

# Genetic Influences on Human Capital Development

Philip Monagan

October 2026

## Abstract

This paper explores the channels and mechanisms by which a child's genetic endowment influences the processes of human capital development during adolescence; I investigate the interplay between genes, parental investment, and the environment, and how they shape skill formation and ultimately educational attainment. Using data from the National Longitudinal Study of Adolescent to Adult Health (Add Health), I estimate how differences in genetic predisposition for educational attainment causally influence years of schooling, college attendance, and earnings, and how the causal effects are mediated through parental investment and the environment. A one-standard-deviation increase in a child's polygenic score is associated with a 0.66 year increase in completed years of schooling; when I account for gene-environment correlation by controlling for parental education and household income, the increase is attenuated to 0.42 years; I find this estimate is robust to a specification leveraging strictly random within-family variation. Applying the same controls, the same increase in polygenic score causes four-year college completion to increase by 7.8 percentage points and projected present-discounted lifetime earnings to increase by 9.3 percent; this finding is again robust to the use of strictly exogenous within-family variation: the analogous estimates are 6.8 percentage points and 10.8 percent. Next, I decompose the endowment-education relationship, and its attenuation, by analyzing the reduced-form relationships between the endowment, skill accumulation, and parental investment, and how they change across different stages of adolescence. However, this analysis of parental choices is not policy invariant; further, the reduced-form coefficient obscures the degree to which the investment response represents a change in total parental investment or a within-family sibling reallocation. To deepen understanding of parental choices, disentangle the margins of investment response, and tease out policy implications, I develop a two-child model of parental investment as an optimal resource allocation problem, and jointly estimate the dynamic structural model of skill production and household choice using the Simulated Method of Moments. Finally, I use the structural parameter estimates to evaluate the counterfactual welfare effects of different policy interventions.

# 1 Introduction

Economic analyses of the development of human capital traditionally focus on the input variables commonly observable to researchers: investment, the environment, and their interplay. We know that genes play a significant causal role in human capital formation, both directly and through influence on other inputs. Genetic scores associated with educational attainment (EA) predict realized years of schooling, cognitive and non-cognitive skills, and a variety of other important outcomes (Lee et al., 2018; Barth et al., 2020). While observed outcome-gene associations can reflect noncausal correlations between genes and environmental factors, estimations focused exclusively on random genetic variation between siblings show that genetic effects are indeed causal (Howe et al., 2022; Trejo et al., 2018).

The focus of this paper is not on understanding whether genes matter, but how they matter: genetic influences are embedded in a complex process involving schooling decisions, parental investments, and social context; our question asks how genes, fixed at conception, interact across later stages of human capital development with skill formation, accumulated investments, and household choices. How do parental investments change in response to the genetic endowment, and do investments and the endowment act as complements or substitutes? How does the response function change as the child enters and exits adolescence, and how does it change in response to differences in the endowment of siblings?

To the degree the existing literature has investigated these questions, the emphasis has been on early childhood. Houmark et al. (2024) try to disentangle direct and indirect genetic effects by estimating a model of skill formation in young children, finding that parental investments are responsive to the genetic endowment of the child. We know, however, that the skill production function is not stationary: marginal investment impacts, skill-investment complementarity, and investment-input substitutability change across the cycle of development (Cunha and Heckman, 2007; Cunha et al., 2010). In adolescence, it is not clear the degree to which parents can still influence the trajectory of human capital development, whether investments at this stage are crowded out by external factors like schools and peers, or whether future outcomes have already been largely determined by the time a child enters high school.

The evidence on adolescents is relatively thin and inconclusive: Arold et al. (2025) show that specifically at the high school level, school quality and the genetic endowment substitutes: genes matter less when school quality is high. Muslimova et al. (2026), using birth order as an all-childhood-stage proxy for the family environment variation, suggest the opposite — the environment and the genetic endowment are complementary — but their measure does not emphasize adolescence: they identify parental time investments in early life as the

dominant mechanism. Agostinelli and Weingarten (2025) do investigate adolescents specifically and find evidence of parental reinforcement and negative complementarity, but their outcome of interest is ADHD, not educational attainment and skill formation in general.

Ultimately, while the genetic endowment is fixed at the start of life, its expression and importance can change and differ at every age: the same genetic variant could separately alter adolescent-stage skill development by altering the pre-stage accumulated level of skill, the marginal skill change, the level of investment, and the responsiveness to investment. We treat the genetic-endowment as a time-invariant input that enters the skill production function in every period, allowing loadings to differ at each estimation change and across each of the abovementioned channels. To the extent non-genetic investigations of these forms of complementarities exist, they must rely on some measure taken after birth; using genes directly as the time-invariant input has the relative advantage of their being uncontaminated by the environmental factors necessarily acting on the post-birth measure after birth.

We measure genes with a polygenic score (PGS): an weighted index sum of an individual’s genetic variants, with weights given by the variant’s association with the outcome, as estimated by a genome-wide association study (GWAS). So far, PGS scores have primarily been embedded into economic models of human capital development using a reduced-form methodology (French et al., 2026; Bolt et al., 2025; van Alten et al., 2025), in which the emphasis is on the estimation and measurement of the latent factor and on gaining an understanding of how the polygenic score influences its trajectory. Parents, however, respond to early signs of cognitive and non-cognitive skill development by adjusting their investments and reallocating resources across siblings, and we are particularly interested in whether that response reinforces or compensates for genetically influenced skill gaps between siblings.

The causal effect of genes on parental investment is itself a mechanism by which the genetic endowment causally influences educational attainment and later-life outcomes. Parental choices are themselves determined jointly by the nature of the skill production function, preferences, and resources: this relationship is obscured by the reduced-form correlations we report. A structural model is required to estimate the deeper parameters that enable us to disentangle parental preferences from constraints, and evaluate the welfare effects of different policy counterfactuals. Given the identified parameters, we evaluate each policy intervention with parents re-optimizing in response to the change in constraints: in particular, we find that parental responses depend significantly on the productivity of parental time and on its complementarity with the endowment, both of which are estimated by the model.

Returning to the research questions we outlined above, this paper answers them with a structural model of how parents allocate time and monetary investments across two siblings who can differ in genetic endowments: we structure the model to separate the direct effect of

the endowment from the evoked parental response; we use within and across-family variation to distinguish between the total and across-sibling response margins; and we identify the degree of complementarity by allowing for interactions between the endowment and investment in the skill production technology.

The outcome variable we choose to emphasize is educational attainment. Educational attainment is a significant outcome variable that has been traditionally emphasized in the economics literature: each additional year of schooling is associated with approximately a 10 percent increase in lifetime earnings (Card, 1999), lower unemployment risk, better health, and higher intergenerational mobility for the next generation. As a phenotypic trait, educational attainment is moderately heritable: a meta-analysis of twin studies attributes about 40 percent of its variance to additive genetic factors and a comparable share, about 36 percent, to the shared family environment (Branigan et al., 2013), and a pooled analysis of 28 twin cohorts gives a similar heritability (Silventoinen et al., 2020). The most recent genome-wide association study (GWAS) of educational attainment identified 3,952 independent loci (Okbay et al., 2022), and a score built from its weights explains 12–16 percent of the variance in an independent sample.

This paper makes three primary contributions. First, we estimate the relationship between children’s polygenic EA predisposition and completed schooling. To show that the estimated relationship is indeed causal, we use a within-family Mendelian randomization design, exploiting the conditionally random assignment of genes at conception among full siblings, to identify the direct causal effect of children’s polygenic EA predisposition on completed schooling. This identification strategy, building on the within-family GWAS literature (Howe et al., 2022; Trejo et al., 2018), addresses the possibility of gene-environment correlations between the initial endowment and other inputs into the production function of educational attainment. We refer to this identification approach as the sibling design, and to the estimator as the within-family regression: further details are outlined in Section 2.1.

The within-family estimate has a robust causal interpretation, but the sibling sample is small, making the estimate relatively imprecise. However, the lack of attenuation between the across-family and within-family coefficients functions as evidence that a substantial fraction of the observed pooled gene-outcome relationships is indeed causal.

Second, we build on the reduced-form relationships and create a genetically-informed structural model of human capital development. Parents allocate money and time between their own consumption and leisure and investment in each of two children, who share the household budget and the parents’ time but differ in the alleles they inherited, in the tradition of the separable-earnings-transfers model of Behrman et al. (1982) (see Francesconi and Heckman, 2016). Preferences are defined over consumption, leisure, and each child’s cognitive

and non-cognitive skills, and a post-secondary enrollment decision at age 18 gives skills their terminal value. Cognitive and non-cognitive skill development is specified by a skill production function that is increasing in time investment, financial investment, and genetic endowment. The model is estimated using the Simulated Method of Moments (SMM). In particular, rather than discarding the pooled (as opposed to within-family) moments as confounded noise, the structural model estimation leverages the delta between the pooled, controlled OLS and within-family moments as additional identifying variation.

Third, we use the parameters identified by the structural model to evaluate the counterfactual welfare effects of three potential policy interventions. The structural estimates are invariant to the policy regimes we vary, and we can use the model to account for the ways responsive parents will alter their planned investment trajectories when accounting for different policy environments. In reality, policy impacts will be context-dependent and rely on factors – such as tutoring efficacy relative to parental time – that would typically be understood by the local policymaker; however, despite relying on externally calibrated parameters, we feel that at minimum the exercises serve to illustrate how the structural model we estimate can be used to inform well-thought-out policy.

The three parts are linked by the objects they share in common. Section 4 estimates the pooled and within-family schooling gradients,  $\beta^{\text{pool}}$  and  $\beta^{\text{wf}}$ , and the within-family parental investment response to sibling-sibling genetic differences, denoted  $\pi_H$ . As mentioned above, the structural model of Section 5 is estimated against the pooled and within-family moments: the gap  $\beta^{\text{pool}} - \beta^{\text{wf}}$  identifies the covariance between the genetic endowment and the unobserved family environment (genetic nurture), denoted  $\sigma_{g\psi}$ , and the skill transitions identify the skill production function, including the level of complementarity between endowment and parental investment. We exclude the investment response gradient  $\pi_H$  from estimation to utilize it as a test of the implied model behavior. The analysis is enabled by the sibling-sibling structure of the model: we simulate the across-sibling genetic variation and let the parents allocate resources across siblings jointly: a one-child model could only reproduce population-level patterns and not the between-sibling moments and relationships in which we are interested.

Having outlined the questions that frame our analysis and main contributions, we turn to a summary of our findings. In the full sample, a one-standard-deviation increase in the EA polygenic score is associated with 0.66 additional years of completed schooling in a regression with demographic, cohort and wave controls only, and with 0.42 years once parental education and household income are added (standard errors 0.03). In the within-family regression, comparing full siblings, the estimate is 0.47 years (standard error 0.12; 697 respondents in 339 families). We reproduce the same pattern when comparing the different

analyses on only the sibling sample itself: the pooled OLS estimate decreases from 0.78 to 0.44 when controlling for parental education and income; again, the controlled OLS estimate is statistically indistinguishable from the within-family estimate. The reduction from the uncontrolled pooled estimate is at the lower end of the 50–65 percent shrinkage that Howe et al. (2022) report. However, the key finding relative to Howe et al. (2022) and the existing literature is that observable parental information absorbs the relevant family-level gene-environment confounding; this result supports a plausible causal interpretation that does not require siblings: perhaps cross-family OLS with suitable controls is sufficient. However, our sibling-only study is too underpowered to conclude this from our data alone: the 95 percent confidence interval on the within-family estimate runs from 0.22 to 0.71 years, and the sibling design could only have detected, with 80 percent power, an attenuation of at least 83 percent of the controlled pooled estimate, so the data are consistent both with no attenuation, negative attenuation, and of course with substantial further attenuation.

The same two findings hold beyond years of schooling, and for these outcomes too we report the controlled pooled estimates, which use the full sample of adolescents, as the magnitudes. A one-standard-deviation increase in the score raises the probability of four-year college completion by 11.9 percentage points with demographic controls and by 7.8 points once parental education and household income are added, and projected present-discounted lifetime earnings, built from Wave 5 earnings and CPS age-earnings profiles, by 15.8 and 9.3 percent (all four  $p < 0.001$ ); at the median of \$1.11 million, the controlled earnings gradient is about \$103,000, and the projection rests on early-career earnings, as Section 4 explains. Parental education and income again take a third to two-fifths off each gradient, and comparing siblings takes nothing more: the within-family estimates, 6.8 percentage points and 10.8 percent, are statistically indistinguishable from the controlled pooled ones and several times less precise. Within-family associations between the EA score and adult attainment have been documented before (Belsky et al., 2018), and pooled associations with wealth are known from the HRS (Barth et al., 2020); what we add is the comparison of pooled and within-family estimates for these outcomes in this cohort. Within families, the probability of having a partner with above-median educational attainment, conditional on having a partner at Wave III, rises by 19.0 percentage points per standard deviation ( $p = 0.02$ ).

Turning to parental investment responses, in the full sample, a one-standard-deviation higher score is associated with 3.7 percent of a standard deviation more parental homework help ( $p = 0.08$ ). The within-family response is reinforcing in sign but imprecise: 9.2 percent of a standard deviation more homework help per standard deviation of the score ( $p = 0.29$ ), the sign the two-child model implies when the endowment and parental time are comple-

ments ( $\alpha_g^k > 0$ ), and statistically indistinguishable from the controlled pooled estimate. We therefore do not read the reduced-form evidence as showing reinforcement or compensation; the role of the within-family coefficient is as a held-out check on the investment response the structural model implies. There is also a reason to expect a raw-score coefficient to be weak: Agostinelli and Weingarten (2025) find, for ADHD, that the within-family investment response visible with a latent measure of genetic risk disappears when the conventional polygenic score is used in its place, which is what measurement error in the score predicts (Section 4).

Looking ahead to the structural model, we preview our findings by noting an effect of particular relevance to the literature: positive complementarity in the skill production function between genetic endowment and the skill-formation returns to parental investment. This estimate is consistent with existing findings documenting how parental investment serves as a channel of intergenerational inequality amplification, even within-family (Houmark et al., 2024).

We summarize the returns on three possible policy implementations, finding they each fail the cost-benefit test. A tutoring subsidy targeted at bottom-quartile children is almost fully crowded out: parents withdraw their own time nearly hour for hour, so post-secondary enrollment increases by 0.21 percent and completion by 0.14 percent: we estimate a welfare gain of 0.63 percent of annual resources (in household leisure from withdrawn parental time), which is paltry relative a fiscal cost of \$7,874 per family per year. A need-based scholarship raises completed schooling by 1.21 percentage points and post-secondary enrollment by 1.36 percentage points, but the welfare gain is only 0.08 percent of resources. A universal early academic intervention raises skills at school entry that largely dissipate by age 18 (0.02 percentage points on both the intensive and extensive margins), but its welfare gain is 5.82 percent of resources, which accrues as the flow value parents place on their children’s skills during childhood. None of the three returns more than it costs in the income-equivalent metric (\$0.05, \$0.01 and \$0.38 per dollar). Again, however, we caveat these results by noting they are highly sensitive to parameters taken from external sources, and note that the exercise should be interpreted as illustrative, rather than hard fact.

The main findings of the paper relate to several literatures. The approach extends the foundational frameworks for the determinants of educational attainment (Cunha and Heckman, 2007; Carneiro and Heckman, 2003; Heckman, 2006) where we incorporate genetic information that substitutes for unobservable endowments. The paper is most directly in conversation with Houmark et al. (2024), who embed a polygenic score in a dynamic latent-factor model in ALSPAC and separate the direct effect of a child’s genes from the indirect effect that runs through parental investment (the “nurture of nature”) and the effect of

parental genes on investment (the “nature of nurture”). We compare the results of that approach to early childhood to adolescence, replacing their parental-genotype controls with sibling variation, and derive the parental investment response from a household choice problem rather than estimating it in reduced form, which is what allows us to evaluate counterfactual policies. Two further papers bear directly on our question. Arold et al. (2025) use Add Health to show that high-school quality moderates the EA polygenic index: a one-standard-deviation increase in school quality reduces the effect of a one-standard-deviation increase in the index by about a fifth, so that school inputs substitute for genetic endowment in adolescence. Muslimova et al. (2026) find the opposite pattern for the family environment in the UK Biobank: the educational advantage of being first-born is larger for higher-PGS siblings, which they interpret as gene-environment complementarity, although the environmental change induced by birth order is difficult to characterize: one could interpret its function purely as an early-environment proxy, or instead as a total-childhood effect, since birth order persists throughout the lifetime. Our estimate of the interaction between genetic endowment and parental investment during adolescence,  $\alpha_g^k$ , speaks directly to this tension; we find a positive but small effect, landing our estimate directly in between between the substitutability that Arold et al. (2025) find for school quality and the complementarity documented by Muslimova et al. (2026). The structural model builds on the child human capital literature (Attanasio et al., 2020; Caucutt et al., 2020; Agostinelli and Wiswall, 2025) by showing how genetic endowments shape parental decisions and outcomes, and the findings contribute to the intergenerational mobility literature (Chetty et al., 2014; Solon, 1999) by investigating an under-explored source of educational inequality, showing how it is mediated by parental decisions, and exploring how policy can respond.

Section 2 of this paper provides background on genetics and genomics, discusses how to interpret empirically estimated genetic “effects”, and shows how those concerns inform our methodological choices. As the discussion becomes technical, we append a glossary (Appendix A) containing definitions of relevant genetic terms. Section 3 describes the Add Health survey data and reports descriptive statistics. Section 4 reports reduced-form empirical results. Section 5 outlines the structural model and reports estimated parameter values. Section 6 reports the effects of three illustrative policy interventions. Section 7 concludes with a summary of our findings and a discussion of limitations that could be ameliorated by additional work and access to more data.

## 2 Polygenic Scores

### 2.1 Genetics Primer and Polygenic Scores

All cellular life contains DNA, which has a double helix shape and is composed of nucleotides, which are pairs of nitrogenous molecules called bases: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). Adenine always bonds with Thymine, and Cytosine always bonds with Guanine, so a complete description of DNA composition can be inferred from a description of a single side of the base pairs. The human genome contains approximately 3 billion nucleotide loci (Nurk et al., 2022) distributed across 23 chromosomes. Human beings are diploid, meaning they retain two haploid copies of each chromosome, one inherited from the mother and one from the father. Therefore, at a given genetic locus, the genetic information for a particular human can be represented with two letters: the base (A, T, C, or G) at that genetic address in the maternal chromosome, and the base at that same address in the paternal chromosome. Recall that the bonding rules make the information regarding the base pair implicit in the identification of a base on one side, so two letters are a sufficient characterization.

The most common way by which human beings vary genetically is by retaining a different base at a particular genetic locus: we call this type of variant a single-nucleotide variant (SNV), and the particular variant base an allele. If the less-common allele is present in at least one percent of the population, we call the SNV a single-nucleotide polymorphism (SNP); otherwise, the SNV is referred to as a rare variant. Other forms of genetic variation include insertions/deletions, multiple adjacent substitutions, repeat-number variation, and larger rearrangements known as structural variants. Two haploid genome sequences selected from two random human beings will have approximately 3 million SNVs between them (The International HapMap Consortium, 2003), meaning there will exist a difference at approximately 0.1% of the genome loci. When we account for both chromosome copies and different types of variants, a given person’s genome is approximately 99.6% identical to a reference genome (National Human Genome Research Institute, 2023).

An important aspect of genetic inheritance is that each parental chromosome is itself a random combination of the maternal and paternal chromosomes present in the parent. The process of meiosis generates each gamete by taking each of the two homologous chromosome copies, duplicating, cutting, and recombining them, yielding four haploid chromosomes, each a random mixture of the original two copies. The relevant implication is that genetic variation across siblings is randomly generated, and that siblings share around 50% of their genes (not accounting for possible genetic similarities between mother and father).

Most of the traits and outcomes, or phenotypes, relevant to social science are polygenic: any effects caused by genetic variation across human beings are engendered by small additive effects of SNVs at many different genetic loci. In other words, for a given trait like cognition, there is no specific “intelligence gene” that explains more than a tiny fraction of variance in intelligence (Lee et al., 2018; Okbay et al., 2022). In order to identify the many different genes that might contribute to a particular trait, geneticists run multiple regressions across each possible variant, known as Genome-Wide Association Studies (GWAS). Using the estimates obtained empirically, we can summarize the trait-relevant genetic information for a given individual into an index score; we add up the frequency of allele occurrences weighted by the GWAS-derived regression coefficient on the trait of interest: the aggregate is known as a polygenic score (PGS).

Perhaps the best-studied phenotype in the genetics of social science is years of educational attainment (EA). Unlike traits that are more difficult to measure, such as cognitive skills or personality, years of educational attainment is easy to survey; as a result, sample sizes are sufficiently large to identify genetic variants that have even a tiny effect among the millions of possibilities. For this reason, despite the availability of polygenic scores for more granular traits of interest like intelligence and personality, this paper will use the polygenic score for educational attainment as a proxy for genetic endowment. Nevertheless, it is important to note that the genes and weights estimated by the most recent EA GWAS (Okbay et al., 2022), which had a sample size of three million people, could account for only 12–16 percent of the variance in educational attainment in an independent sample. While this figure is significantly larger than the variance accounted for by GWAS estimation for most other behavioral traits, it is significantly smaller than the 41–44 percent of EA variance that twin-study pedigree-based methods estimate to be accounted for by variation across the entire genome (Silventoinen et al., 2020).

## 2.2 Polygenic Scores and Methodological Concerns

In addition to the “missing heritability” that remains undetected by GWAS for educational attainment, there are two major categories of concerns regarding the implementation of polygenic scores.

First, we notice that the GWAS procedure outlined briefly above only investigates statistical correlations between genetic variants and phenotypic realizations. Naturally, we cannot naively interpret the genes identified by GWAS as causal, because of the possibility that environmental factors contributing to EA systematically covary with the genes identified as having significant correlations with EA. For example, we know that family socioeconomic

status and parental education contribute significantly to educational attainment (Hauser, 1973; Sewell and Hauser, 1975).

Geneticists conducting population-level GWAS control for genetic structure that corresponds to geography, as well as a variety of natural demographic fixed effects, but the possibility of spurious correlations (from the perspective of causal interpretation) remains. In particular, to the extent that parents both pass down their genes and also shape the environments of their children, we would expect this form of gene-environment correlation, known as genetic nurture, to be a significant contributor to the gene-phenotype correlation represented by the polygenic score. Howe et al. (2022) find that when family fixed effects are applied to GWAS analyses, estimated effects decrease by approximately 50–65 percent; this finding indicates that despite the application of plausible controls, PGS scores estimated by traditional GWAS are significantly confounded by non-causal sources of gene-environment correlation. Section 4.1 shows both faces of this in our data: the uncontrolled pooled gradient is attenuated by about a third once parental education and household income enter, and conditional on them a family fixed effect attenuates it no further.

A second set of concerns regards the nature of the correlation itself. Suppose, for the time being, that we can safely interpret the EA polygenic score as representing the true causal effect of the individual genotype on years of educational attainment. The polygenic score is a weighted aggregate of genetic variants, not a pure measure of “genetic endowment”. In other words, the mechanisms by which the identified genes cause outcome changes could be through biological changes to brain cognition, or could be through causing additional parental investment relevant to educational attainment, or could be through causing preference changes engendering different choices the child herself makes during schooling. While each of these mechanisms is a genuine aspect of causality under the counterfactual framework, if we wish to investigate the interaction between genes and the mechanisms themselves, which we do, we should note that the polygenic score represents an expression of genes that itself is a reflection of an unknown mix of mechanistic environmental responses.

This issue is perhaps best demonstrated by considering a regression of parental investment on the PGS score (Agostinelli and Weingarten, 2025). Since the weights used to generate the polygenic score are themselves functions of parental investment, such a regression includes parental investment on both sides, generating simultaneity bias. To clarify, this issue is independent of whether or not we are using within-family or pooled estimates, and therefore whether or not the PGS coefficients are causal with respect to the outcome phenotype by which they are identified (educational attainment).

To formalize both of the methodological concerns, consider the following. Let  $y_{it}$  denote educational attainment for child  $i$ . Let  $g_{it}$  denote the genetic endowment. Let  $e_{it}$  represent

environmental inputs provided by the parent. The true structural relationship of interest is:

$$y_{it} = f(g_{it}, e_{it}, \varepsilon_{it}) \quad (2.1)$$

where  $\varepsilon_{it}$  represents the set of causal non-genetic factors that are not shaped by the parents, also referred to as the unshared environment.

The problem with the pooled OLS GWAS approach is that the observed correlation between  $g_{it}$  and  $y_{it}$  cannot be interpreted as the causal effect  $df/dg$ , because the correlation will be confounded by correlation between  $g_{it}$  and  $e_{it}$ ,  $\varepsilon_{it}$ . In the language we have introduced above,  $\text{Corr}(g_{it}, e_{it})$  is referred to as genetic nurture, and  $\text{Corr}(g_{it}, \varepsilon_{it})$  is referred to as other gene-environment correlation.

Suppose that we obtained uncontaminated estimates of the PGS weights that yielded a true causal interpretation. The resulting PGS is a weighted sum of SNPs (SNPs are SNVs where the frequency of the minor allele is at least one percent):  $\text{PGS}_{it} = \sum_j \omega_j \cdot \text{SNP}_{ijt}$ , where the weights  $\omega_j$  are causal, representing  $d\mathbb{E}[y_{it}]/d\text{SNP}_{ijt}$ . The outcome causally depends on the genetic endowment both directly, and through mediation via environmental inputs, since environmental inputs are themselves causally dependent on the genetic endowment. We rewrite (2.1) to make this clear:

$$y_{it} = f(g_{it}, e_{it}(g_{it}), \varepsilon_{it}(g_{it})) \quad (2.2)$$

In particular, we would expect that parents endogenously choose environmental inputs  $e_{it}$  in response to early signals of skill, which are themselves functions of both  $g$  and  $e$ ; the weights estimated by GWAS  $\omega_j$  indicate both the direct effect of each SNP and the indirect effect operating through parental investment response. Formally:

$$\begin{aligned} \omega_j \equiv \frac{d\mathbb{E}[y_{it} \mid \text{SNP}_{ijt}]}{d\text{SNP}_{ijt}} &= \underbrace{\frac{\partial f}{\partial g_{it}} \cdot \frac{\partial g_{it}}{\partial \text{SNP}_j}}_{\text{Direct Effect}} + \underbrace{\frac{\partial f}{\partial e} \cdot \frac{\partial \mathbb{E}[e_{it} \mid \text{SNP}_j]}{\partial \text{SNP}_j}}_{\text{Parental Response Mediation}} \\ &+ \underbrace{\frac{\partial f}{\partial \varepsilon} \cdot \frac{\partial \mathbb{E}[\varepsilon_{it} \mid \text{SNP}_j]}{\partial \text{SNP}_j}}_{\text{Other Mediation}} \end{aligned} \quad (2.3)$$

## 2.3 Mendelian Randomization Within Families

Even when controlling for relevant sources of gene-environment correlation, a simple population-level comparison of genes and educational attainment will necessarily be confounded by shared genes between parents and children: at minimum, we would expect that parents

with genes associated with greater educational attainment will be themselves more highly educated, which influences family environment, parental investment decisions, and parental resources through the education premium in the labor market. In addition, we might expect that the genes associated with educational attainment would also be associated with traits that, even controlling for realized parental educational attainment, would positively influence the human capital development of the child.

Mendelian randomization refers to the random genetic differences across siblings (and dizygotic twins) that are generated by meiosis and inherited from each parent. In other words, conditional on parental (maternal and paternal) genotype, variation in the genotype of the child is truly exogenous.

It is often noted that spousal coupling that selects on traits, referred to as assortative mating, increases the correlation across parental genotypes and therefore increases the correlation between the genotypes and PGS of siblings. While assortative mating affects GWAS estimation in a pooled setting, Mendelian randomization exploits meiosis-induced randomness conditional on parental genotype, so assortative mating does not preclude its exogeneity. Meyer et al. (2012) report no evidence of systematic deviations from Mendelian transmission among humans; regardless of the degree of relatedness of the parental genotypes, variation conditional on parental genes can be validly interpreted as random (Howe et al., 2022; Young et al., 2022).

Conditioning on family fixed effects absorbs the shared variation generated by family-level confounders; any phenotypic variation downstream of genetic variance conditioned on parental genotype can be interpreted as causal. However, there exists some subtlety: while implementing a family fixed effect means we can interpret remaining genotypic variation as exogenous, the polygenic score itself aggregates the exogenous genotypic variation using weights that were themselves generated using the pooled OLS method.

To illustrate, note that the Mendelian randomization structure follows this pattern for any outcome of interest  $y$ :

$$y_{ift} = \gamma \cdot \text{PGS}_{ift} + \varphi_f + \eta_{ift}, \quad (2.4)$$

where  $\varphi_f$  is a family fixed effect and  $\gamma$  captures the causal effect of a change in the child's EA polygenic score on educational attainment. Again, the variation in PGS, conditional on parental genotype, is generated by the random Mendelian segregation of alleles, but the way that variation is weighted depends on the pooled, population-level GWAS, and may therefore reflect population stratification, assortative mating, and other indirect genetic effects. Ideally, we would have large enough sample sizes of siblings to estimate the weights

themselves within-family, thereby identifying the causal effect of a given gene in our data on the outcome of interest.

That being said, the utilization of population-derived weights does not preclude us from applying a causal interpretation to the post-weight-estimation family fixed-effects design. As mentioned above, and noting that the weights are estimated in an external sample and can therefore be treated as fixed, sibling PGS differences remain functions of quasi-random Mendelian segregation. To the degree that weights are contaminated, this reflects an issue of measurement quality; we treat pooled-OLS-estimated weights as imperfectly estimated, potentially attenuating the observed relationship between the PGS and an outcome of interest, but not introducing a source of systematic bias.

We proceed by interpreting the within-family results as causal, and the pooled results as reflective of environmental confounds, most notably parental genetic nurture. To preview our empirical results, we find that this indirect component is the main contributor to the population-level coefficient: the within-family estimate is approximately one-third of the pooled OLS estimate, indicating that roughly two-thirds of what the pooled polygenic score effectuates at the population level reflects environmental correlation rather than causal effects.

## 3 Data

### 3.1 The National Longitudinal Study of Adolescent to Adult Health

Our data source is the National Longitudinal Study of Adolescent to Adult Health (Add Health), which is a panel dataset reflecting a longitudinal study of a nationally representative cohort of 90,000 adolescents in grades 7–12, beginning in 1994–1995.

The Add Health study conducted an in-home interview with a stratified subsample of approximately 20,745 respondents and their parents. Data were collected from five subsequent waves of the study, following the cohort across 28–29 years: W2 = 1996; W3 = 2001–02; W4 = 2007–08; W5 = 2016–18; W6 = 2022–23. Most importantly for our questions, saliva samples were collected and genotyped for approximately 12,200 respondents during the Wave 4 study. We outline the specific details pertaining to the collection of genetic data and corresponding polygenic scores in the following section.

As a longitudinal study that includes genotypes and tracks skill development, education, parental investment, and early-career outcomes, the Add Health study is a natural choice to investigate the interactions between genetic endowment and the life cycle of human capital development. However, it is worth highlighting specific aspects of the dataset that can be

leveraged to answer novel questions.

Most importantly, the Add Health study oversamples siblings, which enables the implementation of the within-family design we introduced in Section 2.3 as a response to the methodological critique outlined in Section 2.2.

Second, the emphasis on adolescents specifically is not common across genotyped longitudinal datasets. The wave-structure of the data, specifically repeated sampling during the teenage years, allows us to ask specific questions about the expression of genetic endowments and their interaction with parental investments during adolescence.

Third, later waves allow us to follow members of the study cohort into early adulthood: in addition to observing realized years of education, data on early labor-market outcomes enable us to investigate how investments in education and skill development earn returns in later life. These data motivate our structural model of human capital investment by informing return parameters.

## 3.2 The Add Health EA Polygenic Score

We use the pre-computed polygenic score for educational attainment generated by the Add Health Sequencing and Genotyping Resource Center. The score is created using the Lee et al. (2018) GWAS meta-analysis (EA3). Polygenic scores are generated by the Carolina Population Center, using the procedure outlined in Dudbridge (2013), which applies PRSice within PLINK (Chang et al., 2015).

We choose to implement the Lee et al. (2018) weights instead of the more recent Okbay et al. (2022) weights to avoid sample-overlap bias: the Okbay et al. (2022) GWAS includes Add Health participants in its discovery sample, when the same people are in the sample used to estimate the weights and in the sample where the score is used, the weights partly fit those people’s own outcomes, and the score is mechanically correlated with the outcome (sample-overlap bias). The Lee et al. (2018) GWAS, by contrast, did not include Add Health participants, allowing us to utilize the genotyped Add Health sample as an independent prediction sample.

The major restriction on the use of the Lee et al. (2018) weights is that they are not predictively valid for populations of non-European ancestry. Therefore, following standard practice in the geneoconomics literature (Barth et al., 2020; Houmark et al., 2024), we restrict our study and analysis sample to only European-ancestry subjects for which the Lee et al. (2018)-derived polygenic scores are considered valid. Polygenic scores estimated on samples of people of European ancestry are not portable to non-European groups because different population structures have different linkage disequilibrium patterns (as well as

different minor allele frequencies). In other words, the structure, or gene-gene correlations, that geneticists rely on to tag individual variants differs across populations, so a variant that might be tagged to a causal section of the genome in Europeans might not tag the same section in non-Europeans.

Finally, it is relevant to note that the application of the pre-computed polygenic score eliminates the possibility for researcher discretion over construction aspects like SNP selection, LD-pruning thresholds, and the p-value threshold for inclusion (polygenic score construction terms are defined in Appendix A). This removes this element of “researcher degrees of freedom” as well as making the results of the paper reproducible “out of the box” relative to the restricted-use Add Health data releases.

### **3.3 Measures of Educational Attainment, Economic Outcomes, Skills, and Parental Investment**

Our measure of educational attainment is completed years of schooling, observed at Wave 5 (2016–18) when respondents were aged 32–42 and the vast majority of educational decisions had been made. The survey data lists the highest level of educational credential completed; we map each credential to a continuous years-of-schooling measure using the standard ISCED-aligned coding listed in Appendix B. The resulting variable ranges from 8 (eighth grade or less) to 20 (doctoral degree). We also use indicators for particular educational milestones (graduation achievements) to capture milestone-based non-linearities in the returns to educational attainment.

For long-run economic outcomes, we use cumulative labor earnings reported through Wave 5. Further, we can project earnings forward to age 60 by referencing age-earnings profiles from the Current Population Survey data. Specifically, for each respondent we observe earnings at age  $\text{age}_{w5}$  (between 32 and 42, depending on respondent birth year and Wave 5 interview date), and project earnings forward to age 60 using the CPS age-earnings profile within the education group of the respondent. We then calculate the present-discounted value of earnings from age 25 through to age 60, using the respondent’s actual Wave 5 earnings up to the observed age, and using the CPS-derived conditional mean for all other ages up to age 60.

To create a longitudinal panel of measures of cognitive skills, we used the following: the Add Health Picture Vocabulary Test and self-reported grades in English and mathematics from Wave 1; a grade-point average across academic subjects from Wave 2; a vocabulary assessment from Wave 3; and a cognitive battery that was administered in Waves 4 and 5.

For non-cognitive skills, we use: two psychosocial scale scores from Wave 1; a psychoso-

cial composite from Wave 2; a conscientiousness measure from Wave 3; and the Big Five personality inventory from Waves 4 and 5.

We combine these measures and can extract a single latent factor for each skill type within-wave to form Bartlett scores (Bartlett, 1937); we normalize each score to have mean zero and unit variance, such that  $\theta_{i,t}^C$  and  $\theta_{i,t}^N$  are denoted in terms of within-wave standard deviations. Waves are then mapped to model periods corresponding to the relevant years between sampling.

For parental investment, we rely primarily on the Wave 1 in-home parent questionnaire, which generated direct measures of parental educational expectations, homework supervision, parental involvement in school activities, private K-12 enrollment, and family resources for educational support. We used these data to extract a latent homework-help factor from the three Wave 1 items measuring parent-reported supervision frequency and school involvement. We also used parent-reported school discussion time (in minutes per week), private K-12 enrollment status, and self-reported total household educational expenditure as investment variables. Finally, the structural model required a measure of college tuition: we used an external measure for college costs, but were able to obtain parental contributions to college costs from Wave 3.

### 3.4 Analysis Sample and Descriptive Statistics

The sibling subsample has 826 sibling pairs from 403 families. We use the full sibling pairs identified by the Add Health pre-constructed sibling pair file. Full-sibling and dizygotic twins are included in the sibling subsample, while monozygotic twins are excluded, as they do not have genetic variation between them. Table 3.1 outlines descriptive statistics.

Table 3.1: Descriptive Statistics

|                                                               | Full Sample |         | Siblings Sample |         | Low-PGS |
|---------------------------------------------------------------|-------------|---------|-----------------|---------|---------|
|                                                               | Mean        | S.D.    | Mean            | S.D.    | Mean    |
| <i>Panel A: Educational Attainment Measures</i>               |             |         |                 |         |         |
| Completed Years of Schooling                                  | 14.34       | (2.40)  | 14.36           | (2.45)  | 13.51   |
| High School Completion                                        | 0.942       | (0.234) | 0.936           | (0.245) | 0.911   |
| Any Post-Secondary                                            | 0.774       | (0.418) | 0.769           | (0.422) | 0.664   |
| 4-Year College Completion                                     | 0.359       | (0.480) | 0.367           | (0.482) | 0.209   |
| Graduate Degree                                               | 0.127       | (0.333) | 0.122           | (0.328) | 0.059   |
| <i>Panel B: Genetic and Phenotypic Measures</i>               |             |         |                 |         |         |
| Wave 1 AHPVT                                                  | 104.9       | (11.8)  | 104.1           | (11.1)  | 101.1   |
| Self-Rep. English Grade                                       | 2.88        | (0.97)  | 2.86            | (0.98)  | 2.65    |
| Self-Rep. Math Grade                                          | 2.76        | (1.05)  | 2.75            | (1.05)  | 2.55    |
| Self-Rep. History Grade                                       | 2.93        | (1.01)  | 2.98            | (1.00)  | 2.69    |
| Self-Rep. Science Grade                                       | 2.88        | (1.00)  | 2.85            | (1.00)  | 2.63    |
| EA Polygenic Score (Lee et al.)                               | 0.000       | (1.000) | 0.017           | (0.992) | −1.273  |
| <i>Panel C: Parent and Household Characteristics (Wave 1)</i> |             |         |                 |         |         |
| Maternal Education (years)                                    | 13.34       | (2.34)  | 13.36           | (2.11)  | 12.76   |
| Paternal Education (years)                                    | 13.40       | (2.52)  | 13.46           | (2.39)  | 12.80   |
| Household Income (\$1,000s)                                   | 82.5        | (80.2)  | 84.5            | (97.8)  | 69.5    |
| Parental Educ. Expectations                                   | 0.361       | (0.480) | 0.351           | (0.478) | 0.307   |
| <i>Panel D: Gender</i>                                        |             |         |                 |         |         |
| Female                                                        | 0.530       | (0.499) | 0.517           | (0.500) | 0.544   |
| Observations                                                  | 5,730       |         | 826             |         | 1,433   |

The full sample comprises European-ancestry Add Health subjects with valid EA polygenic scores, completed educational attainment observed at Wave 5, and Wave 1 parent and demographic controls. The low-PGS sample refers only to respondents in the bottom quartile of the EA polygenic score distribution. Standard deviations in parentheses. All dollar amounts converted to 2015 USD.

We find descriptive statistics matching other work that focuses on siblings in the Add Health study (Arold et al., 2025). As expected, the low-PGS quartile exhibits lower educational attainment outcomes; in addition, we find that the low-PGS quartile demonstrates lower indicators of cognitive skill development – AHPVT scores and self-reported grades – at the outset of the study. Finally, family background factors like parental education and resources differ systematically between the overall sample and low-PGS quartile. These differences motivate the within-family methodology we emphasized in Section 2.3, and to the extent that these background factors influence educational attainment, their effects will be reflected in the difference between the pooled and within-family estimates.

## 4 Empirical Results

### 4.1 Reduced-Form Estimation

To start, we present a reduced-form estimation of the relationship between completed years of schooling and the polygenic score for educational attainment. We estimate the regression using both pooled OLS and a within-family (family fixed-effects) design.

Let  $S_{if}$  signify completed years of schooling for individual  $i$  in family  $f$ , and let  $z_{if}$  signify the standardized educational attainment polygenic score. Let  $x_{if}$  denote individual-level controls (e.g., gender, a quadratic in birth year, birth order, and wave indicators),  $q_f$  denote family-level controls (Wave 1 household income and parental education), and let  $PC_{if}$  denote the ten ancestry principal components. Notation is shared with Section 5.

The equation for the pooled OLS is:

$$S_{if} = \mu + \beta^{\text{pool}} z_{if} + x'_{if} \delta + q'_f \gamma + PC'_{if} \zeta + \varepsilon_{if}, \quad (4.1)$$

with clustered standard errors at the family level. The within-family specification replaces controls with a family fixed effect,

$$S_{if} = \beta^{\text{wf}} z_{if} + x'_{if} \delta + PC'_{if} \zeta + \varphi_f + \varepsilon_{if}, \quad (4.2)$$

The family-level controls  $q_f$  are removed from equation (4.2) because they are collinear with  $\varphi_f$ .

Panel A of Table 4.1 shows the pooled OLS estimates. In Column 5, our preferred specification, a one-standard-deviation decrease in the EA polygenic score is associated with a decrease of around 0.422 years of completed schooling ( $p$ -value  $< 0.001$ ).

Panel B shows our preferred within-family Mendelian randomization estimates. Note how the coefficient attenuates significantly: in Column 5 of the within-family specification, a one-standard-deviation decrease in the EA polygenic score is associated with a 0.466-year decrease in years of completed schooling ( $p$ -value  $= 0.001$ ). The attenuation between Panels A and B is approximately  $-10$  percent, and this finding both motivates and helps identify the structural model that represents the core of the paper.

The following equations illustrate the theoretical objects being estimated. Let  $\beta_g$  represent the causal effect of the child's own endowment and  $\beta_\psi$  represent the effect of the family environment  $\ln \psi_f$ , which (4.1) omits. Since  $\text{Var}(z) = 1$  under (5.1) and  $\text{Cov}(\xi_{if}, \ln \psi_f) = 0$

by (5.5), we obtain:

$$\text{plim } \hat{\beta}_g^{\text{pool}} = \beta_g + \beta_\psi \sigma_{g\psi}, \quad \text{plim } \hat{\beta}_g^{\text{wf}} = \beta_g, \quad \Delta_g \equiv \beta_g^{\text{pool}} - \beta_g^{\text{wf}} = \beta_\psi \sigma_{g\psi}. \quad (4.3)$$

The delta between the two is determined by the importance of the family environment to educational attainment and its correlation with the genetic endowment, which is how we can identify  $\sigma_{g\psi}$  from  $\Delta_g$  in Section 5.

Importantly, as we explicated in Section 2.2, even under the causal interpretation for which we argued at the conclusion of Section 2.3, (4.3) does not mean that  $\beta_g$  is a pure biological effect. Indeed, it will reflect the (causal) evocative environmental response manifesting in the environment of the child:

$$\beta_g = \underbrace{\beta_{\text{bio}}}_{\text{direct}} + \underbrace{\beta_{\text{evoc}}}_{\text{evoked response}}. \quad (4.4)$$

For example, a child who has a genetic predisposition for educational attainment might exhibit high cognitive skill at an early age; environmental responses could include parents reading to the child more, taking a greater interest in educational progress, or the school moving the child to a different classroom. To the extent those environmental responses themselves influence educational attainment, that will be captured by  $\beta_{\text{evoc}}$  in (4.4), which we label evocative response. Its presence in addition to  $\beta_g$  is the reason why the within-family coefficient, while reflective of the causal variation attributed to the genes, should not be interpreted as a purely biological effect, and why the parental investment responses to genetic endowment that the following section investigates are of significant importance.

The level of attenuation is similar to findings from the within-family GWAS literature (Howe et al., 2022; Trejo et al., 2018). In Howe et al. (2022), within-family EA-PGS effects are typically attenuated by 50 to 65 percent; this includes educational attainment. Our  $\sim 50$  percent attenuation fits this pattern.

The attenuation we find when applying parental controls and robustly confirmed by the within-family estimates can be used to investigate the degree of gene-environment covariance; in Section 5, we will try to reproduce precisely this delta between coefficients through a structural model.

Table 4.2 repeats our analysis for four-year college completion, projected lifetime earnings (see Section 4.3) and the homework-help index (see Section 4.2).

Table 4.1: Respondent’s EA Polygenic Score and Completed Years of Schooling

|                                                      | (1)              | (2)              | (3)              | (4)              | (5)              |
|------------------------------------------------------|------------------|------------------|------------------|------------------|------------------|
| <i>Panel A: Pooled OLS</i>                           |                  |                  |                  |                  |                  |
| EA Polygenic Score                                   | 0.655<br>(0.030) | 0.655<br>(0.030) | 0.655<br>(0.030) | 0.585<br>(0.034) | 0.422<br>(0.033) |
| Respondents                                          | 5,730            | 5,730            | 5,730            | 4,709            | 4,312            |
| <i>Panel B: Within-Family (Family Fixed Effects)</i> |                  |                  |                  |                  |                  |
| EA Polygenic Score                                   | 0.465<br>(0.125) | 0.466<br>(0.125) | 0.466<br>(0.125) | 0.466<br>(0.125) | 0.466<br>(0.125) |
| Respondents                                          | 697              | 697              | 697              | 697              | 697              |
| Families                                             | 339              | 339              | 339              | 339              | 339              |
| <i>Panel C: Pooled OLS, Sibling Subsample</i>        |                  |                  |                  |                  |                  |
| EA Polygenic Score                                   | 0.778<br>(0.095) | 0.777<br>(0.095) | 0.777<br>(0.095) | 0.588<br>(0.100) | 0.435<br>(0.086) |
| Respondents                                          | 826              | 826              | 826              | 719              | 697              |
| Families                                             | 403              | 403              | 403              | 350              | 339              |
| Demographic Controls                                 | ✓                | ✓                | ✓                | ✓                | ✓                |
| Birth Order/Cohort                                   |                  | ✓                | ✓                | ✓                | ✓                |
| Controls                                             |                  |                  |                  |                  |                  |
| Wave Fixed Effects                                   |                  |                  | ✓                | ✓                | ✓                |
| Household Income Control                             |                  |                  |                  | ✓                | ✓                |
| Parental Education Control                           |                  |                  |                  |                  | ✓                |

The dependent variable is completed years of schooling observed at Wave 5 (ages 32–42). The EA Polygenic Score is the Add Health pre-computed score (Braudt and Harris, 2020), using weight estimates from Lee et al. (2018). PGS scores are normalized to mean zero, unit variance. Panel B applies family fixed effects. Standard errors are clustered at the family level and presented in parentheses. Columns 4–5 add controls observed for fewer respondents, so they change the sample as well as the controls. Panel C reports the Panel A specifications estimated on the sibling subsample.

Table 4.2: Respondent’s EA Polygenic Score and Other Outcomes

|                                                              | Pooled OLS, full sample |                      | Within family           |
|--------------------------------------------------------------|-------------------------|----------------------|-------------------------|
|                                                              | Column 1<br>controls    | Column 5<br>controls | Family fixed<br>effects |
| <i>Completed years of schooling (Table 4.1)</i>              |                         |                      |                         |
| EA polygenic score                                           | 0.655<br>(0.030)        | 0.422<br>(0.033)     | 0.466<br>(0.125)        |
| Respondents                                                  | 5,730                   | 4,312                | 697                     |
| <i>Four-year college completion (percentage points)</i>      |                         |                      |                         |
| EA polygenic score                                           | 11.9<br>(0.59)          | 7.8<br>(0.68)        | 6.8<br>[ $p = 0.011$ ]  |
| Respondents                                                  | 5,730                   | 4,312                |                         |
| <i>Projected lifetime earnings (percent)</i>                 |                         |                      |                         |
| EA polygenic score                                           | 15.8<br>(1.69)          | 9.3<br>(1.95)        | 10.8<br>[ $p = 0.26$ ]  |
| Respondents                                                  | 4,241                   | 3,233                |                         |
| <i>Homework-help index (percent of a standard deviation)</i> |                         |                      |                         |
| EA polygenic score                                           | 3.7<br>(2.10)           | 0.3<br>(2.11)        | 9.2<br>[ $p = 0.29$ ]   |
| Respondents                                                  | 2,417                   | 1,993                |                         |

Effects are standardized. Controls are the same as in Table 4.1. Standard errors are clustered by family

## 4.2 Parental Investment Responses

We are interested in a reduced-form analysis of whether parental investments respond to sibling differences in genetic endowment with reinforcement or compensation. We start by empirically estimating – using a within-family specification – the relation between the EA PGS score and five investment outcomes: log Wave 1 educational expenditure, private K-12 enrollment, log Wave 3 parental college expenditure, the Wave 1 homework-help index, and Wave 1 parental time spent discussing school. For each investment measure  $m$ , we obtain for each sibling  $i$  in family  $f$  at age  $t$ :

$$\ln I_{ift}^m = \pi_m z_{if} + x'_{ift} \delta_m + \varphi_{fm} + u_{ift}^m, \quad (4.5)$$

with errors clustered at the family level. As a binary indicator, private K-12 enrollment is estimated as a linear probability model.

Recalling that  $z_{if} = \bar{z}_f + e_{if}$ , by (5.2),  $\pi_m$  is identified from the variation  $e_{if}$  that is orthogonal to shared-sibling endowment. If the parental response to the relative increase in the genetic endowment of one sibling is to increase investment equally across children, that empirical response will not be captured by  $\pi_m$ : the coefficient only captures the differential

parental response to sibling differences:

$$\pi_m \approx \underbrace{\partial I_1^m / \partial z_1}_{\text{own}} - \underbrace{\partial I_2^m / \partial z_1}_{\text{cross}}, \quad (4.6)$$

where the response is averaged across siblings.

As (4.6) illustrates, the own-child response  $\pi_m$  depends on the cross-response: compensating other-sibling investment reduces the observed coefficient, while reinforcing own-child investment generated through reallocation away from the other sibling increases the observed coefficient. A small observed coefficient could therefore reflect a small overall response, or a large response that applies to both children: the other-sibling response will be differenced out by the empirical strategy.

This ambiguity further motivates the structural approach we develop in Section 5. A parent plausibly allocates investment resources such that the marginal return to each dollar, or hour, is equalized across siblings. If one sibling exogenously obtains a more favorable endowment, the relative (net) change in marginal return will depend on whether (at the margin) a given investment and endowment exhibit dynamic complementarity or substitutability: in other words, the cross-response sign will depend on the interaction term in the skill production function. Total (gross) investment response will depend on how costly (in terms of forgone consumption or leisure) investment is at the margin. The sign of  $\pi_m$  relative to the true own-child response thus depends on the skill-production and parental budget constraint: if the budget constraint is soft at the relevant margin, then the other-sibling effect will be small, and the own-child response is preserved; if the constraint is hard, the complementarity in the production function amplifies the effect; substitutability reverses the sign. The structural model allows us to separately identify the two objects that are being reflected by the reduced-form regression coefficient representing the net differential response.

To summarize, the point estimates are small in each case, and the two time measures are of opposite sign relative to the financial measures. Three interpretations remain plausible: (1) complementarity/substitutability might be weak in either direction, so a sibling-specific favorable draw does not change the relative marginal return across siblings; (2) it might be that the parental allocation decision is not responsive enough to the endowment itself: it might operate on noisy observations of skill development, or on observations that are orthogonal to the genetic endowment; (3) the measures we use for time investment might be too coarse to identify the proposed effects.

Each interpretation yields a subtly different implication we can tease out using the structural model. If the early measures of skill development are too noisy to inform parental allocation, we should see an increase in investment response during later model periods

(corresponding to later Waves). Both weak complementarity/substitutability, and coarse measures, imply a flat response profile: we can separate these two mechanisms by estimating the skill production function directly, using the skill outcome measures we derive from the data (see Section 3.3).

Once we estimate the underlying parameters using the structural approach – incorporating the additional necessary data – we revisit the different estimates  $\pi_m$  across different  $m$ , where the set of signs functions as a (weak) testable implication of the model.

### 4.3 Long-Run Earnings and Partner Selection

Recalling Section 3.3, we only have access to early-stage career earnings; to leverage the earnings variation we do observe in the data, while also accounting for different career trajectories that depend on years of education, we combine the Wave 5 earnings data with data from the Current Population Survey. We discretize completed educational attainment for subject  $i$  into four bins:  $b(S_i) \in \{1, 2, 3, 4\}$ . We further subdivide each educational attainment bin by demographic characteristic  $X_i$ , which reflects sex, and age at earnings observation  $a_i^o$ . We discretize the earnings distributions corresponding to each of the resulting cells into terciles  $q_i \in \{1, 2, 3\}$ :

$$q_i = q(w_i; a_i^o, b(S_i), X_i). \quad (4.7)$$

We then generate projected present discount values for each subject  $i$  by using the average CPS earnings trajectories corresponding to the relevant tercile of the age-education-demographic cell:

$$Y_{if}^{\text{PDV}} = \sum_{a=25}^{60} (1+r)^{-(a-25)} \bar{w}(a, b(S_{if}), q_{if}, X_{if}), \quad (4.8)$$

where  $\bar{w}(a, b(S_{if}), q_{if}, X_{if})$  denotes mean CPS earnings at age  $a$  corresponding to the cell to which subject  $i$  belongs. In Wave 5, earnings are observed at ages 32–42. Haider and Solon (2006) find that the association between an earnings observation from before age 30 and lifetime earnings is severely attenuated; the strength of the association increases through the 30s and peaks near age 40; therefore, an earnings observation from Wave 5 is suitable for our purpose.

We then estimate:

$$\ln Y_{if}^{\text{PDV}} = \beta^Y z_{if} + x'_{if} \delta + \text{PC}'_{if} \zeta + \varphi_f + \varepsilon_{if}. \quad (4.9)$$

The result is a 10.8 percent reduction in present-discounted lifetime earnings per standard

deviation ( $p$ -value = 0.263). At the median of the lifetime earnings distribution, which is approximately \$1.11 million in present-discounted value, this is a total difference of roughly \$119,000.

This figure represents a speculative projection: most of the variation is generated by very early-career earnings, which are unlikely to reflect true differences in skills, educational attainment, and human capital development overall that are causally downstream of the genetic endowment.

Finally, we do have some suggestive, though not statistically significant, evidence of a change in partner selection related to the genetic endowment: the probability that a respondent is partnered with someone of above-median educational attainment falls by 19.0 percentage points per standard deviation of genetic disadvantage ( $p$ -value = 0.024).

## 5 A Structural Model of Parental Investment and Human Capital Formation

Each of the sections of this paper has spoken to the notion that reduced-form estimates using polygenic scores, even when retaining causal interpretation, depend on an equilibrium parental investment (and broader environment) response function. To the extent that we might wish to derive policy implications from our results, it is prudent to embed parental investment itself into a model of human capital development; we wish to represent equilibrium investment levels as themselves responses to constrained optimization and functions of deeper parameters. Table 5.1 provides a road map of how the reduced-form objects from Section 4 are used to identify aspects of the structural model.

Table 5.1: Reduced-Form Objects and Their Structural Counterparts

| Reduced-form object<br>(Section 4)                   | Structural counterpart (Section 5)                                                                                                                    | Role in estimation                                              |
|------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|
| $\beta^{\text{pool}}$ , Eq. (4.1)                    | $\beta_g + \beta_\psi \sigma_{g\psi}$ , Eq. (4.3)                                                                                                     | Targeted moment                                                 |
| $\beta^{\text{wf}}$ , Eq. (4.2)                      | $\beta_g = \beta_{\text{bio}} + \beta_{\text{evoc}}$ , Eq. (4.4); generated by $\alpha_3^k$ , $\alpha_6^k$ , $\alpha_8^k$ and the investment response | Targeted moment                                                 |
| $\Delta_g = \beta^{\text{pool}} - \beta^{\text{wf}}$ | $\beta_\psi \sigma_{g\psi}$                                                                                                                           | Identifies $\sigma_{g\psi}$                                     |
| Skill transitions, pooled and within-family          | Technology, Eq. (5.8)                                                                                                                                 | Targeted moments                                                |
| Post-secondary enrollment and completion             | College choice, Eq. (5.10)                                                                                                                            | Targeted moments                                                |
| Between-family skill variance                        | $\alpha_9^k \ln \psi_f$                                                                                                                               | Targeted moment                                                 |
| $\pi_H$ , homework help, Eq. (4.5)                   | $a_{1,t}/\Phi_{1,t}^H$ , Eq. (5.18); governed by $\alpha_8^k$ and $\gamma_2$                                                                          | Held out; test of fit                                           |
| Monetary investment response                         | $\partial \Phi_{i,t}^M / \partial \ln g_i$ through the shadow values ( $\alpha_6^k$ )                                                                 | Not estimable within families in these data; model-implied only |

Targeted moments are used by the SMM estimator (Section 5.5); the held-out moment is used to test the model output (Section 5.8).

## 5.1 Setup

Motivated by the reduced-form evidence in Section 4, we develop a structural model that jointly captures parental investment, household consumption, and children’s human capital accumulation in the presence of genetic endowments.

The model comprises a family consisting of two parents and two children, who differ in their genetic endowments for EA but share a common family-level unobservable  $\psi_f$ . Parents allocate money and time between the two children under shared constraints, in the tradition of the separable-earnings-transfers model of Behrman et al. (1982) (see Francesconi and Heckman, 2016, for the connection to the modern skill-formation literature); the within-family investment coefficient of Section 4.2 is the reduced-form trace of that allocation. The two-child structure generates the within-family variation that identifies the parameters that the reduced-form analysis isolates within the model, and that is what allows the model to exploit the same pooled-versus-within contrast that Section 4 illustrates as the effect of gene-environmental confounding.

The model follows a child from age  $a_i = 6$  to age  $a_i = 18$ , at which point the child makes

a post-secondary education choice and the model terminates. Each period  $t$  covers one year and is indexed to age, such that  $t = a_i - 5$ ; the first period  $t = 1$  therefore corresponds to age 6, and the terminal period  $t = T = 13$  corresponds to age 18.

Note that each wave of the Add Health survey reaches subjects at different ages, so we obtain different empirical child-wave observation trajectories for different age cohorts. Additionally, the survey data only covers children beginning at age  $a_i = 12$ : we consider earlier periods in the model as unobserved, but amenable to counterfactual policy interventions, of which we can estimate downstream effects. Survey data covering periods after age  $a_i = 18$  enter into the model only indirectly, by informing the returns to education.

At each age and period  $t$ , the state variables are represented by the skill vector corresponding to the child:  $\boldsymbol{\theta}_{i,t} = (\theta_{i,t}^C, \theta_{i,t}^N)$ . At the outset, the child also draws a genetic endowment  $g_i$ , which does not change. Following the literature, we define the equations governing skill production and investment in logs; empirical analogs of each model object are delineated in terms of effect sizes; the polygenic score is standardized to mean zero; the skill measures are estimated using standardized Bartlett factor scores. Therefore, the model objects are defined by:

$$g_i \equiv \exp(z_i), \quad \theta_{i,t}^k \equiv \exp(f_{i,t}^k), \quad k \in \{C, N\}, \quad (5.1)$$

where  $z_i$  is the EA polygenic score standardized to mean zero and  $f_{i,t}^k$  is the standardized Bartlett factor score for each skill type  $k$ . Note that by construction, the model logarithms are defined to be analogous to the standardized empirical objects, so by equation (5.1) we have:  $\ln g_i = z_i$  and  $\ln \theta_{i,t}^k = f_{i,t}^k$ , and the model object parameter values can be read directly as standardized effect sizes.

As explained by Section 2.1, the genetic endowment of child  $i$  is generated via the process of meiosis: each child draws a random set of alleles from each parent:

$$\ln g_{if} = \ln \bar{g}_f + \xi_i, \quad \text{Cov}(\xi_1, \xi_2) = 0, \quad (5.2)$$

where  $\ln \bar{g}_f \equiv E[\ln g_{if} \mid \text{parental genotypes}]$  is the expected value of the genetic endowment conditional on parental genes, and  $\xi_i$  is the random component drawn by each of the two siblings  $i$ . If we assume random mating, and applying the variance normalization from (5.1), we have:

$$\text{Cov}(\ln g_{1f}, \ln g_{2f}) = \text{Var}(\ln \bar{g}_f) \quad \text{Var}(\ln \bar{g}_f) = \text{Var}(\xi_{if}) = \frac{1}{2} \quad (5.3)$$

such that  $\text{Corr}(\ln g_{1f}, \ln g_{2f}) = 1/2$ . If we allow for assortative mating, the genetic endowments of the two parents are positively correlated, so the covariance between  $\ln g_{1f}$  and  $\ln g_{2f}$

exceeds the individual gene variance, and the genetic correlation between siblings will exceed 1/2. Rather than impose the random-mating split, we set  $\text{Var}(\ln \bar{g}_f) = \hat{r}$  and  $\text{Var}(\xi_i) = 1 - \hat{r}$ , where  $\hat{r}$  is the sibling correlation of the standardized score estimated directly in the sibling sample.

Let initial skills of each type, the (standardized) genetic endowment, and the family-fixed effect be distributed normally:

$$\begin{bmatrix} \ln \theta_0^C, \ln \theta_0^N, \ln g, \ln \psi \end{bmatrix}' \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad (5.4)$$

where  $\psi_f$  represents unobservable family-level heterogeneity: in the behavioral genetics literature, this is analogous to the unobservable “shared environment”. The skills  $\theta_0^C$  and  $\theta_0^N$  at age 6 enter as a latent initial condition and are not observed: Section 5.4 explains how we recover their distribution. Again drawing from our reduced-form analysis, the key aspect of  $\boldsymbol{\Sigma}_0$  we wish to highlight for investigation is the covariance between the genetic endowment and unobserved family-level factor,

$$\sigma_{g\psi} \equiv \text{Cov}(\ln \bar{g}_f, \ln \psi_f), \quad \text{Cov}(\xi_i, \ln \psi_f) = 0. \quad (5.5)$$

The covariance  $\sigma_{g\psi}$  represents parental genetic nurture: there is a natural correlation between the genes parents pass to their children and the family-level environment  $\psi_f$  they engender. The statement of zero covariance between child-level genetic variance conditional on parental genes and the family-level fixed effect is again a simple reflection of the exogenous nature of meiosis and Mendelian randomization: it is not an assumption but a consequence of biology.

Note that the reduced-form equation from Section 4 isolates the sibling-level difference  $\xi_i$  as the regressor, while the coefficient estimate reflects both the direct effect as well as the evocative response to  $e_i$  differences.

The object  $\psi_f$  is latent and its scale cannot be separately identified from the data moments. We impose the normalization  $\mathbb{E}[\ln \psi_f] = 0$  and  $\text{Var}(\ln \psi_f) = 1$ .

Finally, the model has no financial state. Money resources are given by:

$$Y_{f,t} = w_{f,t} L_{f,t} + \bar{a}_f, \quad (5.6)$$

As the equation illustrates, money is exogenous and there are no savings or borrowing: money is allocated within each period. While credit constraints are an important part of the literature, this model abstracts from that potential determinant of parental investment.

## 5.2 The Parent's Problem

### 5.2.1 Resource Allocation and Utility

At each period, the parents optimize across allocating resources to consumption  $C_{f,t}$ , educational expenditures ( $M_{i,t}$ ) on each sibling  $i$ , and time investment ( $H_{i,t}$ ) for each sibling  $i \in \{1, 2\}$ . Resources are exogenously drawn: parental labor income is denoted  $w_{f,t}$  and labor supply is denoted  $L_{f,t}$ : we abstract from the decision to supply labor to focus on the tradeoff between consumption and investment, and in particular the investment-type and sibling-sibling-investment tradeoffs. We take hours available to parents after supplying their labor as exogenous: this is consistent with the literature that investigates time spent with children specifically. (Guryan et al., 2008; Aguiar and Hurst, 2007). Without savings, consumption is determined by wages less investment expenditures:  $C_{f,t} = Y_{f,t} - M_{1,t} - M_{2,t}$ ; leisure is determined by total time less labor time less time investment:  $R_{f,t} = \bar{T} - L_{f,t} - H_{1,t} - H_{2,t}$ . Utility in each period is log-additive:

$$u_{f,t} = \gamma_1 \ln C_{f,t} + \gamma_2 \ln R_{f,t} + \sum_{i \in \{1,2\}} (\gamma_3 \ln \theta_{i,t}^C + \gamma_4 \ln \theta_{i,t}^N), \quad (5.7)$$

where  $\gamma_1$  denotes the weight on consumption,  $\gamma_2$  denotes the weight on leisure, and  $\gamma_3, \gamma_4$  denote the weights on each the cognitive and non-cognitive skills of each child: we assume symmetric weighting across children.

Here we can begin to formalize the statement from Section 4.2 regarding factors that determine sibling-sibling allocation decisions that inform the within-family investment coefficient: the degree to which reallocation occurs will depend on the parameters  $\gamma_1$  and  $\gamma_2$ , which govern the opportunity cost of increasing total investment rather than reallocating across siblings.

### 5.2.2 The Skill Formation Production Function

Cognitive and non-cognitive skills follow a trans-log production function (Houmark et al., 2024; Agostinelli and Weingarten, 2025). For skill type  $k \in \{C, N\}$ , we have:

$$\begin{aligned} \ln \theta_{i,t+1}^k &= \alpha_0^k + \alpha_1^k \ln \theta_{i,t}^C + \alpha_2^k \ln \theta_{i,t}^N + \alpha_3^k \ln g_i + \alpha_4^k \ln H_{i,t} + \alpha_5^k \ln M_{i,t} \\ &\quad + \alpha_6^k (\ln \theta_{i,t}^k \cdot \ln g_i) + \alpha_7^k (\ln \theta_{i,t}^k \cdot \ln H_{i,t}) + \alpha_8^k (\ln g_i \cdot \ln H_{i,t}) + \alpha_9^k \ln \psi_f + \eta_{i,t+1}^k, \end{aligned} \quad (5.8)$$

where  $H_{i,t}$  is time investment,  $M_{i,t}$  is financial investment, and  $\ln \psi_f$  represents the unobserved shared environmental factors contributing to skill production, drawn from equation

(5.4), with the assumed normalization discussed above. For each skill type, the family-level object  $\ln \psi_f$  functions as a Hicks-neutral parameter, acting as a shifter without changing the marginal returns to investment.

Perhaps the most notable aspect of the skill production specification is that time investment  $H_{i,t}$  is allowed to interact with the genetic endowment of the child, but monetary investment is not. This asymmetry is intentional: the survey data on within-family variation in monetary expenditure is limited, and we found no statistically significant results in the reduced-form analysis in Section 4.2. There is a tradeoff between separating time and monetary investments, motivated by different directional responses obtained in the data, and the degree to which within-family variation can be leveraged to identify the parameter. We compromise here by making the separation, but only allowing the time investment, which has richer sibling-specific data and variation, to interact with the genetic endowment and generate our estimates of the magnitude and importance of dynamic complementarity.

We also specify a process of per-period independent shocks given by:  $\eta_{i,t+1}^k \sim \mathcal{N}(0, \sigma_{\eta k}^2)$ ; each shock is independent across children, time period, and skill type. The uncertainty generated by the shock process is what causes the parent's problem to be dynamic; the shocks also account for the empirical variation in skill production that remains after regressing on lagged observables.

With reference to Section 4.1, we can now show how the structural model splits the original within-family fixed effect into three categories: the family component  $\ln \bar{g}_f$  of the genetic endowment, with coefficient  $\alpha_3^k$  and the interaction terms; the common monetary and time investments, with coefficients  $\alpha_4^k$  and  $\alpha_5^k$ , and  $\alpha_9^k \ln \psi_f$ . In the reduced-form setting, we could not separate these aspects of skill formation and educational attainment, however, they are relevant to our understanding and potential counterfactual policy analysis: in Section 6, for example, we will consider the effects of altering resources, while keeping (by necessity) the genetic endowment and  $\psi_f$  fixed.

Finally, recall from Section 4.2 our discussion of how the dynamic complementarity or substitutability in the skill production process informs whether parents reinforce or compensate for sibling differences: the degree of gene-investment complementarity is specified by the interaction term  $\alpha_8^k$ . As we will see, the structural approach elucidates this logic formally by showing how  $\alpha_8^k$  enters into the optimal time investment level.

### 5.2.3 Choice of Educational Attainment

In the final period  $T$ , corresponding to age  $a_i = 18$ , the parents choose whether to enroll their child in post-secondary education. The returns to enrolling are dependent on the terminal-period levels of cognitive and non-cognitive skills, the tuition costs, and a preference shock

$\nu_{i,T}$  which is distributed according to a standard logistic:

$$S_{i,T} \equiv V_T^{\text{college}} - V_T^{\text{no college}} = \kappa_0 + \kappa_1 \ln \theta_{i,T}^C + \kappa_2 \ln \theta_{i,T}^N - \kappa_3 \cdot \text{tuition}_f + \nu_{i,T}, \quad (5.9)$$

with  $D_{i,T}^{\text{college}} = 1$  when the value function corresponding to post-secondary education exceeds that of not-enrolling. Therefore, the enrollment probability is distributed according to the logistic CDF:

$$P_{i,T}^{\text{college}} = \Lambda\left(\kappa_0 + \kappa_1 \ln \theta_{i,T}^C + \kappa_2 \ln \theta_{i,T}^N - \kappa_3 \cdot \text{tuition}_f\right). \quad (5.10)$$

Of key importance is the way skill development enters into the return to college enrollment: this is an additional mechanism by which parental investments in skill development during periods  $t < T$  generate positive utility.

### 5.3 Model Solution

We can express the parent's problem recursively as the following Bellman equation:

$$V_t(\boldsymbol{\theta}_{1,t}, \boldsymbol{\theta}_{2,t} \mid g_1, g_2, \psi_f, Y_f) = \max_{\{M_{i,t}, H_{i,t}\}_{i \in \{1,2\}}} \left\{ u_{f,t} + \beta \mathbb{E}_t[V_{t+1}(\cdot)] \right\}, \quad (5.11)$$

subject to the two budget constraints:

$$\begin{aligned} C_{f,t} &= Y_{f,t} - M_{1,t} - M_{2,t} \\ H_{1,t} + H_{2,t} &\leq \bar{T} - L_{f,t} \end{aligned} \quad (5.12)$$

At time  $T$ , the terminal condition is given by:

$$V_T = \sum_{i \in \{1,2\}} \left[ \gamma_3 \ln \theta_{i,T}^C + \gamma_4 \ln \theta_{i,T}^N + \mathbb{E}[\max\{0, S_{i,T}\}] \right], \quad (5.13)$$

where, recalling Equation (5.9), we can take the expectation over the  $\nu_{i,T}$ , obtaining in closed form:  $E[\max\{0, S_{i,T}\}] = \ln\left(1 + \exp\{\kappa_0 + \kappa_1 \ln \theta_{i,T}^C + \kappa_2 \ln \theta_{i,T}^N - \kappa_3 \text{tuition}_f\}\right)$ .

#### 5.3.1 Parental Investment Shares

Since utility is logarithmic and the skill production function is trans-log, we can derive closed-form expressions for the share of resources devoted to time and monetary investment for each child. Letting

$$\lambda_{i,t+1}^k \equiv \frac{\partial V_{t+1}}{\partial \ln \theta_{i,t+1}^k}, \quad k \in \{C, N\}, i \in \{1, 2\}. \quad (5.14)$$

we obtain the following closed-form solutions for the first-order conditions for monetary and time expenditures:

$$M_{i,t}^* = \frac{\beta \Phi_{i,t}^M}{\gamma_1 + \beta \Phi_{1,t}^M + \beta \Phi_{2,t}^M} Y_{f,t}, \quad \Phi_{i,t}^M \equiv \sum_k \lambda_{i,t+1}^k \alpha_5^k, \quad (5.15)$$

$$H_{i,t}^* = \frac{\beta \Phi_{i,t}^H}{\gamma_2 + \beta \Phi_{1,t}^H + \beta \Phi_{2,t}^H} (\bar{T} - L_{f,t}), \quad \Phi_{i,t}^H \equiv \sum_k \lambda_{i,t+1}^k (\alpha_4^k + \alpha_7^k \ln \theta_{i,t}^k + \alpha_8^k \ln g_i). \quad (5.16)$$

We can see directly that the expenditure shares of each investment type for each child are determined by shadow-value-weighted own-child investment productivity, as a fraction of the sum of parental opportunity cost plus own-child productivity plus sibling productivity.

Note that the genetic endowment only enters optimal decision-making directly through the complementarity term  $\alpha_8^k \ln g_i$ : if  $\alpha_8^k > 0$ , then investments into the higher-endowment sibling are more efficient, and parents reallocate on that basis. The genetic endowment only influences optimal monetary expenditure  $M_{i,t}^*$  through the shadow values. Finally, since total available time is determined exogenously, parents who work longer hours will invest less time into the human capital development of their children, conditional on optimal share and endowment.

### 5.3.2 Own and Cross-Sibling Investment Responses

We can connect the expressions we have delineating the share rules to our question from Section 4.2. Equation (4.6) combined two structural objects: the genetic endowment-investment interaction in the skill production function, and the opportunity cost of increasing total investment (instead of reallocating). The first object is given by  $\alpha_8^k$  as it enters into Equation (5.16) through  $\Phi_{i,t}^H$ ; the second object is given by  $\gamma_2$ , which enters into the equation through the denominator alongside the sum of the productivity of investment. We can separate each object as follows. Let  $a_{i,t} \equiv \sum_k \lambda_{i,t+1}^k \alpha_8^k$ , representing the marginal effect of an increase in the genetic endowment of child  $i$  on the shadow-weighted marginal return to time investment. If we differentiate Equation (5.16), noting the denominator is common across siblings, we obtain:

$$\frac{\partial \ln H_{1,t}^*}{\partial \ln g_1} = \frac{a_{1,t} (\gamma_2 + \beta \Phi_{2,t}^H)}{\Phi_{1,t}^H (\gamma_2 + \beta \Phi_{1,t}^H + \beta \Phi_{2,t}^H)}, \quad \frac{\partial \ln H_{2,t}^*}{\partial \ln g_1} = \frac{-\beta a_{1,t}}{\gamma_2 + \beta \Phi_{1,t}^H + \beta \Phi_{2,t}^H}. \quad (5.17)$$

The across-sibling response will be reinforcing if  $\alpha_8^k > 0$ ; the magnitude of the cross-response declines as  $\gamma_2$  grows, which formalizes the notion of the soft-constraint case introduced in Section 4.2. Intuitively, if a parent weighs leisure heavily, at equilibrium their stock of leisure

will be relatively high; with log utility, the marginal cost of reallocation away from leisure (as opposed to reallocation away from investing in the sibling) will be relatively low.

If we subtract the cross-sibling response from the own-sibling response, the shared denominator falls out, and we obtain the elasticity:

$$\frac{\partial \ln H_{1,t}^*}{\partial \ln g_1} - \frac{\partial \ln H_{2,t}^*}{\partial \ln g_1} = \frac{a_{1,t}}{\Phi_{1,t}^H} = \frac{\partial \ln \Phi_{1,t}^H}{\partial \ln g_1}, \quad (5.18)$$

The denominator  $\Phi_{1,t}^H = \sum_k \lambda_{1,t+1}^k (\alpha_4^k + \alpha_7^k \ln \theta_{1,t}^k + \alpha_8^k \ln g_1)$  represents the value of one hour of time investment into child 1. The numerator  $a_{1,t} = \sum_k \lambda_{1,t+1}^k \alpha_8^k$  represents the value derivative with respect to the genetic endowment of child 1.

This object is the structural analog to the reduced-form coefficient  $\pi_m$  (where  $m$  is time investment) from Equation (4.5): however, note that the coefficient from the within-family regression represents one half of the sibling-difference averaged across siblings and averaged across ages.

Since there is no interaction term between monetary expenditure and the genetic endowment,  $a_{i,t}$  is replaced by  $\partial \Phi_{i,t}^M / \partial \ln g_i$ . Therefore, a shock to the genetic draw only alters parental monetary investment indirectly through the shadow values.

Recall that Section 4.2 found a small within-family investment response to sibling-sibling genetic variation, yielding three potential interpretations: (1) Parents invest in response to the dynamic complementarity/substitutability engendered by the genetic endowment, but the response is small; (2) Parents invest in response to realized skill, but the response is small or skill observation is noisy; (3) The parental response is large, but the investment measures obtained from the survey data are noisy. We can distinguish between interpretation (1) and (2) because the degree of complementarity is denoted by  $\alpha_8^k$ , while interaction with previous-period skill is governed by  $\alpha_7^k$ . We can distinguish between (1) and (3) by examining  $\alpha_8^k$  as estimated from skill outcomes alone: if the skill production function entails a small  $\alpha_8^k$ , then interpretation (1) is correct; if a large  $\alpha_8^k$  is entailed, then that is an indication that our investment measure isn't capturing the true magnitude of parental response.

## Recursive Solution

We solve following a backward recursion. We can differentiate (5.11) with respect to  $\ln \theta_{i,t}^k$ , which produces the return to time- $t$  skill  $k$ , and the change in utility generated by the change in optimal parental investments  $(H^*, M^*)$ , in response to the skill change. By the envelope theorem, at the optimal level of parental investments the marginal utility change is equal to zero, so only the direct value-function return to skill remains. For  $t < T$  and each child  $i$ ,

we obtain:

$$\lambda_{i,t}^C = \gamma_3 + \beta \mathbb{E}_t \left[ \lambda_{i,t+1}^C \left( \alpha_1^C + \alpha_6^C \ln g_i + \alpha_7^C \ln H_{i,t}^* \right) + \lambda_{i,t+1}^N \alpha_1^N \right], \quad (5.19)$$

$$\lambda_{i,t}^N = \gamma_4 + \beta \mathbb{E}_t \left[ \lambda_{i,t+1}^N \left( \alpha_2^N + \alpha_6^N \ln g_i + \alpha_7^N \ln H_{i,t}^* \right) + \lambda_{i,t+1}^C \alpha_2^C \right], \quad (5.20)$$

At time  $T$ , we obtain the terminal shadow values as directly specified functions of the returns to education:  $\lambda_{i,T}^C = \gamma_3 + \kappa_1 P_{i,T}^{\text{college}}$ , and  $\lambda_{i,T}^N = \gamma_4 + \kappa_2 P_{i,T}^{\text{college}}$ .

Note that the marginal effects of cross-sibling reallocation changes are zero because of the envelope theorem, so all cross-sibling terms drop out, aside from determining the equilibrium levels common across the shared denominators. Also note that we can integrate across the distribution of the  $\eta$  shocks, such that the expectation leaves every term deterministic at each stage.

The structure of the recursion is clear. At each stage, the object  $\Phi_{i,t}^H$  can be expressed as a function of  $t + 1$  objects, so we can iterate backwards from the terminal choice. The only interaction created by simultaneous optimization across siblings is created by the shared denominators in (5.15) and (5.16), for which we have closed-form expression in terms of  $t + 1$  objects. Therefore, solving one stage of the model requires only evaluating each of the four share rules (closed-form), then evaluating (5.19) and (5.20) for each child. We iterate backwards inductively on a simple four-dimensional grid tracking each of the four sibling-skill state variables  $(\theta_1^C, \theta_1^N, \theta_2^C, \theta_2^N)$ .

The Bellman equation yields the optimal parental investment decisions as functions of the four-dimensional state vector  $(\theta_1^C, \theta_1^N, \theta_2^C, \theta_2^N)$  at any period  $t$ . We can therefore imagine simulating the counterfactual effects of any number of fine-tuned policy interventions, examples of which we introduce in Section 6.

## 5.4 Calibration

We augment the model with the following objects that are not estimated jointly as structural parameters:

1. The discount factor  $\beta$  is fixed at 0.95 (Del Boca et al., 2014).
2. The parental time endowment  $\bar{T}$  is fixed at 112 waking hours per week; for the labor supply path  $L_{f,t}$ , we use reported parental work hours from the Wave 1 parent questionnaire, and we apply this level of weekly labor supply for all periods  $t$ .
3. The variances of the skill-production shocks  $\sigma_{\eta^C}^2, \sigma_{\eta^N}^2$  are calibrated prior to estimation by running auxiliary skill-production regressions and equating the shock variance to

the residual variance after regression of the skill-production factors.

We choose to calibrate rather than estimate the variances alongside the other parameters can be estimated consistently directly from the data; additionally, we want to discipline the important parameters of the model, as well as evaluate model fitness on their basis, rather than optimizing for fit by assigning noise to the shock variances. Note that for every bootstrap sample we re-calibrate the shock variances and re-estimate the model using the newly calibrated variances.

4. Finally, we calibrate the pre-sample parental investment trajectories using average parental investment data taken from the PSID Child Development Supplement (Panel Study of Income Dynamics, 1997; Hofferth and Sandberg, 2001). Investment at each period prior to the age corresponding to Wave 1 is given by the average age-specific investment for a household of similar demographics and resources. We can run Equation (5.8) backwards starting from the first skill observation in the Add Health data (using estimated skill-production parameters) to obtain the initial condition  $(\ln \theta_0^C, \ln \theta_0^N)$ . Note that each estimated set of technology parameters must generate a re-evaluated initial skill distribution, so we iterate until they are consistent with one another. While this process is relatively cumbersome, this story explains and endogenizes the sudden appearance of the first observed skill distribution later in the life-cycle, and more importantly, enables the counterfactual evaluation of targeted early childhood interventions we explore in Section 6.

## 5.5 Estimation

We implement the Simulated Method of Moments (SMM) to estimate the model. For each simulated trajectory, the family draws one  $\psi_f$ , a monetary trajectory  $Y_f$  taken from the empirical distribution of household income and net worth, and the two sibling endowments  $(g_1, g_2)$ , generated by (5.2). Applying the optimized equilibrium parental investments using the recursive solution described in Section 5.3, we simulate skills forward through Equation (5.8), and at the terminal period  $T$  the value functions (5.9) determine the choice of post-secondary education. On the basis of each simulated trajectory we compute simulated moments for matching.

The differences between the pooled and within-family moments are opened to investigation by the two-child structure of the model: with a one-child model, we would not be able to compare the simulated differences within-family as well as across-family. Note also that identifying the monetary and time resource moments requires variation in the distribution of  $Y_f$  across families, which we obtain directly from the data.

The moment vector comprises five categories of moments: (1) pooled and within-family educational attainment regressions; (2) pooled and within-family skill-production regressions of each of the two skill types; (3) post-secondary-margin moments; (4) the levels of parental hours and of annual money investment at ages 12–17, matched to the PSID-CDS profiles – the moments that discipline the utility weights; and (5) between-family skill variances, controlling for family resources and genetic endowment, thereby isolating the unobservable family-level factor. The vector has 43 elements; the money level (PSID-CDS prediction \$506 per year, s.e. 67) identifies the consumption weight  $\gamma_1$  jointly with the two money coefficients. We note that each simulated child can only match a moment at the age for which we have a corresponding Wave-observation for the counterpart child.

With relation to Section 4, note in particular that the covariance  $\sigma_{g\psi}$  is identified by the attenuation gap  $\Delta_g = \beta_g^{\text{pool}} - \beta_g^{\text{wf}}$  between the pooled and within-family estimates that motivates much of this analysis. Note also that we purposely exclude the within-family parental investment regression coefficients  $\pi_m$  from Equation (4.5): as previewed in Section 4.2, we can identify the relevant parameters from the skill-production moments, and use the implied values of  $\pi_m$  as a test of model fit.

Identification parameter by parameter is outlined in more detail in in Table 5.2.

Table 5.2: Identification: Parameter Blocks and Their Identifying Moments

| Parameter block                                         | Identifying moments                                                                                                                                         |
|---------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Technology levels $\alpha_1^k - \alpha_5^k, \alpha_0^k$ | Skill-production regressions, pooled and within family, at observed ages                                                                                    |
| Interactions $\alpha_6^k, \alpha_7^k, \alpha_8^k$       | Covariances of skill growth with endowment and investment; $\alpha_8^k$ from the within-family covariance of skill growth with the sibling score difference |
| Utility weights $\gamma_1, \gamma_2$                    | Mean money and time investment shares, Eqs. (5.15)–(5.16)                                                                                                   |
| Utility weights $\gamma_3, \gamma_4$                    | Covariance of enrichment time with lagged cognitive and non-cognitive skill                                                                                 |
| College choice $\kappa_0 - \kappa_2$                    | Post-secondary enrollment and completion moments                                                                                                            |
| Tuition sensitivity $\kappa_3$                          | In-state posted tuition by state and entry cohort                                                                                                           |
| Gene-environment covariance $\sigma_{g\psi}$            | Pooled–within gap $\Delta_g$                                                                                                                                |
| Family loading $\alpha_9^k$                             | Between-family skill variance net of endowment and investment                                                                                               |

A key point regarding identification combined with calibration is that conditional on genetic endowment, skill variance across-families could be created by two factors: the effects of differences in the unobserved family-environment  $\alpha_9^k \ln \psi_f$ ; or the covariance between

initial skills and unobserved family-factors:  $\text{Cov}(\ln \theta_0^k, \ln \psi_f)$ . We can distinguish these two stories by noting that a difference in initial skills decays at the rate  $\alpha_1^k$  per period, while the fixed factor  $\psi_f$  yields a persistent advantage. Therefore, the process of estimating  $\alpha_9^k$  while calibrating  $\text{Cov}(\ln \theta_0^k, \ln \psi_f)$  can be accomplished without bias.

We implement the Nelder-Mead method: we start with an identity weighting matrix and re-estimate with the inverse value of the bootstrapped variance for two-step efficiency; we calculate standard errors by implementing a 200-sample non-parametric bootstrap. We are over-identified: the number of independent moments we take from the data exceeds the 29 estimated parameters, yielding the opportunity to evaluate the model with a  $J$ -statistic test.

## 5.6 Parameter Estimates

We report the structural parameter estimates in Table 5.3.

Panel A reports parental utility parameters. The relative magnitudes  $\gamma_3$ ,  $\gamma_4$ , which correspond to the relative degree to which parents value cognitive and non-cognitive skills, respectively are identified by their role in the shadow values, delineated in Equations (5.19) and (5.20). The weights on consumption and leisure time are identified using the first-order conditions generated by (5.15) and (5.16).

Panel B reports the parameters governing the skill production function. The key parameter is  $\alpha_8^C > 0$ : the positive value indicates dynamic complementarity between time investment and genetic endowment, which motivates increased parental investment: this confirms a source of parental reinforcement and compounding cognitive skill inequality.

Panel C compares  $\alpha_8^N$  with  $\alpha_8^C$ : we investigate the differences in dynamic complementarity across skill type; Panel C also compares  $\alpha_3^N$  with  $\alpha_3^C$ , which investigates the relative degree to which the EA polygenic score influences cognitive and non-cognitive skills.

Panel D reports the parameters relevant to the post-secondary education decision:  $\kappa_1$ ,  $\kappa_2$  capture the degrees to which different realized terminal-period skill types covary with enrollment, and  $\kappa_3$  captures the degree to which tuition variation influences the enrollment decision.

Panel E reports  $\sigma_{g\psi}$ , which is a measure of genetic nurture, and its significance motivates the adoption of family fixed effects. Using the reported estimate of .03328, we obtain (under random mating), a correlation between the genetic endowment and the unobserved family-environment factor:  $\text{Corr}(\ln \bar{g}_f, \ln \psi_f) = .047065$ , using the fact that  $\text{Var}(\ln \bar{g}_f) = 1/2$ , from Equation (5.2), and using the fact that we imposed the variance of  $\psi_f$  to equal one.

We can test our model by using the estimated parameters to generate implied regression coefficients on parental investments. With reference to Equation (5.18), the skill-production

parameter estimates  $\alpha_8^C > 0$  and  $\alpha_8^N > 0$  generate higher time investment returns to sibling-specific endowment increases via complementarity; in turn, we should expect that parental time investments reinforce within-family genetic endowment differences. However, as reported in Section 4.2, we find that parental time investments respond insignificantly and to the extent there is a directional response, it is compensatory.

Although the structural model does not generate as strong a prediction for monetary expenditure due to the lack of interaction between monetary investment and genetic endowment, since we obtain  $\alpha_6^k > 0$  for both skill types, the shadow value Equation (5.19) implies a modest prediction that monetary parental investments should reinforce within-family endowment differences: indeed, we find this to be the case.

Returning to our three possible interpretations of the modest coefficients, we note:  $\alpha_8^k > 0$ , indicating dynamic complementarity and parental responsiveness to genetic endowments, ruling out interpretation (1);  $\alpha_7^k > 0$ , but only modestly, indicating moderate responsiveness to realized skills across time periods. Given these two facts in conjunction with our reduced-form results, the appropriate conclusion is interpretation (3): parental investment, motivated by the complementarity we found by estimating the skill production function, does respond to the genetic endowment, but our measures of parental investment, particularly time investment, are too coarse to clearly identify this responsiveness from the reduced-form regressions.

## 5.7 Identification Tests

Table 5.4 outlines the results of Hansen’s J test of overidentifying restrictions:

Table 5.3: Structural Parameter Estimates

| Parameter                                                               | Estimate | Std. Error |
|-------------------------------------------------------------------------|----------|------------|
| <i>Panel A: Utility Parameters</i>                                      |          |            |
| Consumption ( $\gamma_1$ )                                              | 8.190    | (93.390)   |
| Leisure ( $\gamma_2$ )                                                  | 0.530    | (0.316)    |
| Child Cognitive Skills ( $\gamma_3$ )                                   | 0.371    | (0.393)    |
| Child Non-Cognitive Skills ( $\gamma_4$ )                               | 2.956    | (2.675)    |
| <i>Panel B: Cognitive Skill Formation Technology</i>                    |          |            |
| Lagged Cognitive Skills ( $\alpha_1^C$ )                                | 0.565    | (0.025)    |
| Lagged Non-Cog. Skills ( $\alpha_2^C$ )                                 | 0.036    | (0.026)    |
| Genetic Endowment ( $\alpha_3^C$ )                                      | 0.073    | (0.079)    |
| Parental Enrichment Time ( $\alpha_4^C$ )                               | 0.043    | (0.029)    |
| Educational Expenditure ( $\alpha_5^C$ )                                | 0.017    | (0.068)    |
| Skills $\times$ Genetic Endowment ( $\alpha_6^C$ )                      | -0.037   | (0.026)    |
| Skills $\times$ Enrichment Time ( $\alpha_7^C$ )                        | 0.001    | (0.010)    |
| Genetic Endow. $\times$ Enrichment ( $\alpha_8^C$ )                     | 0.016    | (0.012)    |
| Constant ( $\alpha_0^C$ )                                               | -0.255   | (0.502)    |
| Family Unobservable ( $\alpha_9^C$ )                                    | 0.174    | (0.141)    |
| <i>Panel C: Non-Cognitive Skill Formation Technology</i>                |          |            |
| Lagged Cognitive Skills ( $\alpha_1^N$ )                                | 0.050    | (0.044)    |
| Lagged Non-Cog. Skills ( $\alpha_2^N$ )                                 | 0.557    | (0.032)    |
| Genetic Endowment ( $\alpha_3^N$ )                                      | -0.015   | (0.109)    |
| Parental Enrichment Time ( $\alpha_4^N$ )                               | 0.022    | (0.017)    |
| Educational Expenditure ( $\alpha_5^N$ )                                | 0.008    | (0.247)    |
| NC $\times$ Genetic Endowment ( $\alpha_6^N$ )                          | 0.005    | (0.030)    |
| NC $\times$ Enrichment Time ( $\alpha_7^N$ )                            | 0.000    | (0.006)    |
| Genetic Endow. $\times$ Enrichment ( $\alpha_8^N$ )                     | 0.014    | (0.006)    |
| Constant ( $\alpha_0^N$ )                                               | -0.045   | (0.173)    |
| Family Unobservable ( $\alpha_9^N$ )                                    | 0.170    | (0.230)    |
| <i>Panel D: College Choice Parameters</i>                               |          |            |
| Constant ( $\kappa_0$ )                                                 | 2.482    | (8.832)    |
| Cognitive Skills Loading ( $\kappa_1$ )                                 | 2.207    | (1.627)    |
| Non-Cognitive Skills Loading ( $\kappa_2$ )                             | -0.170   | (1.362)    |
| Tuition Sensitivity ( $\kappa_3$ )                                      | 0.046    | (1.708)    |
| <i>Panel E: Family Unobservable</i>                                     |          |            |
| Endowment-Unobservable Cov. ( $\sigma_{g\psi}$ )                        | -0.157   | (0.345)    |
| Implied Corr( $\ln \bar{g}_f, \ln \psi_f$ ) = $\sigma_{g\psi} \sqrt{2}$ | -0.222   |            |

Table 5.4: Overidentification Test

|                                           | Step 1 (identity $W$ ) | Step 2 (optimal $W$ ) |
|-------------------------------------------|------------------------|-----------------------|
| Moments, $\dim(m)$                        | 43                     | 43                    |
| Parameters, $\dim(\theta)$                | 29                     | 29                    |
| Degrees of freedom                        | 14                     | 14                    |
| Criterion at optimum, $Q(\hat{\theta})$   | 0.342                  | 0.020                 |
| $J$ -statistic, $S \cdot Q(\hat{\theta})$ | 1,923                  | 56.0                  |
| $p$ -value, $\chi^2_{14}$                 | —                      | $1 \times 10^{-7}$    |

Table 5.5 displays a sensitivity matrix connecting parameters to data moments.

Table 5.5: Sensitivity of Key Parameters to Targeted Moments

|                  | $\beta_g^{\text{pool}}$ | $\beta_g^{\text{wf}}$ | $\hat{\alpha}_8^C$ (wf) | $\hat{\alpha}_7^C$ (wf) | $\text{Var}_{\text{bf}}(\theta^C)$ | $\bar{H}$ | $P_{\text{college}}$ |
|------------------|-------------------------|-----------------------|-------------------------|-------------------------|------------------------------------|-----------|----------------------|
| $\alpha_8^C$     | −0.001                  | 0.001                 | 0.000                   | −0.003                  | −0.003                             | 0.000     | 0.000                |
| $\alpha_8^N$     | 0.001                   | 0.000                 | 0.001                   | 0.001                   | 0.001                              | 0.000     | 0.000                |
| $\alpha_9^C$     | 0.017                   | 0.013                 | 0.022                   | 0.010                   | −0.091                             | 0.000     | 0.001                |
| $\sigma_{g\psi}$ | −0.069                  | 0.129                 | 0.013                   | −0.027                  | −0.064                             | −0.002    | 0.013                |
| $\gamma_2$       | 0.005                   | 0.013                 | 0.032                   | −0.163                  | 0.014                              | 0.035     | −0.015               |
| $\gamma_3$       | −0.017                  | −0.019                | −0.029                  | −0.197                  | 0.022                              | 0.000     | 0.007                |
| $\kappa_3$       | −0.027                  | 0.035                 | 0.035                   | 0.722                   | −0.302                             | 0.004     | 0.001                |

Each entry is an element of  $\Lambda = -(G'WG)^{-1}G'W$ , evaluated at  $\hat{\theta}$ , where  $G = \partial m_{\text{sim}}/\partial \theta'$  is calculated using central finite differences, and  $W$  is the Step 2 weighting matrix. Each entry represents the change in the row parameter corresponding to a unit change in the column moment, holding all other moments fixed. Moments are standardized using their bootstrap standard errors.

Table 5.6 outlines a numerical check that parameters are locally pinned down: we create a numerical Jacobian matrix of sensitivity to simulated (not data) moments.

Table 5.6: Numerical Jacobian and Local Identification

|                                              |                                                                                        |
|----------------------------------------------|----------------------------------------------------------------------------------------|
| <i>Panel A: Rank</i>                         |                                                                                        |
| Rank of $G$                                  | 29                                                                                     |
| Required for local identification            | 29                                                                                     |
| Condition number of $G'G$                    | $3.9 \times 10^9$                                                                      |
| Smallest singular value of $G$               | 0.014                                                                                  |
| <i>Panel B: Weakly Identified Parameters</i> |                                                                                        |
|                                              | $\frac{\ \partial m / \partial \theta_p\ }{\max_q \ \partial m / \partial \theta_q\ }$ |
| Consumption weight ( $\gamma_1$ )            | 0.0023                                                                                 |
| Non-cognitive college loading ( $\kappa_2$ ) | 0.0024                                                                                 |
| Cognitive college loading ( $\kappa_1$ )     | 0.0033                                                                                 |

$G$  denotes the  $\dim(m) \times 29$  Jacobian of simulated moment derivatives with respect to the parameters, evaluated at  $\hat{\theta}$ . Each derivative is calculated using central finite differences, with the same simulation seed used for the positive and negative perturbations of each parameter. Panel B reports parameters with a column norm less than one percent of the largest column norm, indicating that the targeted moments are relatively insensitive to those parameters.

## 5.8 Model Fit

Table 5.7 shows model fit by comparing distance between the targeted and simulated moments. Table 5.8 investigates fit specifically by using the moments that were held out of estimation.

Table 5.7: Model Fit: Targeted Moments, Simulated Minus Data (Provisional)

|    | Moment  | Miss (SE)          |    | Moment  | Miss (SE)          |    | Moment  | Miss (SE)          |
|----|---------|--------------------|----|---------|--------------------|----|---------|--------------------|
| 1  | b_pool  | -0.95              | 16 | Cw_H    | -2.76 <sup>†</sup> | 31 | Nw_z    | -0.03              |
| 2  | b_wf    | +0.76              | 17 | Cw_thz  | +0.19              | 32 | Nw_H    | -2.07 <sup>†</sup> |
| 3  | m_yrs   | -0.23              | 18 | Cw_thH  | +1.86              | 33 | Nw_thz  | -1.67              |
| 4  | m_ps    | -3.31 <sup>†</sup> | 19 | Cw_zH   | -1.76              | 34 | Nw_thH  | +0.92              |
| 5  | Cp_thC  | -0.26              | 20 | Cw_cons | +0.24              | 35 | Nw_zH   | -2.84 <sup>†</sup> |
| 6  | Cp_thN  | +0.16              | 21 | Np_thC  | -0.20              | 36 | Nw_cons | +0.99              |
| 7  | Cp_z    | +2.70 <sup>†</sup> | 22 | Np_thN  | -1.77              | 37 | ps_z    | -3.91 <sup>†</sup> |
| 8  | Cp_H    | -0.62              | 23 | Np_z    | -0.65              | 38 | ps_thC  | -0.23              |
| 9  | Cp_thz  | -0.47              | 24 | Np_H    | +0.11              | 39 | ps_wf   | +1.15              |
| 10 | Cp_thH  | -1.07              | 25 | Np_thz  | +0.22              | 40 | vbf_C   | -1.31              |
| 11 | Cp_zH   | -0.01              | 26 | Np_thH  | -1.22              | 41 | vbf_N   | +1.08              |
| 12 | Cp_cons | -3.01 <sup>†</sup> | 27 | Np_zH   | +1.39              | 42 | m_H     | +1.04              |
| 13 | Cw_thC  | +3.28 <sup>†</sup> | 28 | Np_cons | +2.03 <sup>†</sup> | 43 | m_M     | +0.13              |
| 14 | Cw_thN  | -1.47              | 29 | Nw_thC  | -1.37              |    |         |                    |
| 15 | Cw_z    | -0.25              | 30 | Nw_thN  | +1.84              |    |         |                    |

Table 5.8: Held-Out Investment Response: Estimated and Model-Implied

|                                         | Estimated<br>(Section 4.2) | Model-implied |
|-----------------------------------------|----------------------------|---------------|
| Homework-help index, $\pi_H$ (% per SD) | 9.2 ( $p = 0.29$ )         | 54            |
| Monetary investment (% per SD)          | not estimable              |               |

The homework-help coefficient is held out of the SMM sample of moment targets.

## 6 Counterfactual Policy Analysis

We present three examples of using the structurally-informed estimated parameters to evaluate the counterfactual impacts of plausible policy interventions. Importantly, the precise fiscal and welfare evaluations of each policy will necessarily depend on an external parameter governing the effectiveness of the policy itself: naturally, plausible values of the governing parameters will often vary with context, and in most case be better understood by the local policymaker. Nevertheless, we present these policy evaluations as, at minimum, an exercise in illustrating the broad effectiveness of using our genetically-informed structural estimates for evaluating educational and developmental policies.

Two choices regarding the structural model inform the interpretation of each of the policy effects.

There is no financial state variable tracking cross-period borrowing or savings in the model, so any policy that alleviates a resource constraint does not operate in part by relaxing a credit constraint. The effects of each policy intervention operate directly on the relevant period- $t$  constraint or on the returns to skill production. Nevertheless, to the degree that parents draw financial resources and leisure time exogenously, equilibrium marginal wage will not necessarily be equated to equilibrium investment, so we cannot use that condition as informative of the marginal value of parental investment: we must consult the model parameters directly.

Second, that the future earnings of the child do not enter into time- $t$  parental utility (5.7) or time- $T$  terminal utility (5.13). While we do report related earnings consequences motivated by some of the later-Wave data that did not enter into the structural model, child earnings data do not enter into the main counterfactual welfare estimates, and are instead reported as auxiliary checks on plausibility and supplementary effect estimates.

Third, policies are not cost-equalized at their natural scales. Policies differ in natural scale, so Table 6.1 reports each at its natural scale and, in a second panel, at a common

fiscal cost per family per year; returns per dollar are reported at a marginal cost of public funds of 1.0, 1.3 and 1.5.

## 6.1 Targeted Tutoring Subsidy

We consider a tutoring subsidy, where a policymaker arranges for a non-parent tutor to provide time investment into a child for 6.06 hours a week per family. The family obtains this tutoring at no cost, and the tutoring functions as an alternative parental time investment  $H^{\text{sub}}$ , which substitutes for true parental time investment at a factor rate  $\rho$ . Total time investment into the child is therefore given by:  $H_{i,t} + \rho H^{\text{sub}}$ . In other words, tutoring is a linear substitute for parental time enrichment, although the degree to which the time is additive or substitutive depends on the parental equilibrium response.

Motivated by the frequent political desire to ameliorate inequality, we consider the effects targeting the tutoring program towards the 1,406 children in the bottom quartile of PGS score for educational attainment (equal to the genetic endowment  $\ln g_i$ ). We generate estimates using a value of  $\rho = 1$ , motivated in part by the existing literature (Nickow et al., 2020; Guryan et al., 2023), but also because we want to conspicuously not take a particular stance on the relative value of  $\rho$ , reinforcing the notion that this exercise is illustrative of how our model could be used.

We report the results as follows: we estimate completed years of schooling to increase among the targeted recipients by approximately .0018172 years; at the post-secondary enrollment margin, we estimate an increase in enrollment probability of 0 percentage points. Combining these two elements, we obtain an estimated welfare gain of approximately 7.647705 percent of yearly household money resources.

Much of Section 5 emphasized the importance and extent of dynamic complementarity in the production function of skills. This form of inequality-attenuating targeting is therefore less efficient than neutral or high-PGS targeting: we found that the skill (and therefore welfare) returns on time investment were increasing in genetic endowment. Nevertheless, this example of targeting illustrates how a policymaker might use the structurally-informed estimates of the complementarity parameter to trade off desire for equity against the expected decrease in efficiency resulting from targeting low-PGS students.

One of the major motivations for estimating a structural model is that the estimates are policy-invariant: we can use the parameters to directly account for how parents will respond to the free tutoring. If child 1 is the recipient, and letting  $F_t \equiv \gamma_2 + \beta\Phi_{1,t}^H + \beta\Phi_{2,t}^H$  denote the shared denominator from Equation (5.16), we can solve for the new equilibrium level of

parental time investment:

$$H_{1,t}^* = \frac{\beta\Phi_{1,t}^H(\bar{T} - L_{f,t}) - (\gamma_2 + \beta\Phi_{2,t}^H)\rho H^{\text{sub}}}{F_t}, \quad \frac{\partial H_{1,t}^*}{\partial(\rho H^{\text{sub}})} = -\frac{\gamma_2 + \beta\Phi_{2,t}^H}{F_t}, \quad (6.1)$$

$$H_{2,t}^* = \frac{\beta\Phi_{2,t}^H(\bar{T} - L_{f,t} + \rho H^{\text{sub}})}{F_t}, \quad \frac{\partial H_{2,t}^*}{\partial(\rho H^{\text{sub}})} = \frac{\beta\Phi_{2,t}^H}{F_t}, \quad (6.2)$$

$$R_{f,t}^* = \frac{\gamma_2(\bar{T} - L_{f,t} + \rho H^{\text{sub}})}{F_t}, \quad \frac{\partial R_{f,t}^*}{\partial(\rho H^{\text{sub}})} = \frac{\gamma_2}{F_t}. \quad (6.3)$$

Each hour of tutoring causes parental time investment to decrease by  $(\gamma_2 + \beta\Phi_{2,t}^H)/F_t$ ; total time invested into the child increases to  $(\beta\Phi_{1,t}^H(\bar{T} - L_{f,t}) + \beta(\Phi_{1,t}^H + \Phi_{2,t}^H)\rho H^{\text{sub}})/F_t$ , and the return to increased parental leisure is given by  $\gamma_2 \ln(1 + \rho H^{\text{sub}}/(\bar{T} - L_{f,t}))$ . Finally, a share of the substituted hours are reallocated to the untargeted sibling:  $\beta\Phi_{2,t}^H/F_t$ .

We find that the “crowd-out” effect increases when the parent values leisure more, and when the productivity term  $\Phi_{i,t}^H$  is smaller: recall from Equation (5.16) that  $\Phi_{i,t}^H$  is increasing in  $\alpha_g^k > 0$ . This is sensible and matches our intuitions: the greater the degree of complementarity, the lower the relative return of time investment into a low-endowment child, the greater the degree to which the parent will allow the tutor to substitute instead of add to time enrichment of a particular child with a relatively low PGS score.

## 6.2 Need-Based Scholarship

Again drawing on a potential motivation to reduce inequality across families, we consider the effects of a need-based scholarship. This policy applies a 50 percent subsidy to total college tuition for families in the bottom income quartile, as opposed to the bottom PGS quartile; this approach better reflects how need-based aid is actually administered (and is certainly more politically feasible at this time than a policy targeting by PGS).

We report the results as follows: we estimate a welfare increase equivalent to 1.300049 percent of annual household resources among the targeted recipients; using the estimated state-level college tuition costs from the Add Health data, we find that the policy costs approximately \$3246.214 per family per year, and generates returns of \$.019328 per dollar. We report that for targeted recipients, the probability of post-secondary enrollment increases by 2.212705 percentage points.

Note that increasing the returns to college for low-income families increases the probability of enrollment, which in turn increases the returns to skill investment and production. The results we obtain correspond to the perspective of a family with a child aged 6 at the

time of the policy announcement, such that parents can alter their investment trajectories at each time period in response to the new information.

### 6.3 Universal Early Academic Intervention

Third, we consider implementing a universal early (pre-age-6) intervention analogous to a pre-kindergarten intervention on all families (Heckman et al., 2010). We operationalize the effects of such an intervention by applying an upward level shift to the initial skill vector  $\ln \theta_0^C, \ln \theta_0^N$  at age 6 for all students.

Applying the calibrated values, we report the results as follows: post-secondary enrollment and four-year completion each rise by 0.02 percentage points and completed schooling by 0.002 years; the welfare gain is 5.82 percent of annual resources at a fiscal cost of \$10,440 per family, a return of 38 cents per dollar.

Applying the calibrated values, we report the results as follows: final cognitive skills increase .0014086 standard deviations; final non-cognitive skills increase .0014666 standard deviations; post-secondary enrollment increases by .0533239 percentage points; and we estimate a welfare increase equivalent to 10.08911 percent of annual household resources. Using the calibrated cost data, equal to \$10440 per family, we estimate a fiscal return of \$.6606186 per dollar invested in the policy intervention.

There are two effects: the initial effect is multiplied through by the skill persistence parameters  $\alpha_1^k, \alpha_2^k$  across each of the  $T$  periods: most of the direct skill effects are ultimately dissipated across repeated time-period multiplications. Additionally, there is the effect of increased parental time investment engendered by the dynamic complementarity ( $\alpha_7^k > 0$ ) of parental investment with previous-period skills, increasing the optimal level of parental investment. The investment effect, in the presence of lagged-skill-investment complementarity, is what accounts for most of the long-term effect.

Table 6.1 summarizes each of the counterfactual results.

Table 6.1: Summary of Counterfactual Policy Analysis

| Outcome                           | Tutoring<br>Subsidy | Need-Based<br>Scholar. | Early<br>Intervention |
|-----------------------------------|---------------------|------------------------|-----------------------|
| Post-secondary enrollment         | +0.21 pp            | +1.36 pp               | +0.02 pp              |
| 4-Year college completion         | +0.14 pp            | +1.21 pp               | +0.02 pp              |
| Years of schooling gain           | +0.012 yrs          | +0.085 yrs             | +0.002 yrs            |
| Lifetime earnings gain (%)        | +0.14%              | +1.14%                 | +0.02%                |
| Child cognitive skills<br>(SDs)   | +0.004              | +0.001                 | +0.001                |
| Child non-cog. skills (SDs)       | +0.001              | +0.000                 | +0.001                |
| Welfare Gain (%)<br>resources)    | +0.63%              | +0.08%                 | +5.82%                |
| Fiscal Cost (per<br>family/yr)    | \$7,874             | \$3,167                | \$10,440              |
| Return per \$ spent               | \$0.045             | \$0.0012               | \$0.381               |
| Concentrated in<br>Low-Wealth HH? | Strong              | Strong                 | Moderate              |

Lifetime earnings changes are estimated using an extrapolation of CPS age-earnings profiles. Return per dollar is equal to the ratio of the income-equivalent welfare gain to the estimated fiscal cost of the policy intervention. All dollar amounts converted to 2015 USD.

A clear limitation of our policy analysis is the extent to which it depends on parameter values or outside estimates not present in our data. However, we believe the exercise serves to illustrate how the parameters identified by the structural model are valuable: given parameter values that are more accurate or precise, we can “plug and play” the deep parameters to trace out a plausible counterfactual: one that explicitly accounts for parental responses to the changed policy environment.

## 7 Discussion and Conclusion

This paper has developed and estimated a structural model that investigates how children’s polygenic predisposition for educational attainment shapes completed schooling and college attendance, and specifically how parental investment decisions both respond to and amplify these genetic differences. Our core empirical contribution is a within-family Mendelian ran-

domization estimate of the effect of the child’s own polygenic predisposition on educational outcomes, exploiting Mendelian randomization among full siblings to account for the effects of genetic nurture and gene-environment confounding. We find an effect of 0.422 years of completed schooling per standard deviation in the pooled OLS specification, attenuating to 0.466 years in the within-family design. The  $-10$  percent attenuation between the two specifications is the empirical signature of family-level confounding: it reflects the differences in parenting, parental resources, investment, and other aspects of the environment that systematically covary with the genetic endowment.

A central concern is the extent to which the polygenic score is reflective of a true latent genetic factor, and the degree to which its contamination with equilibrium environmental responses makes it unsuitable as a starting object for analysis. While we cannot definitively overcome all questions, we take the degree to which even within-family GWAS-derived weights reflect equilibrium responses seriously: our structural model endogenizes parental response and illuminates the degree to which responses are reinforcing or compensatory.

Finally, we use the structural model and the unique time setting of our data to investigate both the direction and significance of the interaction between the initial genetic predisposition and parental investment during the teenage years specifically. Our model finds that positive complementarity dominates, which informs the relative efficiencies of different forms of policy targeting regimes.

Further research should fully incorporate raw genetic data to better understand the relationship between genetic variants and outcomes of interest to social scientists; relying on the polygenic score, even stripped of confounding, should not be considered satisfactory in our quest to understand the deep nature of human capital development and its relation to our genetic endowments.

## Appendix A. Glossary of Genetic Terms

**Allele.** An allele is one of two or more versions of DNA sequence (a single base or a segment of bases) at a given genomic location.

**Locus.** a physical site or location within a genome.

**Single-nucleotide polymorphism (SNP).** A single nucleotide polymorphism (abbreviated SNP, pronounced snip) is a genomic variant at a single base position in the DNA.

**Rare variant.** Typically, a rare variant is a genetic variant with minor allele frequency (MAF)  $< 1\%$ .

**Structural variant.** all genomic rearrangements involving at least 50bp of DNA.

**Diploid / haploid.** A diploid cell has two complete sets of chromosomes. Haploid refers to the presence of a single set of chromosomes in an organism's cells.

**Meiosis.** Meiosis is a type of cell division in sexually reproducing organisms that reduces the number of chromosomes in gametes.

**Polygenic.** A polygenic trait is a characteristic, such as height or skin color, that is influenced by two or more genes.

**Genome-wide association study (GWAS).** a research approach used to identify genomic variants that are statistically associated with a risk for a disease or a particular trait.

**Polygenic score (PGS).** A polygenic score (PGS) aggregates the effects of many genetic variants into a single number which predicts genetic predisposition for a phenotype.

**Ancestry principal components.** PC analysis creates new variables from SNP data, the principal components, corresponding to axes of genetic variation.

**Linkage disequilibrium (LD).** Linkage disequilibrium (LD) refers to the non-random association of alleles at different loci in a population.

**LD pruning / clumping.** The clumping step prunes redundant correlated effects caused by linkage disequilibrium (LD) between variants. In contrast to clumping, pruning does not take the p value of a SNP into account.

**p-value threshold.** Thresholding (T) consists in removing SNPs that are under a certain level of significance (P-value threshold to be determined).

**Sample overlap.** individuals or samples are included in both the discovery (training) dataset and the target (validation or application) dataset used for PRS analysis.

**Discovery vs. target sample.** the training GWAS is referred to as the discovery sample, and the testing dataset is referred to as the target sample.

**Genetic nurture.** Parental genes influencing offspring outcomes through environmental pathways.

**Evocative (evoked) gene–environment correlation.** An evocative (or reactive) genotype–environment correlation is evident when an individual’s genetically influenced traits or behaviors evoke a response from those around them.

**Mendelian randomization.** Mendelian randomisation is an analytical method that uses genetic variants as instrumental variables for modifiable risk factors that affect population health.

## Appendix B. Data Appendix

### B.1 Samples

Table B1: Analysis Samples

|                                                                      | Full sample | Sibling sample |
|----------------------------------------------------------------------|-------------|----------------|
| Analysis sample: respondents                                         | 5,730       | 826            |
| households (full sample) / families (sibling sample)                 | 5,307       | 403            |
| With household income (Table 4.1, Column 4): respondents             | 4,709       | 719            |
| families                                                             |             | 350            |
| With household income and parental education (Column 5): respondents | 4,312       | 697            |
| families                                                             |             | 339            |
| Bottom quartile of the EA score: respondents                         | 1,433       |                |

### B.2 Educational Attainment and Other Outcomes

Table B2: Wave 5 Credentials: Years of Schooling, ISCED Level, and Milestones

| Highest credential                                                 | Years | ISCED 2011 | Milestone indicators |               |                |                 |
|--------------------------------------------------------------------|-------|------------|----------------------|---------------|----------------|-----------------|
|                                                                    |       |            | High school          | Any post-sec. | 4-year college | Graduate degree |
| Eighth grade or less                                               | 8     | 1          |                      |               |                |                 |
| Some high school, no diploma                                       | 10    | 2          |                      |               |                |                 |
| High school diploma or GED                                         | 12    | 3          | ✓                    |               |                |                 |
| Vocational, technical or apprenticeship training after high school | 13    | 4          | ✓                    | ✓             |                |                 |
| Some college, no degree                                            | 13    | 3          | ✓                    | ✓             |                |                 |
| Associate degree                                                   | 14    | 5          | ✓                    | ✓             |                |                 |
| Bachelor's degree                                                  | 16    | 6          | ✓                    | ✓             | ✓              |                 |
| Some graduate school                                               | 17    | 6          | ✓                    | ✓             | ✓              |                 |
| Master's degree                                                    | 18    | 7          | ✓                    | ✓             | ✓              | ✓               |
| Professional degree (e.g., JD, MD)                                 | 19    | 7          | ✓                    | ✓             | ✓              | ✓               |
| Doctoral degree                                                    | 20    | 8          | ✓                    | ✓             | ✓              | ✓               |

Table B3: Outcome Measures

| Measure                      | Source                                   | Coding                                                    | Full  | Siblings | Enters                                                                                               |
|------------------------------|------------------------------------------|-----------------------------------------------------------|-------|----------|------------------------------------------------------------------------------------------------------|
| Completed years of schooling | Wave 5 (2016–18), ages 32–42; respondent | Highest credential mapped to years (Table B2); range 8–20 | 5,730 | 826      | Table 4.1; moments <b>b_pool</b> , <b>b_wf</b> , <b>m_yrs</b>                                        |
| High school completion       | As above                                 | Diploma, GED or higher                                    | 5,730 | 826      | Table 3.1                                                                                            |
| Any post-secondary education | As above                                 | Any education after high school                           | 5,730 | 826      | Within-family margin (Section 4.1); moments <b>m_ps</b> , <b>ps_z</b> , <b>ps_thC</b> , <b>ps_wf</b> |
| Four-year college completion | As above                                 | Bachelor’s degree or higher                               | 5,730 | 826      | Within-family margin; counterfactual outcome (Table 6.1)                                             |
| Graduate degree              | As above                                 | Master’s, professional or doctoral degree                 | 5,730 | 826      | Table 3.1                                                                                            |

## B.3 Skills

Table B4: Skill Indicators by Wave

| Wave                               | Skill     | Indicators (respondent)                                                                                                                                                      | Loadings | Full   | Siblings |
|------------------------------------|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|--------|----------|
| Wave 1<br>(1994–95), ages<br>12–19 | Cognitive | Add Health Picture Vocabulary Test<br>(administered, age-standardized);<br>self-reported most recent grades in<br>English and mathematics, coded so<br>that higher is better | estA16   | estA6  | estA11   |
|                                    | Non-cog.  | Two psychosocial scale scores<br>(self-report)                                                                                                                               | estA17   |        |          |
| Wave 2 (1996),<br>ages 13–20       | Cognitive | Grade-point average across<br>academic subjects (self-report)                                                                                                                | single   | estA7  | estA12   |
|                                    | Non-cog.  | Psychosocial composite (self-report)                                                                                                                                         | single   |        |          |
| Wave 3<br>(2001–02), ages<br>18–26 | Cognitive | Vocabulary assessment<br>(administered)                                                                                                                                      | single   | estA8  | estA13   |
|                                    | Non-cog.  | Conscientiousness measure<br>(self-report)                                                                                                                                   | single   |        |          |
| Wave 4<br>(2008–09), ages<br>24–32 | Cognitive | Cognitive battery (administered)                                                                                                                                             | estA18   | estA9  | estA14   |
|                                    | Non-cog.  | Big Five personality inventory<br>(self-report)                                                                                                                              | estA19   |        |          |
| Wave 5<br>(2016–18), ages<br>32–42 | Cognitive | Cognitive battery (administered)                                                                                                                                             | estA20   | estA10 | estA15   |
|                                    | Non-cog.  | Big Five personality inventory<br>(self-report)                                                                                                                              | estA21   |        |          |

## B.4 Genotype and Controls

Table B5: Genotype and Control Variables

| Variable           | Source                             | Coding                                                                                                                                                  | Full               | Siblings         | Enters                                                |
|--------------------|------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------|------------------|-------------------------------------------------------|
| EA polygenic score | Genotype (DNA collected at Wave 4) | Released score, Add Health PGS Release 2 (Braudt and Harris, 2020), weights from Lee et al. (2018); standardized in the analysis sample                 | 5,730              | 826              | All specifications ( $z$ )                            |
| Genetic ancestry   | Genotype                           | European-ancestry classification from genetic principal components against reference panels (Braudt and Harris, 2020); principal components as controls | 5,730              | 826              | Sample restriction; controls                          |
| Female             | Wave 1; respondent                 | Indicator                                                                                                                                               | 5,730              | 826              | Table 3.1; demographic controls                       |
| Birth order        | Wave 1                             | Position among the respondent's siblings                                                                                                                | 5,730              | 826              | Table 4.1, Columns 2–5                                |
| Birth year         | Wave 1                             | Quadratic                                                                                                                                               | 5,730              | 826              | Table 4.1, Columns 2–5                                |
| Wave indicators    | –                                  |                                                                                                                                                         | 5,730              | 826              | Table 4.1, Columns 3–5                                |
| Household income   | Wave 1 parent questionnaire        | Annual, \$1,000s                                                                                                                                        | 4,709              | 719              | Table 4.1, Columns 4–5; resources $Y_f$ (Section 5.5) |
| Parental education | Wave 1 parent questionnaire        | Years, mother and father                                                                                                                                | 4,312 <sup>a</sup> | 697 <sup>a</sup> | Table 4.1, Column 5; Table 3.1                        |

## B.5 Data from Outside Add Health

Table B6: External Data

| Source                                                                                                | Object                                                                             | Value           | Enters                                                 |
|-------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------|-----------------|--------------------------------------------------------|
| PSID Child Development Supplement (Panel Study of Income Dynamics, 1997; Hofferth and Sandberg, 2001) | Mean weekly hours of parental time with children aged 12–17                        | 13.48 hours     | Moment $m_H$ (Table 5.7)                               |
|                                                                                                       | Mean annual money investment in the child                                          | \$506 (s.e. 67) | Money-level moment                                     |
|                                                                                                       | Investment by age, ages 6–11, for households of similar demographics and resources | Age profile     | Pre-sample profile and initial condition (Section 5.4) |
| Current Population Survey                                                                             | Mean earnings by age, education bin, sex and earnings tercile                      | Ages 25–60      | Earnings projection (Equations (4.7)–(4.8))            |
| American Time Use Survey                                                                              | Parental weekly work hours by age and education                                    |                 | Robustness                                             |
| Posted in-state tuition                                                                               | Tuition by state and entry cohort                                                  |                 | Tuition sensitivity $\kappa_3$ (Table 5.2)             |

## References

- Francesco Agostinelli and Zach Weingarten. Modeling and measuring the genetic determinants of child development. NBER Working Paper 34427, National Bureau of Economic Research, October 2025.
- Francesco Agostinelli and Matthew Wiswall. Estimating the technology of children’s skill formation. *Journal of Political Economy*, 133(3):846–887, 2025. doi:10.1086/733396.
- Mark Aguiar and Erik Hurst. Measuring trends in leisure: The allocation of time over five decades. *The Quarterly Journal of Economics*, 122(3):969–1006, 2007. doi:10.1162/qjec.122.3.969.
- Benjamin W. Arold, Paul Hufe, and Marc Stoeckli. Genetic endowments, educational outcomes, and the moderating influence of school quality. *Journal of Political Economy Microeconomics*, 2025. doi:10.1086/738878.
- Orazio Attanasio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina. Estimating the production function for human capital: Results from a randomized controlled trial in Colombia. *American Economic Review*, 110(1):48–85, 2020. doi:10.1257/aer.20150183.
- Daniel Barth, Nicholas W. Papageorge, and Kevin Thom. Genetic Endowments and Wealth Inequality. *Journal of Political Economy*, 128(4):1474–1522, 2020.
- Maurice S. Bartlett. The statistical conception of mental factors. *British Journal of Psychology*, 28(1):97–104, 1937. doi:10.1111/j.2044-8295.1937.tb00863.x.
- Jere R. Behrman, Robert A. Pollak, and Paul Taubman. Parental preferences and provision for progeny. *Journal of Political Economy*, 90(1):52–73, 1982. doi:10.1086/261039.
- Daniel W. Belsky, Benjamin W. Domingue, Robbee Wedow, Louise Arseneault, Jason D. Boardman, Avshalom Caspi, Dalton Conley, Jason M. Fletcher, Jeremy Freese, Pamela Herd, Terrie E. Moffitt, Richie Poulton, Kamil Sicinski, Jasmin Wertz, and Kathleen Mullan Harris. Genetic analysis of social-class mobility in five longitudinal studies. *Proceedings of the National Academy of Sciences*, 115(31):E7275–E7284, 2018. doi:10.1073/pnas.1801238115.
- Uta Bolt, Eric French, Varun Warriar, and Qianyu Yang. Genetic endowments and lifetime earnings: Understanding the mechanisms. Working paper, this version March 2025, 2025.
- Amelia R. Branigan, Kenneth J. McCallum, and Jeremy Freese. Variation in the heritability of educational attainment: An international meta-analysis. *Social Forces*, 92(1):109–140, 2013. doi:10.1093/sf/sot076.
- David B. Braudt and Kathleen Mullan Harris. Polygenic scores (PGSs) in the national longitudinal study of adolescent to adult health (Add Health) — release 2. Technical report, Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, 2020.
- David Card. The causal effect of education on earnings. In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 3, chapter 30, pages 1801–1863. Elsevier, Amsterdam, 1999. doi:10.1016/S1573-4463(99)03011-4.

- Pedro Carneiro and James J. Heckman. Human capital policy. In James J. Heckman, Alan B. Krueger, and Benjamin M. Friedman, editors, *Inequality in America: What Role for Human Capital Policies?*, pages 77–239. MIT Press, Cambridge, MA, 2003.
- Elizabeth M. Caucutt, Lance Lochner, Joseph Mullins, and Youngmin Park. Early and late human capital investments, borrowing constraints, and the family. *Journal of Political Economy*, 128(3):1065–1147, 2020. doi:10.1086/704759.
- Christopher C. Chang, Carson C. Chow, Laurent C. A. M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4:7, 2015. doi:10.1186/s13742-015-0047-8.
- Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. Where is the land of opportunity? the geography of intergenerational mobility in the United States. *Quarterly Journal of Economics*, 129(4):1553–1623, 2014. doi:10.1093/qje/qju022.
- Flavio Cunha and James Heckman. The technology of skill formation. *American Economic Review*, 97(2):31–47, 2007. doi:10.1257/aer.97.2.31.
- Flavio Cunha, James J. Heckman, and Susanne M. Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931, 2010. doi:10.3982/ECTA6551.
- Daniela Del Boca, Christopher Flinn, and Matthew Wiswall. Household choices and child development. *Review of Economic Studies*, 81(1):137–185, 2014. doi:10.1093/restud/rdt026.
- Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, 9(3):e1003348, 2013. doi:10.1371/journal.pgen.1003348.
- Marco Francesconi and James J. Heckman. Child development and parental investment: Introduction. *The Economic Journal*, 126(596):F1–F27, 2016. doi:10.1111/ecoj.12388.
- Eric French, Jamie Hentall MacCuish, and Cormac O’Dea. The intergenerational elasticity of earnings: Exploring the mechanisms. *Journal of Political Economy*, 2026. Accepted for publication.
- Jonathan Guryan, Erik Hurst, and Melissa Kearney. Parental education and parental time with children. *Journal of Economic Perspectives*, 22(3):23–46, 2008. doi:10.1257/jep.22.3.23.
- Jonathan Guryan, Jens Ludwig, Monica P. Bhatt, Philip J. Cook, Jonathan M. V. Davis, Kenneth Dodge, George Farkas, Roland G. Fryer, Jr., Susan Mayer, Harold Pollack, Laurence Steinberg, and Greg Stoddard. Not too late: Improving academic outcomes among adolescents. *American Economic Review*, 113(3):738–765, 2023. doi:10.1257/aer.20210434.
- Steven Haider and Gary Solon. Life-cycle variation in the association between current and lifetime earnings. *American Economic Review*, 96(4):1308–1320, 2006. doi:10.1257/aer.96.4.1308.
- Robert M. Hauser. Disaggregating a social-psychological model of educational attainment. In Arthur S. Goldberger and Otis Dudley Duncan, editors, *Structural Equation Models in the Social Sciences*, pages 255–284. Seminar Press, New York, 1973.
- James J. Heckman. Skill formation and the economics of investing in disadvantaged children. *Science*, 312(5782):1900–1902, 2006. doi:10.1126/science.1128898.

- James J. Heckman, Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz. The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, 94(1–2): 114–128, 2010. doi:10.1016/j.jpubeco.2009.11.001.
- Sandra L. Hofferth and John F. Sandberg. How american children spend their time. *Journal of Marriage and Family*, 63(2):295–308, 2001.
- Mikkel Aagaard Houmark, Victor Ronda, and Michael Rosholm. The nurture of nature and the nature of nurture: How genes and investments interact in the formation of skills. *American Economic Review*, 114(2):385–425, 2024. doi:10.1257/aer.20220456.
- Laurence J. Howe, Michel G. Nivard, Tim T. Morris, Ailin F. Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, Rafael Ahlskog, Penelope A. Lind, Teemu Palviainen, et al. Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nature Genetics*, 54(5):581–592, 2022. doi:10.1038/s41588-022-01062-7.
- James J. Lee, Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, Peter Bowers, Julia Sidorenko, Richard Karlsson Linnér, et al. 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, 2018. doi:10.1038/s41588-018-0147-3.
- Wynn K. Meyer, Barbara Arbeithuber, Carole Ober, Thomas Ebner, Irene Tiemann-Boege, Richard R. Hudson, and Molly Przeworski. Evaluating the evidence for transmission distortion in human pedigrees. *Genetics*, 191(1):215–232, 2012. doi:10.1534/genetics.112.139576.
- Dilnoza Muslimova, Hans van Kippersluis, Cornelius A. Rietveld, Stephanie von Hinke, and S. Fleur W. Meddens. Gene-environment complementarity in educational attainment. *Journal of Labor Economics*, 44(3), 2026. doi:10.1086/734087.
- National Human Genome Research Institute. Human genomic variation. <https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genomic-variation>, February 2023. Last updated February 1, 2023.
- Andre Joshua Nickow, Philip Oreopoulos, and Vincent Quan. The impressive effects of tutoring on PreK-12 learning: A systematic review and meta-analysis of the experimental evidence. Working Paper 27476, National Bureau of Economic Research, 2020.
- Sergey Nurk, Sergey Koren, Arang Rhie, et al. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022. doi:10.1126/science.abj6987.
- Aysu Okbay, Yeda Wu, Nancy Wang, Hariharan Jayashankar, Michael Bennett, Seyed Moeen Nezhati, Julia Sidorenko, Hyeokmoon Kweon, Grant Goldman, Tamara Gjorgjieva, et al. Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics*, 54(4):437–449, 2022. doi:10.1038/s41588-022-01016-z.
- Panel Study of Income Dynamics. Child development supplement, CDS-I. Public use dataset. Produced and distributed by the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, 1997. URL <https://psidonline.isr.umich.edu/CDS/>. Accessed [date].

- William H. Sewell and Robert M. Hauser. *Education, Occupation, and Earnings: Achievement in the Early Career*. Academic Press, New York, 1975.
- Karri Silventoinen, Aline Jelenkovic, Reijo Sund, Antti Latvala, Chika Honda, Fujio Inui, Rie Tomizawa, Norio Sakai, et al. Genetic and environmental variation in educational attainment: An individual-based analysis of 28 twin cohorts. *Scientific Reports*, 10:12681, 2020. doi:10.1038/s41598-020-69526-6.
- Gary Solon. Intergenerational mobility in the labor market. In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 3, chapter 29, pages 1761–1800. Elsevier, Amsterdam, 1999. doi:10.1016/S1573-4463(99)03010-2.
- The International HapMap Consortium. The international hapmap project. *Nature*, 426(6968): 789–796, 2003. doi:10.1038/nature02168.
- Sam Trejo, Daniel W. Belsky, Jason D. Boardman, Jeremy Freese, Kathleen Mullan Harris, Pam Herd, Kamil Sicinski, and Benjamin W. Domingue. Schools as moderators of genetic associations with life course attainments: Evidence from the WLS and Add Health. *Sociological Science*, 5: 513–540, 2018. doi:10.15195/v5.a22.
- Sjoerd van Alten, Silvia H. Barcellos, Leandro Carvalho, Titus J. Galama, and Marina Aguiar Palma. A chip off the old block? genetics and the intergenerational transmission of socioeconomic status. Technical Report 34208, National Bureau of Economic Research, 2025. URL <https://www.nber.org/papers/w34208>.
- Alexander I. Young, Seyed Moeen Nehzati, Stefania Benonisdottir, Aysu Okbay, Hariharan Jayashankar, Chanwook Lee, David Cesarini, Daniel J. Benjamin, Patrick Turley, and Augustine Kong. Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nature Genetics*, 54(6):897–905, 2022. doi:10.1038/s41588-022-01085-0.