Causal identification in social science genetics

Paul Hufe*

*University of Bristol

January 6, 2025

Outline

Outline

Causality, potential outcomes ... a quick recap

Identifying genetic effects

Identifying gene-environment interactions

An example: Genes and Schools

Causation vs. correlation

- In social science, we are (often) interested in causal effects, not correlations. Why?

Causation vs. correlation

- In social science, we are (often) interested in causal effects, not correlations. Why?

→ Because we are scientists: “We do not have knowledge of a thing until we have grasped its why, that is to say, its cause.” — Aristotle

→ Because it matters for normative evaluations: Is inequality due to native talent or due to its correlates ? — e.g., Arneson (2018)

→ Because it matters for policy: Do schools address the inequity of birth? — e.g., Coleman et al. (1966)

Causation vs. correlation

- In social science, we are (often) interested in causal effects, not correlations. Why?

→ Because we are scientists: “We do not have knowledge of a thing until we have grasped its why, that is to say, its cause.” — Aristotle

→ Because it matters for normative evaluations: Is inequality due to native talent or due to its correlates ? — e.g., Arneson (2018)

→ Because it matters for policy: Do schools address the inequity of birth? — e.g., Coleman et al. (1966)

- For example:
  - Does a higher PGI^{EA} cause higher educational attainment?
  - Do investments into schools amplify/mitigate the effects of PGI^{EA} on educational attainment?
  - ...

A framework to think about causality

- Our phenomenon of interest: Some people have higher educational attainment than others. It is a perennial question in education economics what causes these differences.
- Our research question: Does a high PGI^{EA} cause higher educational attainment?
- Our data:
  - Y_i indicates the education of i.
  - $D_i = \begin{cases} 1 \text{ if individual } i \text{ has a high } PGI^{EA}, \\ 0 \text{ if individual } i \text{ has a low } PGI^{EA}. \end{cases}$

Potential outcomes

- Each individual i has two potential outcomes.

- Both potential outcomes are defined, but only one is realized:

$$Y_i = \begin{cases} Y_{1i} \text{ if } D_i = 1 \\ Y_{0i} \text{ if } D_i = 0 \end{cases}$$

Potential outcomes

- Each individual i has two potential outcomes.

- Both potential outcomes are defined, but only one is realized:

$$Y_i = \begin{cases} Y_{1i} \text{ if } D_i = 1 \\ Y_{0i} \text{ if } D_i = 0 \end{cases}$$

- For any i, the causal effect of treatment D on outcome Y is:

$$\tau_i = Y_{1i} - Y_{0i}$$

Potential outcomes

- Each individual i has two potential outcomes.

- Both potential outcomes are defined, but only one is realized:

$$Y_i = \begin{cases} Y_{1i} \text{ if } D_i = 1 \\ Y_{0i} \text{ if } D_i = 0 \end{cases}$$

- For any i, the causal effect of treatment D on outcome Y is:

$$\tau_i = Y_{1i} - Y_{0i}$$

- We cannot estimate τ_i unless we solve the "missing data" problem:

$$\begin{aligned} \tau_i &= Y_{1i} - ? \text{ if } D_i = 1, \\ &= ? - Y_{0i} \text{ if } D_i = 0 \end{aligned}$$

A naive comparison

- Let's assume homogeneous treatment effects, i.e., that high PGI^{EA} improves everyone's education by $\tau_i = \tau$:

$$Y_{1i} = Y_{0i} + \tau$$

- We can then write our naive comparison as:

$$\begin{aligned}
&\mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\
&= \mathbb{E}[Y_{0i} + \tau|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\
&= \mathbb{E}[\tau|D_i = 1] + \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\
&= \tau + \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0]
\end{aligned}$$

A naive comparison

- Let's assume homogeneous treatment effects, i.e., that high PGI^{EA} improves everyone's education by $\tau_i = \tau$:

$$Y_{1i} = Y_{0i} + \tau$$

- We can then write our naive comparison as:

$$\begin{aligned}
&\mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\
&= \mathbb{E}[Y_{0i} + \tau|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\
&= \mathbb{E}[\tau|D_i = 1] + \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\
&= \tau + \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\
&= \text{ATE + Selection bias}
\end{aligned}$$

- Selection bias: Potential outcomes of treated and non-treated are often not identical, i.e., we are not comparing apples to apples.

More formally ...

- We can identify the ATE if Strong Ignorability holds:

→ D_i is strongly ignorable conditional on $\mathbf{X}_i$ if

1. $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i | \mathbf{X}_i$
2. $\exists \epsilon > 0$ s.t. $\epsilon < \Pr(D_i = 1 | X_i) < 1 - \epsilon_i$

   - The first condition asserts independence of the treatment from the "potential" outcomes.
   - The second condition asserts that there are both treated and untreated individuals.

- We often also say "D_i is conditionally randomly assigned" or "D_i is exogeneous".

- We need to look for a research design such that strong ignorability is satisfied.

Randomized control trials (RCTs)

- RCTs (usually) comply with strong ignorability without conditioning on $\mathbf{X}_i$.
  $\rightarrow$ Selection bias vanishes.

Randomized control trials (RCTs)

- RCTs (usually) comply with strong ignorability without conditioning on $\mathbf{X}_i$.
  $\rightarrow$ Selection bias vanishes.

- RCTs are possible in challenging settings, i.e., health insurance coverage of individuals.
  $\rightarrow$ Oregon Health Insurance Experiment.

Randomized control trials (RCTs)

- RCTs (usually) comply with strong ignorability without conditioning on $\mathbf{X}_i$.
  $\rightarrow$ Selection bias vanishes.

- RCTs are possible in challenging settings, i.e., health insurance coverage of individuals.
  $\rightarrow$ Oregon Health Insurance Experiment.

- For the identification of genetic effects, it is inherently impossible to run an RCT.
  $\rightarrow$ Can we allocate alleles across individuals?

$\rightarrow$ We have to think about research designs that allow us to mimic a random allocation of alleles.

Outline

Causality, potential outcomes ... a quick recap

Identifying genetic effects

Identifying gene-environment interactions

An example: Genes and Schools

Outline

Causality, potential outcomes ... a quick recap

Identifying genetic effects
- Addressing gene-environment correlation
- Addressing measurement error

Identifying gene-environment interactions

An example: Genes and Schools

Setting the stage

- Our research question: Does a high PGI^{EA} cause higher educational attainment?
- Our data:
  - Y_i indicates the education of i.
  - $PGI_i^{EA} \in \mathcal{N}(0,1)$
- Our naive comparison:

$$Y_i = \tau PGI_i^{EA} + \epsilon_i$$

$\rightarrow$ What can go wrong?

Strong ignorability is violated

- We cannot identify causal effects if Strong Ignorability is violated.
- In our naive comparison, the treatment PGI^{EA} is not assigned independently of potential outcomes:

$$Y_i = \tau PGI_i^{EA} + \underbrace{\alpha_p PGI_{p(i)}^{EA} + \alpha_m PGI_{m(i)}^{EA} + \xi_i}_{=\epsilon_i}$$

- Estimates are confounded by gene-environment correlations (or “selection bias” in more general parlance) ...

Strong ignorability is violated

- We cannot identify causal effects if Strong Ignorability is violated.
- In our naive comparison, the treatment PGI^{EA} is not assigned independently of potential outcomes:

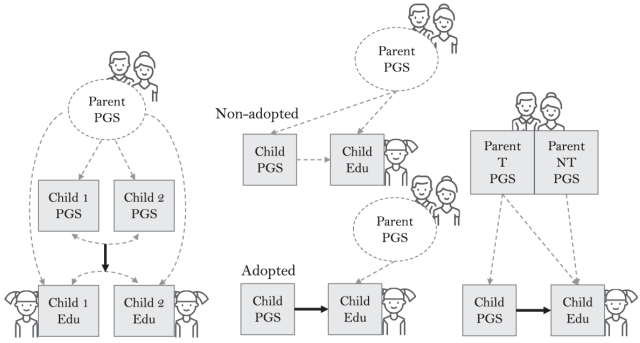
$$Y_i = \tau PGI_i^{EA} + \underbrace{\alpha_p PGI_{p(i)}^{EA} + \alpha_m PGI_{m(i)}^{EA} + \xi_i}_{=\epsilon_i}$$

- Estimates are confounded by gene-environment correlations (or “selection bias” in more general parlance) ...

... unless we choose $\mathbf{X}_i$ wisely,

... unless we choose our sample wisely.

Research designs (Demange et al., 2022)



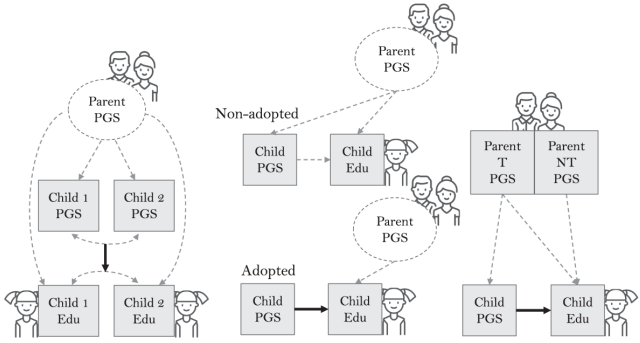
	Sibling design	**Adoption design**	**Non-transmitted PGS design**
Population genetic effect	$\beta_{Population}$	$\beta_{Non\text{-}adopted}$	$\beta_{Transmitted}$
Direct genetic effect	$\beta_{Within\text{-}siblings}$	$\beta_{Adopted}$	$\beta_{Transmitted} - \beta_{Non\text{-}transmitted}$
Indirect genetic effect	$\beta_{Population} - \beta_{Within\text{-}siblings}$	$\beta_{Non\text{-}adopted} - \beta_{Adopted}$	$\beta_{Non\text{-}transmitted}$

Threads to validity of estimates

- Internal validity: Can we estimate the treatment effect for our particular sample (i.e., do we address the problem of selection bias)?
- External validity: Can we extrapolate the estimated treatment effect to other populations or settings?

$\rightarrow$ How should we rate these three designs in terms of their validity?

Research designs (Demange et al., 2022)



	Sibling design	**Adoption design**	**Non-transmitted PGS design**
Population genetic effect	$\beta_{Population}$	$\beta_{Non\text{-}adopted}$	$\beta_{Transmitted}$
Direct genetic effect	$\beta_{Within\text{-}siblings}$	$\beta_{Adopted}$	$\beta_{Transmitted} - \beta_{Non\text{-}transmitted}$
Indirect genetic effect	$\beta_{Population} - \beta_{Within\text{-}siblings}$	$\beta_{Non\text{-}adopted} - \beta_{Adopted}$	$\beta_{Non\text{-}transmitted}$

Sibling design

- Sibling designs address gene-environment correlations by using within-family variation only:

$$
\begin{aligned}
Y_{i1} &= \tau PGI_{i1}^{EA} + \eta PGI_{i2}^{EA} + \alpha_p PGI_{p(i)}^{EA} + \alpha_m PGI_{m(i)}^{EA} + \xi_{i1} \\
Y_{i2} &= \tau PGI_{i2}^{EA} + \eta PGI_{i1}^{EA} + \alpha_p PGI_{p(i)}^{EA} + \alpha_m PGI_{m(i)}^{EA} + \xi_{i2} \\
\Delta Y_i &= (\tau - \eta) \Delta PGI_i^{EA} + \Delta \xi_i
\end{aligned}
$$

Sibling design

- Sibling designs address gene-environment correlations by using within-family variation only:

$$
\begin{aligned}
Y_{i1} &= \tau PGI_{i1}^{EA} + \eta PGI_{i2}^{EA} + \alpha_p PGI_{p(i)}^{EA} + \alpha_m PGI_{m(i)}^{EA} + \xi_{i1} \\
Y_{i2} &= \tau PGI_{i2}^{EA} + \eta PGI_{i1}^{EA} + \alpha_p PGI_{p(i)}^{EA} + \alpha_m PGI_{m(i)}^{EA} + \xi_{i2} \\
\Delta Y_i &= (\tau - \eta) \Delta PGI_i^{EA} + \Delta \xi_i
\end{aligned}
$$

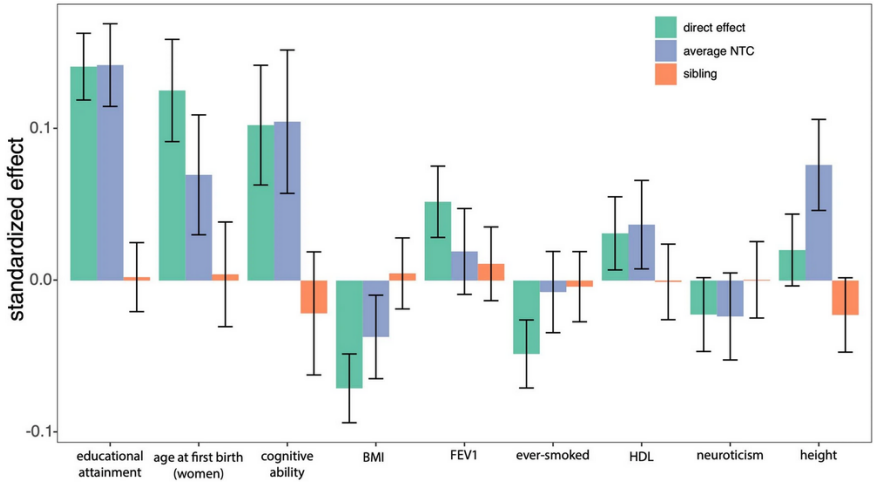
→ Provide valid estimates of τ in the absence of sibling spillovers ($\eta = 0$, see next slide for evidence).

→ External validity limited to multi-child families.

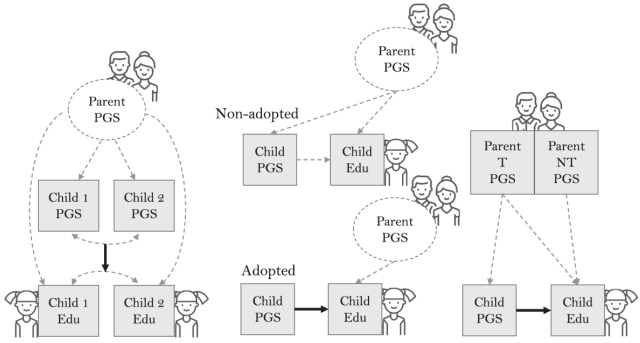
→ Drawbacks in terms of statistical power.

→ Some limitations for gene-environment interplay analyses (need variation across siblings).

Sibling design, cont'd (Young et al., 2022)



Research designs (Demange et al., 2022)



	Sibling design	**Adoption design**	**Non-transmitted PGS design**
Population genetic effect	$\beta_{Population}$	$\beta_{Non\text{-}adopted}$	$\beta_{Transmitted}$
Direct genetic effect	$\beta_{Within\text{-}siblings}$	$\beta_{Adopted}$	$\beta_{Transmitted} - \beta_{Non\text{-}transmitted}$
Indirect genetic effect	$\beta_{Population} - \beta_{Within\text{-}siblings}$	$\beta_{Non\text{-}adopted} - \beta_{Adopted}$	$\beta_{Non\text{-}transmitted}$

Adoption design

- Adoption designs address gene-environment correlations by leveraging the random allocation of adoptees to families (independent of genotypes):

$$Y_i = \tau PGI_i^{EA} + \epsilon_i$$

where $Cov(PGI_i^{EA}, PGI_{p(i)}^{EA}) = Cov(PGI_i^{EA}, PGI_{m(i)}^{EA}) = 0$.

Adoption design

- Adoption designs address gene-environment correlations by leveraging the random allocation of adoptees to families (independent of genotypes):

$$Y_i = \tau PGI_i^{EA} + \epsilon_i$$

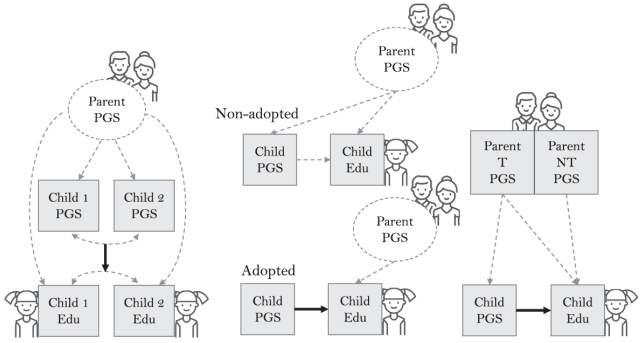
where $Cov(PGI_i^{EA}, PGI_{p(i)}^{EA}) = Cov(PGI_i^{EA}, PGI_{m(i)}^{EA}) = 0$.

→ Provide valid estimates of τ in the presence of random allocation.

→ External validity limited to adoptees.

→ Drawbacks in terms of statistical power due to small samples.

Research designs (Demange et al., 2022)



	Sibling design	**Adoption design**	**Non-transmitted PGS design**
Population genetic effect	$\beta_{Population}$	$\beta_{Non\text{-}adopted}$	$\beta_{Transmitted}$
Direct genetic effect	$\beta_{Within\text{-}siblings}$	$\beta_{Adopted}$	$\beta_{Transmitted} - \beta_{Non\text{-}transmitted}$
Indirect genetic effect	$\beta_{Population} - \beta_{Within\text{-}siblings}$	$\beta_{Non\text{-}adopted} - \beta_{Adopted}$	$\beta_{Non\text{-}transmitted}$

Genetic trios

- Genetic trios address gene-environment correlations by explicitly conditioning on parental genotypes:

$$Y_i = \tau PGI_i^{EA} + \alpha_p PGI_{p(i)}^{EA} + \alpha_m PGI_{m(i)}^{EA} + \xi_i$$

Genetic trios

- Genetic trios address gene-environment correlations by explicitly conditioning on parental genotypes:

$$Y_i = \tau PGI_i^{EA} + \alpha_p PGI_{p(i)}^{EA} + \alpha_m PGI_{m(i)}^{EA} + \xi_i$$

→ Provide valid estimates of τ.

→ Capture all children (incl. singleton children).

→ Very, very data demanding but can be emulated by imputation techniques (Young et al., 2022).

Outline

Causality, potential outcomes ... a quick recap

Identifying genetic effects
- Addressing gene-environment correlation
- Addressing measurement error

Identifying gene-environment interactions

An example: Genes and Schools

Measurement error in PGI

- Constructed PGI are noisy measures of the “true” PGI:
  $\rightarrow$ GWAS are based on finite samples.
  $\rightarrow$ GWAS may be based on different populations than the estimation sample.

Measurement error in PGI

- Constructed PGI are noisy measures of the “true” PGI:
  $\rightarrow$ GWAS are based on finite samples.
  $\rightarrow$ GWAS may be based on different populations than the estimation sample.

- PGIs are usually standardized on the estimation sample such that:

$$\frac{PGI_i^{EA,\text{true}} + \nu_i}{\sqrt{Var(PGI_i^{EA,\text{true}} + \nu_i)}} = \frac{PGI_i^{EA,\text{true}} + \nu_i}{\sigma_{PGI+\nu}}$$

Measurement error in PGI, cont'd

- What we would like to estimate:

$$Y_i = \tau \left(\frac{PGI_i^{EA,\text{true}}}{\sigma_{PGI}} \right) + \epsilon_i$$

$$\tau = \frac{Cov(Y_i, \frac{PGI_i^{EA,\text{true}}}{\sigma_{PGI}})}{Var(\frac{PGI_i^{EA,\text{true}}}{\sigma_{PGI}})} = \frac{Cov(Y_i, PGI_i^{EA,\text{true}})}{\sigma_{PGI}}$$

Measurement error in PGI, cont'd

- What we would like to estimate:

$$Y_i = \tau \left(\frac{PGI_i^{EA,\text{true}}}{\sigma_{PGI}} \right) + \epsilon_i$$

$$\tau = \frac{Cov(Y_i, \frac{PGI_i^{EA,\text{true}}}{\sigma_{PGI}})}{Var(\frac{PGI_i^{EA,\text{true}}}{\sigma_{PGI}})} = \frac{Cov(Y_i, PGI_i^{EA,\text{true}})}{\sigma_{PGI}}$$

- What we can estimate:

$$Y_i = \hat{\tau} \left(\frac{PGI_i^{EA,\text{true}} + \nu_i}{\sigma_{PGI+\nu}} \right) + \epsilon_i$$

$$\hat{\tau} = \frac{Cov(Y_i, \frac{PGI_i^{EA,\text{true}} + \nu_i}{\sigma_{PGI+\nu}})}{Var(\frac{PGI_i^{EA,\text{true}} + \nu_i}{\sigma_{PGI+\nu}})} = \frac{Cov(Y_i, PGI_i^{EA,\text{true}})}{\sigma_{PGI+\nu}} = \tau \times \underbrace{\frac{\sigma_{PGI}}{\sigma_{PGI+\nu}}}_{=\phi \text{ Attenuation factor}}$$

Obviously-related instrumental variables (ORIV)

- Relies on well-established method in the literature, e.g., Gillen et al. (2019).
- Use alternative (mismeasured) $PGI_i^{EA,\mathsf{IV}}$ to re-scale attenuated estimates via IV:

$$PGI_i^{EA,\mathsf{IV}} = \frac{PGI_i^{EA,\mathsf{true}} + \nu_i^{\mathsf{IV}}}{\sigma_{PGI+\nu}}$$

Obviously-related instrumental variables (ORIV)

- Relies on well-established method in the literature, e.g., Gillen et al. (2019).
- Use alternative (mismeasured) $PGI_i^{EA,\mathsf{IV}}$ to re-scale attenuated estimates via IV:

$$PGI_i^{EA,\mathsf{IV}} = \frac{PGI_i^{EA,\mathsf{true}} + \nu_i^{\mathsf{IV}}}{\sigma_{PGI+\nu}}$$

- This can be done by re-estimating the PGI weights in a different discovery sample, e.g., by splitting the original GWAS sample.

Obviously-related instrumental variables (ORIV), cont'd

- First stage:

$$\theta = \frac{Cov(\frac{PGI_i^{EA,\text{true}}+\nu_i^{\text{IV}}}{\sigma_{PGI+\nu}}\times X, \frac{PGI_i^{EA,\text{true}}+\nu_i^{\text{IV}}}{\sigma_{PGI+\nu}}\times X)}{Var(\frac{PGI_i^{EA,\text{true}}+\nu_i^{\text{IV}}}{\sigma_{PGI+\nu}}\times X)} = \underbrace{\frac{\sigma^2_{PGI}}{\sigma^2_{PGI+\nu}}}_{=\phi^2}$$

Obviously-related instrumental variables (ORIV), cont'd

- First stage:

$$\theta = \frac{Cov(\frac{PGI_i^{EA,\text{true}}+\nu_i^{\text{IV}}}{\sigma_{PGI+\nu}}\times X, \frac{PGI_i^{EA,\text{true}}+\nu_i^{\text{IV}}}{\sigma_{PGI+\nu}}\times X)}{Var(\frac{PGI_i^{EA,\text{true}}+\nu_i^{\text{IV}}}{\sigma_{PGI+\nu}}\times X)} = \underbrace{\frac{\sigma^2_{PGI}}{\sigma^2_{PGI+\nu}}}_{=\phi^2}$$

- Reduced form:

$$\kappa = \frac{Cov(Y_i, \frac{PGI_i^{EA,\text{true}}+\nu_i^{\text{IV}}}{\sigma_{PGI+\nu}}\times X)}{Var(\frac{PGI_i^{EA,\text{true}}+\nu_i^{\text{IV}}}{\sigma_{PGI+\nu}}\times X)} = \frac{Cov(Y_i, PGI_i^{EA,\text{true}})}{\sigma_{PGI+\nu}} \times \frac{1}{X} = \tau \times \frac{\sigma_{PGI}}{\sigma_{PGI+\nu}} \times \frac{1}{X}$$

Obviously-related instrumental variables (ORIV), cont'd

- First stage:

$$\theta = \frac{Cov(\frac{PGI_i^{EA,\text{true}}+\nu_i^{\text{IV}}}{\sigma_{PGI+\nu}}\times X, \frac{PGI_i^{EA,\text{true}}+\nu_i^{\text{IV}}}{\sigma_{PGI+\nu}}\times X)}{Var(\frac{PGI_i^{EA,\text{true}}+\nu_i^{\text{IV}}}{\sigma_{PGI+\nu}}\times X)} = \underbrace{\frac{\sigma^2_{PGI}}{\sigma^2_{PGI+\nu}}}_{=\phi^2}$$

- Reduced form:

$$\kappa = \frac{Cov(Y_i, \frac{PGI_i^{EA,\text{true}}+\nu_i^{\text{IV}}}{\sigma_{PGI+\nu}}\times X)}{Var(\frac{PGI_i^{EA,\text{true}}+\nu_i^{\text{IV}}}{\sigma_{PGI+\nu}}\times X)} = \frac{Cov(Y_i, PGI_i^{EA,\text{true}})}{\sigma_{PGI+\nu}} \times \frac{1}{X} = \tau \times \frac{\sigma_{PGI}}{\sigma_{PGI+\nu}} \times \frac{1}{X}$$

- Wald estimate:

$$\begin{aligned}\tau^{IV} = \frac{\kappa}{\theta} &= \tau \times \frac{\sigma_{PGI+\nu}}{\sigma_{PGI}} \times \frac{1}{X} \\ &= \tau \quad \text{if } X = \frac{\sigma_{PGI+\nu}}{\sigma_{PGI}} = \frac{1}{\sqrt{Corr(PGI_i^{EA,\text{true}}+\nu_i, PGI_i^{EA,\text{true}}+\nu_i^{\text{IV}})}}.\end{aligned}$$

Analytical correction

- First proposed by Becker et al. (2021), and recently extended by Sanz-de-Galdeano and Terskaya (forthcoming).
- Scales the estimated effects ex-post by the attenuation factor ϕ.
- The attenuation factor can be calculated from the estimation sample or by invoking prior knowledge from the literature.

Analytical correction, cont'd

- We know that the attenuation factor can be expressed as follows:

$$\phi = \frac{\sigma_{PGI}}{\sigma_{PGI+\nu}}$$

Analytical correction, cont'd

- We know that the attenuation factor can be expressed as follows:

$$\phi = \frac{\sigma_{PGI}}{\sigma_{PGI+\nu}}$$

- Expanding and re-arranging, we get:

$$\begin{aligned}
\phi^2 &= \frac{\sigma^2_{PGI}}{\sigma^2_{PGI+\nu}} \\
&= \frac{\sigma^2_{PGI} \times Var(Y_i) \times Cov(Y_i, PGI_i^{EA,\text{true}})^2}{\sigma^2_{PGI+\nu} \times Var(Y_i) \times Cov(Y_i, PGI_i^{EA,\text{true}})^2} \\
&= \frac{Cov(Y_i, PGI_i^{EA,\text{true}})^2 / [\sigma^2_{PGI+\nu} \times Var(Y_i)]}{Cov(Y_i, PGI_i^{EA,\text{true}})^2 / [\sigma^2_{PGI} \times Var(Y_i)]} \\
&= \frac{Cov(Y_i, PGI_i^{EA,\text{true}} + \nu)^2 / [\sigma^2_{PGI+\nu} \times Var(Y_i)]}{Cov(Y_i, PGI_i^{EA,\text{true}})^2 / [\sigma^2_{PGI} \times Var(Y_i)]} \\
&= \frac{R^2}{\hat{h^2}}, \text{ where } \hat{h^2} \text{ is an estimate of SNP heritability.}
\end{aligned}$$

ORIV vs. analytical correction

- (Dis)advantages of ORIV:
  - \+ Straightforward extension to within-family designs.
  - \+ Econometric properties, incl. standard errors well-understood.
  - \- Requires access to molecular data.
  - \- Loss of power in GWAS sample due to splitting.

ORIV vs. analytical correction

- (Dis)advantages of ORIV:
  - \+ Straightforward extension to within-family designs.
  - \+ Econometric properties, incl. standard errors well-understood.
  - \- Requires access to molecular data.
  - \- Loss of power in GWAS sample due to splitting.

- (Dis)dvantages of analytical correction:
  - \+ Straightforward and easy implmentation without access to molecular data.
  - \- Standard errors likely (upward) biased.
  - \- Harder to implement in within-family designs due to absence of estimates of SNP heritability in within-family GWAS.

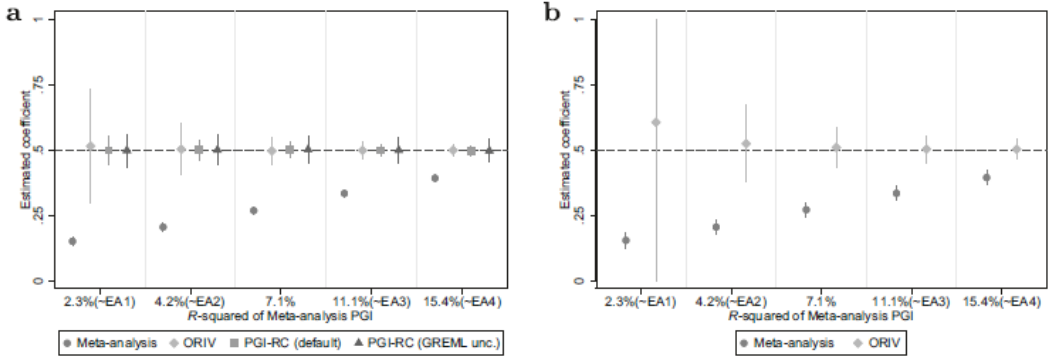
ORIV vs. analytical correction

- (Dis)advantages of ORIV:
  - \+ Straightforward extension to within-family designs.
  - \+ Econometric properties, incl. standard errors well-understood.
  - \- Requires access to molecular data.
  - \- Loss of power in GWAS sample due to splitting.

- (Dis)dvantages of analytical correction:
  - \+ Straightforward and easy implmentation without access to molecular data.
  - \- Standard errors likely (upward) biased.
  - \- Harder to implement in within-family designs due to absence of estimates of SNP heritability in within-family GWAS.

→ See Kippersluis et al. (2023) for a methodological comparison.

Research designs (Kippersluis et al., 2023)



Outline

Causality, potential outcomes ... a quick recap

Identifying genetic effects

Identifying gene-environment interactions

An example: Genes and Schools

Why should we care about G $\times$ E?

1. Test seminal theories on parental investments and skill formation.
   $\rightarrow$ Becker and Tomes (1979) and Cunha et al. (2010), and many others ...

Why should we care about G $\times$ E?

1. Test seminal theories on parental investments and skill formation.
   $\rightarrow$ Becker and Tomes (1979) and Cunha et al. (2010), and many others ...
2. Assess inequality-reducing/increasing effects of policy reforms.
   $\rightarrow$ Chetty et al. (2014), Clark and Royer (2013), and Jackson et al. (2024), and many others ...

What do we need?

- **Estimation**:

$$Y_i = \alpha PGI_i^{EA} + \beta E_i + \kappa(PGI_i^{EA} \times E_i) + \mathbf{X}_i\gamma + \epsilon_i$$

- **Identification**:

Requirement	Potential bias	Affected parameters	Solutions

What do we need?

- **Estimation**:

$$Y_i = \alpha PGI_i^{EA} + \beta E_i + \kappa(PGI_i^{EA} \times E_i) + \mathbf{X}_i\gamma + \epsilon_i$$

- **Identification**:

Requirement	Potential bias	Affected parameters	Solutions	
Exogenous PGI^{EA}	Gene-environment correlation	α, κ	Sibling design Adoption design Genetic trios	✓

What do we need?

- **Estimation**:

$$Y_i = \alpha PGI_i^{EA} + \beta E_i + \kappa(PGI_i^{EA} \times E_i) + \mathbf{X}_i\gamma + \epsilon_i$$

- **Identification**:

Requirement	Potential bias	Affected parameters	Solutions	
Exogenous PGI^{EA}	Gene-environment correlation	α, κ	Sibling design Adoption design Genetic trios	✓
Exogenous E	"Selection bias"	β, κ	Application-specific	✓

What do we need?

- **Estimation**:

$$Y_i = \alpha PGI_i^{EA} + \beta E_i + \kappa(PGI_i^{EA} \times E_i) + \mathbf{X}_i\gamma + \epsilon_i$$

- **Identification**:

Requirement	Potential bias	Affected parameters	Solutions	
Exogenous PGI^{EA}	Gene-environment correlation	α, κ	Sibling design Adoption design Genetic trios	✓
Exogenous E	"Selection bias"	β, κ	Application-specific	✓
Exogenous $PGI^{EA} \times E$	Spurious interaction terms	κ	Full interaction	?

The econometric argument

- **The “naive approach”**:

$$Y_i = \alpha PGI_i^{EA} + \beta E_i + \kappa(PGI_i^{EA} \times E_i) + \mathbf{X}_i\gamma + \epsilon_i$$

The econometric argument

- **The "naive approach"**:

$$Y_i = \alpha PGI_i^{EA} + \beta E_i + \kappa(PGI_i^{EA} \times E_i) + \mathbf{X}_i\gamma + \epsilon_i$$

- **The fully interacted model**:

$$\begin{aligned} Y_i = \alpha PGI_i^{EA} + \beta E_i + \hat{\kappa}(PGI_i^{EA} \times E_i) + \mathbf{X}_i\gamma \\ + (\mathbf{X}_i \times PGI_i^{EA})\gamma_{PGI^{EA}} + (\mathbf{X}_i \times E_i)\gamma_E + \hat{\epsilon}_i \end{aligned}$$

The econometric argument

- **The "naive approach"**:

$$Y_i = \alpha PGI_i^{EA} + \beta E_i + \kappa(PGI_i^{EA} \times E_i) + \mathbf{X}_i\gamma + \epsilon_i$$

- **The fully interacted model**:

$$\begin{aligned} Y_i = \alpha PGI_i^{EA} + \beta E_i + \hat{\kappa}(PGI_i^{EA} \times E_i) + \mathbf{X}_i\gamma \\ + (\mathbf{X}_i \times PGI_i^{EA})\gamma_{PGI^{EA}} + (\mathbf{X}_i \times E_i)\gamma_E + \hat{\epsilon}_i \end{aligned}$$

- **Under which conditions $\kappa = \hat{\kappa}$?**

The econometric argument

- **The "naive approach"**:

$$Y_i = \alpha PGI_i^{EA} + \beta E_i + \kappa(PGI_i^{EA} \times E_i) + \mathbf{X}_i\gamma + \epsilon_i$$

- **The fully interacted model**:

$$\begin{aligned} Y_i = \alpha PGI_i^{EA} + \beta E_i + \hat{\kappa}(PGI_i^{EA} \times E_i) + \mathbf{X}_i\gamma \\ + (\mathbf{X}_i \times PGI_i^{EA})\gamma_{PGI^{EA}} + (\mathbf{X}_i \times E_i)\gamma_E + \hat{\epsilon}_i \end{aligned}$$

- **Under which conditions $\kappa = \hat{\kappa}$?**
  - $Cov(PGI_i^{EA} \times E_i, \mathbf{X}_i \times PGI_i^{EA}) = Cov(PGI_i^{EA} \times E_i, \mathbf{X}_i \times E_i) = 0$
    $\rightarrow$ Very unlikely since $Cov(PGI_i^{EA}, \mathbf{X}_i) \neq 0$ or $Cov(E_i, \mathbf{X}_i) \neq 0$.

The econometric argument

- **The "naive approach"**:

$$Y_i = \alpha PGI_i^{EA} + \beta E_i + \kappa(PGI_i^{EA} \times E_i) + \mathbf{X}_i\gamma + \epsilon_i$$

- **The fully interacted model**:

$$\begin{aligned} Y_i = \alpha PGI_i^{EA} + \beta E_i + \hat{\kappa}(PGI_i^{EA} \times E_i) + \mathbf{X}_i\gamma \\ + (\mathbf{X}_i \times PGI_i^{EA})\gamma_{PGI^{EA}} + (\mathbf{X}_i \times E_i)\gamma_E + \hat{\epsilon}_i \end{aligned}$$

- **Under which conditions $\kappa = \hat{\kappa}$?**
  - $Cov(PGI_i^{EA} \times E_i, \mathbf{X}_i \times PGI_i^{EA}) = Cov(PGI_i^{EA} \times E_i, \mathbf{X}_i \times E_i) = 0$
    $\rightarrow$ Very unlikely since $Cov(PGI_i^{EA}, \mathbf{X}_i) \neq 0$ or $Cov(E_i, \mathbf{X}_i) \neq 0$.
  - $Cov(Y_i, \mathbf{X}_i \times PGI_i^{EA}) = Cov(Y_i, \mathbf{X}_i \times E_i) = 0$
    $\rightarrow$ Needs to be tested empirically.

The pros and cons of full interaction

- Fully interacted models may be necessary to estimate heterogeneous treatment effects (Feigenberg et al., forthcoming; Keller, 2014).

The pros and cons of full interaction

- Fully interacted models may be necessary to estimate heterogeneous treatment effects (Feigenberg et al., forthcoming; Keller, 2014).

- However, researcher face a variance-bias trade-off. Standard errors of κ may increase substantially if
  1. The loss in degrees of freedom is large,
  2. The increase in R^2 is small,
  3. The collinearity of $PGI_i^{EA} \times E_i$ and the additional interaction terms is large.

- The strength of the variance-bias trade-off depends on the specific application.

Additional considerations for gene-environment studies

- $G \times E$ studies need to be powered adequately.
  $\rightarrow$ $G \times E$ are usually 2-3 times smaller than main effects.

- $G \times E$ studies need to defend functional form assumptions.
  $\rightarrow$ Ex ante it is unlclear that $G \times E$ should only operate via linear interaction effects.

$\rightarrow$ See Biroli et al. (2022) for an excellent review of the current state of the $G \times E$ literature.

Outline

Causality, potential outcomes ... a quick recap

Identifying genetic effects

Identifying gene-environment interactions

An example: Genes and Schools

The genetic lottery goes to school: evidence from Norway

Nicolai Borgen, Rosa Cheesman, Paul Hufe & Astrid Sandsor

- Education is a core determinant of life outcomes (Acemoglu and Autor, 2011; Hanushek and Woessmann, 2008; Krueger and Lindahl, 2001).

- Equity of education systems as a central policy goal:

  Most fundamental, of course, is the question of how well schools reduce the inequity of birth by providing children an equitable foundation of mental skills and knowledge [...].

  Coleman Report, p.36

- Education is a core determinant of life outcomes (Acemoglu and Autor, 2011; Hanushek and Woessmann, 2008; Krueger and Lindahl, 2001).

- Equity of education systems as a central policy goal:

  Most fundamental, of course, is the question of how well schools reduce the inequity of birth by providing children an equitable foundation of mental skills and knowledge [...].
  Coleman Report, p.36

- Effective education policies require understanding of the production function:

$$Y = f(\underbrace{G}_{\text{Nature}}, \underbrace{I^F, I^S}_{\text{Nurture}}).$$

This paper in a nutshell

Research question

Do better schools increase or decrease the effect of genes on educational attainment?

This paper in a nutshell

Research question

Do better schools increase or decrease the effect of genes on educational attainment?

- **Empirical approach**
  - We use the universe of Norwegian students in grades 8-9 to measure school value added ($N \approx 1,300$ schools).
  - We link the VA measures to a sample of genotyped trios (children, mothers, fathers) ($N \approx 32,000$ families).
  - We use exogenous variation in PGI^{EA} and school VA to causally estimate $G \times E$ for reading and numeracy test scores in grade 9.

This paper in a nutshell

Research question

Do better schools increase or decrease the effect of genes on educational attainment?

- **Empirical approach**
  - We use the universe of Norwegian students in grades 8-9 to measure school value added ($N \approx 1,300$ schools).
  - We link the VA measures to a sample of genotyped trios (children, mothers, fathers) ($N \approx 32,000$ families).
  - We use exogenous variation in PGI^{EA} and school VA to causally estimate $G \times E$ for reading and numeracy test scores in grade 9.

- **Findings**
  - We find causal evidence for substitutability of PGI^{EA} and school quality in reading (but not numeracy):

  $\rightarrow$ 1 SD increase of school quality decreases the impact of PGI^{EA} on reading test scores by 4%.

  - Substitutability arises through gains of students with lower PGI^{EA}.

Data sources

- MoBa:
  - Initial information for a sample of mothers ($N > 114,000$) from 1999-2008.
  - 44,017 genotyped father-mother-child trios.
  - Linked to Norwegian register data.
  - We restrict the sample to birth cohorts 2002-2008 and students of European descent.
  - Effective sample size $N \approx 32,000$.

- Norwegian registers:
  - Population of students in Norway ($N \approx 60,000$ per cohort).
  - Information on standardized tests in reading and numeracy in grades 5, 8, and 9.
  - We restrict the sample to birth cohorts 1997-2007.
  - Effective sample size $N \approx 670,000$.

Data inputs

- Recall our estimation model:

$$Y_i = \alpha PGI_i^{EA} + \beta Q_i^S + \kappa(PGI_i^{EA} \times Q_i^S) + \mathbf{X}_i\gamma + \epsilon_i$$

▸ Educational outcomes Y_i ▸ Genetic endowments PGI^{EA} ▸ School quality Q^S ▸ Controls $\mathbf{X}_i$ (a)

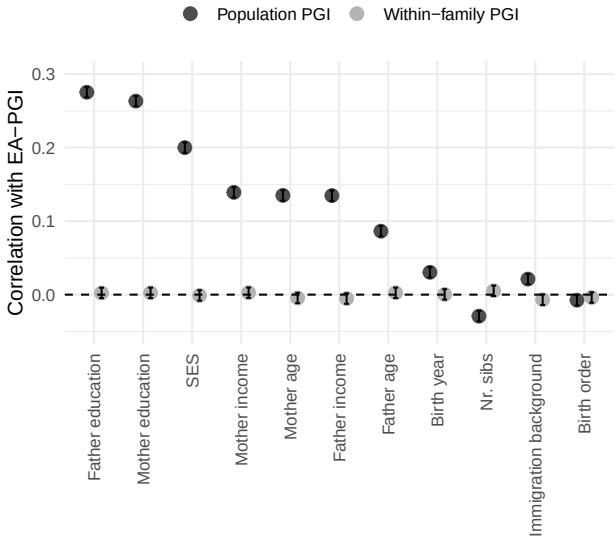
Summary statistics

	Mean	SD	Min	Max
a) Child characteristics				
Sex	0.50	0.50	0.00	1.00
Parity	1.70	0.80	1.00	13.00
Migration status	0.10	0.30	0.00	1.00
Birth year	2004.90	1.60	2002.00	2008.00
b) Parental characteristics				
PGI (Mother)	0.00	1.00	-4.30	4.10
Education in years (Mother)	15.10	2.30	9.00	21.00
PGI (Father)	0.00	1.00	-4.30	3.90
Education in years (Father)	14.60	2.60	7.00	21.00
c) Treatment variables				
PGI	0.00	1.00	-3.80	3.70
School VA	0.00	1.00	-4.10	4.50

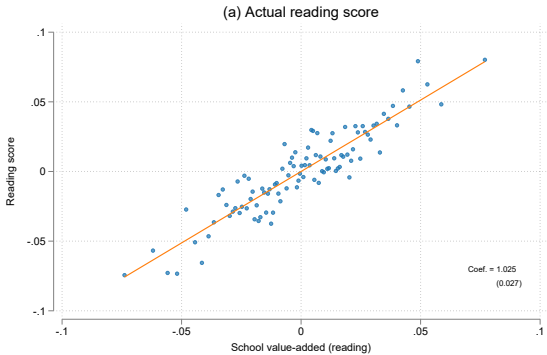
Recap on identifying assumptions

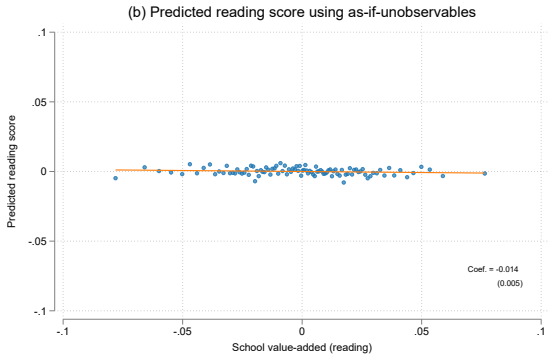
- ✓ No gene-environment correlations (α, κ).
- ✓ No selection into schools (β, κ).
- ✓ No spurious interaction effects (κ).

Identification of genetic effects

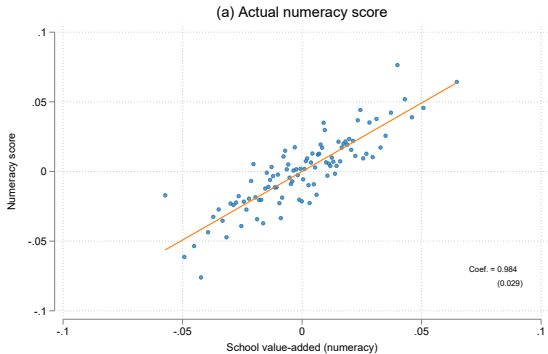


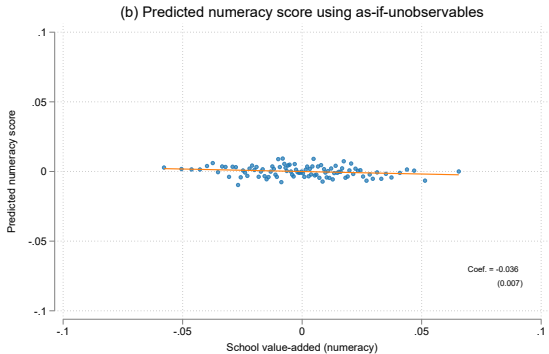
Identification of school effects (Reading)





Identification of school effects (Numeracy)





Gene-environment interaction (Reading)

Outcome: Reading (Grade 9)	(1)
PGI^{EA}	0.302***
	(0.006)
Q^S (Reading)	0.064***
	(0.009)
$PGI^{EA} \times Q^S$ (Reading)	-0.014**
	(0.005)
Parental PGI	×
Genotyping controls	×
Child controls	×
School controls	×
Saturation controls	×
N	$32,262$

Note: Own calculations. Standard errors (in parentheses) are clustered at the family level. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Gene-environment interaction (Reading)

Outcome: Reading (Grade 9)	(1)	(2)
PGI^{EA}	0.302***	0.227***
	(0.006)	(0.008)
Q^S (Reading)	0.064***	0.063***
	(0.009)	(0.009)
$PGI^{EA} \times Q^S$ (Reading)	-0.014**	-0.014**
	(0.005)	(0.005)
Parental PGI	×	✓
Genotyping controls	×	✓
Child controls	×	×
School controls	×	×
Saturation controls	×	×
N	$32,262$	$32,262$

Note: Own calculations. Standard errors (in parentheses) are clustered at the family level. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Gene-environment interaction (Reading)

Outcome: Reading (Grade 9)	(1)	(2)	(3)
PGI^{EA}	0.302***	0.227***	0.231***
	(0.006)	(0.008)	(0.005)
Q^{S} (Reading)	0.064***	0.063***	0.036***
	(0.009)	(0.009)	(0.005)
$PGI^{EA} \times Q^{S}$ (Reading)	-0.014**	-0.014**	-0.009**
	(0.005)	(0.005)	(0.003)
Parental PGI	×	✓	✓
Genotyping controls	×	✓	✓
Child controls	×	×	✓
School controls	×	×	✓
Saturation controls	×	×	×
N	$32,262$	$32,262$	$32,262$

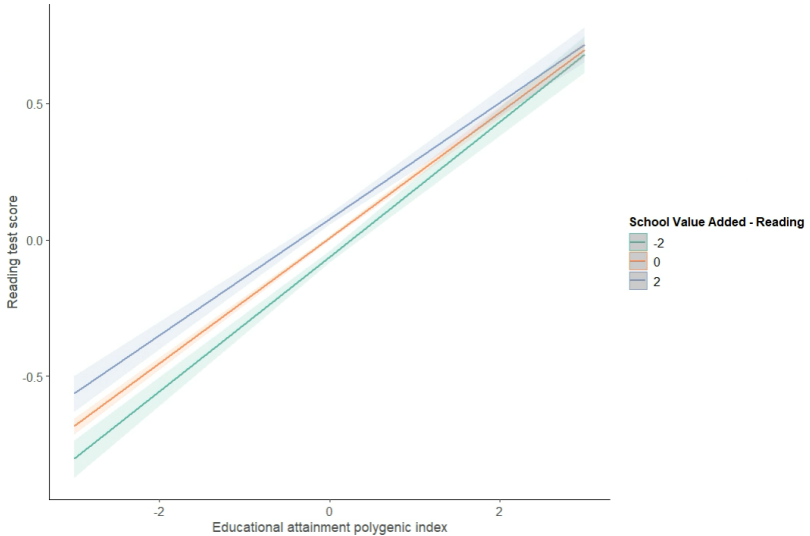
Note: Own calculations. Standard errors (in parentheses) are clustered at the family level. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Gene-environment interaction (Reading)

Outcome: Reading (Grade 9)	(1)	(2)	(3)	(4)
PGI^{EA}	0.302***	0.227***	0.231***	0.230***
	(0.006)	(0.008)	(0.005)	(0.005)
Q^{S} (Reading)	0.064***	0.063***	0.036***	0.034***
	(0.009)	(0.009)	(0.005)	(0.005)
$PGI^{EA} \times Q^{S}$ (Reading)	-0.014**	-0.014**	-0.009**	-0.008
	(0.005)	(0.005)	(0.003)	(0.005)
Parental PGI	×	✓	✓	✓
Genotyping controls	×	✓	✓	✓
Child controls	×	×	✓	✓
School controls	×	×	✓	✓
Saturation controls	×	×	×	✓
N	$32,262$	$32,262$	$32,262$	$32,262$

Note: Own calculations. Standard errors (in parentheses) are clustered at the family level. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Gene-environment interaction (Reading)

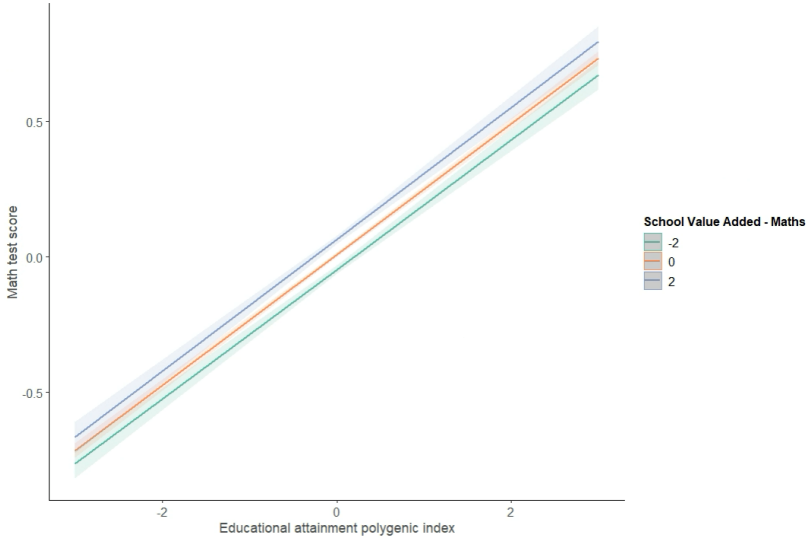


Gene-environment interaction (Numeracy)

Outcome: Reading (Grade 9)	(1)	(2)	(3)	(4)
PGI^{EA}	0.315***	0.237***	0.241***	0.241***
	(0.005)	(0.008)	(0.004)	(0.004)
Q^S (Numeracy)	0.056***	0.055***	0.027***	0.028***
	(0.010)	(0.010)	(0.003)	(0.003)
$PGI^{EA} \times Q^S$ (Numeracy)	-0.005	-0.005	-0.001	-0.001
	(0.005)	(0.005)	(0.003)	(0.004)
Parental PGI	×	✓	✓	✓
Genotyping controls	×	✓	✓	✓
Child controls	×	×	✓	✓
School controls	×	×	✓	✓
Saturation controls	×	×	×	✓
N	$32,262$	$32,262$	$32,262$	$32,262$

Note: Own calculations. Standard errors (in parentheses) are clustered at the family level. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Gene-environment interaction (Numeracy)



Contextualizing effect sizes

- Estimates pertain to a low inequality country. ▸ Inequality in VA
  - Assuming cross-country portability of effects, substitutability would be 10% for grade 9 in Chicago high schools.

- Estimates pertain to one year of schooling.
  - Assuming linear additive effects, substitutability increases to 12% over the course of lower secondary school (grades 8-10) in Norway.

- Estimates can be compared to substituability in other dimensions of advantage:
  - Latent family SES ($\Delta 1SD$): 2.87% (Jackson et al., 2024).

Take-aways

- Genetic effects and gene-environment interactions relate to fundamental questions in research on socioeconomic inequality.
- Causal studies require careful identification strategies (and excellent data) to avoid bias.
- To date, causal studies are constrained by data availability.
- Proliferation of new genetic data will lift current constraints and open new avenues for research on socioeconomic inequality.

Thank you for your attention! Questions?

✉ paul.hufe@bristol.ac.uk

www.paulhufe.net

paulhufe

References I

Acemoglu, D. and D. Autor (2011). “Skills, Tasks and Technologies: Implications for Employment and Earnings”. Ed. by D. Card and O. Ashenfelter. Vol. 4. Handbook of Labor Economics. Elsevier, pp. 1043–1171.

Ainsworth, R., R. Dehejia, C. Pop-Eleches, and M. Urquiola (2023). “Why Do Households Leave School Value Added on the Table? The Roles of Information and Preferences”. *American Economic Review* 113 (4), pp. 1049–82.

Angrist, J., P. Hull, P. A. Pathak, and C. Walters (2024). “Credible School Value-Added with Undersubscribed School Lotteries”. *Review of Economics and Statistics* 106 (1), pp. 1–19.

Angrist, J., P. Hull, and C. Walters (2023). “Methods for measuring school effectiveness”. Ed. by E. A. Hanushek, S. Machin, and L. Woessmann. Vol. 7. Handbook of the Economics of Education. Elsevier. Chap. 1, pp. 1–60.

Arneson, R. J. (2018). “Four Conceptions of equal opportunity”. *Economic Journal* 128 (612), F152–F173.

Becker, G. S. and N. Tomes (1979). “An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility”. *Journal of Political Economy* 87 (6), pp. 1153–1189.

Becker, J. et al. (2021). “Resource profile and user guide of the Polygenic Index Repository”. *Nature Human Behaviour* 5 (12), pp. 1744–1758.

Biroli, P., T. J. Galama, S. von Hinke, H. van Kippersluis, C. A. Rietveld, and K. Thom (2022). “The Economics and Econometrics of Gene-Environment Interplay”. *arXiv* 2203.00729.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). “Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates”. *American Economic Review* 104 (9), pp. 2593–2632.

Clark, D. and H. Royer (2013). “The Effect of Education on Adult Mortality and Health: Evidence from Britain”. *American Economic Review* 103 (6), pp. 2087–2120.

Coleman, J. S., E. Campbell, C. Hobson, J. McPartland, A. Mood, F. Weinfeld, and R. York (1966). *Equality of Educational Opportunity*. Washington D.C.: National Center for Educational Statistics.

References II

Cunha, F., J. J. Heckman, and S. M. Schennach (2010). “Estimating the Technology of Cognitive and Noncognitive Skill Formation”. *Econometrica* 78 (3), pp. 883–931.

Demange, P. A. et al. (2022). “Estimating effects of parents’ cognitive and non-cognitive skills on offspring education using polygenic scores”. *Nature Communications* 13 (1), p. 4801.

Feigenberg, B., B. Ost, and J. A. Qureshi (forthcoming). “Omitted Variable Bias in Interacted Models: A Cautionary Tale”. *Review of Economics and Statistics*.

Gillen, B., E. Snowberg, and L. Yariv (2019). “Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study”. *Journal of Political Economy* 127 (4), pp. 1826–1863.

Hanushek, E. A. and L. Woessmann (2008). “The Role of Cognitive Skills in Economic Development”. *Journal of Economic Literature* 46 (3), pp. 607–668.

Jackson, C. K. (2013). “Match quality, worker productivity, and worker mobility: direct evidence from teachers”. *Review of Economics and Statistics* 95 (4), pp. 1096–1116.

Jackson, C. K., S. C. Porter, J. Q. Easton, and S. Kiguel (2024). “Who Benefits From Attending Effective High Schools?” *Journal of Labor Economics* 42 (3), pp. 717–751.

Jackson, C. K., S. C. Porter, J. Q. Easton, A. Blanchard, and S. Kiguel (2020). “School Effects on Socioemotional Development, School-Based Arrests, and Educational Attainment”. *American Economic Review: Insights* 2 (4), pp. 491–508.

Keller, M. C. (2014). “Gene × Environment Interaction Studies Have Not Properly Controlled for Potential Confounders: The Problem and the (Simple) Solution”. *Biological Psychiatry* 75 (1). Temperament: Genetic and Environmental Factors, pp. 18–24.

Kippersluis, H. van, P. Biroli, R. Dias Pereira, T. J. Galama, S. von Hinke, S. F. W. Meddens, D. Muslimova, E. A. W. Slob, R. de Vlaming, and C. A. Rietveld (2023). “Overcoming attenuation bias in regressions using polygenic indices”. *Nature Communications* 14 (1), p. 4473.

References III

Kirkebøen, L. J. (2022). “School Value-Added and Long-Term Student Outcomes”. *CESifo Working Paper* 9769.

Krueger, A. B. and M. Lindahl (2001). “Education for Growth: Why and for Whom?” *Journal of Economic Literature* 39 (4), pp. 1101–1136.

Okbay, A. 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), pp. 437–449.

Sanz-de-Galdeano, A. and A. Terskaya (forthcoming). “Sibling Differences in Educational Polygenic Scores: How Do Parents React?” *Review of Economics and Statistics*.

Young, A. I., S. M. Nehzati, S. Benonisdottir, A. Okbay, H. Jayashankar, C. Lee, D. Cesarini, D. J. Benjamin, P. Turley, and A. Kong (2022). “Mendelian imputation of parental genotypes improves estimates of direct genetic effects”. *Nature Genetics* 54 (6), pp. 897–905.

Back

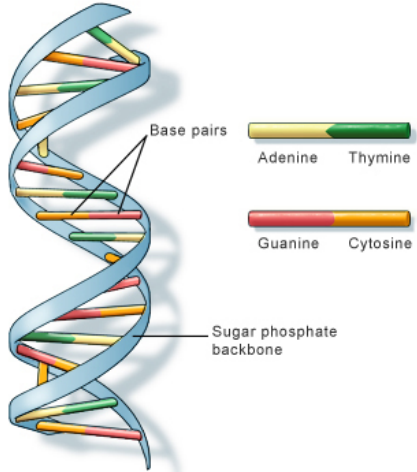
Educational outcomes Y_i

Standardized national tests in reading and numeracy (grade 9)

- Low stakes
  - $\rightarrow$ Communicated to parents and teachers but mostly used to track student development.
- Computer corrected
  - $\rightarrow$ Not affected by teacher biases.
- Taken at beginning of the school year
  - $\rightarrow$ Measure skills accumulated until grade 9.
- Same test as in grade 8
  - $\rightarrow$ Allow mapping for VA calculation.
- Highly predictive of later life-outcomes
  - $\rightarrow$ 1 SD $\uparrow$ in numeracy, increases high school graduation at age 21 by 9.5 p.p.

Genetic endowments PGI^{EA}

- We use the polygenic index (PGI) for educational attainment from Okbay et al. (2022):
  - Discovery sample of 3 mn people of European descent.
  - Explains 16% of variation in years of education.
  - $\approx$ 56% of explanatory power due to direct genetic effects.



Back

School quality Q^S

1. We construct school VA for reading and numeracy in grade 8 (Angrist et al., 2023).

2. We model educational outcomes Y of student i attending school j in cohort c for subject d:

$$Y^d_{ijc} = \beta^d Z_{ijc} + \underbrace{Q^d_{jc} + \epsilon^d_{ijc}}_{=e^d_{ijc}}$$

3. We estimate school effects in subject d by averaging over residuals in school-cohort cells:

$$Q^d_{jc} = \sum e^d_{ijc}/N_{jc}$$

4. We apply the Bayesian Shrinkage estimator à la Chetty et al. (2014).

5. Highly predictive of later life-outcomes
   → 1 SD ↑ in VA, increases years of schooling by 0.5-0.8 years (Kirkebøen, 2022).

Back

Controls $\mathbf{X}_i(a)$

Child controls

- Lagged test scores in numeracy, reading, English
- Parental years of education
- Migration status
- Age of arrival in Norway
- # of siblings
- Gender
- Year of birth
- Birth order

School controls

- School-cohort averages of all child background variables

Parental PGI

- PGI^{EA} mother
- PGI^{EA} father

Genotyping controls

- Genotyping center
- Genotyping batch
- Genotyping plate
- Imputation batch

Saturation controls

- Interaction of child background controls, school controls, and parental PGIs with PGI^{EA} and Q^S

Back

Inequality in VA