Gene-environment interactions with essential heterogeneity *

Johannes Hollenbach
RWI &
Paderborn University

Hendrik Schmitz
Paderborn University &
RWI

Matthias Westphal
University of Hagen &
RWI

September 2024

Abstract

We show that two-stage least squares (2SLS) estimates of interactions can be misleading in settings with essential heterogeneity (e.g., selection into gains) and where complier status to the instrument depends on the interaction variable. The 2SLS estimator cannot disentangle interaction effects from shifts in complier groups. Estimating marginal treatment effects addresses this problem by fixing the underlying population and unobserved heterogeneity. We illustrate this using the example of gene-environment studies, where the central parameter is the interaction effect between an endogenous, instrumented measure of environment or behavior and a predetermined measure of genetic endowment. Our application examines the effect of education on cognitive performance in old age. The results show complementarities between education and genetic predisposition in determining cognitive abilities. The marginal treatment effect estimates reveal a substantially larger gene-environment interaction, exceeding the 2SLS estimate by a factor of at least 2.5.

JEL Classification: *C31, J14, J24*
Keywords: Two-stage least squares estimation, marginal treatment effects, gene-environment interactions, cognitive decline

---

Johannes Hollenbach: RWI - Leibniz Institute for Economic Research, Hohenzollernstr. 1-3, 45128 Essen, Germany (johannes.hollenbach@rwi-essen.de); Hendrik Schmitz: Paderborn University, Warburger Str. 100, 33098 Paderborn, Germany (hendrik.schmitz@uni-paderborn.de); Matthias Westphal: FernUniversität in Hagen, Universitätsstraße 47, 58097 Hagen, Germany (matthias.westphal@fernuni-hagen.de).

*We thank Pietro Biroli for insightful comments and suggestions, Kristina Strohmaier, Martin Fischer and Souvik Banerjee for thoughtful discussions, the participants of the Initiative in Social Genomics research group meetings at UW-Madison, the CINCH-dggö Academy in Health Economics, the EuHEA PhD conferences in Bologna and Lucerne as well as the Brown Bag Seminar in Wuppertal for valuable comments. Financial support from Deutsche Forschungsgemeinschaft (DFG, project number 437564156) is gratefully acknowledged. This paper uses data from the English Longitudinal Study of Ageing (ELSA). ELSA is funded by the National Institute on Aging (R01AG017644), and by UK Government Departments coordinated by the National Institute for Health and Care Research (NIHR).

1 Introduction

The recent availability of genetic data has revived the old debate in the social sciences about nature versus nurture in determining success over the life course (see, e.g., Behrman and Taubman, 1989; Plug and Vijverberg, 2003; Björklund and Salvanes, 2011). The focus is on estimating gene-environment interactions to assess how the effects of environmental exposures or individual decisions vary by genetic endowment. These interaction models typically take the form of

$$Y_i = \beta_0 + \beta_1 E_i + \beta_2 G_i + \beta_3 G_i \times E_i + X'\gamma + \varepsilon_i, \tag{1}$$

where Y_i measures (long-run) outcomes, E_i indicates an endogenous environmental exposure or individual decision, G_i denotes pre-determined genetic endowment, and X is a vector of controls. Recent studies revolve around the causal identification of β_1, β_2, and β_3, for example, by instrumenting E_i and $G_i \times E_i$ or by removing factors correlated with the environment from G_i (see, e.g., Schmitz and Conley, 2017; Barcellos et al., 2018; Biroli et al., 2022; Pereira et al., 2022; Barcellos et al., 2021).

Our paper focuses on how the central parameter in the gene-environment literature, the interaction coefficient β_3, is estimated. In its intended interpretation, it measures how the causal effect of the environment varies with genetic endowment, all else being equal. However, as we show here, the widely used two-stage least squares (2SLS) approach may not provide a reliable estimate of this effect, even with a valid instrumental variable. This is the case when two conditions hold simultaneously. First, compliers to the instrument for E_i have different unobserved characteristics between different values of G_i. Second, the (individual) treatment effects of E_i exhibit essential heterogeneity. This occurs when the propensity to take the treatment correlates with the unobserved effect heterogeneity (Heckman et al., 2006). A prominent example of essential heterogeneity is self-selection into treatment based on unobserved gains. These conditions frequently occur in real-world settings that are investigated with causal methods. As a result, 2SLS conflates two different changes when estimating the $G \times E$ coefficient: first, how the local average treatment effect (LATE) of E_i on Y_i changes with G_i, which is the interaction we aim to estimate. Second, how the complier subpopulation of this LATE shifts as G_i varies.

In this paper, we (1) comprehensively describe the problem, (2) propose a solution, and (3) apply it to a real-world setting. Using a numerical example, we show that relying on 2SLS estimates of β_3 to provide evidence on how genes and the environment interact can be misleading in a setting with essential heterogeneity and a substantial gradient in the first-stage coefficients across different G_i. In our simulation example, the 2SLS coefficient even has the opposite sign of the actual interaction effect. We propose a solution that maintains a fixed underlying population when comparing the effect of E_i on Y_i for different values

of G_i. Estimating marginal treatment effects (MTEs) offers a suitable approach to achieve this (Heckman and Vytlacil, 2005). Finally, we apply this method to the long-term effect of education E_i on cognitive abilities Y_i using data from English Longitudinal Study of Aging (ELSA). We select our sample around the pivotal cohort of a compulsory schooling reform, which extended the minimum school-leaving age from 14 to 15 for individuals born after 1933. Cognitive ability is measured by the word recall test in six waves between 2002 and 2012 when individuals in our sample were between 65 and 80 years old. To measure genetic endowment, we use a polygenic score (PGS), a summary measure that predicts educational attainment based on genetic makeup. When estimating MTEs, we rely on a recently developed partial identification method by Mogstad et al. (2018), also used by Rose and Shem-Tov (2021).

The problem we describe is not limited to the gene-environment literature and is, in principle, relevant to any interaction effects between an endogenous (and instrumented) and a pre-determined variable. However, gene-environment applications are a natural choice to illustrate it since the target parameter is the interaction coefficient. Our contribution is mainly methodological, but we also contribute substantively to the literature on gene-environment interactions – a highly dynamic research field with many recent papers in areas related to ours. Possible problems with 2SLS estimation of the gene-environment interaction effects are also briefly mentioned in Barcellos et al. (2021). They find differences in returns to schooling between individuals with different genetic endowments in terms of socioeconomic status and use a linear MTE estimation to check whether unobservable factors can explain these differences, which is not the case in their study. We are unaware of any study estimating the causal effects of education and its interaction with genetic makeup on cognition in later life. Ding et al. (2019) study the relationship between genes/educational attainment and cognition using data from the Health and Retirement Study (HRS) but do not use exogenous variation in education. Anderson et al. (2020) estimate a positive bi-directional relationship between educational attainment and intelligence using genetic variants as instruments. Schmitz and Conley (2017) study whether the effect of the Vietnam War draft lottery on schooling outcomes differs by a genetic predisposition for education. Going beyond educational outcomes, Barcellos et al. (2018) estimate whether genetic predisposition to obesity moderates the effect of education on health using the UK compulsory schooling reform for the 1957 birth cohort as an exogenous variation but a different data set and outcomes (health). Besides Schmitz and Conley (2017) and Barcellos et al. (2018), the earliest study in economics on how education can compensate for the effects of genetic differences is, probably, Papageorge and Thom (2020), who study the impact on labor market outcomes.

Our results are as follows: Applying a benchmark 2SLS estimator, we find a zero effect of education on cognition for individuals in the lowest quintile of G_i, that is, those with the lowest genetic propensity for education. On average, moving to a higher quintile of G_i

goes along with an increase in the effect of E_i on Y_i by an insignificant 0.16 words correctly recalled. Using marginal treatment effects, we still find a zero effect of E_i in the lowest quintile of G_i. However, the interaction effect is much larger. Here, moving from one quintile of G_i to a higher one increases the effect of E_i on Y_i by 0.39–0.46 words, on average. This corresponds to roughly 10–15 percent of the standard deviation of the outcome variable. While education does not improve cognitive abilties in the group with the lowest genetic endowment, it increases word recall by 1.6–1.8 words in the highest quintile of G_i. 2SLS would considerably underestimate the gene-environment complementarity. In our application, genetic endowment is correlated with the complier status: The share of compliers to the education reform is highest in the lowest quintile of G_i (65 percent) and monotonically decreases to 35 percent in the highest quintile. Moreover, there is evidence of selection into gains. Overall, the 1947 UK compulsory schooling reform we use has increased schooling, especially for those with lower genetic propensity for schooling (first stage results). However, there are no returns to schooling in terms of cognitive abilities for these individuals. Instead, the large returns are seen for those with a higher genetic propensity.

The paper proceeds as follows: Section 2 outlines the challenges in identifying the gene-environment interplay from an econometric perspective and presents our suggested solution. Section 3 describes the institutional setting of our application and the data used. Section 4 presents 2SLS estimates of gene-environment interactions in our application and tests if the necessary conditions that interfere with the interpretation of 2SLS estimates apply. Section 5 gives an overview of the partial identification approach to estimate MTEs and presents our main results. Section 6 concludes.

2 Potential identification problems of interaction effects

2.1 The problem

We are interested in the effect of a particular environment or life decision (here, education), E_i, on an outcome Y_i (here, old-age cognitive abilities) and how this effect interacts with genetic endowment G_i. For simplicity, first assume that E_i and G_i are binary variables and that the potential outcomes of individual i are defined by the following functions: $Y_i^{jg} = \mu^{jg}(X_i) + \varepsilon_i^{jg}$, $j \in \{0,1\}$, $g \in \{0,1\}$, where j denotes potential outcomes for education status and G_i for the genetic endowment, $\mu^{jg}(X_i)$ is a function of observable characteristics and ε_i^{jg} is an unobservable part. Each individual has four potential outcomes, e.g., Y_i^{10}

with a high educational level ($E_i = 1$) and a low genetic propensity ($G_i = 0$). However, only one of the four is realized and observed by the researcher. The observation rule is

$$\begin{aligned} Y_i &= E_i \cdot G_i \cdot Y_i^{11} + E_i \cdot (1 - G_i) \cdot Y_i^{10} + (1 - E_i) \cdot G_i \cdot Y_i^{01} + (1 - E_i) \cdot (1 - G_i) \cdot Y_i^{00} \\ &= Y_i^{00} + (Y_i^{10} - Y_i^{00})E_i + (Y_i^{01} - Y_i^{00})G_i + (Y_i^{11} - Y_i^{01} - (Y_i^{10} - Y_i^{00}))E_i \cdot G_i \end{aligned}$$

The second equality represents the link to Eq. (1), the workhorse estimating equation used in the literature. The gene-environment interaction effect is given by $Y_i^{11} - Y_i^{01} - (Y_i^{10} - Y_i^{00})$, that is, the difference in the effect of E_i on Y_i when $G_i = 1$ (which is $Y_i^{11} - Y_i^{01}$) and the effect of E_i on Y_i when $G_i = 0$ (which is $Y_i^{10} - Y_i^{00}$).

Assume that G_i is pre-determined while E_i is a choice variable and, therefore, endogenous.[1] We model the choice E_i in a generalized Roy framework (Roy, 1951), where individuals choose E_i if the (expected) returns to education exceed monetary and/or non-monetary costs $C_i = \mu_C(X_i, G_i, Z_i) + U_{C,i}$. Costs depend on G_i, the observable characteristics X_i, and an instrumental variable Z_i, in our case, being born in or after the pivotal cohort of an education reform. Note that Z_i does not directly affect Y_i^{jg}. C_i also includes an unobservable term $U_{C,i}$. The decision rule for E_i (depending on the realization of $G_i = g$) reads:

$$\begin{aligned} E_i(G_i = g) = 1 \quad &\Leftrightarrow \quad Y_i^{1g} - Y_i^{0g} > C_i \\ &\Leftrightarrow \quad \mu^{1g}(X_i) - \mu^{0g}(X_i) - \mu_C^g(X_i, Z_i) > -(\varepsilon_i^{1g} - \varepsilon_i^{0g} - U_{C,i}^g) \\ &\Leftrightarrow \quad \mu_E^g(X_i, Z_i) > V_i^g \end{aligned}$$

While not necessary for any theoretical result, $\mu_E^g(X_i, Z_i) = \mu^{1g}(X_i) - \mu^{0g}(X_i) - \mu_C^g(X_i, Z_i)$ can be represented as a linear index, such as:

$$\mu_E^g(X_i, Z_i) = \pi_0 + \pi_1 G_i + \pi_2 Z_i + \pi_3 Z_i \times G_i + \pi X_i + V_i^g$$

where $V_i^g = -(\varepsilon_i^{1g} - \varepsilon_i^{0g} - U_i^C)$ is the unobservable term. The decision rule implies that E_i correlates with ε_i^{1g} and ε_i^{0g} (and V_i^g), which renders E_i endogenous.

As a common approach to solve the endogeneity problem and to estimate, among other parameters, β_3 from Eq. (1), researchers usually use Z_i and $Z_i \times G_i$ as instrumental variables in two-stage least squares regressions for the endogenous variables E_i and $E_i \times G_i$. We set up a simple simulation model to visualize potential problems with this approach. Assume

[1] The extension of our framework to an endogenous G_i entails the same kind of problems. Our proposed solution applies to this case but is not straightforward in applications as it requires an instrumental variable for G_i. In Schmitz and Westphal (2024), we apply marginal treatment effect (MTE) estimation with two endogenous variables in a different context, namely causal mediation analysis. However, the estimation of interaction effects with two endogenous variables is beyond the scope of this paper.

the following arbitrary parameterizations of the potential outcomes, where, for simplicity, we leave out observable variables X_i:

$$\begin{aligned} Y_i^{11} &= 2.3 + \varepsilon_i^1, \qquad Y_i^{10} = 0.5 + \varepsilon_i^1, \qquad Y_i^{01} = 0.3 + \varepsilon_i^0, \qquad Y_i^{00} = 0 + \varepsilon_i^0 \\ E_i &= \mathbb{1}\{0.23 + 2.5G_i - 4Z_i + 3Z_i \times G_i > -(\varepsilon_i^1 - \varepsilon_i^0)\} \\ Z_i, G_i &= \text{Bernoulli distributed with } p = 0.5, \end{aligned}$$

where $\varepsilon_i^0 = \varepsilon_i^{0g}$ and $\varepsilon_i^1 = \varepsilon_i^{1g}$ are error terms, here (not in the application later) assumed to follow a bivariate normal distribution with the following parameters:[2]

$$\begin{pmatrix} \varepsilon_i^1 \\ \varepsilon_i^0 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0.4 \\ 0.4 & 2 \end{pmatrix}\right].$$

This simplified model generates a constant gene-environment interaction of $Y_i^{11} - Y_i^{01} - (Y_i^{10} - Y_i^{00}) = 1.5$ for each individual. Yet, 2SLS estimation of parameter β_3 in Eq. (1) returns an estimate of -1.3. The reason is that 2SLS identifies

$$\widehat{\beta}_3 = \mathbb{E}\left[Y_i^{11} - Y_i^{01} | C(G_i = 1)\right] - \mathbb{E}\left[Y_i^{10} - Y_i^{00} | C(G_i = 0)\right],$$

where $C(G_i = 1)$ represents the subgroup of individuals who are compliers to the education instrument, i.e., who take $E_i = 1$ if and only if $Z_i = 1$ when $G_i = 1$. $C(G_i = 0)$ stands for the subgroup of individuals who are compliers when $G_i = 0$. Therefore, the first part of the equation is identified for a group complying when $G_i = 1$, while the second part is identified for a group complying when $G_i = 0$. See Appendix A for a more formal derivation. Researchers who use 2SLS in this setting implicitly make the assumptions about the counterfactuals, namely: $\mathbb{E}[Y_i^{11} - Y_i^{01} | C(G_i = 1)] = \mathbb{E}[Y_i^{11} - Y_i^{01} | C(G_i = 0)]$ and/or $\mathbb{E}[Y_i^{10} - Y_i^{00} | C(G_i = 1)] = \mathbb{E}[Y_i^{10} - Y_i^{00} | C(G_i = 0)]$. If these assumptions hold, 2SLS estimates a well-defined causal effect.

There are two conditions under which these assumptions do not hold. For at least two values of G_i, $g' \neq g''$:

1. Complying types must differ between G_i such that $C(G_i = g') \neq C(G_i = g'')$. This is the necessary condition.
2. There must be unobserved effect heterogeneity such that $\mathbb{E}[Y_i^{1g'} - Y_i^{0g'} | C(G_i = g')] \neq \mathbb{E}[Y_i^{1g'} - Y_i^{0g'} | C(G_i = g'')]$, the sufficient condition (as the first condition is nested).

[2] In the simulation study we make the simplifying assumption that $\varepsilon^{11} = \varepsilon_i^{10} = \varepsilon_i^1$ and $\varepsilon_i^{01} = \varepsilon_i^{00} = \varepsilon_i^0$. This does not affect the main line of argumentation in this section and is merely for a simple exposition. It restricts the gene-environment effect to 1.5 for each complier type. Assuming four different error terms allows for a different gene-environment interaction effect by complier type. Our argument is not affected by that, and neither does our solution need this restriction, nor do we make this assumption in the application in Sections 3 to 5.

In our example data-generating process, these conditions hold by design. Thus, the 2SLS comparison is not a well-defined causal effect for this data-generating process, which we illustrate in Figure 1. Set up as an illustrative example, it shows the average effects of E_i on Y_i (depending on G_i) for four groups in the simulated data. Normally, these effects are unobserved by the researcher. The blue circles show $Y_i^{11} - Y_i^{01}$, while the red circles show $Y_i^{10} - Y_i^{00}$. Group 1 are always-takers (AT), irrespective of their realization of G_i. This is because their unobserved gains from E_i (that is, $\varepsilon_i^1 - \varepsilon_i^0$) are so large that they choose more education regardless of Z_i and G_i. The example also produces individuals that are always-takers when $G_i = 1$ but compliers (C) when $G_i = 0$ (Group 2), compliers when $G_i = 1$ and never-takers (NT) when $G_i = 0$ (Group 3) and never-taker, irrespective of G_i (Group 4).

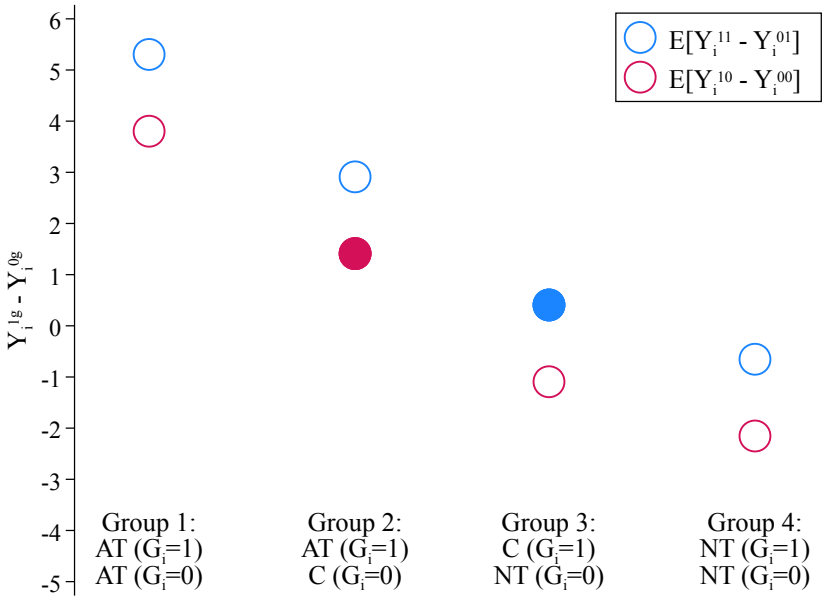

Figure 1: Effects of E_i on Y_i by G_i and complier type in the simulation model

Notes: This figure visualizes stylized potential outcomes from our simulation model. The differences in potential outcomes are defined by the data-generating process outlined above.

We sort these four groups on the horizontal axis according to their willingness to take education. Those on the left are most willing, and those on the right are least willing. According to the data-generating process, our stylized example is set up so that the gene-environment interaction effect (blue circle minus red circle) equals 1.5 for each group. However, 2SLS calculates this effect as the difference between the filled red circle (where $Y_i^{10} - Y_i^{00} = 1.42$) and the filled blue circle (where $Y_i^{11} - Y_i^{01} = 0.13$). Thus, it makes an inadequate comparison with a different estimate. This extreme example yields a negative 2SLS estimate while the true interaction is positive.

It is equally important to understand when this problem will not occur. First, whenever G_i does not affect the complier status of individuals such that groups with $C(G_i = 1)$ and $C(G_i = 0)$ do not differ on average. This is the case when G_i does not affect E_i. Second,

when there is no selection into gains, that is, individuals do not self-select into education based on the unobserved gain from treatment $\varepsilon_{i1} - \varepsilon_{i0}$, leading to $Cov(E_i, \varepsilon_{i1} - \varepsilon_{i0}) = 0$. Then the effects of E_i on Y_i do not differ by complier type. Without selection into gains, all red circles are on a horizontal line, and so are all blue circles. We argue that both conditions are not unlikely to apply in many real-world scenarios. Selection into gains has widely been documented in the context of education (Carneiro et al., 2011, Nybom, 2017, Kamhöfer et al., 2019, Westphal et al., 2022). Moreover, Barcellos et al. (2018) and Barcellos et al. (2021) show differences in first-stage responses to a compulsory schooling reform according to G_i. Biroli et al. (2022) discuss self-selection into environments according to genetic makeup as "active gene-environment correlation" where the environment mediates the effect of genes on the outcome.

2.2 The solution

We suggest going beyond estimating the two points that form the 2SLS estimate. Instead, we propose to estimate the MTE curve (see, e.g., Heckman and Vytlacil, 2005) by genetic endowment G_i. The MTE approach naturally sorts all individuals according to their willingness to take the treatment. It expands the sorting by stylized groups from Figure 1 to the continuous unit interval. In the spirit of Heckman and Vytlacil (2005) we rewrite the choice equation as:

$$
\begin{aligned}
E(G_i) &= \mathbb{1}\{\mu_E^g(X_i, Z_i) \geq V_i^g\} \\
&= \mathbb{1}\{F_{V_i^g}(\mu_E^g(X_i, Z_i)) \geq F_{V_i^g}(V_i^g)\} \\
&= \mathbb{1}\{\Pr\left(V_i^g \leq \mu_E^g(X_i, Z_i)\right) \geq F_{V_i^g}(V_i^g)\} \\
&= \mathbb{1}\{\Pr\left(E(G_i = 1)|X_i, Z_i)\right) \geq U_i^E\} \\
&= \mathbb{1}\{PS(G_i, X_i, Z_i) \geq U_i^E\}
\end{aligned}
$$

The second step applies the monotonic transformation $F_{V_i^g}(\cdot)$ – which is the cumulative density of V_i^g – to both sides of the inequality. $F_{V_i^g}(\cdot)$ evaluated at the point $\mu_E^g(X_i, Z_i)$ is defined as $\Pr(V_i^g \leq \mu_E^g(X_i, Z_i))$ and, referring to the choice equation, the same as $\Pr(E(G_i = 1)|X_i, Z_i)$. This choice probability based on observable characteristics is the propensity score, and we abbreviate it by $PS(G_i, X_i, Z_i)$. Irrespective of the underlying distribution of V_i^g, the unobserved term U_i^E is uniformly distributed on the unit interval and comprises the unobserved heterogeneity correlating with the decision to take E_i. Low values of unobserved resistance to more education U_i^E increase $PS(G_i, X_i, Z_i)$, leading to $E_i = 1$. This corresponds to high unobserved preferences for E_i, whereas large values of U_i^E indicate a high distaste for E_i.

MTEs are estimates of the causal effect of education on the outcome Y_i at certain values of $U_i^E = u$. That is, we estimate $\mathbb{E}[Y^{1g} - Y^{0g}|U_i^E = u]$. The MTEs are identified by those individuals who, at $U_i^E = u$, are indifferent between choosing $E_i = 0$ and $E_i = 1$. Referring to the choice equation, this is the group for whom the realization p of the propensity score $PS(G_i, X_i, Z_i) = p = u$. See Heckman and Vytlacil (2005) for an extensive introduction to MTEs, their derivation, and traditional ways to estimate them with continuous instruments. For our framework, the quantities $\mathbb{E}[Y_i^{1g}|U_i^E = u]$ and $\mathbb{E}[Y_i^{0g}|U_i^E = u]$ are essential (as their difference is the MTE). We follow the literature and call these quantities marginal treatment response curves (MTRs). We provide details on estimating MTEs and MTRs in Section 5. Figure 2 shows the MTE curves based on our simulation example. The interaction effect is calculated as the difference between the blue and red curves. According to our simulation example, the interaction effect is 1.5 and constant over the entire U_i^E range.

In practice, when the two curves are not parallel, the interaction effect differs by U_i^E. Thus, there are several ways to estimate interaction effects. One possibility is to compute the difference at a specific value of U_i^E, say $U_i^E = 0.4$. The advantage of this over a 2SLS estimation is that unobserved heterogeneity is fixed. Thus, it will be a consistent estimate, albeit very local, i.e., only at $U_i^E = 0.4$. MTEs can be used to estimate all treatment parameters, depending on how they are aggregated and how MTEs in different areas of the unit interval are weighted. In principle, it is possible to compute interaction effects using the MTE curves with 2SLS weights either for $C(G_i = 1)$ or $C(G_i = 0)$. In our application below, we use a simpler solution. We will aggregate the MTE results to receive the average interaction effect for all individuals on the U_i^E interval between 0.6 and 0.8, visualized by the two vertical lines in Figure 2. In this interval, all MTEs are weighted uniformly. MTEs from other parts of the U_i^E range are not considered. We choose this interval since most of the compliers to the education instrument in our application are located in this area.[3] Furthermore, we use an area where most compliers are located to maintain comparability to 2SLS/LATE estimates.

2.3 Going beyond a binary representation of G

The problem and its solution are not specific to cases where G_i is binary. On the one hand, our solution requires a discrete G_i because we will estimate separate curves by G_i. On the other hand, generating a binary indicator of genetic endowment from a continuous polygenic score entails a loss of information. In our application below, we transform the continuous PGS into a discrete measure that takes the values $g \in \{1, 2, 3, 4, 5\}$, indicating the PGS quintiles. Consequently, the number of potential outcomes we estimate increases from four to ten. In Table 1, we list these potential outcomes and how to calculate the effect of E_i on Y_i and the $G \times E$ interaction by genetic type, i.e., quintile of the PGS.

[3] We will show robustness checks for larger intervals.

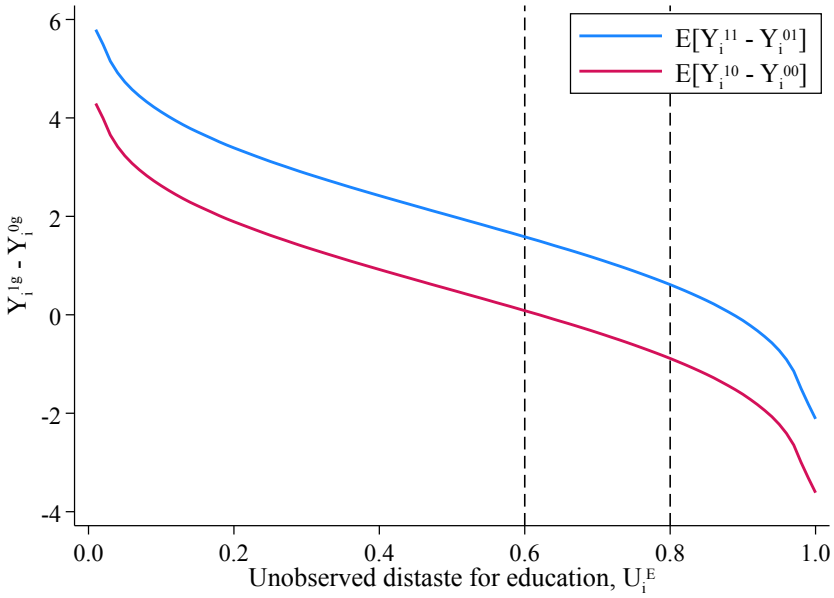

Figure 2: Marginal treatment effects of E_i on Y_i by G_i in the simulation model

Notes: This figure shows stylized marginal treatment effect curves in our simulation model. The differences in potential outcomes are defined by the data-generating process outlined above.

Table 1: Potential outcomes and calculation of MTEs using quintiles of the polygenic score

		$E_i = j$		Individual treatment effects for	
		0	1	the effect of E_i on Y_i	the gene-environment interaction
$G_i = g$	1	Y_i^{01}	Y_i^{11}	$Y_i^{11} - Y_i^{01}$	$(Y_i^{11} - Y_i^{01}) - (Y_i^{11} - Y_i^{01})$
	2	Y_i^{02}	Y_i^{12}	$Y_i^{12} - Y_i^{02}$	$(Y_i^{12} - Y_i^{02}) - (Y_i^{11} - Y_i^{01})$
	3	Y_i^{03}	Y_i^{13}	$Y_i^{13} - Y_i^{03}$	$(Y_i^{13} - Y_i^{03}) - (Y_i^{11} - Y_i^{01})$
	4	Y_i^{04}	Y_i^{14}	$Y_i^{14} - Y_i^{04}$	$(Y_i^{14} - Y_i^{04}) - (Y_i^{11} - Y_i^{01})$
	5	Y_i^{05}	Y_i^{15}	$Y_i^{15} - Y_i^{05}$	$(Y_i^{15} - Y_i^{05}) - (Y_i^{11} - Y_i^{01})$

Notes: This table lists all combinations of potential outcomes when G_i corresponds to quintiles of the PGS such that $G \in \{1, 2, 3, 4, 5\}$ (left panel). The right panels show how to compute different individual treatment effects, including the interaction effects at every quintile we are after.

Extending the setting to a more complex (but still discrete) classification has advantages. We can use more of the rich variation the PGS offers and allow for possible non-linearities in the interaction effects between different sections of the PGS distribution. Of course, the choice to use quintiles is arbitrary. Barcellos et al. (2018) and Barcellos et al. (2021) show differences in their results according to the terciles of the education PGS. This is already considerably less restrictive than using a binary representation. The use of quintiles offers a further improvement over terciles. It strikes a balance between estimating non-linear interaction effects and comparing subsets of the PGS that are relevant in terms of size.

3 Institutional Setting and Data

3.1 Compulsory schooling reform in the UK

In our application, we exploit exogenous variation from a compulsory schooling reform in the UK. Based on the Education Act of 1944, two reforms were enacted to raise the minimum school leaving age in England, Scotland, and Wales. We use the first reform, which took effect on April 1, 1947. This reform raised the minimum age for leaving school from 14 to 15.[4] Given that students in the UK typically enter school at the age of 5, this reform effectively extended compulsory education from nine to ten years. The first birth cohorts to be affected by the change, i.e., the first to be required to attend school for an additional year (the "pivotal cohort"), were those born in April 1933. This reform is presented in detail in Clark and Royer (2013) and has served as an exogenous variation for compulsory schooling in studies on the effect of education on health (Clark and Royer, 2013) and cognitive abilities (Banks and Mazzonna, 2012). Since we are interested in studying cognitive abilities in old age, we use the first reform in 1947. Cohorts affected by the second reform in 1972 are - for the most part - still too young at data collection of the English Longitudinal Study of Ageing.

To demonstrate the strong response to the compulsory schooling reform from 1947, Figure 3 shows aggregated cohort-level data from ELSA. It depicts the share of individuals with different levels of schooling by birth year cohort. The pivotal cohorts of both compulsory schooling reforms are marked with vertical lines. The highest line (circle markers) shows how the 1947 reform caused a significant increase in the share of students leaving school at age 15 or later from about 40% to almost 100%. The middle line shows how the second reform in 1972 lead to a still remarkable but comparably smaller increase in the share of leaving school at 16 or later from 75% to about 90%. The lowest line can be read as a placebo test, showing the general trend in increased years of schooling but no discontinuity at the two reform cut-offs (Clark and Royer, 2013).

3.2 Sample and Variables

Sample

We use data from the English Longitudinal Study of Ageing (ELSA), a large representative microdata set providing information on health and other socioeconomic characteristics of individuals aged 50 and over in England (Banks et al., 2023). ELSA was launched in 2002 and is conducted every two years. It currently comprises ten waves of interviews.[5] We use

[4] The second part was enacted much later, in 1972, raising the school leaving age to 16.

[5] For details of the ELSA sampling procedure, questionnaire content, and fieldwork methodology, see Steptoe et al., 2013.

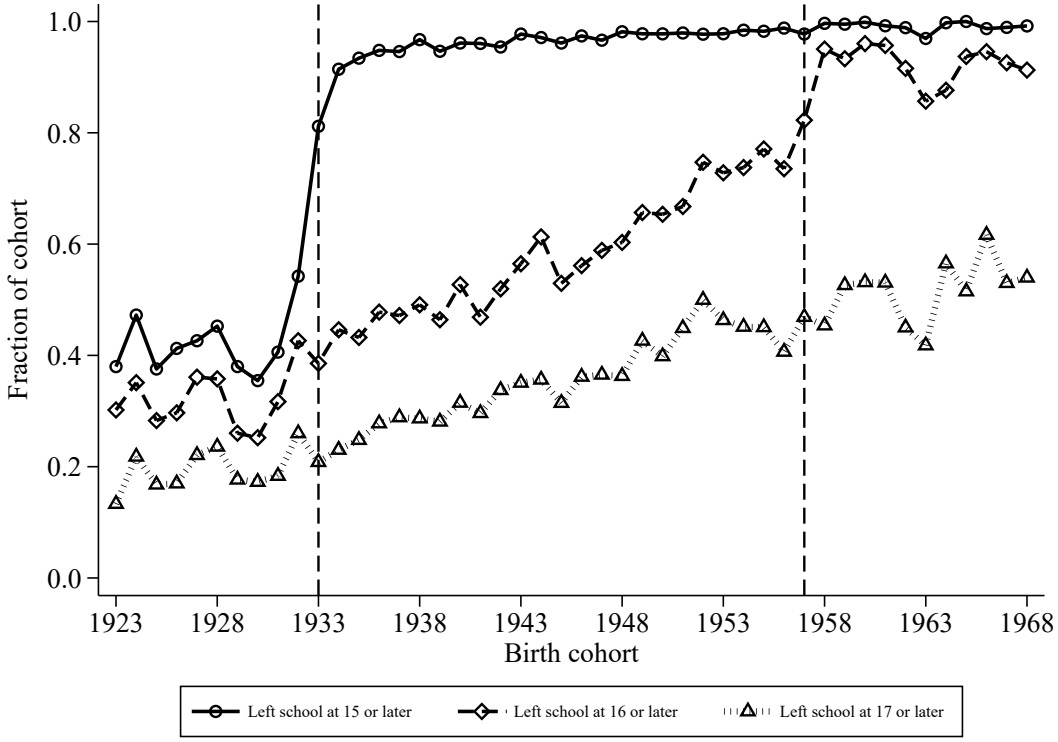

Figure 3: Education by birth cohort

Notes: This figure illustrates the shares of students leaving school at 15 or later, 16 or later, and 17 or later changed over birth cohorts and how these shares were affected by two compulsory schooling reforms. Vertical dashed lines indicate the first affected birth cohorts of two school-leaving age increases. The three groups are not mutually exclusive and do not add up to 100%. The illustration is adapted from Clark and Royer (2013) to fit our definition of educational attainment and uses data from ELSA waves 1–9, totaling 76,829 obs. from 17,063 individuals.

individuals aged 65–80 from waves 1–6 of ELSA. Data collection for wave 6 took place in 2012 and 2013 when individuals born in 1933 – our cutoff – turned 80. Thus, starting with wave 7, only individuals born after the cutoff can theoretically enter the sample. We drop the 1933 birth cohort because we do not have information on birth month and cannot correctly assign this cohort to pre- or post-reform (the cutoff is April 1933). We also restrict the data to birth cohorts ten years before and after the reform cut-off. Finally, for our main analysis, we need to limit the data to individuals for whom genetic data is available. This reduces the number of individuals substantially by about 50 percent. This may introduce a selection bias if the compulsory schooling reform affects the willingness to be genotyped. We find that the sample is selective regarding the outcome variable: Individuals who consent to be genotyped have higher recall scores on average (see Table C.2 in the appendix). However, we do not find evidence of a statistically significant effect of the compulsory schooling reform on the probability of being genotyped. Similarly, the willingness to be genotyped does not interact with the impact of compulsory schooling on the probability of going to school until at least the age of 15 (see Table D.4 in the appendix). Our final sample includes 11,027 observations from 3,009 individuals born between 1923 and 1943 and observed between 2002 and 2013.

Cognitive Abilities

Cognitive abilities – as a broad concept – include "the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings – 'catching on,' 'making sense' of things, or 'figuring out' what to do." (Gottfredson, 1997). The sum of these abilities is called intelligence (Schiele and Schmitz, 2023). A wide range of cognitive tests measure different aspects of cognitive abilities to accommodate this multifaceted notion. ELSA contains several measures, such as memory capacity, temporal orientation, literacy, and numerical ability. We use test scores from the word recall test. In the word recall test, an interviewer reads ten words to the respondent, who is then asked to recall as many words as possible. This test is administered twice: immediately after the words are read (immediate recall) and five minutes later (delayed recall). The total number of words recalled at both times is added together. The total recall score combines the two and can range from 0 to 20. It serves as a measure of episodic memory, susceptible to aging (Rohwedder and Willis, 2010). The test is considered a fluid intelligence component, reflecting the innate cognitive ability to store and retrieve information. It is distinct from crystallized intelligence that people learn over a lifetime (using their fluid intelligence). In our estimation sample, the total recall score, our dependent variable, has a mean of 9.67 correctly recalled words (out of 20) with a standard deviation (SD) of 3.37 words (see Table 2).

Education

ELSA does not offer information on respondents' years of education but on the age at which they finished their continuous full-time education. However, the data is aggregated at the low (finished age 14 or earlier) and high (finished age 19 or later) ends. Our treatment variable E_i is a binary variable equal to one if the individual has left school at 15 or later and zero otherwise. By design, and as observable in Figure 3, the proportion of individuals having left school at 15 or later (i.e., having stayed in school for at least ten years) is affected by the 1947 education reform that raised the minimum school leaving age from 14 to 15.

Genes

We use an Educational Attainment Polygenic Score (PGS) provided by ELSA and based on Lee et al. (2018) to measure genetic makeup. This indicator predicts educational attainment based on differences in genetic variants across individuals. The education PGS we use explains 11-13 percent of the variation in educational attainment in the original discovery sample (Lee et al., 2018). An individual's polygenic score can be considered their genetic propensity (or genetic "risk") for a particular trait – in our case, education (Biroli et al., 2022). For a more detailed explanation of polygenic scores and their construction, see

Appendix B. The education polygenic score is normally distributed. Individuals whose genetic endowment puts them on the left side of this distribution have a lower genetic propensity to pursue education; individuals on the right side have a higher propensity.[6] Our analysis uses the quintiles of the PGS, yielding five equally sized groups.

When estimating gene-environment interactions, researchers usually use a polygenic score of the outcome they are investigating since it is the obvious choice and will produce an effect. However, the choice is not set in stone. As Biroli et al. (2022) point out, "any PGI could be used if warranted by theory or for empirical reasons". We target the polygenic score towards the environmental variable (education) by using an education polygenic score, and the outcome we investigate is cognitive ability. Education polygenic scores are associated with several different outcomes besides educational attainment: wealth at retirement (Barth et al., 2020), labor market earnings (Papageorge and Thom, 2020) and socioeconomic success (Belsky et al., 2018). In our setting, we can use the education PGS to demonstrate heterogeneous responses to the education reform by genetic endowment. At the same time, the effect of education on cognition likely varies with genetic propensity for education – which is what we want to estimate.

As it is established in the gene-environment literature (see, e.g., Barth et al., 2020; Barcellos et al., 2018; Pereira et al., 2022), we include the first ten principal components of the genetic information as controls to make comparisons between "individuals within a common lineage and from the same genetic pool" (Biroli et al., 2022). Principal components are linear combinations of genetic markers that summarize the major patterns of gene variation *across a population* into fewer dimensions. They reflect population stratification, i.e., different frequencies of genetic variants among subpopulations that could be responsible for spurious correlations with outcomes of interest. Price et al. (2006) show that including principal components as controls can mitigate the confounding effects of population stratification, ensuring that differences in ancestry or population structure do not drive observed associations between genetic variants and traits.

3.3 Descriptive Statistics

Table 2 shows descriptive statistics of our main sample of individuals for whom genetic information is available as well as of "treatment" ($E_i = 1$) and "control" ($E_i = 0$) groups separately. Overall, about three-quarters of the observations in the sample are in the treatment group; 66 percent were born in 1933 or later, and 52 percent are female. The treatment group scores significantly higher in recall than the control group. More educated individuals ($E_i = 1$) exhibit a more favorable genetic endowment (significantly less

[6] As the choice equations in Section 2 emphasize, this propensity is not deterministic. Individuals with a high PGS do not necessarily have to be highly educated, and vice versa.

observations in the first and more in the top quintile). Unsurprisingly, individuals in the treatment group are, on average, younger since they are more likely to be born after the compulsory schooling reform. Table C.1 extends the statistics for the principal components and birth year. Table C.3 in the appendix shows the sample means by quintiles of the education PGS. Instrument assignment, age, and proportion of women do not vary across quintiles of the education PGS. However, individuals in higher quintiles perform better on the recall test. The difference between an average person in the lowest PGS quintile and an average person in the highest quintile is 1.24 words, a sizable difference compared to the overall mean of 9.67. Not surprisingly, the probability of having more schooling is also higher in higher education PGS quintiles.

Table 2: Descriptive statistics

	Main sample		By E_i		
	Mean (SD)		E_i=1	E_i=0	Difference (SE)
Outcome Y_i					
Recall score	9.67	(3.37)	10.11	8.08	2.03 (0.07)***
Treatment E_i					
Left school $\geq$ 15	0.78	(0.41)	1.00	0.00	1.00
Polygenic score G_i					
1st PGS quintile	0.20	(0.40)	0.18	0.25	−0.07 (0.01)***
2nd PGS quintile	0.19	(0.40)	0.19	0.21	−0.02 (0.01)**
3rd PGS quintile	0.20	(0.40)	0.21	0.19	0.02 (0.01)**
4th PGS quintile	0.21	(0.41)	0.21	0.20	0.01 (0.01)
5th PGS quintile	0.20	(0.40)	0.22	0.15	0.07 (0.01)***
Instrument Z_i					
Born 1933 or later	0.66	(0.47)	0.82	0.13	0.69 (0.01)***
Selected Controls (for a complete list, see Table C.1)					
Female	0.52	(0.50)	0.52	0.50	0.02 (0.01)**
Age	71.82	(4.29)	70.89	75.10	−4.21 (0.09)***
Observations	11,027		8,590	2,437	

Notes: This table presents descriptive statistics. We include the mean and standard deviation of the main sample and means by E_i, the difference of means, and standard errors of a t-test for equality of means. $^{*}p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

4 Benchmark 2SLS estimation

4.1 Empirical Strategy

We start by estimating the gene-environment interactions using "conventional" methods. Since education is a choice variable, an OLS regression will yield biased estimates. We estimate the following 2SLS regression:

$$E_i = \pi_0 + \pi_1 G_i + \pi_2 Z_i + \pi_3 G_i \times Z_i + X'\gamma + f(t) + u_i \tag{2}$$

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 \widehat{E_i} + \beta_3 \widehat{G_i \times E_i} + X'\delta + f(t) + \varepsilon_i \tag{3}$$

Eq. (2) is the first stage, where we regress education E_i on our instrument Z_i, genetic predisposition G_i and the interaction of G_i and Z_i.[7] Eq. (3) shows the second stage. Here, we regress the outcome variable Y_i (the total recall score for individual i) on the predicted values $\widehat{E_i}$ from the first stage, G_i and the predicted values $\widehat{G_i \times E_i}$. In both stages, we add the same controls X_i that include the first ten principal components of the genetic data (see Section 3.2 for a description) as well as $f(t)$, a function that captures a linear cohort trend and its interaction with the instrument Z_i. This specification estimates a fuzzy regression discontinuity model with the re-centered distance to the reform cohort of 1933 (the cohort trend) as the running variable. Finally, u_i and ε_i capture all unobserved factors that affect outcome variables in their respective stages. We cluster standard errors at the individual level in all analyses.

Besides the potential problems due to essential heterogeneity, this specification linearizes the $G \times E$ effect (and the effect of G itself). This may also mask potentially interesting non-linearities. To be more flexible, we extend our analysis by fully saturating our specification using information on the quintiles of the polygenic score. Therefore, these effects compare better to our MTE approach because we directly estimate effects by quintiles. Accordingly, we estimate the following adapted model:

$$E_i = \sum_{g=1}^{5} \left[\pi^f_{g,0} \mathbb{1}[G_i = g] + \pi^f_{g,\Delta} \mathbb{1}[G_i = g] \times Z \right] + X'\gamma^f + f(t) + \omega_i \tag{4}$$

$$Y_i = \sum_{g=1}^{5} \beta^f_{g,0} \mathbb{1}[G_i = g] + \beta^f_{1,1} \widehat{E_i} + \sum_{g=2}^{5} \beta^f_{g,1} \widehat{\mathbb{1}[G_i = g] \times E_i} + X'\delta^f + f(t) + \eta_i \tag{5}$$

This is the equivalent of the 2SLS model described above in Eqs. (2) and (3), but with sets of indicator variables for the five gene quintiles ($G_i = g$ with $g \in \{1, 2, 3, 4, 5\}$). To

[7] Note that there are technically two first stages, one with the dependent variable E_i and one with the dependent variable $G_i \times E_i$. Depending on how G_i is included, there are more. With G_i as quintiles of the PGS, there are six first stages. We only show one of them here.

distinguish the coefficients from the base model, we add the superscript f. While in the baseline first stage (Eq. 2), π_2 informs about the share of compliers in the data, the $\pi^f_{1,\Delta}$ to $\pi^f_{5,\Delta}$ of Eq. (4) inform about the share of compliers by PGS quintile. In the second stage (Eq. 5), we include $\widehat{E}_i$ as the reference category that captures the LATE for the lowest quintile ($\beta^j_{1,1}$). The coefficients $\beta^f_{2,1}$ to $\beta^f_{5,1}$ inform about gene-environment interactions relative to the first quintile.

4.2 Assumptions

We need to assume that the compulsory schooling reform is a valid instrument to identify the causal effects of extending schooling beyond the age of 14. Specifically, the cutoff at the 1933 birth cohort needs to be exogenous to the individuals in our sample. Clark and Royer (2013) convincingly show that this reform can be a credible instrument (especially for studying the outcomes of older individuals, as they find no effects on mortality). For this reform to be a valid instrument, we assume that only compulsory schooling changes for individuals born after April 1933 and nothing else. Two important events took place roughly when the cohorts in our sample were born: The Great Depression and the Second World War. While they may have been exposed to rationing or evacuations, individuals on either side of the 1933 cutoff were likely affected similarly (Clark and Royer, 2013). Taken together, it seems plausible that the compulsory schooling instrument is valid. Furthermore, we assume that genes are exogenous, i.e., determined before age 14 when individuals were exposed to the reform, depending on their birth cohort. For the gene-environment interaction, our setting also needs the effects of genes to be predetermined before secondary education. Lastly, assessing the graphical evidence on the first stage and reduced form by G_i (Figure D.1 in the Appendix) shows that changes occur at the cutoff, and linear trends approximate the data well.

4.3 Results

OLS and second-stage results

Table 3 presents the OLS and 2SLS regression results of the second stage (Eq. 5). Panel A includes the nonlinear estimates for each quintile; Panel B uses the standardized PGS as a continuous interaction variable to show linear effects. The $G \times E$ interaction in Panel A is computed to compare the bottom quintile (the reference category) and respective higher quintiles. The OLS coefficients of $G_i = 2$ to $G_i = 5$ suggest that, in general, only individuals in the highest PGS quintile score statistically significantly higher on the recall test relative to individuals in the lowest — about 0.83 words higher than individuals in the lowest quintile. This positive relationship between an education PGS and cognitive

performance is also documented by Jeong et al. (2024), who use data from the US Health and Retirement Study. An additional year of education (E_i) is associated with an increase of about 0.74 words in the 20-item recall summary score later in life. This association refers to individuals in the lowest PGS quintile because we control for the interaction between E_i and all higher quintiles. The markups on this association for individuals in the four higher PGS quintiles ($E_i \times (G_i = g)$) are positive across all quintiles. However, their magnitude varies. For individuals in the third quintile, an additional year of schooling results in about 0.47 more words recalled than for individuals in the first quintile. For the fourth quintile, this relative premium increases to 0.77. Except for the interaction coefficient for the fourth quintile, they are not statistically significant, but importantly, they all indicate a positive gene-environment interaction. All in all, this suggests that the association between genes, more education, and cognition are mutually reinforcing. When we include the standardized PGS as a continuous variable (in a separate regression, shown in Panel B), its interaction coefficient suggests that a one standard deviation increase in PGS is associated with an additional rise in recall score by 0.2 words.

This complementarity implies higher returns to education in terms of cognitive abilities later in life for individuals whose genetic markers predict more years of education. This would mean that the education reform exacerbated differences in cognition between those in the bottom quintile and those in higher quintiles. This is in line with what Barcellos et al. (2021) find for socioeconomic status. However, our OLS results only represent correlations.

Our 2SLS estimates are reported in Column (2). For education, the 2SLS regression finds a zero effect of an additional year of schooling on cognition later in life for those in the lowest PGS quintile (E_i coefficient) and a positive estimate for individuals in the second, third and fifth quintile and a slight negative one for those in the fourth quintile. The standard errors are large, so we cannot be certain that these interactions differ from zero, especially in quintiles 3 and 4, where the estimates are very small. The linear interaction effect using a standardized PGS (Panel B) is also close to zero. Based on these results, we would conclude that, after resolving the endogeneity problem with E_i by instrumenting — if anything — there may only be a small positive interaction effect that cannot be precisely estimated. The returns to education are likely not higher for individuals with higher genetic endowment. However, consistent with the problem outlined in Section 2, recall that when comparing the effects between two quintiles, the complier group also changes, which may offset the small and monotonic gene-environment interaction.

First-stage results

We report the coefficients of Z_i by G_i, that is, $\pi^f_{1,\Delta}$ to $\pi^f_{5,\Delta}$ of Eq. (4) in Figure 4.[8] It shows that overall, there is a large share of compliers to the education reform in the data. However, it

[8] Regression results are reported in Table D.5 in the Appendix.

Table 3: OLS and 2SLS estimates of the $G \times E$ interaction

	Dependent variable – total recall score			
	OLS (1)		2SLS (2)	
Panel A: nonlinear $G \times E$ effect with G_i as quintiles				
E_i	0.742	(0.244)***	−0.042	(0.460)
$G_i = 1$	reference category		reference category	
$G_i = 2$	0.182	(0.272)	0.196	(0.403)
$G_i = 3$	0.255	(0.271)	0.631	(0.455)
$G_i = 4$	0.237	(0.271)	0.946	(0.462) **
$G_i = 5$	0.827	(0.316)***	0.819	(0.621)
$E_i \times (G_i = 1)$	reference category		reference category	
$E_i \times (G_i = 2)$	0.391	(0.327)	0.423	(0.504)
$E_i \times (G_i = 3)$	0.469	(0.328)	0.071	(0.555)
$E_i \times (G_i = 4)$	0.774	(0.330)**	−0.043	(0.591)
$E_i \times (G_i = 5)$	0.524	(0.367)	0.649	(0.726)
Panel B: linear $G \times E$ effect with continuous G_i				
$E_i \times G_i$	0.214	(0.111)*	0.066	(0.216)
Controls	Yes		Yes	
Observations	11,027		11,027	

Notes: This table presents OLS and 2SLS estimates of the interaction between a polygenic score for education and staying in school until at least the age of 15 (our treatment) on cognitive abilities later in life. In panel A, we use quintiles of the polygenic score (G_i) and estimate non-linear (interaction) effects. Panel B shows estimates of a linear effect when including the standardized PGS as a continuous variable. Coefficients in both panels are obtained from separate regressions. Controls in each case include a linear cohort trend, its interaction with the instrument (born in 1933 or later), gender, and the first ten principal components of the genetic data. Standard errors clustered at the individual level shown are in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

varies substantially over the quintiles of the PGS. Complier share monotonically decreases along the PGS. In the lowest quintile ($G_i = 1$), 64 percent of all individuals increased their length of education due to the reform. The share of compliers reduces to a still sizable 36 percent in the highest quintile. The compulsory schooling reform had a more substantial impact on individuals in the lowest quintiles of the PGS, who are disadvantaged in terms of the genetic endowment that predicts education. Therefore, the reform was likely effective in targeting disadvantaged children. It drastically increased their probability of staying in school until at least age 15. Our estimates in Table 3 suggest that the reform may not have successfully reduced differences in the cognitive returns to education but may have increased them. This finding is consistent with Barcellos et al. (2021), who document that the UK's 1972 compulsory schooling reform reduced disparities in education and qualifications between children from different backgrounds but ultimately increased differences in socioeconomic status.

This finding provides the first piece of evidence that a 2SLS estimation of the $G \times E$ effect might be problematic in our setting. As mentioned before, two prerequisites need to jointly hold for this to be the case: 1) the individual response to the compulsory schooling

reform depends on G_i, and 2) the instrument response types for $E_i = 1$ (always-takers and treated compliers) and $E_i = 0$ (never-takers and untreated compliers) need to exhibit a specific heterogeneity that implies essential heterogeneity. Our first stage results show that complier status is correlated with genetic type, the first of the two requirements that cause 2SLS to make potentially problematic comparisons. The second one – evidence of essential heterogeneity – will be discussed in the next section, where we estimate marginal treatment effects. This will allow us to make statements about whether this kind of selection occurs in the data.

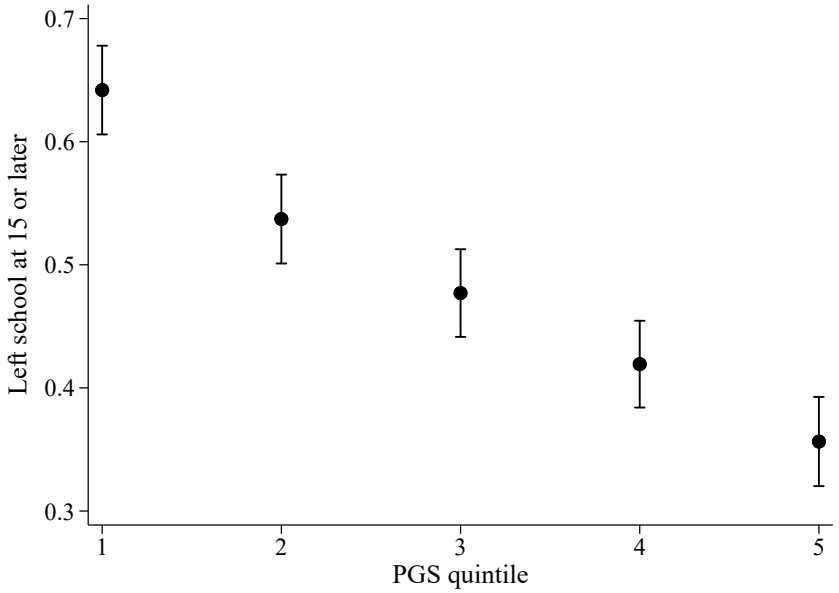

Figure 4: Strength of the first stage by quintiles of the polygenic score

Notes: The figure shows the five estimated coefficients $\pi^f_{1,\Delta}$ to $\pi^f_{5,\Delta}$ of Eq. (4), which correspond to the complier shares by PGS quintile. We add 95% confidence intervals. We report the point estimates and their standard errors in the appendix in Table D.5.

5 MTE estimation of the $G \times E$ interaction

5.1 Setup

There are many different ways to estimate marginal treatment effects, depending on the underlying data, setting (e.g., continuous or binary instrumental variables) and the assumptions the researcher wants to impose (e.g., functional form assumptions for the MTE, separability between observed and unobserved terms). In our case, with a binary instrument, there are three options. The first entails estimating separate potential outcomes $\mathbb{E}[Y_i^{1g}|AT]$, $\mathbb{E}[Y_i^{1g}|C]$, $\mathbb{E}[Y_i^{0g}|C]$, and $\mathbb{E}[Y_i^{0g}|NT]$ for each value of G_i. Plotted on the U_i^E unit interval and assuming linearity, we can fit lines through each pair of points, one for treated and one for untreated potential outcomes (Brinch et al., 2017). As conventional

in the literature, we call these average potential outcomes conditional on U_i^E (and G_i for our specific setting) marginal treatment response curves (MTRs): $\mathbb{E}[Y_i^{jg}|U_i^E = u, G_i = g]$. The linearity ensures that the lines run through the respective (type-specific) midpoints on the U_i^E scale. The difference between the two MTR lines is the linear MTE by G_i, and the four differences between the five G_i-specific MTEs inform about the interaction effects. The second option relaxes the linearity assumption but imposes additive separability between controls X_i and error terms. This allows variation in X_i to parametrically or semi-parametrically identify the MTEs, since binary instrument alone cannot provide this (Brinch et al., 2017). The third option is to allow for a wide range of flexible polynomial shapes of the MTEs and subsequently restrict the shapes. This can be achieved by requiring the curves to reproduce observable sample analogs and imposing further reasonable assumptions derived from theory and the data. The target parameter the researcher aims to identify can be bounded by the two shapes that produce minimum and maximum values (Mogstad et al., 2018).

A linearity assumption is hard to justify a priori. Furthermore, although additive separability is commonly assumed across the entire literature that uses regression models, we do not benefit from it for a semi-parametric identification of the MTEs. This is because we only use a sparse set of control variables that do not add sufficient variation in the propensity score, which would help identify substantially more than the four points from the first approach. Overall, the third approach seems most appropriate for our setting. Nevertheless, we first estimate linear MTEs according to Brinch et al. (2017). They help illustrate the setting and show general trends. They are also informative about underlying shape restrictions. For our main results, we ease this linearity restriction and allow for flexible polynomials

We begin by estimating the type-specific group means $\mathbb{E}[Y_i^{1g}|AT]$, $\mathbb{E}[Y_i^{1g}|C]$, $\mathbb{E}[Y_i^{0g}|C]$, and $\mathbb{E}[Y_i^{0g}|NT]$ as well as the shares of AT, C, and NT for each quintile of the PGS. Appendix F presents the details on generating these 35 values by applying the Imbens and Rubin (1997) method. We visualize the 20 means (circles) as well as the type shares (horizontal lines at the bottom) for the bottom PGS quintile ($G_i = 1$) in Figure 5. Again, we sort the three types according to their willingness to take education on the U_i^E scale. Always-takers have the highest willingness and are located at the left. The share of always-takers $AT(G_i = 1)$ is 22 percent. The share of compliers with $G_i = 1$ is 68 percent. They are located between 0.22 and 0.9 on the U_i^E unit interval. The remaining 10 percent are never-taker. Following Kowalski (2023), we use the midpoints of the range where each type is located to place the dots on the x-axis, while the y-axis measures the estimated potential outcomes. The blue dots denote treated potential outcomes $\mathbb{E}[Y_i^{1g}]$ while the red dots denote untreated potential outcomes $\mathbb{E}[Y_i^{0g}]$. The numbers next to the dots refer to the realization of G_i.

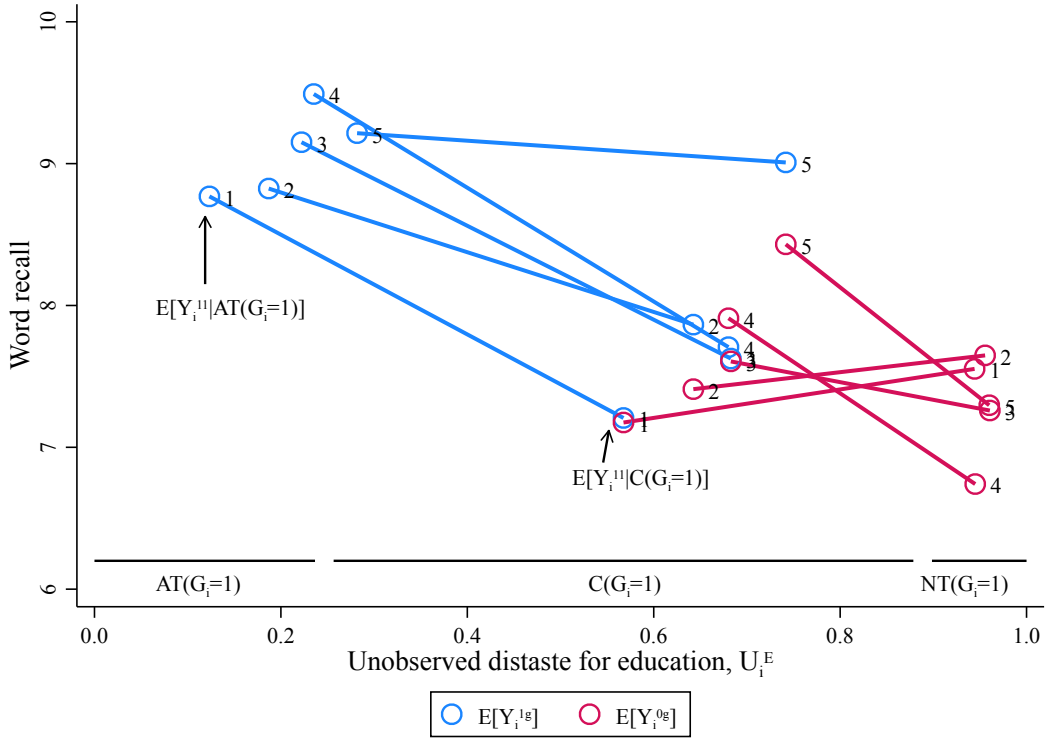

Figure 5: Linear potential outcome curves

Notes: This figure shows the 20 estimated potential outcomes $\mathbb{E}[Y_i^{jg}]$. The lines through them represent linear MTRs. Red color refers to potential outcomes for $E_i = 0$; blue to $E_i = 1$. Thus, for example, the red line labelled "1" shows our estimate of Y_i^{01}; the blue line labelled "3" shows the estimate for Y_i^{13}. Horizontal lines at the bottom show type shares and their location on the unit interval. For readability, we only report type shares for $G_i = 1$

The lines through the points produce the MTRs under a linearity assumption that allows us to identify them by the two points. In principle, the lines can be extrapolated to the full unit interval. Taking differences between $\mathbb{E}[Y_i^{1g}]$ and $\mathbb{E}[Y_i^{0g}]$ yields the MTEs by G_i and the comparison of the resulting 5 MTEs informs about (non-linear effects of) the gene-environment interaction. However, as mentioned above, the linearity assumption, which will drive the final results, is hard to defend a priori. Nevertheless, this analysis has important implications for our bounding approach to compute our main results (Section 5.2). Treated potential outcomes (blue) are higher for always-takers than compliers, causing the $\mathbb{E}[Y_i^{1g}|U_i^E = u]$-MTR curves to have a negative slope. Therefore, there seems to be a correlation between types and our dependent variable, word recall. Moreover, the $\mathbb{E}[Y_i^{1g}|U_i^E = u]$ MTR curves are fairly parallel, with no substantial slope differences. They represent level shifts in treated outcomes by G_i. We base the following assumptions for our bounding approach on these findings: 1) the MTR-curves $\mathbb{E}[Y_i^{1g}|U_i^E = u]$ are declining over a relevant area on the unit interval and 2), the curves for different values of G_i are parallel, i.e., there is additive separability between observed G_i and unobservable factors that materialize in the shapes of the MTRs.

The picture is less clear for the untreated outcomes (red). Here, we see that the outcomes of the untreated compliers are, on average, slightly smaller than those of the treated

compliers, suggesting positive effects of education on cognition. We can replicate the 2SLS finding of a zero effect for $C(G_i = 1)$ and positive effects for $C(G_i = 2)$ and $C(G_i = 5)$. However, the estimated results for never-takers are less clear, as they are above those for the untreated compliers for the first and second quintiles, resulting in positive slopes of the two lowest $\mathbb{E}[Y_i^{0g}|U_i^E = u]$-MTRs. Given this ambiguous result and the small share of never-takers in the data, we include a robustness check of our main result where we estimate MTEs without relying on never-takers in Section 5.4.

5.2 Estimation using Mogstad, Santos, and Torgovitsky (2018)

We now sketch a partial estimation method suggested by Mogstad et al. (2018) and recently applied by Rose and Shem-Tov (2021) that allows a transparent and credible estimation of marginal treatment effects when the instrument is binary or when the variation in the instrument does not sufficiently identify marginal treatment effects over the whole support of the propensity score. This is the third approach we briefly mentioned in Section 5.1. This approach uses the estimated potential outcomes and the type shares from the previous chapter, flexibly approximates possible MTEs, and computes bounds on the gene-environment interaction effect. Using a bounding approach reduces the strength of the assumptions like linearity. Its drawbacks are that the bounds may not be informative, i.e., not tight enough to use them for inference. However, we show that it is possible to estimate informative bounds using reasonable assumptions on the shape of the MTRs.

Possible MTRs are approximated using Bernstein polynomials constructed from a linear combination of simpler "basis" versions of themselves. This provides a natural way to build up more complex curves from simpler ones and allows for representing flexible and complicated forms of unobserved heterogeneity. See Figure F.3 in the Appendix for a graphical representation of Bernstein functions. We denote the polynomials that define the MTRs by $\mathbb{E}[Y_i^{jg}|U_i^E = u, G_i = g] = m^k(u, g)$. The MTEs are computed as $\mathbb{E}[Y_i^{1g} - Y_i^{0g}|U_i^E = u, G_i = g] = m^1(u, g) - m^0(u, g)$. The Bernstein polynomials themselves are defined as

$$m^j(u, g) = \sum_{v=0}^{n} \theta_v^{jg} \binom{n}{v} u^v (1-u)^{n-v},$$

where u is a specific point on the unit interval, G_i is a bin of the interaction variable (quintile of the polygenic score), j refers to the treatment state, and n is the polynomial degree. We choose $n = 5$. Therefore, we have $n + 1 = 6$ parameters $(\theta_0^{jg}, \ldots, \theta_n^{jg})$ that determine each MTR function $m^j(u, g)$. In total, there are 60 parameters: 6 times 2 (treated and untreated cases) times 5 (different values of G_i). Estimating the bounds (i.e., choosing the 60 parameters) involves solving a linear programming problem where constraints on the Bernstein polynomial shapes can be represented as constraints on the parameters θ

(Rose and Shem-Tov, 2021). We find $m^j(u,g)$ that fulfill several restrictions and maximize and minimize our target parameter

$$\beta_{G\times E}(0.6,0.8) := \frac{\int_{0.6}^{0.8}\Big[m^1(u,5)-m^0(u,5)\Big]-\Big[m^1(u,1)-m^0(u,1)\Big]du}{5-1} \tag{6}$$

This is a linearized gene-environment interaction effect on the U_i^E-range that is always covered with compliers from every quintile. The denominator ensures a normalization of the effect to a one-unit increase in G_i. We optimize the interaction effect of the difference between the first and fifth quintile as the natural choice covering the entire PGS distribution. In Section 5.4, we report robustness checks to show this choice is not crucial.

The restrictions on the linear programming problem are:

1. All values of $m^j(u,g)$ are on the support of Y_i, that is between 0 and 20.
2. Averaged over the type-specific U_i^E range, the resulting $m^j(u,g)$ reproduce the type-specific outcome means $\mathbb{E}[Y_i^{1g}|AT]$, $\mathbb{E}[Y_i^{1g}|C]$, $\mathbb{E}[Y_i^{0g}|C]$, and $\mathbb{E}[Y_i^{0g}|NT]$ (the y-coordinates from Figure 5). This also implies that $m^1(u,g)-m^0(u,g)$ reproduces the LATEs for each G_i.
3. Monotone treatment selection (see Manski, 1997): $\mathbb{E}[Y_i^{1g}|U_i^E=u]$ decreases in U_i^E for every G_i quintile. See our discussion of Figure 5 as a justification.
4. No selection into losses: The MTE $m^1(u,g)-m^0(u,g)$ is not allowed to increase in U_i^E, the distaste for the treatment. Suppose the treatment is a choice and the outcome is beneficial (or correlates with such a variable). This is likely in our setting with education as treatment and cognitive abilities as outcome. In that case, we may expect selection into gains (MTEs decrease in U_i^E). The literature on the effect of education on earnings and cognitive skills documents overwhelming empirical evidence of selection into gains (Carneiro et al., 2011; Nybom, 2017; Kamhöfer et al., 2019; Westphal et al., 2022). Note that we allow our MTEs to exhibit no essential heterogeneity (i.e., horizontal MTEs, a setting in which a 2SLS estimation of $G\times E$ is non-problematic). In Appendix E, we provide suggestive evidence that selection into losses is unlikely in our setting.
5. Additive Separability in terms of G_i: The slope of $m^1(u,g)$, $m^0(u,g)$, and $m^1(u,g)-m^0(u,g)$ does not depend on G_i. This assumption is explicitly or implicitly made when estimating MTEs (and 2SLS regressions are specified). While generally a strong assumption, Figure 5 suggests it can be reasonable.

Lastly, we make the problem finite and evaluate u at 20 equidistant grid points (as Rose and Shem-Tov, 2021).

5.3 Results

Our main results are visualized in Figure 6. Each panel compares the bounded marginal treatment effects from the first PGS quintile (in red) to the remaining four (in blue). The MTE curves that produce the minimum possible interaction effect are the dashed curves, and the solid curves are MTEs that produce the maximum. They almost coincide, suggesting that the effects are practically point-identified. Accordingly, the differences between the solid MTE curves for quintiles 2–4 and the reference category produce an estimate of the maximal gene-environment interaction effect. The difference between the blue and red dashed curves in each panel yields an estimate of the minimum interaction effect. For example, in the top panel, the area between the blue and red curves indicates how the effect of education on cognitive ability changes in the population when G_i "moves" from the first to the second quintile. Recall that we set up the linear programming approach to optimize the $G \times E$ effect in the interval $U_i^E \in [0.6, 0, 8]$. This is because the compliers from all quintiles are located in this range. In Section 5.4, we show that our results are robust to variations of this range.

The results have the same sign as our 2SLS estimates. The interaction effect is positive for each quintile comparison. This suggests that individuals with a higher polygenic score for education benefit more from an additional year of education due to the compulsory schooling reform in terms of their cognitive abilities later in life. Our approach also allows us to capture possible nonlinearities in the interaction effect across the PGS. Indeed, the estimated magnitude of the interaction differs across comparisons. Not surprisingly, the highest quintile has the largest interaction effect. However, the size of the interaction for the second quintile is substantial. Those in the third and fourth quintiles have the smallest effects.

We present estimates of the nonlinear interaction effects in Panel A of Table 4. While the previously discussed 2SLS results from Table 3 are reported in column (1) as a benchmark, columns (2) and (3) present the bounds of the marginal treatment effects from Figure 6 aggregated over the U_i^E range from 0.6 to 0.8. As in Table 3, the effect on E_i in the first row indicates the baseline effect in the bottom quintile. The direct effects on G_i in the subsequent rows are not of immediate interest, but we present them for completeness. Our focus is on the interaction effects (which are again relative to the reference category, the bottom quintile). In addition, we construct a linearized interaction effect from the quintile coefficients (Panel B) that is our main measure of the gene-environment interaction effect. This measure is simply the slope of a line through the interaction effect estimates at the lowest and highest quintiles (or the difference standardized to a one-quintile change).[9] This parameter can be calculated for each method, allowing us to compare linear effects

[9]The linear slope is calculated as $(\beta^f_{5,1} - \beta^f_{1,1})/4$. The interaction coefficient for the bottom quintile is zero since this quintile serves as the reference category.

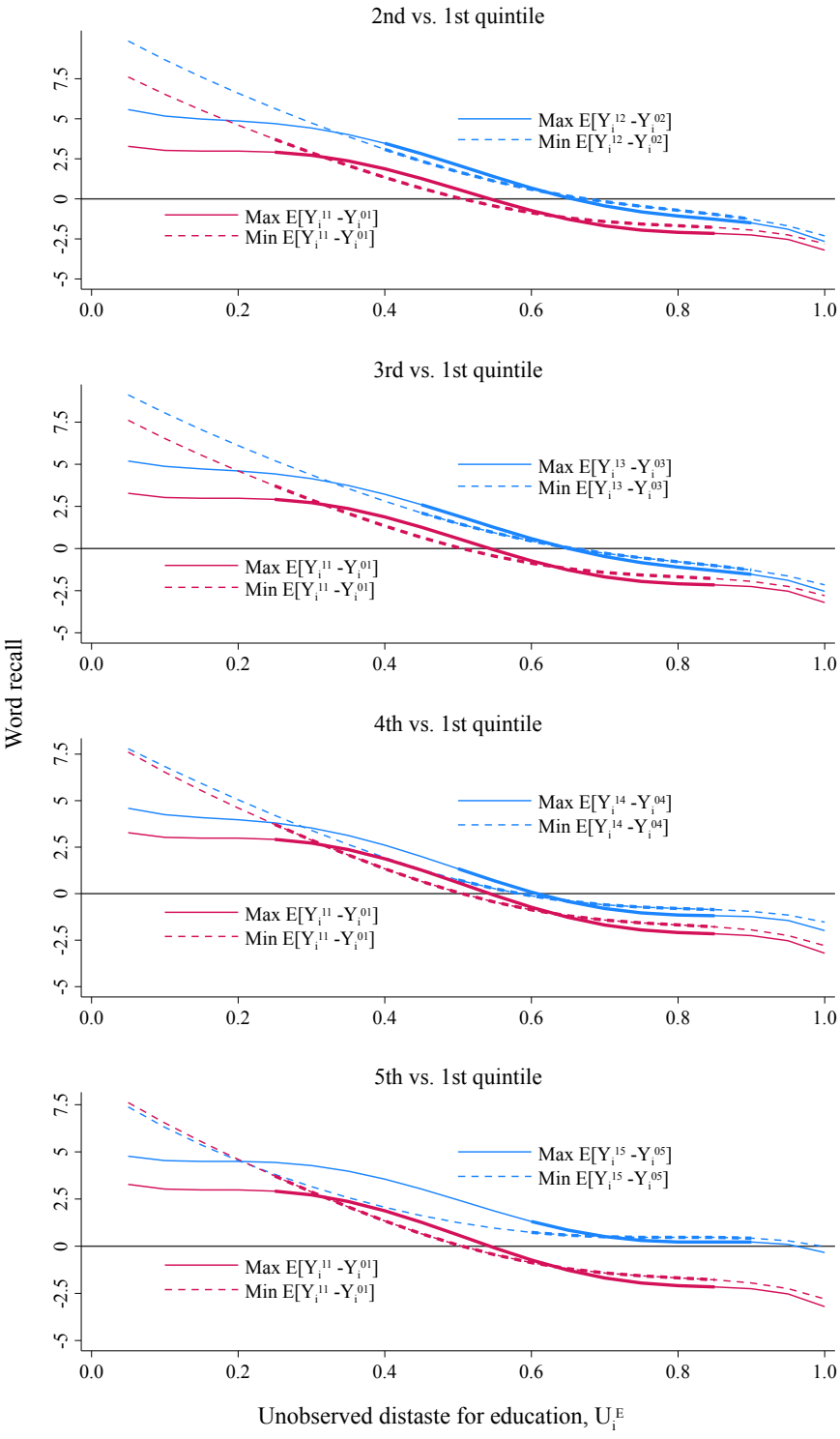

Figure 6: Quintile comparisons of the interaction effect

Notes: This figure shows the four comparisons of gene-environment interactions from our bounding approach. For every PGS quintile, we estimate bounds: maxima (solid lines) and minima (dashed lines) at which the interaction effect is maximized/minimized. The bounds for quintiles 2-4 (in blue) are compared to those of the bottom quintile (in red), our reference category, yielding four comparisons. The gene-environment interaction is the difference between the blue and red curves at $U_i^E \in [0.6, 0, 8]$. The thick part of the curves indicates the size of the complier share and its location on the U_i^E scale, both of which differ by PGS quintile.

across methods to infer whether unobserved effect heterogeneity and different proportions of compliers in G_i — which we fix by estimating marginal treatment effects — affected the 2SLS coefficients.

Table 4: Estimates of the $G \times E$ interaction

	Dependent variable – total recall score					
	2SLS (1)		MTE_{min} (2)		MTE_{max} (3)	
Panel A: nonlinear $G \times E$ effect with G_i as quintiles						
E_i	−0.042	(0.460)	0.137	(0.459)	0.137	(0.459)
$G_i = 1$	reference category		reference category		reference category	
$G_i = 2$	0.196	(0.403)	−0.388	(0.444)	−0.325	(0.436)
$G_i = 3$	0.631	(0.455)	−0.428	(0.592)	−0.425	(0.602)
$G_i = 4$	0.946	(0.462)**	−0.152	(0.470)	−0.170	(0.535)
$G_i = 5$	0.819	(0.621)	0.201	(0.858)	−0.083	(0.913)
$E_i \times (G_i = 1)$	reference category		reference category		reference category	
$E_i \times (G_i = 2)$	0.423	(0.504)	1.294	(0.610)**	1.281	(0.645)*
$E_i \times (G_i = 3)$	0.071	(0.555)	1.180	(0.738)	1.232	(0.734)
$E_i \times (G_i = 4)$	−0.043	(0.591)	0.804	(0.682)	0.866	(0.725)
$E_i \times (G_i = 5)$	0.649	(0.726)	1.834	(0.873)**	2.154	(0.959)**
Panel B: linearized $G_i \times E_i$ effect from quintile coefficients						
$E_i \times G_i$	0.162	(0.182)	0.459	(0.218)**	0.539	(0.240)**
Controls	Yes		Yes		Yes	
Observations	11,027		11,027		11,027	

Notes: This table presents estimates of the effect of staying in school until at least age 15 (E_i), an education PGS (G_i) and their gene-environment interaction ($G \times E$) on cognition later in life. Panel A shows estimates for which we use quintiles of the PGS to estimate possible nonlinear effects across G_i. Estimates that include G_i are computed relative to the reference category, the bottom quintile. Panel B shows a linearized slope of a line through the coefficients for $G = 1$ and $G = 5$ from Panel A. 2SLS estimates from Table 3 are included for reference in Column (1). The MTE estimates in column (2) refer to the minimal effects where the underlying optimization minimizes the linearized interaction effect. Estimates in column (3) are the maximal effects estimated accordingly. The controls in each case include a linear cohort trend, its interaction with the instrument (being born in 1933 or later), gender, and the first ten principal components of the genetic data. Results in different panels are obtained from separate regressions. Standard errors clustered at the individual level shown are in parentheses. For MTE bounds, standard errors are bootstrapped with 100 repetitions. $^{*}p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Overall, four features characterize our results. First, the MTE method yields informative and narrow upper and lower bounds of the interaction MTE, which almost point-identify the effect. Second, in contrast to the 2SLS estimates, even the minimum MTE results indicate substantially higher effect sizes. The linearized lower bound is about 2.5 times larger than the linearized 2SLS coefficient. While we could not detect significant gene-environment effects with 2SLS, we now find significant effects at the 5 percent level. This significance level is more conservative than necessary because it is based on a two-sided test, despite us only being interested in inference of the true MTE, not the identified set. Imbens and Manski (2004) suggests that a one-sided test is sufficient and would lead all interaction coefficients except for $(E \times G_i = 2)$ to shift one significance level (i.e., gain one star). Third, our estimates suggest that the gene-environment interaction is more substantial for individuals with higher PGS while 2SLS estimates suggest a zero or small

and statistically insignificant interaction effect. On average, "moving" to a higher PGS quintile leads to an additional increase of 0.46 words in the impact of compulsory education on recall due to the education reform. This finding reveals substantial heterogeneity and suggests a high complementarity between education and "nature" as measured by the PGS. Individuals with a higher PGS have higher returns to schooling in terms of cognitive ability later in life. This result is independent of observable and unobservable factors, both of which we can fix by estimating marginal treatment effects. Fourth, the interaction effect size does not appear linear along the PGS, as was already indicated by the visual differences in the interaction effects between each panel in Figure 6. The MTE results suggest that individuals in the highest quintile experience a large additional increase in recall of between 1.83 and 2.15 words relative to those in the first quintile. The increases for individuals in quintiles 3 and 4 may not be statistically different from the interaction for individuals in the lowest quintile, although the point estimates are positive and substantial.

5.4 Robustness

We perform several robustness checks and report the linearized estimates in Table 5. First, we estimate the interaction over wider U_i^E ranges. While the range $U_i^E \in [0.6, 0.8]$ covers most compliers from all quintiles well, we show that this particular choice is not critical to our main results (see Panel A). Over wider U_i^E, the range for which we identify MTE interaction effects is only marginally wider. Nevertheless, calculating our estimate over a wider U_i^E range could theoretically pick up local changes in the slope of the MTEs and, ultimately, the interaction effect. Thus, in all subsequent robustness checks, we report results for $U_i^E \in [0.6, 0.8]$, the range over which we compute our main results, and for the larger range $U_i^E \in [0.5, 0.9]$. We do not go beyond the latter since the never-takers are predominantly located to the right of $U_i^E = 0.9$.

Our data set consists of repeated cross-sections (waves) of ELSA. Some individuals are observed only once, others several times. Panel B shows our main result when we use only the most recent observation for each individual. This reduces the number of observations but not the number of individuals in the analysis. For our main U_i^E range, especially the maximum interaction effect is larger, so are standard errors. Over the bigger U_i^E range, the minimum is smaller and the maximum larger, increasing the possible effect range slightly. To further show that the composition of our sample does not change our results, we plot the estimates when we include individuals under the age of 65 in Panel C. The minimum is slightly smaller, but the maximum effect is very similar to our main specification with higher statistical significance. The choice of polynomial to control for cohort trends is not obvious. In Panel D, we show what happens to the results when we use quadratic trends instead of linear ones. The minimum interaction effects are slightly smaller for our main U_i^E range, and the maximum effects slightly larger. The effects for $U_i^E \in [0.5, 0.9]$ are

almost identical to our main result. Next, we remove never-takers from the analysis (see Section 5.1 for a discussion of never-takers). As the robustness check in Panel E shows, their presence helps to tighten the bounds. However, even without them, the minimum and maximum MTEs are informative. The interaction effect's lower and upper bounds are still positive, although the lower bound may not be statistically different from zero. The upper bound is larger than in our main result. This is to be expected, since never-takers have lower expected outcomes. Not including them in the analysis means that the MTE bounds do not have to reproduce these lower means. As a result, the resulting MTE curves will look different if the bounds are computed without relying on these sample moments of the never-takers. We visualize the quintile comparisons when computing interaction effects without never-takers in Figure F.4 in the Appendix.

Table 5: Robustness

	Dependent variable – total recall score	
	MTE_{min} (1)	MTE_{max} (2)
Panel A: Baseline		
– $U_i^E \in [0.6, 0.8]$ (main result, Table 4)	0.459 (0.218)**	0.539 (0.240)**
– $U_i^E \in [0.55, 0.85]$	0.456 (0.216)**	0.537 (0.236)**
– $U_i^E \in [0.5, 0.9]$	0.454 (0.213)**	0.547 (0.227)**
Panel B: Using last panel observation of each individual		
– $U_i^E \in [0.6, 0.8]$	0.470 (0.246)*	0.653 (0.285)**
– $U_i^E \in [0.5, 0.9]$	0.376 (0.243)	0.779 (0.259)***
Panel C: Keeping individuals below age 65		
– $U_i^E \in [0.6, 0.8]$	0.359 (0.214)*	0.576 (0.211)***
– $U_i^E \in [0.5, 0.9]$	0.353 (0.208)*	0.587 (0.202)***
Panel D: Squared trends		
– $U_i^E \in [0.6, 0.8]$	0.387 (0.220)*	0.657 (0.233)***
–$U_i^E \in [0.5, 0.9]$	0.457 (0.216)**	0.538 (0.212)**
Panel E: No never taker		
– $U_i^E \in [0.6, 0.8]$	0.222 (0.156)	1.093 (0.161)***
– $U_i^E \in [0.5, 0.9]$	0.189 (0.155)	1.163 (0.162)***

Notes: This table presents robustness checks for the linearized gene-environment estimates. Panel A shows our main result from Table 4 alongside results when using larger U_i^E ranges. We compute all other robustness checks for $U_i^E \in [0.6, 0.8]$, the range over which we calculate our main results and $U_i^E \in [0.5, 0.9]$. Panel B includes a robustness check where we reduce our dataset (repeated cross-sections) and only keep the most recent observation for each individual. In Panel C, we show results when increasing the age range of our sample to individuals below age 65. Panel D includes estimates controls where we for squared (age) trends. Panel E shows results when we exclude never-takers. Standard errors are bootstrapped with 200 repetitions. $^{*}p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Finally, we show the robustness of our estimation method for the underlying optimization of the interaction effect (see Eq. 6). In our main specification, we maximize/minimize the difference between the first and the fifth quintile. Here, we additionally estimate linearized effects of the final interaction effect when the optimization maximizes/minimizes the difference between the first and each of the other quintiles. We visualize the result in Figure G.5 in the Appendix. The results show that optimizing for different comparisons does not produce substantial changes in the final interaction effect, especially not in the crucial U_i^E range where we estimate our MTEs. The choice of quintile comparisons for the underlying optimization is not critical. The reason is that the MTRs must reproduce the data in form of group means, and given our shape constraints, the range of possible meaningful candidate MTRs that produce maximum and minimum MTEs is limited.

6 Conclusion

The growing gene-environment literature aims to estimate interactions between genetic endowments and environmental exposure (e.g., behavior or choice variables like education) in their effect on an outcome of interest. The goal is to assess whether the effect of the environment varies by genetic endowment (or vice versa) while all else is equal. Since environmental variables are often endogenous, a popular choice is using instruments or natural experiments as a source of exogenous variation. This usually involves estimating a two-stage least squares model. Estimating gene-environment interactions by two-stage least squares regression identifies gene-specific effects of the environment. However, they may not retain the desired interpretation as interaction effects if (1) the first stage is heterogeneous across different values of G_i and (2) the empirical setting entails essential heterogeneity in E_i (the unobserved heterogeneities for the outcome and treatment correlate). If both conditions hold, then two properties differ between gene-specific local average treatment effects: the genetic endowment and the unobserved effect heterogeneity. While the former is precisely what researchers want to isolate (the interaction), 2SLS cannot account for the latter. Thus, 2SLS estimates may not reflect complementarity between genes and the environment. We suggest solving this problem by estimating marginal treatment effects. MTEs allow for the computation of $G \times E$ estimates while accounting for unobserved heterogeneity.

While gene-environment interactions are a natural choice to illustrate this problem, since the central parameter is the instrumented interaction estimate, it theoretically applies to all interactions estimated by 2SLS. The two conditions that generate it, non-overlapping complier groups due to variations in the interaction variable and unobserved effect heterogeneity correlated with treatment propensity, could be present in other real-world scenarios involving choice variables. Nevertheless, there are likely also many settings where they

are not present or the 2SLS comparisons are inconsequential. For example, Barcellos et al. (2021) find no differences between 2SLS and linear MTE estimates of their gene-education interaction. Moreover, in many applications, heterogeneous first stages by the interaction variable are unlikely and studies that estimate only reduced form (gene-environment) interaction effects avoid wrong 2SLS comparisons all together.

Our empirical application studies the long-run effects of education, genetic predisposition for education, and their interaction on old-age cognitive abilities using data from the English Longitudinal Study of Ageing. To identify the effect of education, we use a compulsory schooling reform from 1947 that increased the minimum school-leaving age in the UK to 15. Our baseline 2SLS estimates document a zero effect of education on recalled words (our measure of cognitive abilities) for individuals in the lowest PGS quintile. Effects for higher quintiles are positive, but we lack the precision to estimate them precisely with 2SLS. We find evidence that both conditions for 2SLS to make the wrong comparisons apply in our setting. We see a strong gradient in the first stage across the quintiles of the education polygenic score and essential heterogeneity is present, more precisely, selection into gains. This is well documented for educational decisions. We estimate marginal treatment effects using the partial identification approach from Mogstad et al. (2018). Building on reduced–form evidence, we generate minimal and maximal $G \times E$ effects consistent with the data. We add further benign restrictions (such as additive separability and negative MTE slopes that imply selection into gains) to gain precision and tighten the bounds. The resulting bounds almost point-identify the interaction effect.

Our main finding is that, holding unobserved heterogeneity across G_i fixed, even the lower bound $G \times E$ effect is 2.5 times larger than the corresponding 2SLS estimate. In absolute terms, the gene-environment complementarity is substantial: on average, the effect of education on recalled words increases by 0.46–0.54 with each PGS quintile. This means that the MTE results imply higher returns to education for cognitive abilities later in life for those with a higher polygenic score. The complementarity between education and genetic predisposition that widens existing gaps in returns to education is larger than initially estimated with two-stage least squares. Not accounting for essential heterogeneity limits the usefulness of the 2SLS estimates.

References

Anderson, E. L., Howe, L. D., Wade, K. H., Ben-Shlomo, Y., Hill, W. D., et al. (2020). Education, intelligence and Alzheimer's disease: evidence from a multivariable two-sample Mendelian randomization study. *International Journal of Epidemiology*, 49(4):1163–1172.

Banks, J., Batty, G. D., Breedvelt, J., Coughlin, K., Crawford, R., et al. (2023). English Longitudinal Study of Ageing: Waves 0-9, 1998-2019.

Banks, J. and Mazzonna, F. (2012). The Effect of Education on Old Age Cognitive Abilities: Evidence from a Regression Discontinuity Design. *The Economic Journal*, 122(560):418–448.

Barcellos, S. H., Carvalho, L. S., and Turley, P. (2018). Education can reduce health differences related to genetic risk of obesity. *Proceedings of the National Academy of Sciences*, 115(42).

Barcellos, S. H., Carvalho, L. S., and Turley, P. (2021). The Effect of Education on the Relationship between Genetics, Early-Life Disadvantages, and Later-Life SES. NBER Working Paper No. 28750.

Barth, D., Papageorge, N. W., and Thom, K. (2020). Genetic Endowments and Wealth Inequality. *Journal of Political Economy*, 128(4):1474–1522.

Behrman, J. R. and Taubman, P. (1989). Is Schooling "Mostly in the Genes"? Nature-Nurture Decomposition Using Data on Relatives. *Journal of Political Economy*, 97(6):1425–1446.

Belsky, D. W., Domingue, B. W., Wedow, R., Arseneault, L., Boardman, J. D., et al. (2018). Genetic analysis of social-class mobility in five longitudinal studies. *Proceedings of the National Academy of Sciences*, 115(31):E7275–E7284.

Biroli, P., Galama, T. J., Hinke, S. v., Kippersluis, H. v., Rietveld, C. A., and Thom, K. (2022). The Economics and Econometrics of Gene-Environment Interplay. arXiv preprint.

Björklund, A. and Salvanes, K. G. (2011). Education and Family Background. In *Handbook of the Economics of Education*, volume 3, pages 201–247. Elsevier.

Brinch, C. N., Mogstad, M., and Wiswall, M. (2017). Beyond LATE with a Discrete Instrument. *Journal of Political Economy*, 125(4):985–1039.

Carneiro, P., Heckman, J. J., and Vytlacil, E. J. (2011). Estimating Marginal Returns to Education. *American Economic Review*, 101(6):2754–2781.

Carneiro, P. and Lee, S. (2009). Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. *Journal of Econometrics*, 149(2):191–208.

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

Ding, X., Barban, N., Tropf, F. C., and Mills, M. C. (2019). The relationship between cognitive decline and a genetic predictor of educational attainment. *Social Science & Medicine*, 239:112549.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, 24(1):13–23.

Heckman, J. J., Urzua, S., and Vytlacil, E. (2006). Understanding Instrumental Variables in Models with Essential Heterogeneity. *The Review of Economics and Statistics*, 88(3):389–432.

Heckman, J. J. and Vytlacil, E. (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation1. *Econometrica*, 73(3):669–738.

Imbens, G. W. and Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467.

Imbens, G. W. and Manski, C. F. (2004). Confidence Intervals for Partially Identified Parameters. *Econometrica*, 72(6):1845–1857.

Imbens, G. W. and Rubin, D. B. (1997). Estimating Outcome Distributions for Compliers in Instrumental Variables Models. *The Review of Economic Studies*, 64(4):555–574.

Jeong, Y., Papageorge, N. W., Skira, M., and Thom, K. (2024). Genetic risk for alzheimer's disease and related dementias: Cognition, economic behavior, and actionable information. NBER Working Paper No. 32181.

Kamhöfer, D. A., Schmitz, H., and Westphal, M. (2019). Heterogeneity in Marginal Non-Monetary Returns to Higher Education. *Journal of the European Economic Association*,

17(1):205–244.

Kowalski, A. E. (2023). Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform. *The Review of Economics and Statistics*, 105(3):646–664.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, 50(8):1112–1121.

Manski, C. F. (1997). Monotone Treatment Response. *Econometrica*, 65(6):1311–1334. Publisher: [Wiley, Econometric Society].

Mogstad, M., Santos, A., and Torgovitsky, A. (2018). Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters. *Econometrica*, 86(5):1589–1619.

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

Papageorge, N. W. and Thom, K. (2020). Genes, Education, and Labor Market Outcomes: Evidence from the Health and Retirement Study. *Journal of the European Economic Association*, 18(3):1351–1399.

Pereira, R. D., Rietveld, C. A., and Kippersluis, H. v. (2022). The Interplay between Maternal Smoking and Genes in Offspring Birth Weight. *Journal of Human Resources*, 58(6).

Plug, E. and Vijverberg, W. (2003). Schooling, Family Background, and Adoption: Is It Nature or Is It Nurture? *Journal of Political Economy*, 111(3):611–641.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

Rohwedder, S. and Willis, R. J. (2010). Mental Retirement. *Journal of Economic Perspectives*, 24(1):119–138.

Rose, E. K. and Shem-Tov, Y. (2021). How Does Incarceration Affect Reoffending? Estimating the Dose-Response Function. *Journal of Political Economy*, 129(12):3302–3356.

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

Schiele, V. and Schmitz, H. (2023). Understanding cognitive decline in older ages: The role of health shocks. *European Economic Review*, 151:104320.

Schmitz, H. and Westphal, M. (2024). Early and later-life stimulation: How retirement shapes the effect of education on old-age cognitive abilities. *mimeo*.

Schmitz, L. L. and Conley, D. (2017). The effect of Vietnam-era conscription and genetic potential for educational attainment on schooling outcomes. *Economics of Education Review*, 61:85–97.

Steptoe, A., Breeze, E., Banks, J., and Nazroo, J. (2013). Cohort Profile: The English Longitudinal Study of Ageing. *International Journal of Epidemiology*, 42(6):1640–1648.

Westphal, M., Kamhöfer, D. A., and Schmitz, H. (2022). Marginal College Wage Premiums Under Selection Into Employment. *The Economic Journal*, 132(646):2231–2272.

A What 2SLS is estimating

Assume that E_i and G_i are binary. There are four potential outcomes Y_i^{jg}, $j \in \{0,1\}$, $g \in \{0,1\}$ of individual i . Only one is observed. The observation rule is

$$\begin{aligned} Y_i &= E_i \cdot G_i \cdot Y_i^{11} + E_i \cdot (1 - G_i) \cdot Y_i^{10} + (1 - E_i) \cdot G_i \cdot Y_i^{01} + (1 - E_i) \cdot (1 - G_i) \cdot Y_i^{00} \\ &= Y_i^{00} + (Y_i^{10} - Y_i^{00})E_i + (Y_i^{01} - Y_i^{00})G_i + (Y_i^{11} - Y_i^{01} - (Y_i^{10} - Y_i^{00}))E_i \cdot G_i \end{aligned}$$

The second equation is the individual potential-outcome representation of the workhorse interaction model

$$Y_i = \beta_0 + \beta_1 E_i + \beta_2 G_i + \beta_3 G_i \times E_i + \varepsilon_i$$

Expressing this interaction equation as separate regressions for $G_i = 0$ and $G_i = 1$ yields

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 E_i + \varepsilon && \text{for } G_i = 0 \\ Y_i &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3)E_i + \varepsilon && \text{for } G_i = 1 \end{aligned}$$

Environment E_i is often a choice variable, therefore endogenous and instrumented by Z_i, a binary instrument. In Wald notation, separately estimating 2SLS regressions for $G_i = 0$ and $G_i = 1$ yields:

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\mathbb{E}[Y_i|Z_i = 1, G_i = 0] - \mathbb{E}[Y_i|Z_i = 0, G_i = 0]}{\mathbb{E}[E_i|Z_i = 1, G_i = 0] - \mathbb{E}[E_i|Z_i = 0, G_i = 0]} && \text{for } G_i = 0 \\ \widehat{\beta}_1 + \widehat{\beta}_3 &= \frac{\mathbb{E}[Y_i|Z_i = 1, G_i = 1] - \mathbb{E}[Y_i|Z =_i 0, G_i = 1]}{\mathbb{E}[E_i|Z_i = 1, G_i = 1] - \mathbb{E}[E_i|Z_i = 0, G_i = 1]} && \text{for } G_i = 1 \end{aligned}$$

Using the LATE theorem (Imbens and Angrist, 1994 – 2SLS estimates are average treatment effects for the compliers), we can rewrite these expressions as:

$$\begin{aligned} \widehat{\beta}_1 &= \mathbb{E}[Y_i^{10} - Y_i^{00}|C(G_i = 0)] \\ \widehat{\beta}_1 + \widehat{\beta}_3 &= \mathbb{E}[Y_i^{11} - Y_i^{01}|C(G_i = 1)] \end{aligned}$$

The mechanics of the LATE require that the group-specific effects ($\widehat{\beta}_1$ and $\widehat{\beta}_1 + \widehat{\beta}_3$) are average treatment effects for the G_i-specific compliers. Without further covariates, the joint interaction regression specification is as flexible as the separate ones. The mechanics of interaction models attribute any difference in the causal effects of E_i on Y_i between $G_i = 0$ and $G_i = 1$ to the interaction coefficient. In essence, the interaction model is numerically identical to separate estimations. In the interaction model, any difference between the

G_i-specific LATEs is mechanically attributed to $\widehat{\beta}_3$. Hence, using the expressions above, this difference amounts to:

$$\widehat{\beta}_3 = (\widehat{\beta}_1 + \widehat{\beta}_3) - \widehat{\beta}_1 = \mathbb{E}[Y_i^{11} - Y_i^{01} | C(G_i = 1)] - \mathbb{E}[Y_i^{10} - Y_i^{00} | C(G_i = 0)]$$

This demonstrates that the interaction coefficient reflects differences in G_i-specific LATEs.

B Polygenic scores

The human genome has about 3 billion base pairs, the pairs of nucleic acids that make up the DNA. However, any two people differ by only about 0.1 percent of the base pairs. Most of these genetic differences are substitutions of a single base (adenine, thymine, cytosine, or guanine) for another at a specific location in the genome, called "single nucleotide polymorphisms" (SNPs) that are common across the whole genome. These substitutions result in different genetic variants (alleles) that vary among parts of the population.[10] For example, at a specific SNP location, the DNA sequence might have an adenine base in some individuals, while others may have a thymine base at the same position. One is (arbitrarily) chosen as the reference variant. Then, each SNP can be represented as a count variable of occurrences of the reference variant at this location that can either be 0, 1 or 2, since there are two copies of each chromosome. Large research projects called genome-wide association studies (GWAS), correlate each $j = 1, \ldots, J$ SNPs with a disease or trait, e.g., diabetes, years of education, or smoking. This entails running J regressions of type

$$Y_i = \beta_j S_{ij} + X_i'\delta + \zeta_i \tag{7}$$

where Y_i is the outcome of interest (in our case educational attainment) of individual i, β_j is the individual effect of each SNP j, S_{ij} is the count variable of the reference variant of the SNP with $S_{ij} \in \{0, 1, 2\}$, X_i is a vector of controls that typically include age, gender and principal components of the genetic data, which control for population stratification, i.e., common ancestry[11]. The PGS is then calculated as a weighted sum of all S_{ij}'s, where the weights correspond to the (correlation-adjusted) β_j's obtained in the GWAS:

$$PGS_i = \sum_{j=1}^{J} \tilde{\beta}_j S_{ij} \tag{8}$$

Polygenic scores for various traits or behaviors (personality, mental and physical health, health behaviors, and more) have been calculated for the ELSA sample based on various GWAS and are readily available.

[10] The generally agreed-upon threshold for a substitution to be regarded a SNP is common occurrence in at least one percent of the population.

[11] Principal components are linear combinations of genetic markers that summarize the major patterns of genetic variation *across a population* into fewer dimensions. They reflect population stratification, i.e., different frequencies of genetic variants among subpopulations that could be responsible for spurious correlations with outcomes of interest. Price et al. (2006) show that including principal components as controls can mitigate the confounding effects of population stratification, ensuring that observed associations between genetic variants and traits are not driven by differences in ancestry or population structure.

C Additional sample information

Table C.1: Descriptive statistics (extended)

	Main sample	By E_i		
	Mean (SD)	E_i=1	E_i=0	Difference (SE)
Outcome Y_i				
Recall score	9.67 (3.37)	10.11	8.08	2.03 (0.07)***
Treatment E_i				
Left school ≥ 15	0.78 (0.41)	1.00	0.00	1.00 (0.00)
Polygenic score G_i				
1st PGS quintile	0.20 (0.40)	0.18	0.26	−0.08 (0.01)***
2nd PGS quintile	0.20 (0.40)	0.20	0.21	0.02 (0.01)*
3rd PGS quintile	0.20 (0.40)	0.20	0.19	0.01 (0.01)
4th PGS quintile	0.20 (0.40)	0.20	0.19	0.01 (0.01)
5th PGS quintile	0.20 (0.40)	0.21	0.15	0.07 (0.01)***
Instrument Z_i				
Born 1933 or later	0.66 (0.47)	0.82	0.13	0.69 (0.01)***
Controls				
Female	0.52 (0.50)	0.52	0.50	0.02 (0.01)**
Principal components (standardized):				
– 1 –	0.00 (1.00)	0.00	−0.01	0.02 (0.02)
– 2 –	0.00 (1.00)	0.01	−0.02	0.03 (0.02)
– 3 –	0.00 (1.00)	0.01	−0.04	0.05 (0.02)**
– 4 –	0.00 (1.00)	−0.01	0.02	−0.03 (0.02)
– 5 –	0.00 (1.00)	0.00	0.00	0.00 (0.02)
– 6 –	0.00 (1.00)	0.02	−0.07	0.09 (0.02)***
– 7 –	0.00 (1.00)	0.01	−0.03	−0.04 (0.02)*
– 8 –	0.00 (1.00)	0.00	0.02	−0.02 (0.02)
– 9 –	0.00 (1.00)	0.01	−0.02	0.02 (0.02)
– 10 –	0.00 (1.00)	0.01	−0.02	0.02 (0.02)
Age pattern				
Birth year	1934.89 (5.00)	1936.29	1929.92	6.37 (0.10)***
Age	71.82 (4.29)	70.89	75.10	−4.21 (0.09)***
Observations	11,027	8,590	2,437	

Notes: This table presents extended descriptive statistics including the first 10 principal components of the genetic data. We include mean and standard deviation of the main sample as well as means by E_i, the difference of means and standard errors of a t-test for equality of means. $^{*}p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table C.2: Descriptive statistics by availability of genetic information

	Full sample	By availability of genetic information		
	Mean (SD)	Yes	No	Difference (SE)
Outcome Y_i				
Recall score	9.32 (3.50)	9.67	8.76	0.91 (0.05)***
Treatment E_i				
Left school ≥ 15	0.76 (0.43)	0.78	0.72	0.06 (0.01)***
Instrument Z_i				
Born 1933 or later	0.65 (0.48)	0.66	0.62	0.04 (0.01)***
Controls				
Female	0.52 (0.50)	0.52	0.52	0.00 (0.01)
Age pattern				
Birth year	1934.67 (5.10)	1934.89	1934.32	0.57 (0.08)***
Age	71.90 (4.26)	71.82	72.02	−0.20 (0.07)***
Observations	17,884	11,027	6,857	

Notes: This table presents descriptive statistics. We include mean and standard deviation of the main sample as well as means by E_i, the difference of means and standard errors of a t-test for equality of means. $^{*}p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table C.3: Descriptive statistics by PGS quintiles

	1st quintile	2nd quintile	3rd quintile	4th quintile	5th quintile
Outcome Y_i					
Recall score	8.99	9.48	9.73	9.80	10.33
Treatment E_i					
Left school ≥ 15	0.71	0.77	0.79	0.79	0.84
Instrument Z_i					
Born 1933 or later	0.67	0.65	0.65	0.68	0.67
Controls					
Female	0.53	0.51	0.54	0.49	0.50
Age pattern					
Birth year	1934.87	1934.87	1934.96	1934.77	1934.96
Age	71.72	71.84	71.87	71.86	71.81
Observations	2,206	2,205	2,206	2,205	2,205

Notes: This table presents sample means by quintiles of the education polygenic score.

D Additional regression results

Table D.4: The 1947 UK compulsory schooling reform and providing genetic information to ELSA

	Provided genetic information (1)	Left school at 15 or later (E_i) (2)
Z_i	−0.018 (0.030)	0.453 (0.037)***
Provided genetic information $\times Z_i$		0.035 (0.046)
Controls	Yes	Yes
Observations	17,884	17,884

Notes: In this table we show that our instrument, the 1947 UK compulsory schooling reform did not affect the probability of providing genetic information to ELSA and that providing genetic information does not interact with our first stage, the effect of the reform on staying in school until at least 15. Column 1 shows estimates of a linear regression of the instrument Z_i on the probability to provide genetic information to ELSA. Controls include gender, the running variable (distance to 1933 birth cohort) and its interaction with the instrument. Column 2 shows estimates of the first stage interacted with a dummy for providing genetic information to ELSA. Controls include gender, running variable and interactions with the running variable. Both regressions are estimated in a larger sample that fulfils all criteria outlined in section 3.2 but still includes individuals without genetic data available. Standard errors in both regressions are clustered at the individual level. $^{*}p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table D.5: Estimates of the first stage by PGS quintile

	Left school at 15 or later (E_i)	
	Coefficient (1)	Standard error (2)
$Z_i \times (G_i = 1)$	0.642	(0.018)***
$Z_i \times (G_i = 2)$	0.537	(0.018)***
$Z_i \times (G_i = 3)$	0.477	(0.018)***
$Z_i \times (G_i = 4)$	0.419	(0.018)***
$Z_i \times (G_i = 5)$	0.356	(0.018)***
Controls	Yes	
Observations	11,027	

Notes: This table presents estimates of the effect of the 1947 UK compulsory schooling reform on the probability of attending school until at least age 15 by quintiles of the education polygenic score. These effects are obtained from the coefficients $\pi^f_{1,\Delta}$ to $\pi^f_{5,\Delta}$ of eq. 4, which correspond to the complier shares in the respective quintile. Standard errors clustered at the individual level shown are in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

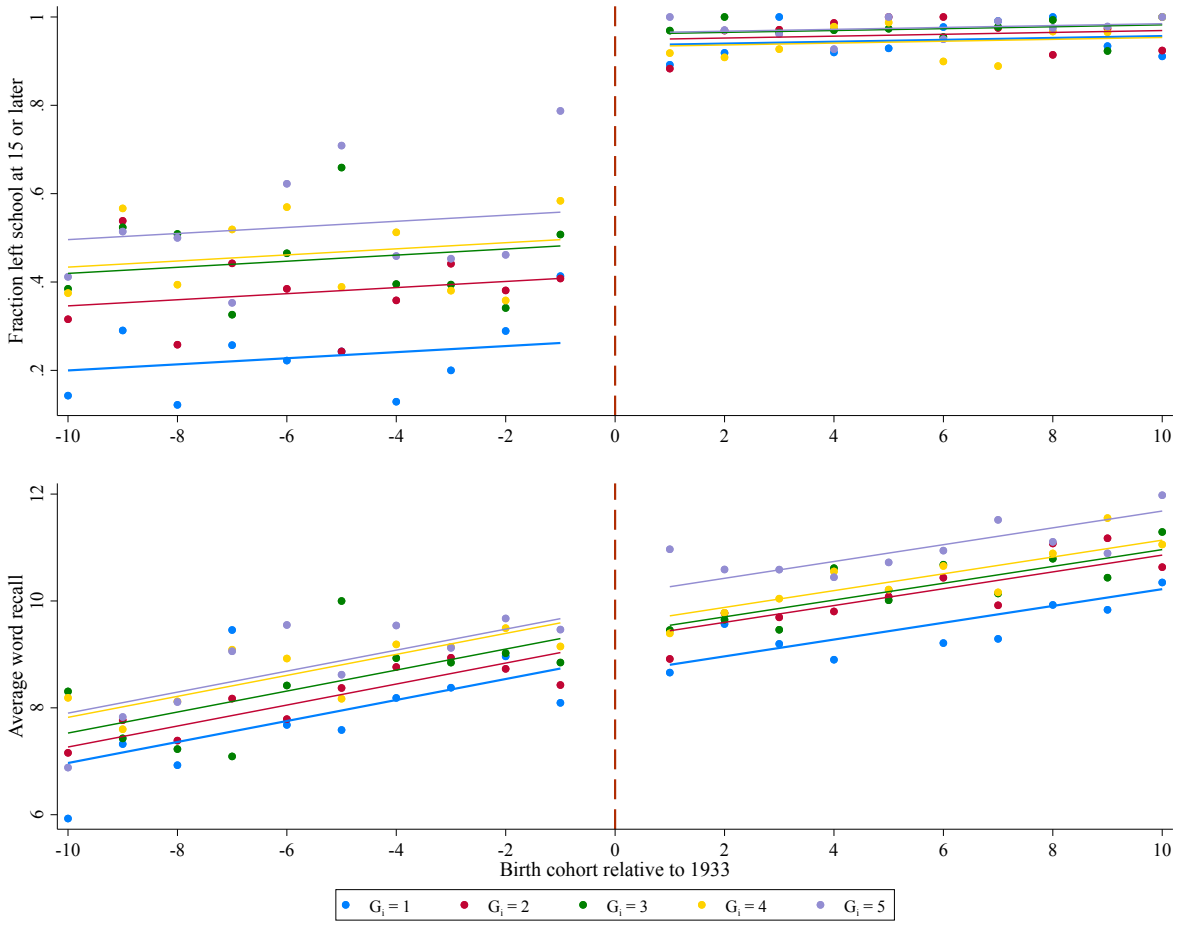

Figure D.1: First Stage and reduced form by G_i

Notes: This figure shows a visualization of the first stage (upper panel) and the reduced form (lower panel) results of our regression discontinuity design by quintiles of the education polygenic score $G_i \in 1, 2, 3, 4, 5$ detailed in Section 4.

E Testing "no selection into losses" (non-positive MTE slopes)

An important constraint we apply in our linear programming approach is "no selection-into-losses", i.e. no MTEs that increase in U_E. To test this in our setting, we follow Imbens and Rubin (1997) and use the instrument to compute mean outcomes for always-taker, treated and untreated complier, and never-taker. For simplicity, we test this condition globally and do not distinguish between cells of G_i (we show the complete G_i-specific means in Figure 5). We present the results in Table E.6. In Panel A, we focus on differences between always-takers and treated compliers (Column 3) and untreated compliers and never-takers (Column 6). The differences are informative about whether the treated outcome $\mathbb{E}[Y_i^1|U_E = u]$ and the untreated outcome $\mathbb{E}[Y_i^0|U_E = u]$ – the difference of which is the MTE – are heterogeneous in U_E.

Column (3) presents the mean recall difference between always-takers and treated compliers. It shows a substantial and statistically significant heterogeneity: Always-taker recall about 1.25 words more. Intuitively, this is unsurprising, as always-taker to a compulsory schooling reform will, on average, have more years of education, will be more likely to hold advanced degrees, or may be positively selected in terms of unobserved characteristics (if we have selection into gains, which we want to argue). Furthermore, this result shows that $\mathbb{E}[Y_i^1|U_E = u]$ has a negative slope. Likewise, we do the same with untreated compliers and never-taker. Here, the heterogeneity is less pronounced and not statistically significant. If we conclude that both groups do not have different outcomes, we can stop as in this case, the difference in the first two groups proves that we have essential heterogeneity. If the insignificant difference is meaningful, things may change. The difference is also negative, contrasting the existing empirical evidence for the slope of the untreated outcome (see, e.g., Carneiro and Lee, 2009; Westphal et al., 2022). However, it is essential to mention that never-taker should not exist with a compulsory schooling reform, where everyone should be forced to stay in school until age 15. If this group has never existed, this might be a measurement error. If these individuals had special exemptions from the rule change (and therefore existed), the difference between never-taker and untreated compliers may not inform about the global course of the curve. Assessing the multiple complier groups that we gain by stratifying by G_i (see Figure 5) indeed suggests that never-taker are different and $\mathbb{E}[Y_i^0|U_E = u]$ indeed increases when $U_E < 0.95$.

Nonetheless, with only a binary instrument and without exploiting covariate heterogeneity together with the additive separability assumption (which we will do below), an additional linearity assumption is necessary (due to the never-taker) to point-identify a marginal treatment effect via the method introduced by Brinch et al. (2017). We document a formal

Table E.6: Mean outcomes by instrument response types and test for essential heterogeneity

	Unobserved heterogeneity					
	in the treated outcome			in the untreated outcome		
	(1) Always-taker	(2) Treated Complier	(3) Difference (2) − (1)	(4) Untreated Complier	(5) Never-taker	(6) Difference (5) − (4)
Panel A:						
Mean recall:	9.500	8.306	−1.245***	8.109	7.679	−0.353
	(0.215)	(0.332)	(0.454)	(0.215)	(0.340)	(0.396)
Share:	0.456	0.489		0.489	0.055	
	(0.035)	(0.036)		(0.036)	(0.011)	
Panel B:						
Test for essential heterogeneity:						
(sufficient condition, may be uninformative if heterogeneity is nonlinear)						
Slope of $\mathbb{E}[Y_i^1 \mid U_E = u]$			−2.631***			
			(0.961)			
Slope of MTE $\mathbb{E}[Y_i^1 - Y_i^0 \mid U_E = u]$			−1.326			
			(1.423)			

Notes: This table presents estimates of mean outcomes for always-taker, treated and untreated complier, and never-taker (panel A) as well as results of a test for essential heterogeneity (panel B). We compute the type-specific shares using the specification of Eq. (2) without G_i. The complier share is the coefficient on Z_i, the always-taker share is the constant (as all variables are demeaned), and the never-taker share is the remainder. For the type-specific outcome means, we compute means by E_i and Z_i (and their interaction) using a reduced-form specification to regress recall on the same controls and full interactions of E_i and Z_i. As compliers never appear alone in these means, we use the formula provided in Imbens and Rubin (1997) and the type-specific shares. Standard errors are computed using 200 bootstrap replications and are shown in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ indicate significance levels for the differences.

test of the slope of $\mathbb{E}[Y_i^1 | U_E = u]$ and $\mathbb{E}[Y_i^1 - Y^0 | U_E = u]$ in Panel B.[12] It shows that the slope of the treated outcome is negative and statistically significant (as shown in Panel A). The slope of the linear MTE is also negative and still large in magnitude. However, likely due to the concerns about never-taker outlined above, it is not statistically significant, albeit with a negative sign. Again, evidence from the G_i-specific complier groups strongly suggests that the $\mathbb{E}[Y_i^0 | U_E = u]$ increases at least for a relevant range when $U_E < 0.95$. We conclude that we likely face essential heterogeneity in our setting. Combined differences in the first stage induced by G_i, the result may be that 2SLS cannot recover the true interaction parameter. We would need to make accurate statements about the interaction effect.

[12]The exact formula reads

$$\frac{\partial \mathbb{E}[Y_i^1 | U_E = u]}{\partial U_E} = \frac{Y_i^{CT} - Y_i^{AT}}{\frac{\pi^C + \pi^{AT}}{2}}, \qquad \frac{\partial \mathbb{E}[Y_i^1 - Y_i^0 | U_E = u]}{\partial U_E} = \frac{Y_i^{CT} - Y_i^{AT}}{\frac{\pi^C + \pi^{AT}}{2}} - \frac{Y_i^{NT} - Y_i^{CU}}{\frac{\pi^C + \pi^{NT}}{2}},$$

where Y_i^{AT}, Y_i^{CT}, Y_i^{CU}, and Y^{NT} are means from Columns (1), (2), (4), and (5), respectively and π^{AT}, π^C, π^{NT} are the corresponding shares (compliers do not need to be differentiated).

F Details on the MTE estimation

We run the following two regressions:

$$E_i = \sum_{g=1}^{5}\sum_{k=0}^{1}\left[\pi^f_{g,k}\mathbb{1}[G_i = g]\times[Z_i = k]\right] + \text{controls} + \omega_i \tag{9}$$

$$Y_i = \sum_{g=1}^{5}\sum_{j=0}^{1}\sum_{k=0}^{1}\left[\delta^f_{g,j,k}\mathbb{1}[G_i = g]\times[E_i = j][Z_i = k]\right] + \text{controls} + \eta_i. \tag{10}$$

The first equation estimates G_i-specific first-stage from which the complier types can be inferred. The second equation estimates conditional means of Y_i, conditional on G_i, Z_i, and Y_i when covariates are fixed. On these estimates, we apply the Imbens and Rubin (1997) formula to compute G_i-specific outcome means for always-taker (AT), never-taker (NT), and (treated on untreated) compliers (C) the are plotted in Figure 5:

$$\begin{aligned}
\mathbb{E}[Y_i^{1g}|C, G_i = g] &= \frac{\delta_{g,1,1}\pi_{g,1} - \delta_{g,1,0}\pi_{g,0}}{\pi_{g,1} - \pi_{g,0}} \\
\mathbb{E}[Y_i^{0g}|C, G_i = g] &= \frac{\delta_{g,0,0}\pi_{g,0} - \delta_{g,0,1}\pi_{g,1}}{\pi_{g,1} - \pi_{g,0}} \\
\mathbb{E}[Y_i^{0g}|NT, G_i = g] &= \delta_{g,0,1} \\
\mathbb{E}[Y_i^{1g}|AT, G_i = g] &= \delta_{g,1,0}
\end{aligned}$$

These linear potential outcome curves could already solve the problems associated with 2SLS estimation of interactions while using richer variations of the polygenic score. Based on them, we can calculate the (interaction) effects according to Table 1 in the interval $0.6 \leq U_D \leq 0.8$. Graphically, this would entail subtracting the blue from the red lines for each quintile. However, this would require extrapolating the lines with $E_i = 0$ to the left or the lines with $E_i = 1$ to the right, demonstrating the general extrapolation problem that we could solve here by a linearity restriction. If we are willing to make this extrapolation, it yields five MTE curves for the effect of E_i on Y_i, one for each quintile, which can then be used to calculate the interaction effects.

In the paper, we are unwilling to make such an assumption and apply the partial identification method by Mogstad et al. (2018). As one input, the method uses the conditional means that the coefficients ($\delta^f_{g,j,k}$ and the corresponding $\pi^f_{g,k}$) reflect. These are the "moments" for the linear programming method by Mogstad et al. (2018). Figure F.2 plots the results of this approach, where the slightly transparent, horizontal lines are the "moments" (G_i-specific outcome means and their placement on the unit-interval, which we derive from the complier shares). The blue (for the treated outcome) and red (for the untreated) lines

are the output of this linear programming approach. They reflect the minimal (the dashed lines) and maximal (the solid lines) possible interaction effect (defined in the main text) that the MTR lines (Bernstein polynomials, see Figure F.3) produce while being consistent with the shape restrictions and matching the moments.

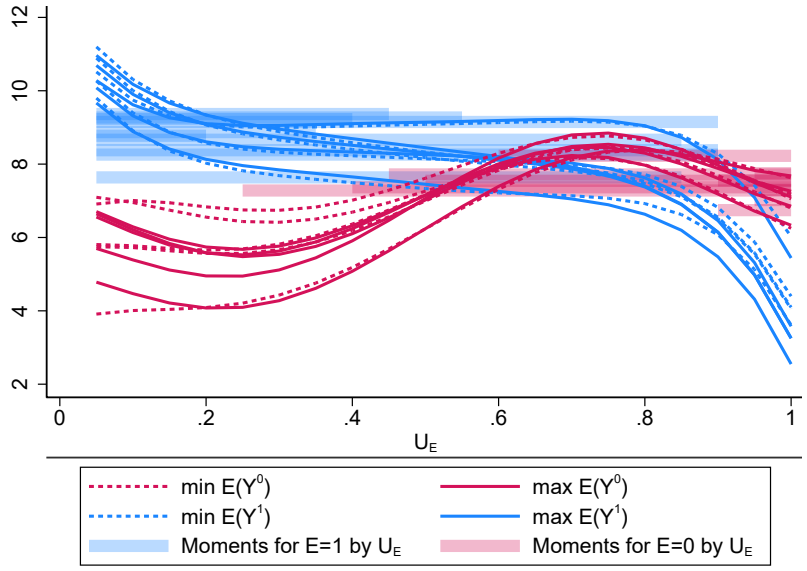

Figure F.2: Potential outcome curves estimated with Bernstein polynomials

Notes: This figure shows the minima and maxima of the ten potential outcome curves estimated via linear program with Bernstein polynomials. Blue indicates curves and moments for $E_i = 1$, and red indicates $E_i = 0$. Solid lines are maxima; dashed lines are minima of the potential outcome curves. There are five pairs of curves for $E_i = 1$ and five for $E_i = 0$, one pair for every PGS quintile. Every pair consists of a minimum and a maximum that bound the potential outcome curve for its respective quintile. The vertical bars indicate the moments the curves must pass and the U_E ranges of individuals contributing to these means.

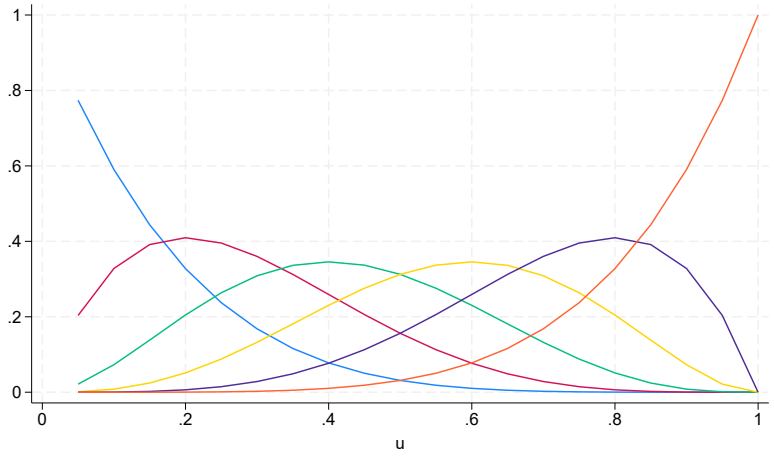

Figure F.3: Graphical representation of the Bernstein base functions

Notes: This figure depicts the six Bernstein base functions that compose a Bernstein polynomial of degree five. The formula for each base function reads $b_{v,n}(u) = \binom{n}{v} u^v (1-u)^{n-v}$, where $n = 5$ is the degree, v denotes the specific base function and u is a specific grid point on the unit interval. The formula that obtains the MTE by the sum of all base functions weighted by the corresponding parameter θ_v^{jg} reads $m^j(u, g) = \sum_{v=0}^{n} \theta_v^{jg} b_{v,n}(u)$, where G_i is the genetic bin, j the treatment state (as defined above) in addition to the variables and parameters defined above.

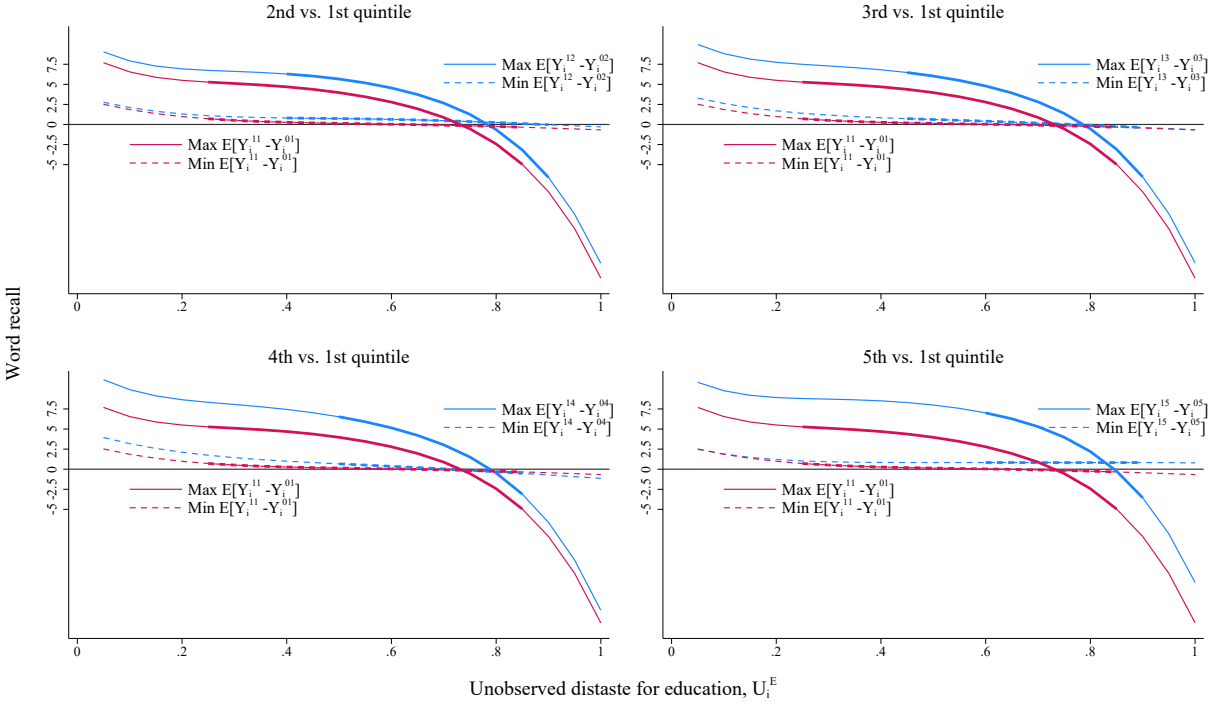

Figure F.4: Quintile comparisons of the interaction effect without never-taker

Notes: This figure shows the quintile comparisons of the interaction effect from Figure 6 when never-taker (their sample moments) are not used to construct the MTE bounds. For every PGS quintile, we estimate bounds: maxima (solid lines) and minima (dashed lines) at which the interaction effect is maximized/minimized. The bounds for quintiles 2-4 (in blue) are compared to those of the bottom quintile (in red), our reference category, yielding four comparisons. The gene-environment interaction is the difference between the blue and red curves at $U_E \in [0.6, 0, 8]$. The thick part of the curves indicates the size of the complier share and its location on the U_E scale, both of which differ by PGS quintile.

G Robustness checks for the linear programming approach

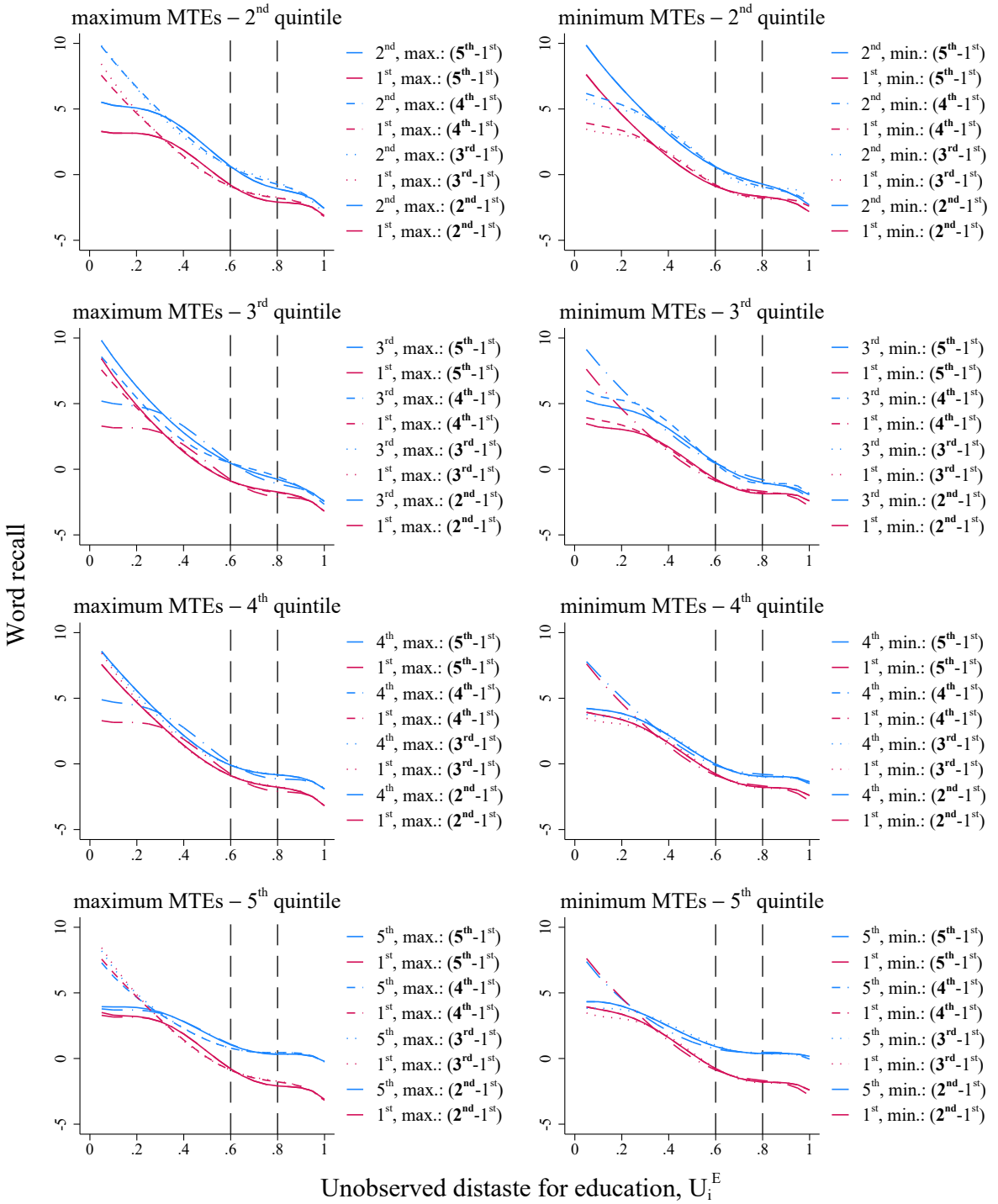

Figure G.5: MTEs when the target $G \times E$ parameter is adjusted to specific quintiles

Notes: This figure shows robustness checks for our main result in Figure 6. Here, we optimize the interaction effect for different comparisons. Whereas our preferred specification optimizes the difference between the first and the fifth quintile (see Eq. 6), we generalize this approach and optimize differences between the first any other quintile such that $\beta_{G\times E}(0.6, 0.8, g) = \frac{1}{g-1}\int_{0.6}^{0.8}[m^1(u,g) - m^0(u,g)] - [m^1(u,1) - m^0(u,1)]du \quad \forall g \in \{2,3,4,5\}$. The solid lines correspond to optimizing $g = 5$, our main result. The dashed lines show the optimization for $g = 4$, the dotted for $g = 3$, and the dashed-dotted line for $g = 2$. The respective quintile G_i used for the target parameter $\beta_{G\times E}(0.6, 0.8, g)$ is highlighted in bold. Maximized and minimized MTEs are shown separately, maximized MTEs in the left and minimized MTEs in the right column. The rows present pairwise comparisons between the first and another PGS quintile (the second quintile in the first row, the third in the second row, ...).