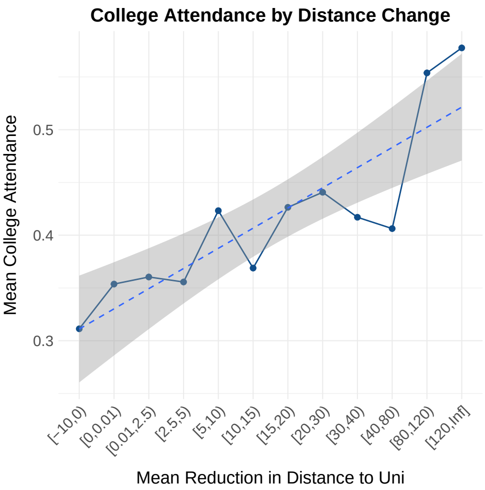
	(1) Overall	(2) Positive Reduction	(3) Negative Reduction	(4) No Reduction
Mean Reduction	4.721	18.949	-0.382	0
Std. Deviation	13.531	21.552	0.582	0
Min	-3.912	0.002	-3.912	0
Q1	0.000	2.606	-0.351	0
Q2 (Median)	0.000	11.928	-0.202	0
Q3	0.000	30.133	-0.135	0
Max	180.798	180.798	-0.005	0
Num. Obs.	158263	39467	1595	117201
Frequency	100%	24.93%	1.01%	74.06%

Summary statistics for the reduction in distance to university (measured in kilometers). The reduction is calculated as the difference between the distance to the nearest university active at birth and the distance to the nearest university active by age 15.

Reduction in Distance and College Attainment

Quadratic Impact Test

Predictive of PGI Test



Test for Quadratic Impact of the Reduction in Distance

Back

	(1) College	(2) College
Reduction in Distance to University	0.012***	0.011***
	(0.002)	(0.001)
Reduction in Distance Squared	-0.000	
	(0.000)	
Universities (pre-birth) within 50km		-0.001***
		(0.000)
UR average (1950-1960)	0.369***	-0.341*
	(0.058)	(0.145)
UR average (1950-1960) squared		0.163***
		(0.040)
Population Relative Growth		0.013
		(0.026)
Population Density	-0.007***	-0.007***
	(0.001)	(0.001)
Num.Obs.	158263	158263
R2	0.019	0.019
AIC	261100.8	261074.3
BIC	264251.9	264235.4
Region FE	Yes	Yes
Cohort FE	Yes	Yes

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses. UKB Weights have been employed (van Alten et al., 2024)

Reduction in Distance on EA PGI

Back	(1) EA PGI	(2) EA PGI	(3) EA PGI
Reduction in Distance to University	0.003	0.003	0.005
	(0.003)	(0.003)	(0.003)
Universities (pre-birth) within 50km		0.000	0.000
		(0.000)	(0.000)
UR average (1950-1960)		-0.910**	-0.677*
		(0.303)	(0.305)
UR average (1950-1960) squared		0.259**	0.194*
		(0.079)	(0.080)
Population Relative Growth		-0.022	-0.011
		(0.052)	(0.051)
Population Density		-0.003	-0.009***
		(0.002)	(0.002)
Num.Obs.	158263	158263	158263
R2	0.015	0.015	0.032
R2 Adj.	0.013	0.013	0.030
Region FE	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes
Genetic PCs	No	No	Yes

* p <0.05, ** p <0.01, *** p <0.001. Robust standard errors in parentheses. UKB Weights have been employed (van Alten et al., 2024)

Timing Robustness and Placebo Tests

	Alternative childhood windows					Placebo windows			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Birth→11	Birth→12	Birth→13	Birth→14	Birth→15	20→24	24→30	30→35	35→40
Reduction in distance (Z)	0.01***	0.01***	0.01***	0.01***	0.01***	-0.02	0.07	0.012	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.006)	(0.008)	(0.003)
Num. Obs.	158223	158223	158223	158223	158223	158223	158223	158223	158223
R2	0.086	0.086	0.086	0.086	0.086	0.086	0.086	0.086	0.086
First-stage F	92.17	93.09	90.31	94.73	94.71	0.55	2.01	3.83	0.00
Region / Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls (X)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. All specifications use the same sample, controls, fixed effects, and UKB weights (van Alten et al., 2024). Panel A varies the childhood exposure window used to construct Z; Panel B reports placebo windows occurring after typical college-entry ages.

First Stage by Subgroups

	EA PGI Tertiles			Sex		SES	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Low	Medium	High	Female	Male	Low	High
Reduction in Distance (Z)	0.008***	0.011***	0.013***	0.012***	0.009***	0.009***	0.008***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Num. Obs.	52741	52742	52740	82865	75358	78572	79651
R2	0.023	0.034	0.043	0.037	0.039	0.030	0.036
First-stage F	19.72	29.12	40.20	65.44	33.99	22.66	39.07
Region FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls (X)	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. UKB weights have been applied. (van Alten et al., 2024).

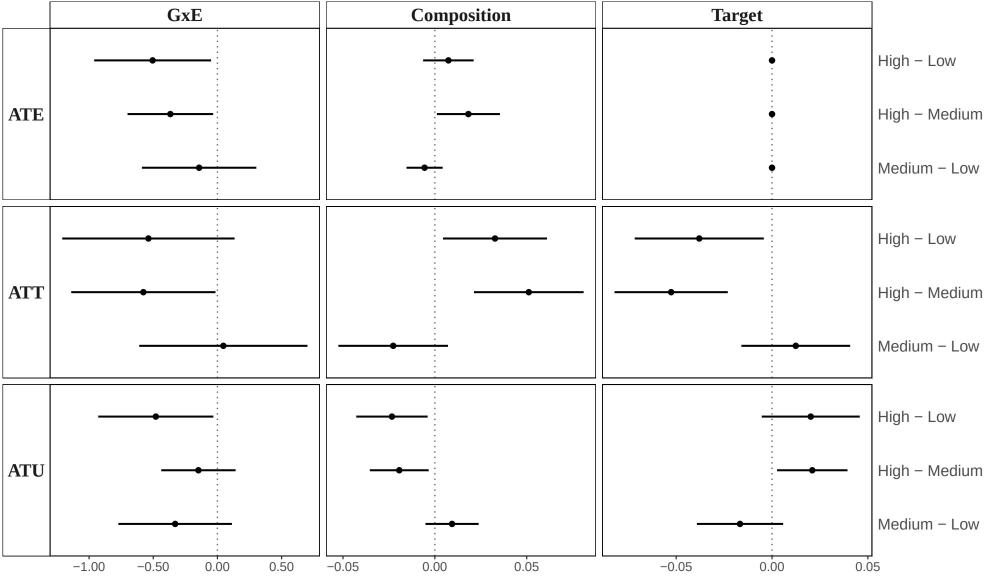
Generalisation of the decomposition of differences in TE

- Let $A_g(u) \equiv \mathrm{MTE}^g(u^g(u))$, $W_g(u) \equiv B_g(u)\, C_g(u)$
  - where $B_g(u) \equiv w_\tau^g(u^g(u))$ and $C_g(u) \equiv \omega_g(u)$
- Then, the g-specific TE is:

$$\tau^g = \int_0^1 A_g(u)\, B_g(u)\, C_g(u)\, du$$

- **Proposition.** The difference $\tau^g - \tau^{g'}$ admits the following exact decomposition:

$$\begin{aligned}
\tau^g - \tau^{g'} = & \int_0^1 \left(\frac{W_g + W_{g'}}{2} \right) (A_g - A_{g'})\, du \quad \text{(MTE channel)} \\
& + \int_0^1 \left(\frac{A_g + A_{g'}}{2} \cdot \frac{C_g + C_{g'}}{2} \right) (B_g - B_{g'})\, du \quad \text{(Target-weight channel)} \\
& + \int_0^1 \left(\frac{A_g + A_{g'}}{2} \cdot \frac{B_g + B_{g'}}{2} \right) (C_g - C_{g'})\, du \quad \text{(Composition channel)}
\end{aligned}$$


GxE
Composition
Target
ATE
ATT
ATU
High − Low
High − Medium
Medium − Low
High − Low
High − Medium
Medium − Low
High − Low
High − Medium
Medium − Low
−1.00
−0.50
0.00
0.50
−0.05
0.00
0.05
−0.05
0.00
0.05