

Middle Childhood Development: Parental Investments, School Quality, and Genetic Influences*

Sarah Cattan Qianyao Ye

[Click Here for the Latest Version](#)

March 15, 2024

Abstract

In this paper, we examine how parental investments, school quality, genetics, and their interactions influence child development. Specifically, we estimate the skill production functions for both cognitive and socio-emotional skills. We implement an instrumental variable approach and leverage information from school application portfolios to address the potential endogeneity of parental investments and school quality. We use polygenic scores to capture an individual's genetic propensity for educational attainment. Using data from the Millennium Cohort Study in the UK, we find distinct patterns for cognitive skills and socio-emotional skills. Cognitive skills at age 7 are significantly influenced by parental investments, school quality, genetics, and lagged skills at age 5. Notably, school quality and polygenic scores are substitutes, indicating that better schools can mitigate skill disparities related to genetic predisposition for educational attainment. In contrast, socio-emotional skills at this stage are predominantly affected by previous skills and are less sensitive to investments.

*Cattan: Institute for Fiscal Studies, IZA Bonn, HCEO. sarah_c@ifs.org.uk. Ye: Yale University. qianyao.ye@yale.edu. *Acknowledgments:* We thank Joseph Altonji, Orazio Attanasio, and Costas Meghir as well as seminar participants at the Yale Labor/Public Prospectus workshop and the IFS EdSkills seminar for their valuable comments. We also thank the UK Data Service and UCL Centre for Longitudinal Studies for facilitating data access. Qianyao Ye gratefully acknowledges financial support from the MacMillan Center for International and Area Studies and Cowles Labor/Public Funds. All mistakes are our own.

1 Introduction

Child development displays substantial inequality, with children from high socioeconomic backgrounds more likely to accumulate higher human capital. Such inequality has significant consequences for labor market success and lifetime well-being. Therefore, it is important to understand the determinants of the development process and how investments can mitigate skill inequality.

It is well acknowledged that child development is shaped by both genetic factors and environmental factors. Yet, these factors are often treated as separate factors with their interaction effects ignored. In this paper, we study genetic factors and two key environmental factors during middle childhood: parental investments and school quality within a unified framework. We investigate whether genetic factors interact with parental investments and school quality to shed light on how public policies can intervene and reduce skill disparities due to genetic variations. We estimate a dynamic skill production function for both cognitive and socio-emotional skills during primary school.

Our measure of genetic variation is the *polygenic score*, a linear index of genetic markers correlated with educational attainment. Polygenic scores have been widely used to assess the risk of developing particular diseases, behavioral outcomes, and more recently, educational outcomes (Belsky et al., 2018; Lee et al., 2018; Okbay et al., 2018). The polygenic score for educational attainment assigns heavy weights to genetic markers related to brain development processes and neuron-to-neuron communication.

Papageorge and Thom (2020) find that the polygenic score for educational attainment predicts higher rates of college graduation and labor earnings. Barth et al. (2020) document the relationship between the polygenic score and wealth accumulation, highlighting a better understanding of complex financial decision-making as an important channel underlying the gene-wealth gradient. Bolyard and Savelyev (2019) show that the education polygenic score has a positive effect on several health-related outcomes. Houmark et al. (2020) document the direct effects of genes on skill development and its effects via parental investments. We contribute to this literature by studying the impacts of the education polygenic score on two dimensions of human capital: cognitive and socio-emotional skills.

This paper is among the first to study the gene-by-environment interaction effects. Bolyard and Savelyev (2019) and Papageorge and Thom (2020) use measures reflecting a family's socio-economic status to represent environmental effects. Biroli and Zünd (2020) and Barcellos et al. (2021) exploit alcohol licensing policy and compulsory schooling reform in the UK, respectively, to understand how these policy changes interact with genetic predispositions. Our paper is the first to examine how family envi-

ronment, as captured by parental investments, and public investments, as represented by school quality, interact with genetic factors together.

It is a concern that the education polygenic score may capture confounding factors in the family environment since children inherit genetics from their parents. For example, children with a higher genetic predisposition for educational attainment may also have parents with higher education polygenic scores. Since the education polygenic score predicts higher college graduation, labor earnings, and wealth, these parents may be better equipped to provide more resources and support to their children. We minimize these concerns by explicitly incorporating parental investments into our framework and by controlling for parents' cognitive skills, mental health, educational attainments, and a rich set of background variables.¹

We use data from the Millennium Cohort Study (MCS), linked with the National Pupil Database (NPD) in the UK. This dataset follows a cohort of children born around 2000 at ages 9 months, 3, 5, 7, 11, 14, and 17. Our study focuses on middle childhood after they enter primary school at age 7 and age 11. In addition to the DNA data for both parents and cohort members, it also provides rich measurements of child development, parenting activities, and the economic circumstances of households. These data allow us to construct the polygenic score and parental investments, and measure two dimensions of human capital: cognitive skills and socio-emotional skills. We rely on two sources of data to measure school quality: the Ofsted inspection results and the value-added (VA) measure from key stage 2 to key stage 1 from the NPD. Ofsted evaluates various aspects of a school, including leadership and management, value for money, teaching quality, and pupil behavior. The VA measure captures gains in English and maths. We use a latent factor model to measure parental investments, school quality, and skills, adapting the approach of [Cunha and Heckman \(2008\)](#) and [Cunha et al. \(2010\)](#).

It is challenging to identify the causal impacts of school quality and parental investments because parents' decisions regarding school and investments can be endogenous. Parents' school choice reflects their preferences, which can be correlated with unobserved factors that impact their child's development. Their investment decisions can respond to shocks to the development process. To address the endogeneity issue, we exploit the information in primary school application portfolios. Specifically, we consider an approach similar to the "matched-applicant" approach proposed by [Dale and Krueger \(2002, 2014\)](#) and [Mountjoy and Hickman \(2021\)](#) in the context of post-secondary enrollment. Students reveal their unobserved types by their applica-

¹We also have access to the education polygenic scores of the parents, but the sample size significantly decreases when including parents' polygenic scores. Our preliminary results suggest that once including the background variables mentioned above, parents' polygenic scores have no significant impact on skill development. Meanwhile, we are working with imputation methods to obtain a larger sample.

tion and admission portfolios. School assignment is as good as random for students sharing the same portfolio.

We focus on pupils who attend state-funded primary schools, which make up about 95% of all pupils of primary school age in England (Burgess et al., 2015). Admission to these schools is not merit-based. Instead, priority is given to pupils with special education needs, children with siblings in the same school, and those who live close to the school.² Taking into account the known admission rule,³ we modify the "matched-applicant" approach by focusing on the application portfolio only.

While we have information on the application portfolios, there is limited overlap among them. This is because households are distributed across diverse localities in our sample, and primary school applications are localized. Therefore, instead of comparing students with the same application portfolio, we focus on the characteristics of schools in the application portfolios and argue that these characteristics reveal households' preferences or types. We use school-level information from the school census and Edubase to construct a set of preference control variables and include them in the production function. The preference control variables include school academic performance, the share of students eligible for free school meals, school types, school denomination, whether siblings attended the same school, and home-school distance. The assumption is that conditional on these preference control variables (and the observed characteristics of households), the factors that lead to different school enrollments are unrelated to the potential outcomes of students. We present evidence that supports this assumption in section 4.3.

We also consider a complementary approach using the birthplace school quality as an instrument.⁴ The birthplace school quality is the school quality of the closest school to children's locations when they were just born at 9 months old. The birthplace school quality is a relevant instrument because the closer a child lives to a school, the more likely the child is to be admitted to that school. This instrument provides exogenous variation because parents' residential choices when their child was just born cannot respond to shocks in the child's development during primary school ages. While parents' residential choice might reflect their characteristics such as income or education, the assumption is that conditional on a very rich set of household characteristics and the child's previous skill development, birthplace school quality affects child development only through the actual school quality the child experiences.

To identify the impacts of parental investments, we use labor market shocks, prox-

²Admission priority is also given to children who are looked after by the state, but our analysis does not include this group of children.

³We have information on whether a child has special education needs, whether their siblings attend the same school, and their distances to schools applied.

⁴We only have birthplace school quality with the value-added measure and not the Ofsted rating for now. Therefore, we employ this strategy only for production function estimates at age 11 when we use the value-added measure.

ied by the female employment rate by local authority as an instrument. A positive shock likely induces parents to increase time at work and reduce time and effort devoted to their child, conditional on household incomes. By incorporating both parental investments and school quality into the production function, we also contribute to and bridge the child development literature and the education literature.

We find distinct results of cognitive skills and socio-emotional skills across different ages. For cognitive skills at age 7, they are significantly influenced by parental investments, school quality, genetics, and skills at age 5. A 10% increase in parental investments and school quality leads to a 1.34% and 0.21% increase in cognitive skills, respectively. A one standard deviation increase in the polygenic score increases cognitive skills by 3.2%. Notably, school quality and polygenic score are substitutes, indicating that better schools can mitigate skill disparities related to genetic predisposition for educational attainment.

Compared to the estimates at age 7, there are notable differences at age 11. Parental investments no longer have a significant impact on cognitive skills, while the influences of the school quality and the polygenic score have reduced considerably. We do not find evidence of interaction effects between school quality and the polygenic score at age 11. The impacts of previous skill endowments and current skill accumulation become stronger as children grow older. A 10% increase in previous cognitive skills leads to about a 5.6% increase and a 9% increase in current cognitive skills at age 7 and age 11, respectively. The results are consistent using either the preference control approach or the birthplace school quality as an instrument. These results indicate the importance of understanding the changing dynamics of skill development. Mitigating skill disparities related to genetic endowments calls for different public policies at different ages.

In terms of socio-emotional skills, high persistence is already evident at age 7. It is not affected by school quality, as measured by the Ofsted rating, or parental investments. While the polygenic score has a positive impact on socio-emotional skills, it is largely driven by previous socio-emotional skills. A 10% increase in socio-emotional skills at age 5 predicts about a 9% increase at age 7. For socio-emotional skills at age 11, the polygenic score no longer plays a role. While we still do not find impacts of school quality measured by the Ofsted rating, the value-added measure shows positive effects. This could be due to different aspects of school quality captured by different measures. Consistent with results at age 7, the primary determinant of socio-emotional skills at age 11 is previous skill development at age 7. Although earlier studies report positive impacts of parental investments on socio-emotional skills during early childhood ([Cunha et al., 2010](#); [Attanasio et al., 2020a](#)), our findings indicate that the windows of opportunity for parents to improve socio-emotional skills may be limited.

The paper is structured as follows: We first discuss the data and measurement in Section 2. Then we introduce the conceptual framework in Section 3. We present the empirical strategy in Section 4, followed by the estimation results in Section 5. Finally, Section 6 concludes.

2 Data and Measurement

2.1 Data

We use data from the Millennium Cohort Study (MCS), linked with the National Pupil Database (NPD). The MCS has followed a cohort of children born around 2000 in the UK and collected data when the cohort members were 9 months old, 3 years old, 5 years old, 7 years old, 11 years old, 14 years old, and 17 years old. In this study, we focus on children in their middle childhood and mainly use data from age 5, age 7, and age 11, corresponding to waves 3, 4, and 5 respectively. In each wave, multiple measurements of the cohort members' socio-emotional and cognitive development are available. It also contains rich information from both resident parents on their cognitive skills, parental investments, economic circumstances, and other demographics of the household. The NPD is an administrative dataset with information on cohort members' academic performance at schools in England. The DNA data is collected for both parents and the cohort member.

We present the descriptive statistics for the cohort members living in England in Table 1. Parents' cognitive skills are measured by their word activity assessments in wave 6, which is the first time that a cognitive assessment is available for parents. Parents' mental health is measured by the Kessler Psychological Distress Scale in wave 4. Parents' educational attainment is measured in wave 4.

We list the measures used for constructing cognitive skills and socio-emotional skills from age 5 to age 11 in Table 2. For parental investments, we focus on parenting activities that are more for educational purposes. At age 7, we use the measurements on the frequency of someone at home helping with reading, the frequency of someone at home helping with writing or spelling, and the frequency of someone at home helping with maths. At age 11, we have measurements on the frequency of someone at home helping with homework, and the frequency of someone at home making sure the cohort member has finished homework before doing other things such as watching TV or going out with friends.

Table 1: Summary statistics of the MCS sample

Variable	Obs	Mean	Std. Dev.
Child characteristics			
Minority	12414	0.268	0.443
Female	12440	0.488	0.5
First-born	12440	0.411	0.492
Child age (in months), w4	8988	86.723	2.978
Child age (in months), w5	8767	133.826	4.092
Household characteristics			
Mum age (in years), w4	8971	36.112	5.863
Dad age (in years), w4	7313	39.5	6.248
Both parents present, w4	8988	0.723	0.447
Number of children, w4	8987	2.587	1.127
HH. Earnings (\$1,000), w4	8705	22.53	41.321
Mum age (in years), w5	8755	39.978	5.841
Dad age (in years), w5	7121	43.389	6.192
Both parents present, w5	8767	0.655	0.476
Number of children, w5	8767	2.65	1.163
HH. Earnings (\$1,000), w5	8413	23.052	32.808
Mum cognitive skills	6961	10.856	4.57
Mum mental health	9201	20.565	4.062
Dad cognitive skills	4623	11.676	4.646
Dad mental health	6823	20.634	3.715
Mum education			
Above A Level	12082	0.168	0.374
A Level	12082	0.285	0.452
GCSE or below	12082	0.547	0.498
Dad education			
Above A Level	9235	0.212	0.409
A Level	9235	0.344	0.475
GCSE or below	9235	0.445	0.497

Notes: This table shows the summary statistics of the MCS sample in England. 'Minority' refers to children who are not white. 'HH.' stands for household. Parents' cognitive skills are measured by their word activity assessments in wave 6. Parents' mental health is measured by the Kessler Psychological Distress Scale in wave 4. Parents' educational attainment is measured in wave 4. 'w4' refers to wave 4 while 'w5' refers to wave 5.

Table 2: Skill measurements

	Age 5	Age 7	Age 11
Cognitive skills	BAS Naming Vocabulary	BAS Pattern Construction,	BAS Verbal Similarities
	BAS Pattern Construction	BAS Word Reading	Maths national curriculum level achieved
	BAS Picture Similarities	NFER Progress in Maths	English national curriculum level achieved
		Maths national curriculum level achieved	
Socio-emo. skills		English national curriculum level achieved	
	SDQ Emotional Symptoms	SDQ Emotional Symptoms	SDQ Emotional Symptoms
	SDQ Conduct Problems	SDQ Conduct Problems	SDQ Conduct Problems
	SDQ Hyperactivity/Inattention	SDQ Hyperactivity/Inattention	SDQ Hyperactivity/Inattention
	SDQ Peer Problems,	SDQ Peer Problems,	SDQ Peer Problems,
	SDQ Prosocial	SDQ Prosocial	SDQ Prosocial
	CSBQ Independence/Self Regulation	CSBQ Independence/Self Regulation	
	CBSQ Emotional-Dysregulation	CBSQ Emotional-Dysregulation	
	CBSQ Cooperation		

Notes: BAS: British Ability Scales, NFER: National Foundation for Educational Research, SDQ: Strengths and Difficulties Questionnaire, CSBQ: Children's Social Behavior Questionnaire.

Our school quality measures come from two sources. The first measure we consider is from the Ofsted inspection results. The inspection gives ratings on several aspects of a school, covering leadership and management, value for money, teaching quality, and pupil behaviors. We use these measurements to construct a factor score as discussed in Section 2.2. The second measure we use is the school-level value-added (VA) measure from the key stage 2 (KS2) to key stage 1 (KS1) from the NPD. According to the Department for Education (2016), an individual pupil’s estimated KS2 performance is calculated by the average of all pupils’ actual KS2 performance who have similar performance at KS1 in the whole country. The difference between the estimated KS2 performance and the actual performance is an individual’s VA measure. Averaging all pupils’ VA scores at a school gives the school-level VA measure.

To control for school preferences, we first get information on the application portfolio from the MCS. We then merge school characteristics including academic performance, the share of students eligible for free school meals, school types, and school denomination using the Edubase and the school census.

As discussed in Section 4, we also construct a birth-place school quality measure as an instrument. Specifically, we know the children’s location when they were just born (at 9 months old in wave 1) at the Output-Areas (OA) level.⁵ For each OA, we find the closest school and its value-added measure and use this value to construct an OA school quality. If there is more than one school with the same distance, we take the average of the value-added measures of these schools. We then assign this OA-level value-added measure to each child based on their location in wave 1 and name this measure as the birthplace school quality.

One of the key inputs we consider in the production function is the genetic endowment, proxied by the polygenic score of educational attainment. Polygenic scores have been used widely to assess the risks of developing a particular disease, behavioral outcomes, and more recently educational outcomes (Belsky et al., 2018; Lee et al., 2018; Okbay et al., 2018). A genetic score captures the genetic variants that are associated with a specific outcome based on large sample analysis.

Formally, a human genome consists of 23 pairs of DNA molecules called *chromosomes*. An individual inherits one copy of a chromosome from each parent. More than 99% of locations along human chromosomes are identical. Locations where individuals differ by a single genetic marker are called *single nucleotide polymorphisms* (SNPs). People can have one of the two possible generic variants for most SNPs and the variant is called *allele*. One of the two possible alleles is chosen as the *reference allele*. At a given SNP, an individual can have zero, one, or two of the reference allele since we

⁵Output Areas represent the smallest geographical units for which census statistics are compiled. They consist of 40 to 250 households and typically have a resident population ranging from 100 to 625 individuals.

have two copies of each chromosome. We use the number of the reference allele to construct polygenic scores.

The polygenic score is constructed with estimates from *genome-wide association studies* (GWAS). GWAS examines associations between SNP-level data to various outcomes, including height, diseases, or socioeconomic outcomes. Specifically, it regresses the outcome of interest on the number of reference alleles an individual has at each SNP. These regressions are univariate and run for each SNP, one at a time.

A PGS is calculated as follows:

$$PGS_i = \sum_j g_{ij}w_j,$$

where PGS_i is the polygenic score for individual i , g_{ij} is the number of reference allele that individual i has at SNP j , and w_j is the weight for SNP j , derived from estimates from a GWAS.

In this paper, we use GWAS coefficients from [Lee et al. \(2018\)](#), which use a discovery sample of over 1.1 million people to estimate the association between SNPs and educational attainment. They show that the constructed score explains 12.7% and 10.6% of the variation in the years of education in the National Longitudinal Study of Adolescent to Adult Health and the Health and Retirement Study, respectively. We will use PGS to refer to this polygenic score for educational attainment. We control for the first ten principle components of the full matrix of genetic data to control for population stratification in all specifications involving the PGS.

2.2 Measurement System

While we have multiple measurements available for skills, parental investments, and school quality, using any one of these measurements can introduce estimation bias because they are imperfect proxies that often contain measurement errors. Following the approach of [Cunha and Heckman \(2008\)](#) and [Cunha et al. \(2010\)](#), we model skills, parental investments, and school quality as latent factors. We develop a measurement system that links the observed measures to latent factors and estimate the distribution of these factors. This approach allows us to efficiently utilize all available measurements for each latent factor and account for measurement errors.

In this section, we discuss the theory and specification of the measurement system. The measurements are all categorical for parental investments and school quality,⁶ all continuous for socio-emotional skills, and a mixture of continuous and categorical variables for cognitive skills.

⁶We use the Ofsted ratings to construct the factor scores for school quality since the value-added measure is only available at age 11.

Let m_{jki} denote the j th available measurement related to latent factor k for individual i . For continuous measurements, we assume the following semi-log relationship between the measurements m_{jki} and the latent factor $\ln\theta_{ki}$, as we consider the latent factor θ_{ki} to be strictly positive.

$$m_{jki} = \alpha_{jk} + \lambda_{jk}\ln\theta_{ki} + \epsilon_{jki},$$

where α_{jk} is the intercept, λ_{jk} is the factor loading, ϵ_{jki} is the measurement error.

When the observed measurement m_{jki} is categorical, we assume it is a manifestation of a continuous latent item m_{jki}^* . The latent item m_{jki}^* , in turn, has a semi-log relationship with the latent factor θ_{ki} ,

$$m_{jki}^* = \alpha_{jk} + \lambda_{jk}\ln\theta_{ki} + \epsilon_{jki},$$

The threshold model below captures the relationship between the continuous latent item m_{jki}^* and the observed item m_{jki} :

$$m_{jki} = \begin{cases} 1 & \text{if } m_{jki}^* < \tau_{1,jk}, \\ 2 & \text{if } m_{jki}^* \in [\tau_{1,jk}, \tau_{2,jk}], \\ \dots & \\ n & \text{if } m_{jki}^* > \tau_{n-1,jk}, \end{cases}$$

where $\tau_{n,jk}$ is the n^{th} threshold.

Our assumptions are that the measurement errors are mean zero, independent of the latent factors, and independent of each other. The measurement errors are normally distributed, and the latent factors are log-normally distributed.⁷ Due to the latent factors lacking an intrinsic scale or location, we introduce normalization assumptions to establish their scale and location.

First, for the location of the latent factors, we normalize the means of log parental investments and log school quality to be zero. However, it is important to allow the *dynamic* latent factors, i.e. cognitive skills and socio-emotional skills, to grow over time. Restricting the log skills to be mean zero overtime can introduce bias in the production function (Agostinelli and Wiswall, 2016). Consequently, we constrain the intercept of one measurement for each latent factor to be zero, and we denote this measurement

⁷These assumptions are more restrictive than necessary for identification purposes. Measurement errors are allowed to be correlated with each other, provided that at least one measure's error remains independent of those of other measures associated with the same factor. The latent factor can follow a mixture of normal distributions if all measurements are continuous, as demonstrated in Cunha et al. (2010) and Attanasio et al. (2020c).

as the reference measurement m_{1ki} .⁸ The assumption is that the mapping from the reference measurement to the related factors is age-invariant. This assumption allows us to attribute the observed growth in the measurements to the growth of the related factors only.

Second, the scale of the latent factors is normalized to be the unit of the reference measurements. We achieve this by setting the factor loading of m_{1ki} to be one, i.e., $\lambda_{1k} = 1$ for factor k . As pointed out by [Agostinelli and Wiswall \(2016\)](#), we should use a consistent scaling of latent factors so that the *dynamic* latent factors are comparable over time. Ideally, we would like to use the same reference measurements across ages. For socio-emotional skills, we use the "SDQ Conduct Problems" as the reference measurement, and set its factor loading to one. For cognitive skills, there is no single measurement that spans the three ages we study. We follow the approach of [Attanasio et al. \(2020a\)](#) to make use of the measures that overlap at least in one time period.⁹

At age 5 and age 7, we have "BAS Naming Vocabulary" available. At age 7 and age 11, "Maths national curriculum levels achieved" and "English national curriculum level achieved" are available. Such overlap allows us to construct a metric for the factors that can be used through the three ages. Specifically, let's denote "BAS Naming Vocabulary" at age 5 as m_{ac1i} , "BAS Naming Vocabulary" at age 7 as m_{ac2i} , "Maths national curriculum levels achieved" at age 7 as m_{bc2i} , and "Maths national curriculum levels achieved" at age 11 as m_{bc3i} . We normalize the location and scale to be equal to that of "BAS Naming Vocabulary" at age 5 and age 7, i.e., $\alpha_{ac1i} = \alpha_{ac2i} = 0$ and $\lambda_{ac1i} = \lambda_{ac2i} = 1$. Then we use the intercepts and factor loadings of m_{bc2i} to express the location and the scale of cognitive skills at age 11 in the same metric by setting $\alpha_{bc2i} = \alpha_{bc3i}$ and $\lambda_{bc2i} = \lambda_{bc3i}$. As "Maths national curriculum levels achieved" are categorical measures, we also restrict the thresholds to be identical across ages 5 and 7.

We make additional assumptions to identify the measurement system with categorical measures. Given that it is not possible to jointly identify the thresholds and the intercepts, we set all the intercepts to be zero for categorical items. In the absence of an intrinsic scale for both the latent item and the latent factor, we set the variance of the latent items m_{jki}^* to be one for all associated categorical measurements. The residual variances are $V(\epsilon_{jki} = 1 - \lambda_{jk}^2 V(\ln \theta_{ki}))$.¹⁰

For a measurement system with one latent factor, a minimum of three measurements per factor is necessary for identification. The presence of multiple latent fac-

⁸This constraint is equivalent to normalizing the means to be the means of the reference measurements.

⁹There is no overlapping measure for parental investments, so we use different measures at age 7 and age 11 as the reference measures.

¹⁰Alternatively, we can set the residual variances $V(\epsilon_{jki})$ to be one and obtain the variance of latent items as $V(m_{jki}^*) = \lambda_{jk}^2 V(\ln \theta_{ki}) + 1$.

tors within a measurement system allows for a reduction in the required number of measurements per factor. We assume a dedicated measurement system, where each measurement exclusively proxies a single factor. While this assumption is not necessary for identification, it enhances the interpretability of the latent factor.¹¹ Lastly, we assume a separable mapping from the latent factors to the measures. [Cunha et al. \(2010\)](#) allows for a non-separable mapping, and shows that with a minimum of three measures, non-parametric identification of the joint distribution of the latent factors and the measurement errors is feasible.

3 Conceptual Framework

In this section, we first provide the institutional background of the educational system in England before presenting the conceptual framework. We focus on pupils who attend state-funded primary schools, which account for approximately 95% of all primary school-aged children in England ([Burgess et al., 2015](#)). Parents apply to these state schools through their local authority (LA), where they can nominate and rank schools on the common application form. In the MCS, more than half of the families apply to only one school (%57), about 24% families nominate two schools, and around 14% indicate a third choice. Admission to state schools is typically not based on merit, with priority given to pupils with special education needs, children with siblings in the same school, and those who live closer to the school in terms of straight-line distance. While nearly all students receive an offer from one of their top three choice schools, those who do not receive any offer will be allocated to a school by the LA.

We use a two-period model to illustrate the decision process of the parents and the potential issues we face in identification. Period t corresponds to age 7, while period $t - 1$ corresponds to age 5 when parents make application decisions. We start backward from the second period, period t , where parents derive utility from current consumption C_t and future human capital level θ_{t+1} . For simplicity, we assume θ_{t+1} is one-dimensional in the model, although in practice, we consider the multidimensional development of human capital.

Parents make decisions on consumption C_t and parental investments I_t subject to the household budget constraint and the skill production function.

$$U_t = \max_{C_t, I_t} U(C_t, \theta_{t+1}),$$

s.t.

¹¹Provided that there exists at least one measure loading exclusively on one factor, other measures are allowed to be associated with multiple factors.

the household budget constraint:

$$C_t = w_t(1 - I_t) + y_t,$$

and the skill production function:

$$\theta_{t+1} = f(\theta_t, I_t, Q_{R^*,t}, pgs, \epsilon_t, \kappa^\theta),$$

where w_t is the wage rate, y_t is non-labor income, $Q_{R^*,t}$ is school quality, pgs is the polygenic score of educational attainment, ϵ_t is a shock, and κ^θ captures idiosyncratic tastes for skill development. Total time is normalized to be one. School quality Q depends on the school admission outcome R^* , which is a function of school choice parents make in the previous period R_{t-1} .

In period $t - 1$, parents make school application choices. They value a set of school characteristics S_{t-1} , and their valuation W_{t-1} depends on their observed characteristics X_{t-1} and an idiosyncratic preference κ^s for school choices. School characteristics S_{t-1} may include student composition, school location, and other aspects. Parents maximize utility by choosing a school R from a choice set \mathcal{N} that includes all state schools within their LA, accounting for its impacts on future human capital development. We abstract from the multi-nomination nature of the school application process, focusing instead on the most preferred school choice made by parents. This modeling choice does not affect the analysis presented below.

$$\max_{R_{t-1}} W_{t-1} \Gamma_w S'_{t-1} + \beta E(U_t),$$

where $W_{t-1} = [X_{t-1} \ \kappa^s]$ is a $1 \times L$ vector, with $L - 1$ observed household characteristics X_{t-1} and an idiosyncratic taste κ^s , Γ_w is a $L \times K$ matrix of parameters, S_{t-1} is a $1 \times K$ vector with K school characteristics. β represents a discount factor and U_t denotes the utility at period t . The expectation is taken over the shock to the skill production function and school admission outcome realized in the next period t .

The objective of this paper is to estimate the skill production function and this model helps us understand the potential sources of endogeneity in parental investments and school quality. Parental investment is a function of current skill θ_t , school quality $Q_{R^*,t}$, genetic endowment pgs , shocks ϵ_t , idiosyncratic tastes κ^θ , wage rate w_t and non-labor income y_t :

$$I_t^* = l(\theta_t, Q_{R^*,t}, pgs, \epsilon_t, \kappa^\theta, w_t, y_t).$$

In particular, parents may respond to shocks ϵ_t that are not observed by researchers. For example, if parents notice that their child is experiencing adverse shocks, such as

an illness, they may increase investments to support the child. Parental investments can also be correlated with idiosyncratic tastes κ^θ for skill development, as parents may adjust their investments based on their child's motivation. Therefore, the correlation between parental investments and skills can be affected by these confounding factors and may not capture the causal effects of parental investments. We consider an instrumental variable approach to deal with the endogeneity concern.

Regarding school choice R_{t-1} , it depends on characteristics $\{S_{t-1}\}$ of schools in the choice set \mathcal{N} , observed characteristics of households X_{t-1} , and idiosyncratic tastes κ^s . The idiosyncratic tastes κ^s for school can be correlated with the idiosyncratic tastes κ^θ for skill development. More motivated parents may opt for higher-quality schools, and their children might also be more inclined to cultivate their skills. In other words, we need to control for these unobserved preferences or types to have a causal interpretation of the effects of school quality. We discuss our empirical approach in the following section.

4 Empirical Strategy

4.1 Empirical Specification

We consider the following specifications for cognitive skills and socio-emotional skills.

$$\begin{aligned} \ln\theta_{t+1}^c = & \alpha_0 + \alpha_1 \ln\theta_t^c + \alpha_2 \ln\theta_t^s + \alpha_3 \ln I_t + \alpha_4 \ln Q_t + \alpha_5 pgs \\ & + \alpha_6 \ln I_t \times pgs + \alpha_7 \ln Q_t \times pgs + \mathbf{Z}_t \Gamma^c + \eta_t, \end{aligned} \quad (1)$$

$$\begin{aligned} \ln\theta_{t+1}^s = & \beta_0 + \beta_1 \ln\theta_t^c + \beta_2 \ln\theta_t^s + \beta_3 \ln I_t + \beta_4 \ln Q_t + \beta_5 pgs \\ & + \beta_6 \ln I_t \times pgs + \beta_7 \ln Q_t \times pgs + \mathbf{Z}_t \Gamma^s + u_t, \end{aligned} \quad (2)$$

where θ^c and θ^s represent cognitive and socio-emotional skills, respectively, I_t are parental investments, Q_t is school quality, and pgs is the polygenic score of educational attainment. \mathbf{Z}_t include both household characteristics and our preference control variables, as discussed in Section 4.2 below. The household characteristics include the child's race, gender, age, whether the child is the first-born, household earnings, maternal cognitive skills, maternal mental health, mother's age, whether both parents are present in the household, and the number of children in the household.¹²

¹²We include maternal information only for the moment because using paternal information results in a smaller sample size. We are working on getting a larger sample with imputation methods.

4.2 Addressing Endogeneity

We consider an instrumental variable approach to deal with the endogeneity concern about parental investments. The investment function derived above gives us a natural candidate for instruments, labor market shocks, captured as wage rates in the model. A positive shock is a relevant instrument as parents are more likely to increase time at work and reduce time and effort devoted to their child, conditional on household incomes. We use the female employment rate by the local authority as a proxy for labor market shocks and consider this as an instrument for parental investments. For the instrument to be valid, the female employment rate should only affect child development through parental investments conditional on a rich set of control variables we have.

In terms of school quality, we use information on schools that parents applied to to capture and control for parents' preferences. Specifically, we consider an approach similar to the "matched-applicant" approach proposed by [Dale and Krueger \(2002, 2014\)](#) and [Mountjoy and Hickman \(2021\)](#) in the context of post-secondary enrollment. Students reveal their unobserved "types" by their application portfolio and admission portfolio. School assignment is as good as random conditional on the same application and admission portfolio. Consequently, the causal effects of attending more selective colleges are identified by comparing students applying to the same schools.

We focus on pupils who attend state-funded primary schools, which make up about 95% of all pupils of primary school age in England ([Burgess et al., 2015](#)). An important difference between our context and post-secondary education is that admission to state-funded primary schools is not merit-based. Instead, pupils with special education needs, children with siblings in the same school, and those who live closer to the school have admission priority.¹³ Accounting for the fact that the admission rule is known,¹⁴ we modify the "matched-applicant" approach by focusing on the application portfolio only.

While we have information on the application portfolio, there is not much overlapping among these portfolios. This is because households are distributed across diverse localities in our sample and primary school applications are primarily localized. Therefore, instead of comparing students with the same application portfolio, we focus on the characteristics of schools in the application portfolios and argue that these characteristics reveal households' preferences or types. We use school-level information from the school census and the Edubase. The characteristics include academic performance, the share of students eligible for free school meals, school types, school

¹³ Admission priority is also given to children who are looked after by the state, but our analysis does not include this group of children.

¹⁴ We have information on whether a child has special education needs, whether their siblings attend the same school, and their distances to schools applied.

denomination, whether siblings attend the same school and home-school distance. We control these characteristics in the production function and refer to them as the preference control.

The assumption is that conditional on these preference control variables (and the observed characteristics of households), the factors that lead to different school enrollments are unrelated to the potential outcomes of students. Given that most parents nominate one school only, we control for the characteristics of the first-choice school and these variables seem sufficient to address the selection issue. We present evidence that supports this assumption in section 4.3.

The preference control addresses the endogeneity in school quality that results from the correlation between parents' preferences for school and unobserved inputs that affect child development. However, if parents' school choices respond to shocks to child development when they make school applications at age 5 and the shocks are serially correlated, school quality can be correlated with shocks to child development at age 7 or age 11. In this case, the preference control is not sufficient to address the endogeneity issue. As a robustness check, we consider a complementary approach using the birthplace school quality as an instrument.¹⁵

The birthplace school quality is a relevant instrument because the closer a child lives to a school, the more likely that this child will be admitted to that school. This instrument provides exogenous variation because parents' residential choice when their child was just born can not respond to shocks to child development at primary school. While parents' residential choice might reflect their characteristics such as income or education, the assumption is that conditional on a very rich set of household characteristics and the child's previous skill development, birthplace school quality affects child development only through the actual school quality the child experiences. In addition to these observed characteristics, parent's residential choice around childbirth may also be correlated to their time-invariant preferences that could affect child development. We address this concern by combining the preference control approach with the birthplace school quality instrument.

With the endogeneity issues of parental investments and school quality addressed, a remaining issue is the interpretation of the polygenic score. The PGS may capture confounding factors in the family environment since parents and their children share some genetics. For example, children with a higher education PGS may also have parents with a higher PGS. Since the education polygenic score predicts higher college graduation, labor earnings, and wealth, parents with a higher PGS may be able to provide more resources and support to their children. We minimize these concerns by

¹⁵We only have birthplace school quality with the value-added measure and not the Ofsted rating for now. Therefore, we employ this strategy only for production function estimates at age 11 when we use the value-added measure.

explicitly incorporating parental investments into our framework, as well as controlling for parents' cognitive skills, mental health, and educational attainments among a rich set of background variables.¹⁶

Lastly, when investigating the interaction effects between genetic endowment and parental investments or school quality, we use a control function approach. We include the residual obtained from the first stage where we regress the endogenous variable on the instrument and other exogenous variables into the production function. Specifically, we assume that

$$E(\eta_t|X_t, W_t) = \kappa_1 v_t,$$

$$E(u_t|X_t, W_t) = \kappa_2 v_t,$$

where η_t and u_t are the shocks to the cognitive and socio-emotional skill production functions in equations 1 and 2, respectively. X_t includes the variables in the production functions including parental investments and school quality, and W_t is the instrument that is included in the investment function, i.e. the first stage, but not in the production function. We include the estimated residual from the first stage, \hat{v}_t , as a regressor in each of the production functions. The estimates of κ_1 and κ_2 provide a test of endogeneity: investments are exogenous if $\kappa_1 = 0$ and $\kappa_2 = 0$.¹⁷

4.3 Assumption Test

We test the assumption that, conditional on these preference control variables, the factors that lead to different school enrollments are unrelated to the potential outcomes of students in this section. In each of the balance graphs below, there are three panels showing how three outcomes vary by measures of school quality in quantiles. In the leftmost panel, the individual *raw* skill outcome is regressed on the indicators of school quality, as measured by the Ofsted rating in six quantiles or the value-added measure in twenty quantiles.¹⁸ The lowest quantile is omitted as the reference group. In the middle panel, the *covariate-predicted* skills are regressed on the school quality indicators. The covariate-predicted skills are from a separate OLS regression of skills on a set of covariates: cognitive and socio-emotional skills at the previous wave, parents'

¹⁶We also have the education polygenic scores of the parents but the sample size shrinks substantially with parents' polygenic scores included. Our preliminary results suggest that once including the background variables mentioned above, parents' polygenic scores have no impact on skill development. Meanwhile, we are working with imputation methods to obtain a larger sample.

¹⁷We use the female labor market shock as an instrument for parental investments at both ages 7 and age 11. This part is relevant for school quality only when we consider the birthplace school quality instrument. As explained before, we only have birthplace school quality in terms of the VA measure at age 11. At age 11, we find no impacts of parental investments and do not investigate its interaction effects with the PGS due to the limited sample size. Consequently, we only include the residual from the first stage of parental investments at age 7, and only include the residual from the first stage of school quality at age 11.

¹⁸The value-added measure displays greater variation and therefore is sorted into twenty quantiles.

educational attainments, parents' cognitive skills, parents' mental health, and household incomes. The rightmost panel regresses the *covariate-predicted* skills on the school quality indicators and the preference control variables. The preference control variables include school academic performance, home-school distance, whether siblings attend the same school, the share of students eligible for free school meals, school types, and school denominations.

For both cognitive and socio-emotional skills at age 7, we see a positive gradient in the raw outcome in Figure 1 and Figure 2, respectively. The positive gradient can be a combination of treatment effects of going to schools with better quality and a selection effect. Similarly, the covariate-predicted outcomes also display a positive gradient by school quality. This confirms the selection effects: children who are predicted to have better skills sort into good schools. However, once we include the preference control in the third panel, the covariate-predicted outcome no longer has a positive gradient. This provides evidence that including the preference control variables mitigates the selection bias issue.

For cognitive skills at age 11, the positive gradient is less pronounced and we only see a significant positive sorting for children in the highest quantile of the Ofsted rating in terms of both raw and predicted outcome in Figure 3. The sorting pattern is more evident with the value-added measure, as shown in Figure 5. The good news is that controlling for preferences addresses the sorting issue. For socio-emotional skills at wave 5, there is no evidence of sorting in the Ofsted school measure, and the positive gradient in the value-added measure is also not obvious. In both cases, controlling for school preferences alleviates the selection concern.

5 Results

In this section, we first present the production function estimates for cognitive skills and socio-emotional skills at age 7. Then we report the estimates for the production function at age 11 using two different measures of school quality: the Ofsted rating and the value-added measure.

5.1 Production Function Estimates at Age 7

Table 3 presents the production function estimates for cognitive skills and socio-emotional skills at age 7 (in wave 4). The OLS estimates suggest that school quality, measured by Ofsted rating, is positively correlated to cognitive skills. However, parental investments, specifically, parents' educational activities are negatively related to cognitive skills. Cognitive skills at age 7 are also positively correlated to the child's PGS, cognitive skills, and socio-emotional skills at age 5 (in wave 3).

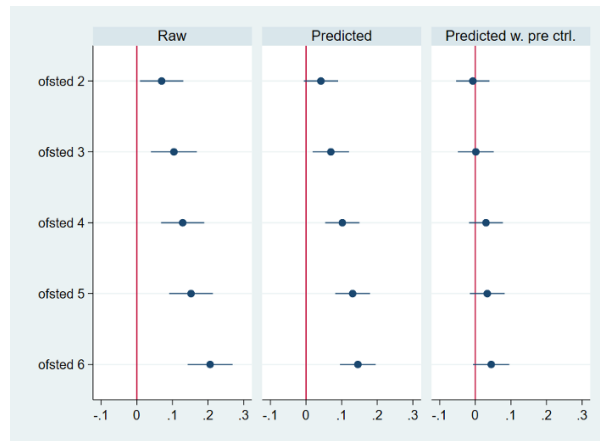


Figure 1: Cognitive skills at age 7

Notes: Each set of point estimates and the 95% confidence intervals come from regressions of individual cognitive skills on the Ofsted rating indicators (in six rankings), omitting the lowest rank as the reference group (signified by the vertical line at zero). The leftmost specification regresses the raw skills. The middle specification regresses the predicted skills. The predicted values are from a separate OLS regression of skills on the following set of covariates: cognitive and socio-emotional skills at age 5, parents' educational attainments, cognitive skills, mental health, and household incomes. The rightmost specification regresses the predicted skills on the Ofsted rating indicators and the preference control variables. The preference control variables include school academic performance, home-school distance, whether siblings attend the same school, the share of students eligible for free school meals, school types, and school denominations.

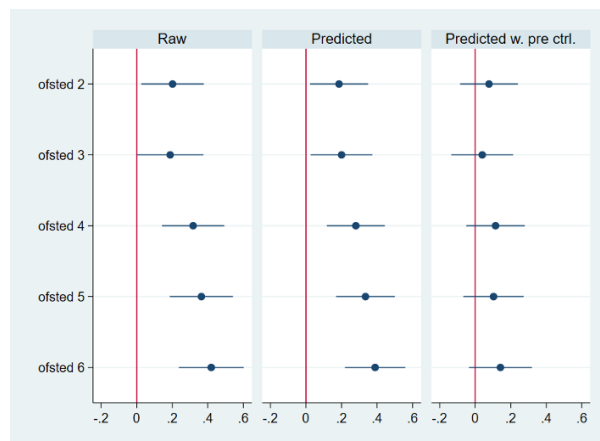


Figure 2: Socio-emotional skills at age 7

Notes: Each set of point estimates and the 95% confidence intervals come from regressions of individual socio-emotional skills on the Ofsted rating indicators (in six rankings), omitting the lowest rank as the reference group (signified by the vertical line at zero). The leftmost specification regresses the raw skills. The middle specification regresses the predicted skills. The predicted values are from a separate OLS regression of skills on the following set of covariates: cognitive and socio-emotional skills at age 5, parents' educational attainments, cognitive skills, mental health, and household incomes. The rightmost specification regresses the predicted skills on the Ofsted rating indicators and the preference control variables. The preference control variables include school academic performance, home-school distance, whether siblings attend the same school, the share of students eligible for free school meals, school types, and school denominations.

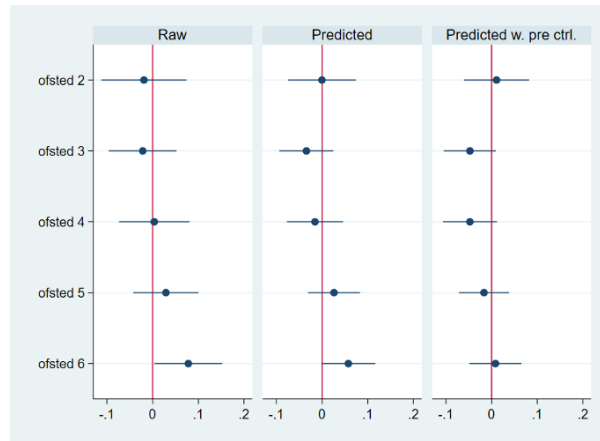


Figure 3: Cognitive skills at age 11 (Ofsted rating)

Notes: Each set of point estimates and the 95% confidence intervals come from regressions of individual cognitive skills on the Ofsted rating indicators (in six rankings), omitting the lowest rank as the reference group (signified by the vertical line at zero). The leftmost specification regresses the raw skills. The middle specification regresses the predicted skills. The predicted values are from a separate OLS regression of skills on the following set of covariates: cognitive and socio-emotional skills at age 7, parents' educational attainments, cognitive skills, mental health, and household incomes. The rightmost specification regresses the predicted skills on the Ofsted rating indicators and the preference control variables. The preference control variables include school academic performance, home-school distance, whether siblings attend the same school, the share of students eligible for free school meals, school types, and school denominations.

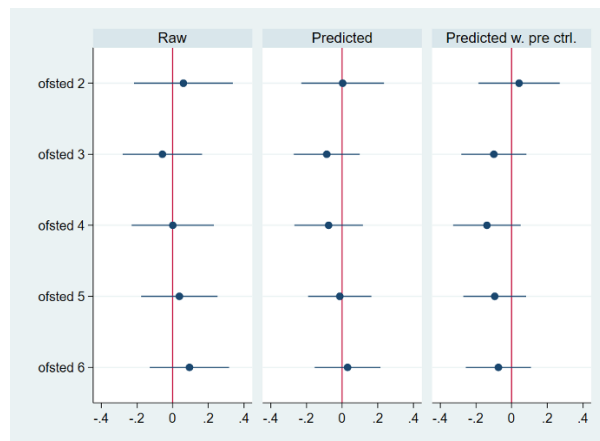


Figure 4: Socio-emotional skills at age 11 (Ofsted rating)

Notes: Each set of point estimates and the 95% confidence intervals come from regressions of individual cognitive skills on the Ofsted rating indicators (in six rankings), omitting the lowest rank as the reference group (signified by the vertical line at zero). The leftmost specification regresses the raw skills. The middle specification regresses the predicted skills. The predicted values are from a separate OLS regression of skills on the following set of covariates: cognitive and socio-emotional skills at age 7, parents' educational attainments, cognitive skills, mental health, and household incomes. The rightmost specification regresses the predicted skills on the Ofsted rating indicators and the preference control variables. The preference control variables include school academic performance, home-school distance, whether siblings attend the same school, the share of students eligible for free school meals, school types, and school denominations.

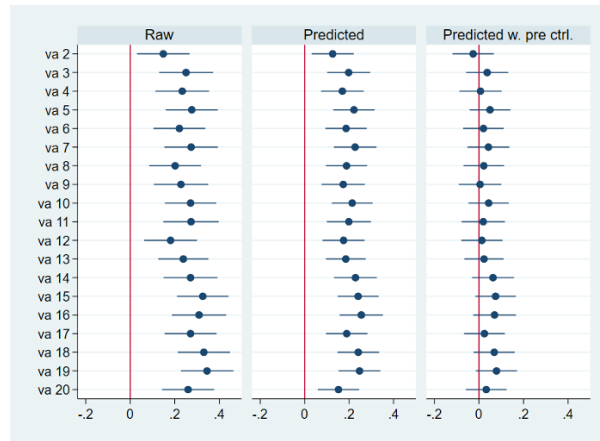


Figure 5: Cognitive skills at age 11 (value-added)

Notes: Each set of point estimates and the 95% confidence intervals come from regressions of individual cognitive skills on the value-added measure indicators (in twenty rankings), omitting the lowest rank as the reference group (signified by the vertical line at zero). The leftmost specification regresses the raw skills. The middle specification regresses the predicted skills. The predicted values are from a separate OLS regression of skills on the following set of covariates: cognitive and socio-emotional skills at age 7, parents' educational attainments, cognitive skills, mental health, and household incomes. The rightmost specification regresses the predicted skills on the value-added measure indicators and the preference control variables. The preference control variables include school academic performance, home-school distance, whether siblings attend the same school, the share of students eligible for free school meals, school types, and school denominations.

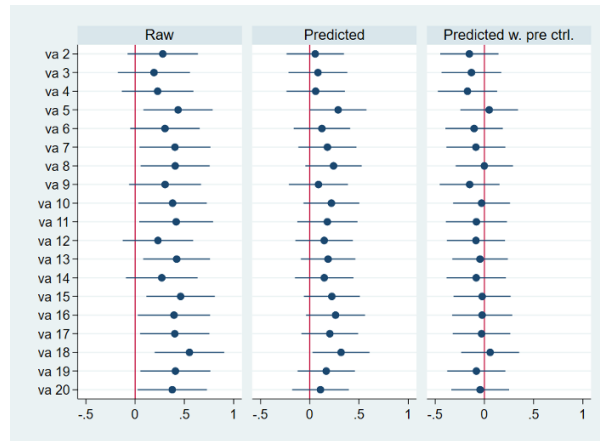


Figure 6: Socio-emotional skills at age 11 (value-added)

Notes: Each set of point estimates and the 95% confidence intervals come from regressions of individual socio-emotional skills on the value-added measure indicators (in twenty rankings), omitting the lowest rank as the reference group (signified by the vertical line at zero). The leftmost specification regresses the raw skills. The middle specification regresses the predicted skills. The predicted values are from a separate OLS regression of skills on the following set of covariates: cognitive and socio-emotional skills at age 7, parents' educational attainments, cognitive skills, mental health, and household incomes. The rightmost specification regresses the predicted skills on the value-added measure indicators and the preference control variables. The preference control variables include school academic performance, home-school distance, whether siblings attend the same school, the share of students eligible for free school meals, school types, and school denominations.

To address the potential endogeneity issue in school quality and parental investments, we use an instrumental variable approach (IV) and include a set of preference control variables (PC). With the endogeneity taken into account, the impacts of educational activities on cognitive skills change turn positive. The pattern that the IV estimates of parental investments are larger than those of the OLS estimates is consistently found in other studies (Cunha et al., 2010; Attanasio et al., 2020b,c). Parents seem to compensate their children for negative shocks in the development process, by increasing time spent with their child. The estimates on the PGS, cognitive skills, and socio-emotional skills at age 5 are still significantly positive and remain similar to their OLS estimates.

We investigate the interaction effects between genetic endowment and school quality as well as parental investments using a control function approach (CF), combined with the preference control (PC). The estimates on Ofsted rating, educational activities, PGS, and skills at age 5 basically remain the same. An increase of 10% in the Ofsted rating results in a 0.2% increase in cognitive skills, whereas a 10% increase in educational activities results in a 1.34% increase. The interaction term between the Ofsted rating and the PGS is negative, suggesting that school quality and genetic endowment are substitutes in the production function. This finding indicates that better school quality can mitigate the skill disparity related to the genetic predisposition of educational attainment, which has significant policy implications. In contrast, the interaction term between educational activities and the PGS is zero and insignificant.

Turning to the estimates for socio-emotional skills, the OLS estimates suggest that the Ofsted rating does not have a significant correlation with socio-emotional skills, while educational activities are negatively correlated with it. The impacts of the PGS and skill development at age 5 are positively correlated to socio-emotional skills at age 7. When we consider the Ofsted rating and educational activities as endogenous variables and use the IV and preference control approach, we find that neither school quality nor educational activities have an impact on socio-emotional skills. On the other hand, PGS and previous skill development, especially socio-emotional skills at age 5 have significantly positive impacts on socio-emotional skills at age 7. There is no evidence of interaction effects on socio-emotional skills. It seems that at this stage, socio-emotional skills are less sensitive to investments either at home or at school, but are largely affected by skill development at the previous stage. Every 10% increase in socio-emotional skills at age 5 predicts about a 9% increase in socio-emotional skills at age 7.

Table 4 presents the estimates of the first stage where we use the female employment rate as an instrument for educational activities. Consistent with our previous hypothesis, educational activities respond negatively to the female employment rate because of a higher opportunity cost of investments at home. Additionally, they re-

Table 3: Production function estimates at age 7 (wave 4)

	Cognitive, w4			Socio-emo., w4		
	OLS	IV + PC	CF + PC	OLS	IV + PC	CF + PC
Ofsted rating, w4	0.012*** (0.003)	0.021*** (0.007)	0.021*** (0.005)	0.007 (0.006)	0.007 (0.011)	0.007 (0.010)
Educational, w4	-0.026*** (0.004)	0.135+ (0.076)	0.134** (0.060)	-0.014** (0.006)	0.115 (0.119)	0.111 (0.110)
PGS	0.031*** (0.004)	0.032*** (0.006)	0.032*** (0.005)	0.017*** (0.006)	0.025** (0.010)	0.024*** (0.009)
Ofsted X PGS			-0.009** (0.005)			-0.014 (0.009)
Edu. X PGS			0.000 (0.004)			0.001 (0.008)
Cognitive, w3	0.485*** (0.008)	0.554*** (0.027)	0.554*** (0.022)	0.040*** (0.015)	0.082+ (0.042)	0.082** (0.040)
Socio-emo., w3	0.031*** (0.004)	0.015** (0.007)	0.015*** (0.005)	0.938*** (0.007)	0.927*** (0.011)	0.927*** (0.010)
Residual			-0.161*** (0.060)			-0.121 (0.110)
Observations	3,772	2,465	2,465	3,774	2,455	2,455

Notes: All models include the same set of control variables: the child's race, the child's gender, the child's age, whether the child was first-born, household earnings, maternal skills, maternal mental health, maternal age, whether both parents are present in the household, the number of children at the household, and the first ten principal components of the genetic data. 'w4' refers to wave 4 (age 7) and 'w3' refers to wave 3 (age 5). 'OLS' refers to estimates from the Ordinary Least Square. 'IV + PC' refers to the instrumental variable approach combined with preference controls. 'CF + PC' refers to the control function approach combined with preference control variables. Preference control variables include the following variables: academic performance, home-school distance, whether siblings attend the same school, the share of students eligible for free school meals, school types, and school denominations. 'Residual' is obtained from the first stage of educational activities. Standard errors are shown in parentheses. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4: Estimates of the first stage of educational activities at age 7

	Educational activities, w4
Female employment, 2008	-0.012*** (0.003)
Ofsted rating, w4	-0.053** (0.022)
PGS	-0.040+ (0.021)
Cognitive, w3	-0.322*** (0.048)
Socio-emo, w3	0.048** (0.022)
F statistics	12.130
Observations	2,465

Notes: The first stage includes the following control variables: the child's race, the child's gender, the child's age, whether the child was first-born, household earnings, maternal skills, maternal mental health, maternal age, whether both parents are present in the household, the number of children at the household, and the first ten principal components of the genetic data, as well as the preference controls. Preference control variables include school academic performance, home-school distance, whether siblings attend the same school, the share of students eligible for free school meals, school types, and school denominations. 'w4' refers to wave 4 (age 7) and 'w3' refers to wave 3 (age 5). Standard errors are shown in parentheses. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

spond negatively to school quality, PGS, and previous cognitive skills, while showing a positive response to previous socio-emotional skills. The F-statistics indicate that female employment is a relevant instrument.

5.2 Production Function Estimates at Age 11

This section presents the production function estimates for cognitive skills and socio-emotional skills at age 11 (in wave 5). At age 11, we have two school quality measures, the Ofsted rating, and the valued-added measure. We first present results using the Ofsted rating in Table 5 and then the value-added measure in Table 6.

The first two columns in Table 5 report the OLS estimates for cognitive skills, with and without the PGS. Despite a significant reduction in sample size when including the PGS, the estimates in these two columns remain very similar. We observe a positive correlation between Ofsted rating, PGS, skill development at age 7, and cognitive skills at age 11, as well as a negative correlation between educational activities and cognitive skills at age 11. This pattern is similar to what we observe in Table 3. Similarly, the endogenous response of educational activities and school quality may be a concern, and we use an IV approach combined with the preference control (IV + PC). As the

sample substantially shrinks with the PGS, the instrument, female employment rate, displays less variation. For the moment, we report results on two of the three key inputs: Ofsted rating and educational activities, or Ofsted rating and PGS.¹⁹

The estimates in the 'IV + PC' column show the impacts of Ofsted rating and educational activities without the PGS. Ofsted rating has a positive impact on cognitive skills. Educational activities no longer show a negative sign and have no impact on cognitive skills at age 11. In the column 'PC', we provide estimates for Ofsted rating and PGS, as well as their interaction using the preference control approach. While both Ofsted rating and PGS have a positive impact, there is no interaction effect at age 11.

In terms of socio-emotional skills, the OLS estimates suggest a positive correlation between the Ofsted rating and socio-emotional skills. However, this correlation is not significant once we control for the PGS. Using the IV approach combined with the preference control, neither Ofsted rating nor educational activities have an impact on socio-emotional skills at age 11. In the last column with the preference control approach, we find no impacts of the Ofsted rating, the PGS, and their interaction.

Compared to the estimates at age 7, there are notable differences at age 11. Educational activities no longer have a significant impact on cognitive skills, while the influences of the Ofsted rating and the PGS have reduced considerably. A 10% increase in the Ofsted rating at age 11 only results in a 0.04% - 0.06% increase in cognitive skills, much smaller than the 0.21% observed at age 7. The impacts of a one standard deviation increase in the PGS decrease from 3.2% at age 7 to 0.4% at age 11. Additionally, we do not find any interaction effects at age 11. Socio-emotional skills at age 11 are primarily influenced by previous skill development. Although the persistent level of lagged skills is already high for socio-emotional skills at age 7, there is a larger increase in cognitive skills from 0.55 at age 7 to 0.9 at age 11.

In addition to the Ofsted rating measure, we have another school quality measure: value-added from key stage 2 to key stage 1. The estimates of the production function at age 11 with the value-added measure are presented in Table 6. For cognitive skills, the estimates on educational activities, PGS, interaction effects, and lagged skills in Table 6 are very similar to results with the Ofsted measure in Table 5. Like the Ofsted rating, the value-added measure also has a positive impact on cognitive skills. Both the value-added measure and PGS have a positive impact on cognitive skills, but there are no interaction effects between these two inputs. Educational activities also have no impact.

For socio-emotional skills, the value-added measure has a positive effect, in contrast to the null effect of the Ofsted rating. While the Ofsted rating captures aspects

¹⁹We are working on obtaining a larger sample using imputation methods to provide credible IV estimates with all inputs included in the production function.

such as school management, leadership, and financing, the results suggest that these factors might not be determinants of socio-emotional skills. Other than the school quality measure, we observe no impact of educational activities or PGS, nor any interaction effects.

The first stage estimates of educational activities are presented in Table 7. Similar to the first stage at age 7, the female employment rate has a negative impact on educational activities. Educational activities respond negatively to previous cognitive skills but positively to previous socio-emotional skills. The F-statistics support the instrument's relevance.

Our previous analysis relies on the preference control approach to address the potential endogeneity of school quality. An alternative or complementary approach is to use a control function approach with the birthplace school quality as an instrument for the actual school quality children experience. We show the estimates in Table 8. The 'CF' column presents the estimates using the control function alone, while the 'CF + PC' column reports estimates using both the control function and the preference control. The residual is obtained from the first stage where we regress the value-added measure on the birthplace school quality instrument.

The estimates in Table 8 demonstrate that whether using the control function approach alone or in combination with the preference control, the results do not change significantly. One exception is the residual. Without the preference control, the residual is significantly negative, indicating endogeneity of school quality is a concern. However, if we include the preference control, the residual becomes insignificant, giving us confidence in our results with the preference control approach.

It is also reassuring to observe that the estimates in Table 8 are similar to estimates in the 'PC' column of Table 6. We find positive impacts of value-added and PGS on cognitive skills at age 11, but no interaction effects. Value-added also matters for socio-emotional skills at age 11. Previous skill development in both dimensions has a persistent effect on current skill development.

Table 5: Production function estimates at age 11 (Ofsted rating)

	Cognitive, w5				Socio-emo., w5			
	OLS	OLS	IV + PC	PC	OLS	OLS	IV + PC	PC
Ofsted rating, w5	0.005*** (0.001)	0.005*** (0.001)	0.006*** (0.001)	0.004*** (0.001)	0.009+ (0.005)	0.008 (0.006)	0.007 (0.007)	0.004 (0.009)
Educational, w5	-0.002*** (0.001)	-0.003*** (0.001)	0.022 (0.021)		0.004 (0.005)	-0.002 (0.007)	-0.057 (0.129)	
PGS		0.005*** (0.001)		0.004*** (0.001)		0.013+ (0.007)		0.009 (0.009)
Ofsted X PGS				0.000 (0.001)				0.007 (0.008)
Cognitive, w4	0.892*** (0.003)	0.894*** (0.004)	0.914*** (0.019)	0.896*** (0.005)	0.156*** (0.018)	0.149*** (0.023)	0.103 (0.123)	0.161*** (0.028)
Socio-emo, w4	0.017*** (0.001)	0.015*** (0.001)	0.015*** (0.002)	0.015*** (0.001)	0.852*** (0.006)	0.847*** (0.007)	0.850*** (0.011)	0.839*** (0.009)
Observations	6,016	3,501	3,847	2,257	5,869	3,435	3,743	2,205

Notes: All models include the same set of control variables: the child's race, the child's gender, the child's age, whether the child was first-born, household earnings, maternal skills, maternal mental health, maternal age, whether both parents are present in the household, the number of children at the household, and the first ten principal components of the genetic data. 'w5' refers to wave 5 (age 11), and 'w4' refers to wave 4 (age 7). 'OLS' refers to estimates from the Ordinary Least Square. 'IV + PC' refers to the instrumental variable approach combined with preference control variables. 'PC' refers to the preference control. Preference control variables include the following variables: school academic performance, home-school distance, whether siblings attend the same school, the share of students eligible for free school meals, school types, and school denominations. Standard errors are shown in parentheses. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 6: Production function estimates at age 11 (value-added measure)

	Cognitive, w5				Socio-emo., w5			
	OLS	OLS	IV + PC	PC	OLS	OLS	IV + PC	PC
Value-added, w5	0.027*** (0.001)	0.027*** (0.001)	0.028*** (0.002)	0.027*** (0.002)	0.043*** (0.006)	0.039*** (0.008)	0.050*** (0.011)	0.033*** (0.011)
Educational, w5	-0.003*** (0.001)	-0.003*** (0.001)	0.013 (0.023)		0.003 (0.005)	-0.003 (0.007)	-0.106 (0.164)	
PGS		0.005*** (0.001)		0.004*** (0.001)		0.008 (0.007)		0.006 (0.009)
VA X PGS				-0.001 (0.002)				-0.004 (0.010)
Cognitive, w4	0.891*** (0.003)	0.893*** (0.004)	0.903*** (0.021)	0.895*** (0.004)	0.138*** (0.017)	0.127*** (0.022)	0.048 (0.153)	0.151*** (0.028)
Socio-emo, w4	0.016*** (0.001)	0.016*** (0.001)	0.015*** (0.002)	0.016*** (0.001)	0.853*** (0.006)	0.852*** (0.007)	0.852*** (0.013)	0.839*** (0.009)
Observations	6,317	3,726	4,036	2,397	6,160	3,650	3,925	2,338

Notes: All models include the same set of control variables: the child's race, the child's gender, the child's age, whether the child was first-born, household earnings, maternal skills, maternal mental health, maternal age, whether both parents are present in the household, the number of children at the household, and the first ten principal components of the genetic data. 'w5' refers to wave 5 (age 11), and 'w4' refers to wave 4 (age 7). 'OLS' refers to estimates from the Ordinary Least Square. 'IV + PC' refers to the instrumental variable approach combined with preference control variables. 'PC' refers to the preference control. Preference control variables include the following variables: school academic performance, home-school distance, whether siblings attend the same school, the share of students eligible for free school meals, school types, and school denominations. Standard errors are shown in parentheses. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 7: Estimates of the first stage of educational activities at age 11

	Educational activities, w5	
Female employment, 2012	-0.008*** (0.002)	-0.007*** (0.002)
Ofsted rating, w5	-0.016 (0.016)	
Value-added, w5		0.038+ (0.020)
Cognitive, w4	-0.916*** (0.052)	-0.903*** (0.050)
Socio-emo, w4	0.066*** (0.018)	0.064*** (0.017)
F statistics	11.55	8.138
Observations	3,847	4,036

Notes: All models include the same set of control variables: the child’s race, the child’s gender, the child’s age, whether the child was first-born, household earnings, maternal skills, maternal mental health, maternal age, whether both parents are present in the household, the number of children at the household, and preference control variables. Preference control variables include the following variables: school academic performance, home-school distance, whether siblings attend the same school, the share of students eligible for free school meals, school types, and school denominations. ‘w5’ refers to wave 5 (age 11), and ‘w4’ refers to wave 4 (age 7). Standard errors are shown in parentheses. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

6 Conclusions

In this paper, we explore how parental investments, school quality, genetics, and their interactions influence child development by estimating the skill production functions for cognitive skills and socio-emotional skills. We implement an instrumental variable approach and exploit information from the school application portfolio to address the potential endogeneity of parental investments and school quality. An education polygenic score is used to capture an individual’s genetic propensity for educational attainment.

Using data from the Millennium Cohort Study, we find different results for cognitive skills and socio-emotional skills. First, cognitive skills at age 7 are significantly influenced by parental investments, school quality, genetics, and skills at age 5. Notably, school quality and the polygenic score act as substitutes, indicating that better schools can mitigate skill disparities related to genetic predisposition for educational attainment. The estimates for cognitive skills at age 11 exhibit significant differences compared to those at age 7. The impact of parental investments on cognitive skills is no longer significant, while the influences of school quality and the polygenic score

Table 8: Production function estimates at age 11 (value-added measure)

	Cognitive, w5		Socio-emo., w5	
	CF	CF + PC	CF	CF + PC
Value-added	0.031*** (0.002)	0.029*** (0.004)	0.058*** (0.016)	0.051** (0.024)
PGS	0.004*** (0.001)	0.004*** (0.001)	0.007 (0.007)	0.005 (0.009)
VA X PGS	-0.001 (0.001)	-0.001 (0.002)	-0.009 (0.008)	-0.004 (0.010)
Cognitive, w4	0.895*** (0.003)	0.894*** (0.004)	0.129*** (0.022)	0.150*** (0.028)
Socio-emo, w4	0.015*** (0.001)	0.016*** (0.001)	0.851*** (0.007)	0.839*** (0.009)
Residual	-0.006** (0.003)	-0.003 (0.004)	-0.028 (0.018)	-0.022 (0.027)
Observations	3,741	2,397	3,662	2,338

Notes: All models include the same set of control variables: the child's race, the child's gender, the child's age, whether the child was first-born, household earnings, maternal skills, maternal mental health, maternal age, whether both parents are present in the household, the number of children at the household, and the first ten principal components of the genetic data. 'w5' refers to wave 5 (age 11), and 'w4' refers to wave 4 (age 7). 'CF' refers to estimates with the control function approach. 'CF + PC' refers to the control function approach combined with preference control variables. Preference control variables include school academic performance, home-school distance, whether siblings attend the same school, the share of students eligible for free school meals, school types, and school denominations. Standard errors are shown in parentheses. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

have substantially diminished. Cognitive skills also display fairly strong persistence at this age. Second, the high persistence of socio-emotional skills is already evident at age 7. The only investment that matters for socio-emotional skills at age 11 is school quality, as measured by the value-added measure.

Understanding the skill development process and how various inputs of this process interact is crucial for policy interventions aimed at improving intergenerational mobility. Our paper underscores the critical role of schools in bridging the skill gap associated with genetic factors. Furthermore, school quality effectively enhances cognitive and socio-emotional skills in middle childhood, with different dimensions of school quality impacting cognitive and socio-emotional skills in diverse ways. These findings carry significant implications for policy design, suggesting a need for targeted efforts to elevate school standards across a broad spectrum of dimensions and to ensure high-quality education access for a wider population.

References

- Agostinelli, F. and Wiswall, M. (2016). Estimating the technology of children’s skill formation. Technical report, National Bureau of Economic Research.
- Attanasio, O., Bernal, R., Giannola, M., and Nores, M. (2020a). Child development in the early years: Parental investment and the changing dynamics of different dimensions. Technical report, National Bureau of Economic Research.
- Attanasio, O., Cattan, S., Fitzsimons, E., Meghir, C., and Rubio-Codina, M. (2020b). Estimating the production function for human capital: results from a randomized controlled trial in colombia. *American Economic Review*, 110(1):48–85.
- Attanasio, O., Meghir, C., and Nix, E. (2020c). Human capital development and parental investment in india. *The Review of Economic Studies*, 87(6):2511–2541.
- Barcellos, S. H., Carvalho, L., and Turley, P. (2021). The effect of education on the relationship between genetics, early-life disadvantages, and later-life ses. Technical report, National Bureau of Economic Research.
- Barth, D., Papageorge, N. W., and Thom, K. (2020). Genetic endowments and wealth inequality. *Journal of Political Economy*, 128(4):1474–1522.
- Belsky, D. W., Domingue, B. W., Wedow, R., Arseneault, L., Boardman, J. D., Caspi, A., Conley, D., Fletcher, J. M., Freese, J., Herd, P., et al. (2018). Genetic analysis of social-class mobility in five longitudinal studies. *Proceedings of the National Academy of Sciences*, 115(31):E7275–E7284.
- Biroli, P. and Zünd, C. L. (2020). Genes, pubs, and drinks: Gene-environment interplay and alcohol licensing policy in the uk. Technical report, Mimeo.
- Bolyard, A. and Savelyev, P. (2019). Understanding the education polygenic score and its interactions with ses in determining health in young adulthood. *Available at SSRN*.
- Burgess, S., Greaves, E., Vignoles, A., and Wilson, D. (2015). What parents want: School preferences and school choice. *The Economic Journal*, 125(587):1262–1289.
- Cunha, F. and Heckman, J. J. (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of human resources*, 43(4):738–782.
- Cunha, F., Heckman, J. J., and Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931.

- Dale, S. B. and Krueger, A. B. (2002). Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *The Quarterly Journal of Economics*, 117(4):1491–1527.
- Dale, S. B. and Krueger, A. B. (2014). Estimating the effects of college characteristics over the career using administrative earnings data. *Journal of human resources*, 49(2):323–358.
- Houmark, M. A., Ronda, V., and Rosholm, M. (2020). The nurture of nature and the nature of nurture: How genes and investments interact in the formation of skills.
- Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics*, 50(8):1112–1121.
- Mountjoy, J. and Hickman, B. R. (2021). The returns to college (s): Relative value-added and match effects in higher education. Technical report, National Bureau of Economic Research.
- Okbay, A., Benjamin, D., and Visscher, P. (2018). Ssgac educational attainment: Gwas and mtag polygenic scores (ver. 1.0).
- Papageorge, N. W. and Thom, K. (2020). Genes, education, and labor market outcomes: evidence from the health and retirement study. *Journal of the European Economic Association*, 18(3):1351–1399.