School, Parents & Genes: Essays on the Determinants of Children's Success in Life

Marc Christian Stöckli

Munich 2022

School, Parents & Genes: Essays on the Determinants of Children's Success in Life

Inaugural-Dissertation

zur Erlangung des Grades

Doctor oeconomiae publicae (Dr. oec. publ.)

an der Ludwig-Maximilians-Universität München

2022

vorgelegt von

Marc Christian Stöckli

Referent: Prof. Dr. Andreas Peichl

Korreferent: Prof. Dr. Ludger Wößmann

Promotionsabschlussberatung: 20. Juli 2022

Datum der mündlichen Prüfung: 11. Juli 2022

Berichterstatter: Prof. Dr. Andreas Peichl, Prof. Dr. Ludger Wößmann,
Prof. Helmut Rainer, Ph.D.

Acknowledgements

I am indebted to many people without whom this thesis would not have been written. First and foremost, I want to express my deepest gratitude to my supervisor, Andreas Peichl. He has given me the opportunity to pursue a research interest that admittedly has little to do with economic forecasting—for which I was initially hired at the ifo Institute. Not only has he wholeheartedly supported me in my research, he has also opened up numerous opportunities for me to grow as a researcher.

I also want to thank my second and third supervisor, Ludger Wößmann and Helmut Rainer, as well as my collaborators Paul Hufe, Benjamin Arold, Eric Hanushek, Marc Piopiunik, Katrin Auspurg, Laila Schmitt, Alain Trannoy, Edwin Fourrier-Nicolai, Florian Dorn, Sahamoddin Khailaie, Sebastian Binder, Stefan Lautenbacher, Patrizio Vanella, Timo Wollmershäuser, Clemens Fuest, Michael Meyer-Hermann, Christian Grimme, Maurizio Bussolo, Mattia Makovec, Carla Krolage, and Christian Wittneben. A special thanks to Paul Hufe, who readily supported me with excellent advice and support.

My thesis has benefited from several research visits. I am indebted to Eric Hanushek, Ravi Kanbur, Richard Blundell, Maurizio Bussolo, and Mattia Makovec, who have kindly hosted me at Stanford, Cornell, the IFS, and the World Bank. I have learned a tremendous amount from you. The thesis has also benefited from financial support from the DFG and the Fritz Thyssen Foundation.

I thank Elisabeth Grewenig, Stefanie Gäbler, Luisa Dörr, Lea Immel, Stefan Lautenbacher, Marcell Göttert, Carla Krolage, Paul Schüle, and Maximilian Blömer for enriching my time at ifo. I fondly remember our visits to the Oktoberfest, Starkbierfest, Sommerfest, and all the Biergardens. I also thank Fritz Espenlaub and Maria Muster for their friendship and the memes.

Last but certainly not least, I am thankful to my partner, Tamina Graber, for her love and support. It was her that kept me sane while writing this thesis. I am now looking forward to moving back to Switzerland and spending more time with her.

Marc Stöckli
Munich, March 2022

Contents

List of Figures

List of Tables

Preface

How does inequality emerge, and why is it (or was it) increasing? These were the two questions I asked myself and tried to tackle in this dissertation. Conceptually, I envision three stages during which inequality can materialize and which require different policy instruments: after, during, and before the labor market (very roughly speaking). *After* the labor market, there is the tax and benefits system. At this stage, policy-makers can address inequality by redistributing resources from those who have to those who haven't. *During* the labor market, policy-makers can intervene in how much people receive for their work, e.g. by introducing a minimum wage. And *before* the labor market, human capital is accumulated, which later becomes relevant for generating income. Investments in education can ensure that skills—and, by extension, incomes—are distributed more equally.

I was and still am fascinated by this *before* stage. From a policy perspective, it is a welcome opportunity to address resource disparities: no one has to be taken something away, and there is broad political support for investing in the next generation. From an economic perspective, this stage is also highly efficient. It's cheaper to make sure that someone doesn't become unemployed than to help an unemployed find a job. Investments in early childhood education can even be self-financing, as e.g. Hendren and Sprung-Keyser (2020) show. Not only is it less likely that someone becomes unemployed, this person also generates revenues for the state via income taxes.

This thesis is titled "School, Parents & Genes". Each chapter discusses one of these (co-)determinants of children's success in life: chapter 1 is on school finance reforms in the U.S. and investigates whether they were able to level the playing field for students (they were not). Chapter 2 is on parenting styles and asks whether child-rearing practices are relevant for the formation of cognitive and non-cognitive skills (they are). Chapter 3 asks whether school investments interact with genetic endowments (they do).

Chapter 1: As a European, I was always perplexed by the difficult standing school financing has in the United States. After all, the U.S. played a pivotal role in advancing mass elementary schooling (Goldin and Katz, 2010). To me, it felt like a no-brainer that public schools ought to be well-funded, and not only because doing so is crucial for distributive justice. But apparently, not everyone sees it this way. Still, I assume that a certain group of legal scholars in the 1960s would have agreed with me. At their time, a district's school finances were tightly linked to the local property tax revenues, such that affluent districts were able to spend more per student. Encouraged by writings such as Coons et al. (1970), Horowitz (1965), or Wise (1968), legal suits challenged these regimes, often successfully so. Courts agreed that the finance regimes created unequal educational opportunities for students and, as such, violated state constitutions. The resulting court-ordered school finance reforms (SFRs) dramatically changed the state education funding formulas.

Since the 1970s, numerous states have introduced such reforms, and in broadly in two waves. During the "equity" era—the 1970s and 1980s—the SFRs aimed at equalizing resources across districts. Their goal was to break the link between property tax revenues and school spending capacity. This worked, albeit sometimes hurting the intended beneficiaries along the way (Hoxby, 2001). The 1976 reform in California is a particularly spectacular case of levelling-down, where spending equality also resulted in fewer resources for low-income districts. Beginning in the late 1980s, reforms during the "adequacy" era aimed at equalizing opportunities instead of resources. To compensate for the disadvantages students in low-income districts face, these districts should receive relatively *more* resources.

Naturally, economists disagree about the reforms' benefits (or harms). On the one hand, Card and Payne (2002), Lafortune et al. (2018), or Rothstein and Schanzenbach (2021) attest to their positive impacts on educational achievement and labor market outcomes. On the other hand, researchers like Caroline Hoxby or Eric Hanushek—the latter a co-author of chapter 1, nota bene—are more sceptical. Personally, I have sympathy for both sides. Reforms mandated by a court presumably pay little attention to the actual usefulness of policies other than restoring constitutional rights (which is, of course, important in and of itself). It is probably not desirable to distribute funds using the "watering can" principle: neither is it guaranteed that school funding is the bottleneck in the first place, nor are there incentives to spend the money efficiently (except if coupled with accountability elements, as Buerger et al., 2021, highlights). Yet, this does not mean that court-ordered SFRs cannot do good. They may not be the most efficient policy, but they may still improve the outcomes of disadvantaged students.

The starting point of chapter 1 is a puzzle in Lafortune et al. (2018): they find that the SFRs during the adequacy era reduced the achievement gap between low- and high-income *districts*, but not between low- and high-income *students*. Intuitively, this is weird. One would expect that the income of a district is a proxy for the income of families living there. Closing the achievement gap by region would then imply a closing of the achievement gap by family income. Yet, as it turns out, the region is an inadequate proxy for the individual in this case: because of within-district inequality, low-income families don't necessarily live in low-income districts. Targeting a policy based on the average income level of a region is then inappropriate as it misses the people who should actually benefit from it. At least this is the explanation of Lafortune et al. (2018) for their puzzling finding.

Alternatively, income gaps may just miss the positive effects of the SFRs. It is possible that, although there is no effect on the income gradient, there are effects on broader measures. This is what Eric Hanushek, Paul Hufe, Marc Piopiunik, and I investigate in chapter 1. We construct a measure of inequality of educational opportunity (IEOp) based on the theoretical concept of Roemer (1998). Intuitively, we distinguish between circumstances and effort as the two determinants of educational achievement. The former are characteristics outside individual control, such as race, assigned gender at birth, or parental income. The latter are the factors that students have control over. Achievement differences due to differences in circumstances are unfair, whereas achievement difference due to differential effort are not (roughly speaking). As such, our measure of IEOp is a generalization of the achievement gap that Lafortune et al. (2018) use. The question that chapter 1 tries to answer is straightforward: did the SFRs of the adequacy era reduce IEOp? Or put differently, we investigate whether the SFRs were able to reduce the importance of circumstances for educational achievement.

Descriptively, we find that IEOp has increased between 1990 and 2015 in most states. This is worrying but doesn't mean that the SFRs were ineffective. It is, of course, possible that IEOp would have increased even more in the absence of the reforms. To assess the causal effect of the SFRs on IEOp, we make use of the fact that court-ordered school finance reforms can be seen as a natural experiment because the exact timing cannot be anticipated by legislatures. This is because of the idiosyncrasies of judicial processes, which make it impossible (or at least very hard) to predict how long a legal suit will take. Under this assumption, the causal effect of the SFRs on IEOp can be estimated in an event-study framework. In particular, we estimate standard two-way fixed effects models and recently proposed alternatives (including Callaway and Sant'Anna, 2020, Sun and Abraham, 2020, and Roth and Sant'Anna, 2021).

Irrespective of the model, we find the same result: the SFRs did not affect IEOp whatsoever. This is a depressing finding, given the massive policy changes the reforms initiated. It is, however, not entirely unexpected (and in line with Lafortune et al., 2018). The natural follow-up question is *why* the SFRs were not able to equalize opportunities. We can only speculate about this, but one possibility is that the benefits of the reforms materialize only later in the students' lives. Maybe, for some reason, attending a post-reform school doesn't matter initially but becomes relevant once the student enters the labor force. This is not entirely implausible, as such phenomena have been documented for other policies as well. But at the end of the day, we have no definite answer to why the SFRs did not level the playing field for students.

Chapter 2: Maybe parents threw a wrench in the works—which brings me to the next chapter. In the summer of 2020, I discovered a phenomenal audio series produced by the New York Times titled "Nice White Parents". In it, reporter Chana Joffe-Walt follows the story of the Boerum Hill School for International Studies, a public school that was built with the explicit goal to improve the educational prospects of Black children. The school was strategically located in a White neighborhood in an effort to desegregate classrooms. Yet, even if they helped build the school, White parents did not send their children there. In interviews with the school's founders, Joffe-Walt finds out that the reason for this is as simple as it is distressing: although the White parents were in support of desegregation, they still didn't want to jeopardize their children's academic career—and sent them to different schools. That is, parents react to policy, thereby potentially counteracting its purpose. This nexus of public policy and parenting sparked an interest in me. Maybe education policy is futile if parents are not taken into account? I'm exaggerating, of course, but still. After listening to the audio series, I decided to add parents to my research agenda.

Around the same time, I read Matthias Doepke's and Fabrizio Zilibotti's fantastic book "Love, Money & Parenting" (Doepke and Zilibotti, 2019). In the book, they argue that parents matter for macroeconomics and should have their place in economic modelling. As a case in point, in Doepke and Zilibotti (2017), they develop a model where the choice of a parenting style is affected by the socioeconomic environment. Roughly speaking, if the environment becomes more competitive, the stakes are higher for children, and parents adjust their parenting accordingly. In particular, they become more demanding. Through the lens of their model, Doepke and Zilibotti connect the rise of authoritative parenting with increasing inequality. Indeed, both within- and across-country comparisons expose a striking correlation between inequality and demandingness. In this sense, tiger moms and helicopter parents may have become so popular because the environment of the 21$^{\text{st}}$ century incentivized them. A competitive world requires a particular skill set, with which parents are willing to equip their children.

Research in developmental psychology proves those parents right: again and again, parenting styles have been shown to be essential for child outcomes. Authoritative parenting in particular, a combination of demandingness and responsiveness, is associated with the most favourable outcomes: authoritative parenting is beneficial for the development of self-regulatory skills, and the verbal give-and-take of responsive parents fosters cognitive and social skills (Steinberg, 2001). Causally identifying the effect of parenting styles on child skills is challenging, however. Early research such as Baumrind (1971, 1978, 1989) relies on observational studies and more or less only reports correlations. More causal evidence based on experimental settings exists but is scarce and leaves plenty of room for contributions. From an economic perspective, we particularly know little about the dynamic processes how parenting styles affect child development and how economic decisions play into them.

In my single-author paper—chapter 2 of this thesis—I address this last point. Earlier research, such as Del Bono et al. (2016), tackle a similar question with respect to parental investments, but for parenting styles, things are trickier. From an econometric point of view, investments are easy to envision as being time-varying (the number of hours spent on childcare can be different in two time periods), but parenting styles are typically not thought of in this way. A parent either *is* authoritative or *is not*. Yet, time-varying measures would allow using dynamic panel estimators that control for the main sources of endogeneity. So this is where I deviate from existing research: I use detailed information from a U.K. cohort study and construct *time-varying* measures of parenting styles. Doing so makes significant improvements in causal identification possible.

Specifically, I construct measures of parental demandingness and responsiveness, two well-established parenting dimensions in psychology that form the basis of much of contemporary research. I construct these measures for child ages 3 to 14, i.e. for a period covering both early and middle childhood. Using value-added models and dynamic panel estimators, I find that demandingness is negatively associated with child cognitive and non-cognitive skills, whereas responsiveness is positively associated. I also show that the effects are heterogeneous across child ages, i.e. that the effects are different at different ages.

An essential contribution of chapter 2 is that I can address feedback effects, i.e. parents reacting to child skills shocks. Well, almost anyway: I can address parents reacting to *past* skill shocks, but contemporaneous reversed causality may still be an issue. Nonetheless, I find little evidence of feedback effects in the first place, which lends credence to the assumption that there are no simultaneous responses either. Still, I am actively working on an extension for chapter 2, which precisely addresses this issue of contemporaneous reversed causality. But these are dreams of the future and not part of this thesis.

Chapter 3: The final chapter is a combination of the first two chapters, in some sense: Benjamin Arold, Paul Hufe, and I investigate how school investments interact with genetic endowments. In a nutshell, we aim at answering the following question: given a student's genetic endowment, does the school matter he or she is going to? Or, put differently, can schools moderate the effects of genetics?

When I first heard questions like these, I was confused. Wasn't the question about nature *versus* nurture rather than nature *with* nurture? As it turns out, it was not. The notion of whether there exists a blank slate or whether we are pre-determined by our genetics is badly outdated. Instead, the relevant question is about gene-environment *interactions* (G×E): how genetic endowments and environments influence each other to produce outcomes (Domingue et al., 2020). G×E can arise if the genetic potential can be better expressed in some environments than others. The Scarr-Rowe hypothesis is a prominent example of this, according to which the heritability of cognitive abilities is higher in more privileged socioeconomic conditions (Tucker-Drob and Bates, 2016). Roughly speaking, having a high genetic potential is useless if you don't have the means to exploit it.

Together with Benjamin and Paul, I tried to shed light on the nexus between genetics and education policy. In particular, in chapter 3, we estimate the complementarity of students' genetic endowments and school investments. Prior research showed that educational attainment is highly heritable, and we are interested in how teacher quality and class size play into this. Is it the "genetically advantaged" students that benefit from good teachers and small classes, or the "genetically disadvantaged" students? (Note that there is no value judgement in the term genetically (dis-)advantaged. By being genetically advantaged, I mean that a student has inherited genetic variations that are positively associated with educational attainment.)

We leverage recent advances in molecular biology to measure genetic endowments and use so-called polygenic scores (PGS). PGSs are individual-level measures for the genetic propensity to an outcome of interest (educational attainment, in our case). Basically, a PGS aggregates the genetic variations associated with the outcome. Their predictive power is impressive: the PGS by Lee et al. (2018) can explain more than 12% of the variation in educational attainment. Some geneticists go even so far as to call them fortune tellers (Plomin, 2018).

Using a between-family design, we find that genetic endowments and teacher quality act as substitutes in the production of educational attainment (on the other hand, for class size, we find no significant interaction effects). Intuitively, the substitution effect we find means that genetically disadvantaged students benefit relatively more from high-quality teachers. This implies that improvements in the quality of teachers may reduce the genetic gradient in educational attainment. A substitution effect is great news because it also implies that there is no trade-off between equity and efficiency concerns: from both perspectives, school investments are desirable.

It is difficult to say what the mechanisms behind the substitution effect are. When we take a closer look at transitions in the educational system, we find that the substitutability of genetic endowments and teacher quality is largest at the stage of college education. One interpretation of this could be that high-quality teachers help disadvantaged students to go to college, e.g. by providing them with information or nudging them to apply.

Now what to do with this result? Although a policy *can* moderate the effect of genetics, this doesn't imply that this policy *should* be implemented because of that. I am sceptical that it would be sensible to base policy decisions on the population's genetic endowments, even when intentions are good. Luckily, at least in the case of school investments, our results suggest that no policy change is required anyway. Because there is no trade-off between equity and efficiency, arguing for investments based on efficiency grounds is already sufficient.

With this, I conclude the preface to my thesis. What I take from my work on the three chapters is this: children's success in life is co-determined by factors outside their responsibility—the schooling system, the household they grow up in, their genetics—but this does not mean that their success is *pre*-determined. Policy still can, and should, aim at increasing equality of opportunity. The question is just which policy is the right one. And as it is so beautifully said in virtually every paper's last paragraph: more research is needed to answer this.

Chapter 1

The Impact of School Finance Reforms on Inequality of Educational Opportunity*

1.1 Introduction

Educational achievement in the U.S. is not only the result of effort but also significantly influenced by factors outside individual responsibility—exemplified by the sizable achievement gap between students from high and low socioeconomic backgrounds (Hanushek et al., 2019). Since the Coleman Report (Coleman et al., 1966), a vast literature has studied policies that try to break the link between these factors and student outcomes. Finding effective policies is of key importance because skills acquired during childhood and adolescence are crucial for economic and social outcomes during adulthood. Inequality of educational opportunity (IEOp) will then translate into unfair inequalities later in life.

In the 1960s, legal scholars argued that the school finance regimes at the time created unequal educational opportunities for students and that these regimes violated federal and state constitutions (Coons et al., 1970; Horowitz, 1965; Wise, 1968). Because public schools were financed primarily via local property taxes, affluent districts were able to spend more per student (Hoxby, 2001). Legal suits followed that were often successful, resulting in court-ordered school finance reforms (SFRs) in many states. In a first wave of reforms, SFRs aimed at equalizing resources across districts by changing the state education funding formulas. In a second wave, beginning in the late 1980s, the goal shifted from equalization of resources to adequate funding. In an effort to create equal opportunities, SFRs during this second wave sought to equalize outcomes instead of inputs, resulting in higher relative

*This chapter is based on joint work with Eric Hanushek, Paul Hufe, and Marc Piopiunik.

spending in low-income districts.

In this study, we focus on the second wave of SFRs and investigate whether they were successful in increasing equality of educational opportunity. We follow a series of recent studies that exploit court-ordered SFRs as a source of exogenous shocks to school finances (Biasi, forthcoming; Jackson et al., 2015; Lafortune et al., 2018; Rothstein and Schanzenbach, 2021). Our framework of IEOp is based on Roemer (1998), who differentiates between effort and circumstances as the two determinants of outcomes. Intuitively, unfair inequality is outcome differences due to circumstances—factors that individuals cannot influence and thus cannot be held responsible for. In this sense, our paper asks whether the SFRs were able to reduce the influence of circumstances on student outcomes.

Our analysis is based on data from the National Assessment of Educational Progress (NAEP). Since 1990, NAEP administers standardized and state-representative tests in math and reading every two years on average. Tests are administered to around 100,000 students in each subject, grade, and year. We use NAEP test scores to measure IEOp along the lines of Bourguignon et al. (2007) and Ferreira and Gignoux (2011), who propose to estimate a linear model as a function of circumstances and effort, and to simulate a counterfactual outcome distribution with the estimated model. The counterfactual distribution is then used to compute the inequality of opportunity measure. We additionally employ a machine learning approach proposed by Brunori et al. (2021) using conditional inference forests. Descriptively, we find that IEOp has increased slightly since 1990 in most states. That is, variation in students' circumstances increasingly explains the variation in test scores.

The question is whether IEOp would have increased even more in the absence of SFRs. We rely on an event-study framework to answer this question. Court-ordered reforms represent natural experiments because of the idiosyncrasies of the judicial process: the exact timing of the reforms cannot be anticipated by legislatures and is thus plausibly exogenous (Lafortune et al., 2018; Rothstein and Schanzenbach, 2021). We employ recently developed methods to address issues related to staggered treatment adoption in event-study designs (see De Chaisemartin and D'Haultfoeuille, 2021, for a review). In particular, we focus on the proposed solution by Callaway and Sant'Anna (2020).

We don't find an effect of SFRs on IEOp, irrespective of the model. This null result is robust to a battery of robustness checks. Although it was the reforms' explicit goal to increase equality of educational opportunity, we find no evidence that they were successful, at least with respect to reducing the influence of circumstances on achievement test scores.

Several other studies exploit court-ordered SFRs for identification. Evidence suggests that the earlier reforms during the 1970s and 1980s narrowed the spending gap between richer and poorer districts as well as the SAT gap between families from different socioeconomic backgrounds (Card and Payne, 2002). Lafortune et al. (2018) and Rothstein and Schanzenbach (2021) are closest to our study from an econometric point of view and provide evidence for the impact of the court-ordered SFRs since the late 1980s. Also using NAEP data, Lafortune et al. (2018) find that the reforms increased the achievement of students in low-income districts, thus reducing the achievement gap between low- and high-income districts. Rothstein and Schanzenbach (2021) additionally find positive effects on educational attainment for students that were affected by the reforms. Interestingly—but consistent with our null result—Lafortune et al. (2018) find no effect of reforms on statewide achievement gaps between high- and low-income students (in contrast to achievement gaps between high- and low-income *districts*). They argue that this is because low-income students are not highly concentrated in low-income districts. That is, because of within-district income inequality, targeting low-income *districts* by reforming funding formulas may fail to benefit low-income *students* that live in high-income districts. This begs the question whether SFRs actually served their intended purpose—after all, their goal was to increase opportunities, which is not the same as the achievement gap between districts (but between groups of individuals).[1]

Our study contributes to the understanding of the effects of SFRs on student outcomes. In particular, we directly address the puzzling finding of Lafortune et al. (2018) that the reforms had no effect on achievement gaps between high- and low-income students. Our measure of IEOp is a broader measure of the distribution of educational opportunities and goes beyond gaps by income or race. Intuitively, although the SFRs may have had no effect on the income gradient, they may have affected other dimensions of IEOp.

We also contribute to the literature on Roemerian inequality of opportunity in at least two ways. First, the focus of the current literature is predominantly on the overall impact of circumstances on either income, wealth, or health-related issues (see e.g. Hufe et al., forthcoming; Hufe and Peichl, 2019). Yet, although we know that circumstances play an important role for life success, we know much less about the dynamic processes that generate these outcomes. Our study sheds light onto the role of the educational system, and whether school finance reforms can mitigate the increasing inequality in educational achievement. Second, our study is one of the few with a causal identification strategy. This allows us to not only describe the evolution of inequality of opportunity, but also to assess whether policy was successful in leveling the playing field.[2]

The remainder of this paper is as follows. In section 1.2, we discuss the school finance reforms since the 1970s and their effects on student outcomes. In section 1.3, we describe our measure of IEOp, and in section 1.4, we describe our data. In section 1.5, we introduce the empirical strategy, and in section 1.6, we discuss the results. Section 1.7 concludes.

[1] On the other hand, Biasi (forthcoming) finds that revenues equalization across school districts had a large effect on intergenerational mobility of low-income students. Note, however, that mobility is measured by comparing outcomes of two generations during adulthood, which is not directly comparable with achievement tests during high school.

[2] Camarero Garcia (2022) also investigates Roemerian inequality of opportunity in education. He exploits a school reform in Germany that shortened the duration of secondary school in several federal states and that increased learning intensity.

1.2 Background: School Finance Reforms

Over the last few decades, school finance reforms dramatically changed the state-level finance regimes in the United States. In a first wave during the 1970s and 1980s, the so-called equity era, SFRs aimed at reducing resource disparities across districts. Because public schools were financed primarily via local property taxes, SFRs during this era explicitly intended to redistribute from districts with high to districts with low per-pupil property values (Hoxby, 2001). At the time, affluent districts were able to spend more per student because the local property tax base is typically higher in those areas. Some of the SFRs during the equity era were implemented in response to court orders: legal scholars brought and won legal suits arguing that the school finance regime violated the responsibility of the state to provide quality education to all children (Jackson et al., 2015). Other SFRs were implemented by legislatures without a court order, sometimes to prevent potential legal battles (Lafortune et al., 2018).

The equity reforms were controversial, not least because the Coleman Report implied that additional revenues for districts with low per-pupil property values were not productive for increasing equality of educational opportunity (Coleman et al., 1966). Also, because redistribution across districts was based on endogenous property values, they have sometimes led to a levelling-down, where greater spending equality was accompanied by lower average spending (Hoxby, 2001). In the extreme case, per-pupil spending even fell in districts that were intended beneficiaries of equalization, as in the 1976 reform in California. Nonetheless—efficiency concerns aside—evidence suggests that the SFRs during the equity era were partly successful in reducing within-state inequality in spending (Murray et al., 1998), and in narrowing the achievement gap between different family background groups (Card and Payne, 2002).

Beginning in the late 1980s, finance regimes were challenged on adequacy instead of equity grounds. The adequacy era began with the 1989 Kentucky Supreme Court ruling that the constitution requires the state to provide each child with an equal opportunity to have an adequate education. To compensate for the disadvantages students in low-income districts face, the Court emphasized that the state ought to provide funding to equalize *outcomes* instead of *inputs*. As a result, many SFRs during the adequacy era aimed at higher spending in low-income than in high-income districts. Indeed, Lafortune et al. (2018) show that since 1990, real per-pupil revenues rose significantly more in the lowest income districts compared to the highest income districts. Specifically, they show that the districts in the bottom quintile of the state-wide income distribution collected about 20 percent less revenues in 1990 than districts in the top quintile, but that they are in parity since the 2000s.

The effects of the SFRs during the adequacy era on educational opportunities are ambiguous, however. On the one hand, Buerger et al. (2021) and Lafortune et al. (2018) find positive effects on educational achievement for students from low-income districts.

Rothstein and Schanzenbach (2021) find that the reforms also lead to increases in educational attainment and earnings, particularly for Black students and men.[3] On the other hand, Lafortune et al. (2018) find no effects on state-level average achievement, nor on the achievement gaps between high- and low-income students. That is, while Rothstein and Schanzenbach (2021) find evidence that the SFRs improved equality of educational opportunity, Lafortune et al. (2018) do not.

There are several explanations for this discrepancy. First, Rothstein and Schanzenbach (2021) have a significantly larger sample, which allows them to better detect state-level effects. Second, they focus on outcomes later in life instead of educational achievement. Not only does this allow them to detect effects of SFRs that materialize with a delay, but also to detect effects mediated by non-cognitive skills. Third, the gradients under consideration by Lafortune et al. (2018) may not be sufficient to capture the full effects of the SFRs on educational opportunities.

Our study addresses this third point. By using a broader measure of IEOp, we may detect effects on the distribution of opportunities that are being missed when focusing on income or racial gradients separately.

1.3 Inequality of Opportunity

1.3.1 Conceptual Framework

Roemer (1998) distinguishes between two determinants of individual outcomes: circumstances and effort. Circumstances are characteristics outside individual control—e.g. race, assigned gender at birth, or parental education—and are the source of unfair inequalities in outcomes. On the other hand, effort includes all factors that are relevant in the production of outcomes that the individual has control over and is personally responsible for. In contrast to circumstances, inequalities that are the result of differential effort are justified.

More formally, assume that an outcome of individual i is produced by that individual's set of P (time-invariant) circumstances, $\Omega_i = \{C_i^1, \dots, C_i^P\}$, and effort, e_i:

$$y_i = f(\Omega_i, e_i). \tag{1.1}$$

[3] Pooling SFRs from the equity and adequacy eras, Biasi (forthcoming) also finds positive effects of revenue equalization on intergenerational mobility. Moreover, in a recent meta-analysis on studies exploring the distribution of the causal effects of public K-12 school spending on student outcomes, Jackson and Mackevicius (2021) find that increases in per-pupil public school spending increases test scores, high school graduation, and college-going. Some but not all of the studies in their meta-analysis use SFRs for identification.

Further assume that a finite population indexed by $i \in \{1, \ldots, N\}$ can be partitioned into k non-overlapping subgroups called *types*, $\prod = \{T_1, \ldots, T_k\}$, such that individuals within each subgroup share the same set of circumstances (Ferreira and Gignoux, 2011). For example, if there are two circumstance variables—assigned gender at birth and parental education—that each have two possible realizations—male/female and high education/low education, respectively—there are $k = 4$ possible types. Each type-specific distribution of the outcome represents the opportunity set which can be achieved by exerting different degrees of effort.[4]

Equality of opportunity (EOp) is achieved if the aggregate values of opportunity sets are equalized across types. We focus on the ex-ante egalitarian approach of EOp, where the value of the opportunity set of a type is given as the expected value of its outcomes, $E[y|\Omega]$ (Ferreira and Gignoux, 2011; Fleurbaey and Peragine, 2013).[5] This implies that the distribution of opportunities in a population can be expressed as follows (Brunori et al., 2021):

$$y^c = (y_1^c, \ldots, y_i^c, \ldots, y_N^c) = (E[y_1|\Omega_1], \ldots, [y_i|\Omega_i], \ldots, [y_N|\Omega_N]), \tag{1.2}$$

where y^c denotes a counterfactual distribution of y. The ex-ante egalitarian approach thus focuses on between-type differences in the value of opportunity sets without paying attention to the specific effort realizations of individual type members.

By applying an inequality measure $I()$ to y^c, one obtains an ex-ante utilitarian measure of inequality of opportunity. We rely on the variance of y^c as our benchmark measure.[6] That is, inequality of opportunity for a population is given as the variance of the counterfactual distribution of that population:

$$\text{IOp} = I(y^c) = \sigma^2(y^c). \tag{1.3}$$

Alternatively, IOp can be expressed in relative terms:

$$\text{IOp}^{\text{rel}} = \frac{I(y^c)}{I(y)}, \tag{1.4}$$

which yields the share of overall inequality that is unfair.

Note that because in empirical applications, there will always be unobserved circumstances, the empirical IOp is weakly smaller than the true value. Measures of IOp are thus interpreted as lower-bound estimates.

[4]Note that "degree of effort" is not the same as "level of effort" for Roemer, 1998. The former refers to the quantile of the type-specific effort distribution, whereas the latter refers to the absolute level of exerted effort. This implies that outcome differences between two individuals may not necessarily be justified by differences in the exerted effort.

[5]For an overview of alternative approaches, see Ramos and Van de Gaer (2016).

[6]In robustness analyses, we also apply the Gini index, the mean log deviation, and percentile ratios.

1.3.2 Measuring Inequality of Educational Opportunity

Implementing Equation (1.3) requires a partitioning of the population into types, $\prod$, and an estimate of the counterfactual outcome, y^c. To obtain the latter, we rely on the parametric approach (Bourguignon et al., 2007; Ferreira and Gignoux, 2011). Specifically, we first estimate the following model:

$$y_i = \beta_0 + \sum_{p=1}^{P} \beta_p C_i^p + \varepsilon_i, \tag{1.5}$$

where y_i is an outcome of individual i, and C_i^p is the p-th circumstance variable. β_0 and β_p are coefficients, and ε_i is an error term. Equation (1.5) is estimated with OLS, and the predicted values yield the counterfactual outcomes:

$$y_i^c = \sum_{p=1}^{P} \widehat{\beta}_p C_i^p. \tag{1.6}$$

The parametric approach has two crucial shortcomings. First, it makes strong assumptions about the functional form of Equation (1.1). Importantly, Equation (1.5) may miss interdependencies or non-linearities. Second, the researcher is forced to make arbitrary choices about which circumstance variables to include in the model. This can be problematic because IOp estimates are sensitive to the number of types considered (Brunori et al., 2021; Ferreira and Gignoux, 2011). We therefore also apply the approach by Brunori et al. (2021), who propose computing y_i^c using conditional inference regression forests. In contrast to the parametric approach, where the researcher has to specify the functional form and manually select the circumstance variables, regression forests rely on machine learning algorithms that make data-driven decisions. Moreover, forests are well-suited to balance the bias-variance trade-off, i.e. the problem that the model might be either over- or underfitted.[7] We describe the approach by Brunori et al. (2021) in section A.2 in the Appendix.

[7] A model with low bias and high variance yields good in-sample predictions, but bad out-of-sample predictions (overfitting). For a model with high bias and low variance, the opposite is the case (underfitting). Regression forests incorporate algorithms that balance the bias and variance of a model.

1.4 Data

1.4.1 NAEP

We construct IEOp measures from restricted-use microdata from the National Assessment of Educational Progress (NAEP). NAEP administers standardized tests in grades 4, 8, and 12 that are comparable both across states and years. The first state-level assessments were conducted in 1990 in 38 states and have been repeated roughly every two years since (see Table 1.1). For each state and year, a stratified sample of approximately 100 grade-eligible public schools is selected within each jurisdiction, plus about 700 private schools across all states. Within each school, about 60 students are selected for assessment. In the most recent assessment in our sample in 2015, 49 states participated, with on average 5,219 students per state. We restrict the sample to 8^{th} grade NAEP scores because important student characteristics are not available for 4^{th} graders, and 12^{th} graders are only assessed in one year. We also restrict our sample to the math and reading assessments because these subjects are available for multiple years.

The math assessments consist of 25-minutes blocks that cover five mathematics content strands: number properties and operations; measurement; geometry; data analysis, statistics, and probability; and algebra. Similarly, the reading assessments consist of 25-minutes blocks that cover three reading targets: locate/recall; integrate/interpret; and critique/evaluate. Assessments take place in a testing environment with fewest possible distractions, such as the students' own classrooms. Due to time constraints, students don't receive all blocks but only a subset. A set of plausible values—imputed values that resemble individual test scores—are therefore provided for each student. We use the first plausible value in our analysis.[8]

Our concept of IEOp is based on the idea that circumstances affect test scores. NAEP collects the following student-level circumstances that are available in all assessment waves (and that we recode as indicated in parentheses to harmonize them across waves): gender (female; male), books at home (more than 25; less or equal than 25), highest degree mother and father (no high school; high school; some college after high school; college), ethnicity (Hispanic; Black; White; Asian; Indian; other), and limited English proficiency (yes; no).[9,10] Information about gender, books at home, parental education, and ethnicity is self-reported by the student. Limited English proficiency is reported by their schools.

[8]We restrict ourselves to the first plausible value due to computation time constraints.

[9]We exclude eligibility for free lunch, a proxy for low-income students, because of missing values. In particular, if a school did not participate in the National School Lunch Program, all sampled students in that school are assigned a missing value ("information not available"). Because of this, information about free lunch eligibility is a bad proxy for low-income status in NAEP.

[10]Limited English proficiency is a borderline case of whether it should be considered a circumstance variable. On the one hand, it is an indicator for migration status. On the other hand, English proficiency is an indicator of effort among immigrants.

Table 1.1: NAEP Overview

	Subject and grade			
Year	Math, 8th grade	Reading, 8th grade	Number of states	Number of students
1990	X		38	97,890
1992	X		42	105,280
1996	X		43	103,880
2000	X		41	87,310
2002		X	45	113,770
2003	X	X	50	288,870
2005	X	X	50	298,200
2007	X	X	50	296,940
2009	X	X	50	306,670
2011	X	X	50	317,240
2013	X	X	50	325,620
2015	X	X	49	255,710

Note: The table reports the NAEP assessment years (math and reading, 8th grade) and the sample sizes. The table excludes Alaska, Utah in 2015, and state-subject-year cells with less than 1,000 observations in NAEP. The table also excludes the 1994 NAEP assessment. In the final column, students are pooled across subjects and rounded to the nearest 10. Data: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), Mathematics and Reading Assessments, Selected Years. Own calculations.

1.4.2 Inequality of Educational Opportunity

Figure 1.1 shows the benchmark estimates for math IEOp in 1992 and 2013, where IEOp is measured using conditional inference forests (see corresponding Figure A.1 in the Appendix for reading IEOp).[11] In section A.2 in the Appendix, we describe how we have implemented the forests. The estimates in Figure 1.1 (Figure A.1) refer to overall IEOp, given as the weighted variance of the counterfactual NAEP math (reading) test score distribution.[12] We compute Equation (1.3) separately for each state-subject-year cell. As circumstance variables we include gender, books at home, highest degree mother, highest degree father, ethnicity, and limited English proficiency.

Figure 1.1 shows that math IEOp increased for the majority of states. Inequality increased most in the South West (Utah and Colorado), Midwest (Minnesota, Iowa, and Wisconsin), and North West (Idaho). Math IEOp decreased between 1992 and 2013 in just eight states (in descending order): New Jersey, Louisiana, Mississippi, Texas, South Carolina, Alabama, Indiana, and Florida. Interestingly, IEOp in the U.S. has converged

[11] Figures A.2 and A.3 in the Appendix also show the relative measure of math and reading IEOp, respectively. The within- and across-state variation in IEOp is similar for the absolute and relative measures.

[12] We also apply the Gini coefficient, mean log deviation (MLD), and three percentile ratios (P95/P5, P90/P10, and P75/P25) as alternatives. Figure A.4 in the Appendix shows that all measures are highly correlated with each other.

between 1992 and 2013 across states: high-inequality states in 1992 either became more equal or only slightly more unequal, whereas low-inequality states became more unequal.

The OLS and tree-based approaches yield similar results to the estimates using conditional inference forests (see Figure A.5 in the Appendix). In section A.3 in the Appendix, we also show that the out-of-sample prediction accuracies of the different approaches are similar.

1.4.3 School Finance Reforms

Some states have implemented more than one SFR (see Table A.1 in the Appendix). Because our empirical framework requires a single reform per state, we follow Lafortune et al. (2018) and select the most consequential reform in each state. Specifically, for each state s and each SFR in that state, Lafortune et al. (2018) estimate the following time series regression:

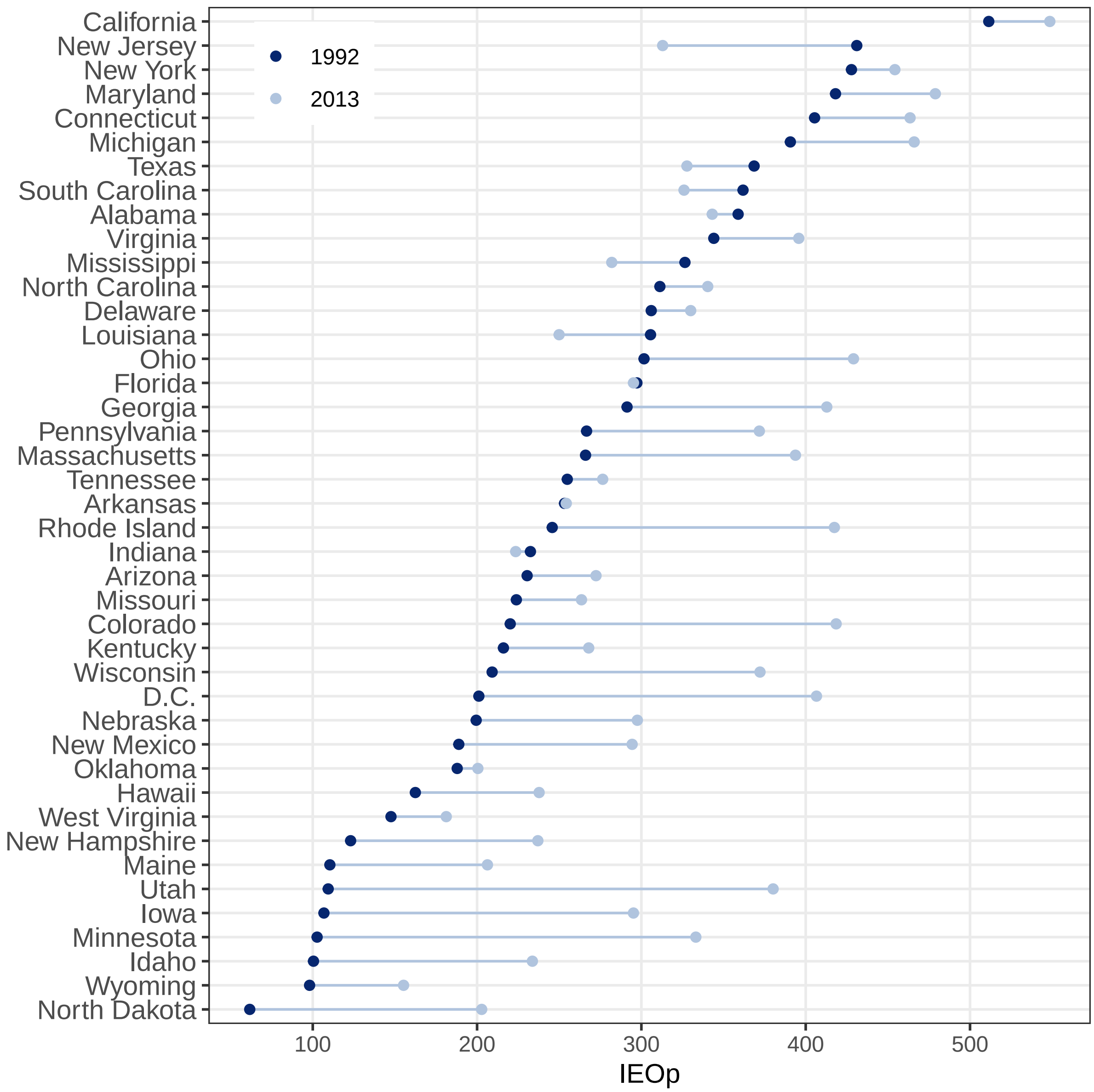
$$E_{st} = \alpha_s + 1(t > t_s^n)\lambda_s + \varepsilon_{st}, \tag{1.7}$$

where $t_s^n \in \{1, \ldots, N_s\}$ is the n-th SFR in state s (out of N_s reforms in total), and E_{st} is a measure of progressivity of state aid. They choose the reform that yields the largest t statistic for λ_s.

Figure 1.2 visualizes the chosen reforms and their timing.[13] The states highlighted in red are states that have implemented test-based accountability reforms. Buerger et al. (2021) show that these reforms were particularly effective in reducing the achievement gap between low- and high-income districts. In a robustness check, we restrict the treatments to these accountability reforms.

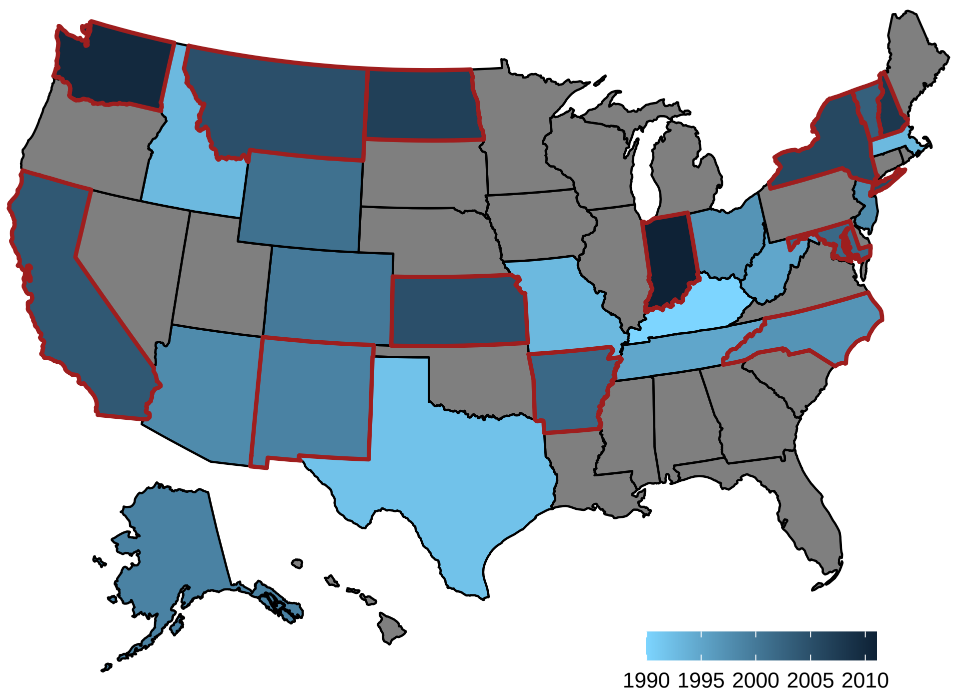
[13] The reforms in Lafortune et al. (2018) are SFRs that are either court-ordered or implemented by legislatures without a court order. In a robustness check, we restrict the sample to court-ordered reforms only.

Figure 1.1: Inequality of Educational Opportunity



Note: The figure shows math IEOp in 1992 and 2013, where IEOp is measured using conditional inference forests. The estimates refer to overall IEOp, given as the weighted variance of the counterfactual NAEP math test score distribution. The included circumstance variables are gender, books at home, highest degree mother, highest degree father, ethnicity, and limited English proficiency. Data: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), Mathematics Assessment, Selected Years. Own calculations.

Figure 1.2: School Finance Reforms



Note: The map shows the school finance reforms chosen by Lafortune et al. (2018). The color-code indicates the reform year. States in grey are defined to have had no school finance reform between 1990 and 2015 (see Table A.1 in the Appendix for an overview). The states highlighted in red are states that have implemented test-based accountability reforms.

1.4.4 Analysis Sample

We remove two outliers from the sample: math IEOp for D.C. in 1996 and math IEOp for New Jersey in 2015. We also remove Utah in 2015 and all of Alaska because of missing values in the circumstance variables, as well as state-subject-year cells with less than 1,000 observations in NAEP. Finally, we remove Kentucky from the sample because it is always treated (reform in 1990).

Figure 1.3 visualizes the treatment periods of the estimation sample. For each state, the first row refers to math IEOp, and the second row refers to reading IEOp. Light-blue (dark-blue) tiles indicate pre-reform (post-reform) periods, and white tiles indicate missing observations, i.e. where no IEOp measure is available.[14]

1.5 Empirical Strategy

Following Lafortune et al. (2018) and Rothstein and Schanzenbach (2021), we leverage variation in the timing of court-ordered and legislative SFRs in an event-study framework to identify their causal effect on IEOp. Our framework assumes that states without a reform in a particular year are valid counterfactuals for states with a reform in that year (after accounting for fixed differences between states and for common time effects). For this validity to hold, a key identifying assumption is that the reforms are exogenous events and not a response to student outcomes. Lafortune et al. (2018) and Rothstein and Schanzenbach (2021) argue that, due to the long judicial process, the exact timing of court-ordered reforms is plausibly exogenous, and that they therefore can be seen as natural experiments.[15,16]

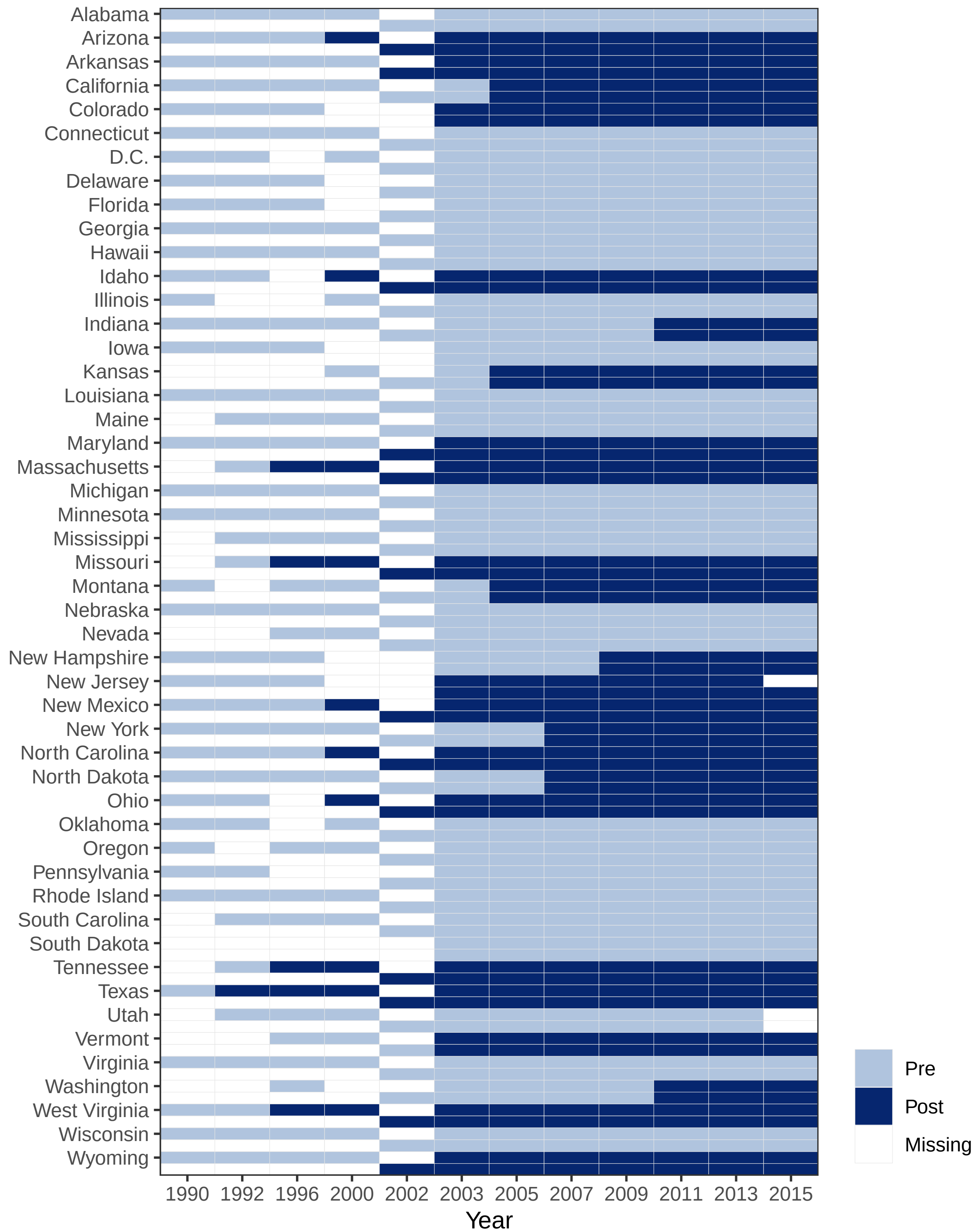
We describe the event-study framework that makes use of this assumption in section 1.5.1. In section 1.5.2, we discuss issues threatening the validity in designs where the treatment adoption is staggered (as in our case).

[14]Note that we have an unbalanced panel. On the one hand, reading IEOp is only available since 2002. On the other hand, the treatment years, i.e. the year when the SFR was implemented, are not necessarily overlapping with the years where we observe IEOp. This implies that not all relative time periods since treatments are observed for all states.

[15]Note that this logic does not apply to *legislative* SFRs. We therefore exclude them in robustness checks (finding that their exclusion barely affects the results).

[16]Card and Payne (2002) and Jackson et al. (2015) are other studies using court-ordered SFRs as exogenous events.

Figure 1.3: Treatment Status



Note: The figure shows the treatment periods of the estimation sample. For each state, the first row refers to math IEOp, and the second row refers to reading IEOp. Light-blue (dark-blue) tiles indicate pre-reform (post-reform) periods, and white tiles indicate missing observations.

1.5.1 Event-study Framework

As a baseline, we estimate the following static two-way fixed effects (TWFE) model:

$$\begin{aligned}\theta_{st} =& \delta_s + \kappa_t + 1(t > t_s^*)\beta^{jump} + 1(t > t_s^*)\min(8, t - t_s^*)\beta^{exposure} + (t - t_s^*)\beta^{trend} + \\ & X_{st}\gamma + \varepsilon_{st},\end{aligned} \tag{1.8}$$

where θ_{st} denotes IEOp in state s at year t. δ_s and κ_t are state and year effects, respectively. t_s^* denotes the year when the reform was implemented in state s, X_{st} is a vector of time-varying controls, γ is a vector of coefficients, and ε_{st} is an error term.

$1(t > t_s^*)$ is a dummy variable indicating the time periods after state s has implemented its SFR. β^{jump} then represents the jump in the outcome immediately following the implementation of a reform. $1(t > t_s^*)\min(8, t - t_s^*)$ is a count variable indicating the number of years state s is exposed to the reform (with a maximum of eight years, and equal to zero for $t \leq t_s^*$). The coefficient $\beta^{exposure}$ captures the idea that the effect of a SFR on IEOp gradually increases because cohorts in s right after t_s^* are not treated for their full school years. Intuitively, the longer a student attends school under a reformed school finance regime, the larger is the impact on his or her outcomes. Capping the count variable reflects that there is an upper bound for how long a student can be treated. Because we use NAEP test scores at 8$^{\text{th}}$ grade to construct θ_{st}, a student that enters 1$^{\text{st}}$ grade in the year after t_s^* can be treated for eight years at maximum. Finally, $(t - t_s^*)$ is a linear trend, and β^{trend} captures trend differences prior to t_s^* between states with and without SFR implemented in year t. $\widehat{\beta}^{trend} = 0$ is typically used as an indication that the timing of the SFRs is exogenous.

Equation (1.8) assumes a linear delayed effect. We additionally estimate the following dynamic TWFE model to also allow the delayed effect to be non-linear:

$$\theta_{st} = \delta_s + \kappa_t + \sum_{\substack{r=k_{min} \\ r \neq 0}}^{k_{max}} 1(t = t_s^* + r)\beta_r + 1(t < t_s^* + k_{min})\beta_{lower} + 1(t > t_s^* + k_{max})\beta_{upper} + \varepsilon_{st}. \tag{1.9}$$

β_r is the effect of the SFRs on IEOp r years after implementation (or prior to, for $r < 0$). We exclude $r = 0$ such that t_s^* is the reference point and all other effects are measured relative to $r = 0$. k_{min} and k_{max} specify the lower and upper ends of the time horizon under consideration, respectively. $1(t < t_s^* + k_{min})$ and $1(t > t_s^* + k_{max})$ indicate periods before or after this time horizon, such that β_{lower} and β_{upper} capture the average effects for the lower and upper endpoints, respectively.

$\beta^{exposure}$ in Equation (1.8) and β_r in Equation (1.9) are our coefficients of interest. For their consistent estimation, we require the standard parallel trends assumption, i.e. that θ_{st} would have moved in parallel in states with and without SFRs. Similar to $\widehat{\beta}^{trend} = 0$ in Equation (1.8), $\widehat{\beta}_r = 0$ for $r < 0$ is typically used as a test for pre-existing trends.

1.5.2 Staggered Adoption of School Finance Reforms

The SFRs during the adequacy era were implemented over a period of almost 20 years. A series of recent studies has highlighted that a TWFE model with staggered adoption can yield biased estimates because the model makes both "clean" comparisons between treated and not (yet) treated units as well as "forbidden" comparisons between units who are both already treated (De Chaisemartin and D'Haultfoeuille, 2021; Roth and Sant'Anna, 2021). The β_r coefficients in Equation (1.9) are then a weighted average of all these comparisons, but when treatment effects are heterogeneous across time or units, the weights may be negative because of the forbidden comparisons (De Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021). In the extreme case, the TWFE coefficients may even have a different sign than every unit's treatment effect.

Several estimators have been proposed to account for this issue. They all have in common that they isolate the clean comparisons between treated and not treated units, and then aggregate the unit-level treatment effects to obtain a parameter of interest. The estimators differ in their identifying assumptions (e.g. the specific assumptions about parallel trends) and their applicability (e.g. the required sample structure such as a balanced panel).

The approach by Callaway and Sant'Anna (2020) is best suited for our application, and we thus implement their estimator. First, they allow for arbitrary heterogeneity of treatment effects. This is important because, *a priori*, we would expect heterogeneity both across time and states: across time because the effect of a SFR is likely to increase as students are progressively exposed to the reformed school finance regime; across states because the environments in which the reforms are implemented may be different (e.g. the business cycles). Second, their approach only requires the parallel trends assumption to hold conditional on covariates, which allows us to condition on observable state-level characteristics. Third, they allow for an unbalanced panel. This is crucial because we don't observe the NAEP reading test scores before 2002.

Callaway and Sant'Anna (2020) is applicable for staggered adoption designs where (i) the treatment is irreversible (i.e once a unit is treated, it stays treated), (ii) there are no always-treated units (or if there were, they have been excluded from the sample), and (iii) treatment anticipation is limited and known.[17] Our application satisfies (i): once implemented, SFRs were not reversed. It also satisfies (ii) by construction because we drop always-treated units. (iii) also holds because the exact timing of the reforms cannot be anticipated by legislatures.

The parallel trends assumption of Callaway and Sant'Anna (2020) further requires

[17] When covariates are included, an overlap condition additionally applies, i.e. that there is common support in the generalized propensity scores between treated and non-treated units. This guarantees that for each treated unit with a particular covariate value, there are at least some untreated units in the population with the same value.

that, conditional on covariates, the average outcomes for units that are first treated in a particular year and for a control group would follow parallel paths in the absence of treatment. The control group can either be the never-treated or not-yet-treated units. We favour the never-treated units because the conditional parallel trends assumption then does not restrict observed pre-treatment trends across groups, whereas it does when using not-yet-treated units as control group (Marcus and Sant'Anna, 2021).

The basic building block of Callaway and Sant'Anna (2020) is the group-time average treatment effect on group g at time t, denoted as $ATT(g,t)$.[18] If the identifying assumptions hold, and covariates play no role, $ATT(g,t)$ is given as the difference between the expected change in the outcome for cohort g between periods $g-1$ and t and the expected change for a control group C:

$$ATT(g,t) = E[Y_t - Y_{g-1}|G_g = 1] - E[Y_t - Y_{g-1}|C = 1], \tag{1.10}$$

where G_g is a binary variable that is equal to one if a unit is in group g ($G_{i,g} = \mathbf{1}\{G_i = g\}$), and C is a binary variable that is equal to one for never-treated units ($C_i = \mathbf{1}\{G_i = \infty\}$).

$ATT(g,t)$ can be estimated by replacing the expectations with their sample analogs:

$$\widehat{ATT}(g,t) = \frac{1}{N_g} \sum_{i:G_i=g} [Y_{it} - Y_{i,g-1}] - \frac{1}{N_C} \sum_{i:G_i=\infty} [Y_{it} - Y_{i,g-1}], \tag{1.11}$$

where N_g and N_C are the numbers of units in group g and the control group, respectively. If there is treatment anticipation, $Y_{i,g-1}$ can be replaced with $Y_{i,g-\delta-1}$, where δ is the anticipation period.

Callaway and Sant'Anna (2020) propose estimating a separate average treatment effect for each group-time combination. In a next step, they aggregate the $\widehat{ATT}(g,t)$ to a parameter of economic interest. For our application, we focus on the event-study representation:

$$\widehat{ATT}(e) = \sum_g \widehat{w}_g \widehat{ATT}(g, g+e), \tag{1.12}$$

where $e = t - g$ denotes the relative time to the treatment adoption, and $\widehat{w}_g$ denotes the (estimated) weight for group g. In the simplest case, $\widehat{w}_g$ is given as the relative group size of g in the treated population.

[18] In our application, g is given by the reform year. Two states share the same g when their respective reforms are implemented in the same year.

1.5.3 Control Variables

We include a battery of state-level control variables to strengthen the parallel trends assumption: the population shares of (i) Whites, (ii) Blacks, and (iii) Hispanics, (iv) the share of unemployed individuals, (v) the share of individuals with at least a four-year college degree, (vi) the poverty rate, and (vii) the mean individual income (in logarithms).

We use data from IPUMS CPS, an integrated data set based on the Current Population Survey (CPS; Flood et al., 2020). The CPS is a monthly household survey conducted jointly by the U.S. Census Bureau and the Bureau of Labor Statistics, and is administered to over 65,000 households. Our variables of interest are taken from the CPS's March Annual Social and Economic Supplement. We use the annual social and economic supplement individual weights to aggregate the variables to the state level.[19]

Table 1.2 presents descriptive statistics of the control variables for states with and without at least one school finance reform (i.e. treatment and control states) in 1990. The values refer to the unweighted averages across all states within a group. Treatment and control states are similar in terms of unemployment, poverty, education, and income. In terms of ethnic composition, treatment states have higher shares of Whites and Hispanics on average, but lower shares of Blacks.

Table 1.2: Control Variables

	States with SFR		States without SFR		
	Mean	S.D.	Mean	S.D.	Δ Mean
White	0.89	0.07	0.81	0.19	0.09^{**}
Black	0.08	0.07	0.14	0.16	-0.07^{*}
Hispanic	0.07	0.10	0.03	0.03	0.04^{*}
Unemployed	0.04	0.01	0.04	0.01	0.00
Below poverty line	0.13	0.03	0.13	0.05	0.00
College degree	0.19	0.04	0.18	0.04	0.01
Log income	9.74	0.15	9.74	0.13	0.00

Note: The table reports unweighted means and standard deviations across states of the shares of (i) Whites, (ii) Blacks, (iii) Hispanics, (iv) unemployed, (v) individuals below the relative poverty line, (vi) individuals with a college degree, as well as the unweighted average of log personal incomes (all values from 1990). The last column reports the differences in means between states with and without SFR, and the statistical significance of these differences ($^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$). Individual-level data is aggregated to the state level using annual social and economic supplement individual weights. All variables are harmonized by IPUMS CPS. Data: IPUMS CPS (Flood et al., 2020).

[19] The state-level sample size of the IPUMS CPS in 1990 ranges from 1,259 (Vermont) to 14,437 (California) individuals.

1.6 Results

1.6.1 Baseline Results

Table 1.3 shows the estimates for the static TWFE model. In the first column, the model is estimated on a pooled sample including both math and reading IEOp (including state-subject fixed-effects). In columns two and three, the model is estimated on separate samples for math and reading IEOp, respectively (including state fixed-effects). All models include the full set of control variables, and standard error are clustered at the state-subject (pooled sample) or state (separate samples) level. None of the estimates reaches statistical significance at the conventional level, suggesting that the SFRs had to effect on IEOp.

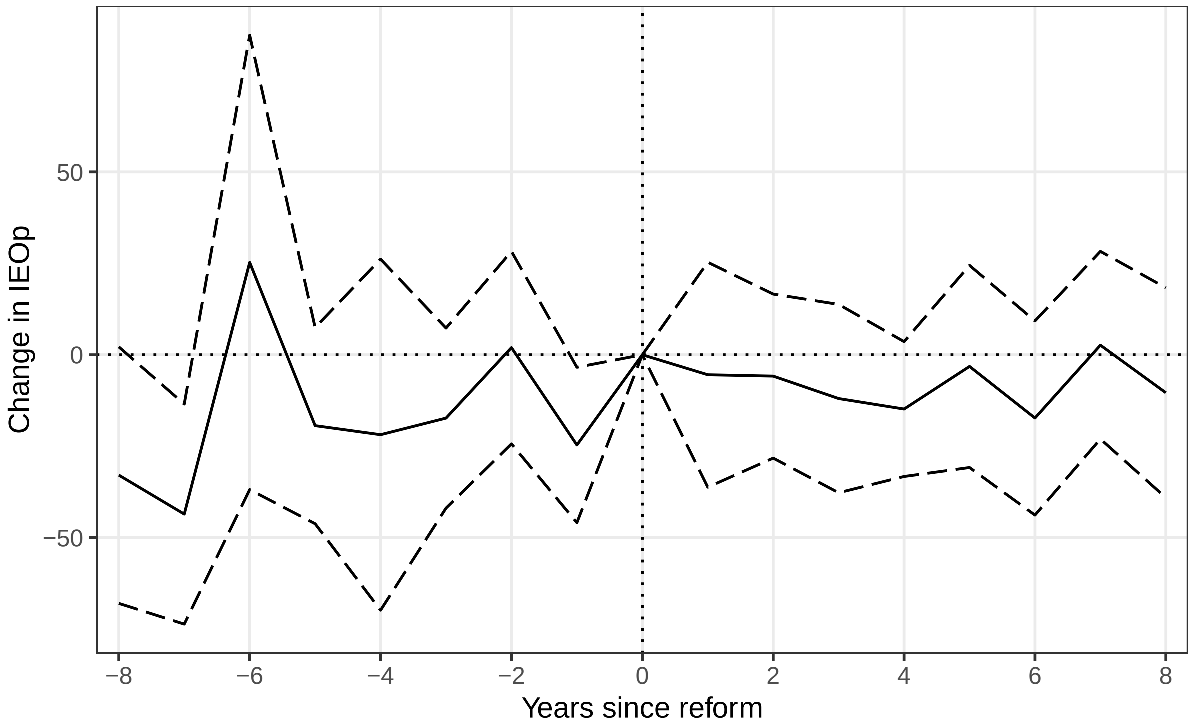
Table 1.3: Static TWFE Estimates

	IEOp		
	Math and reading	Math only	Reading only
β^{jump}	10.75	11.37	13.67
	(9.39)	(13.64)	(12.11)
$\beta^{exposure}$	-0.39	-0.20	-1.28
	(1.49)	(2.32)	(1.53)
β^{trend}	-0.12	-0.44	1.15
	(0.89)	(1.11)	(1.16)
Control variables	Yes	Yes	Yes
Within R^2	0.03	0.04	0.05
N	886	500	386

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$
Note: The table shows the estimates for the static TWFE model. In the first column, the model is estimated on a pooled sample including both math and reading IEOp. In columns two and three, the model is estimated on separate samples for math and reading IEOp, respectively. All models include the following state-level control variables: the population shares of Whites, Blacks, and Hispanics; the share of unemployed individuals; the share of individuals with at least a four-year college degree; the poverty rate; and the mean individual income (in logarithms). The model in the first column includes state-subject and year fixed-effects, the models in columns two and three include state and year fixed-effects. Standard error are clustered at the state-subject (first column) or state (second and third column) level. Regressions are unweighted.

Figure 1.4 visualizes the estimates for the dynamic TWFE model (estimated on the pooled sample). The solid line represents the coefficients, and the dashed lines represent the pointwise 95% confidence intervals using clustered standard errors at the state-subject level. We set the lower and upper end of the time horizon to $k_{min} = -8$ and $k_{max} = 8$, respectively, and the lower and upper endpoint bins to $[-\infty, -9]$ and $[9, \infty]$, respectively. We exclude $r = 0$, i.e. $\beta_0 = 0$ by construction. All estimates of the dynamic model thus indicate changes in IEOp with respect to the reform year. The result of the dynamic model is broadly in line with the static model, but the pre-treatment coefficients reach borderline significance in two periods. Pre-treatment coefficients aside, the dynamic TWFE model supports the conclusion of the static model.

Figure 1.4: Dynamic TWFE Estimates



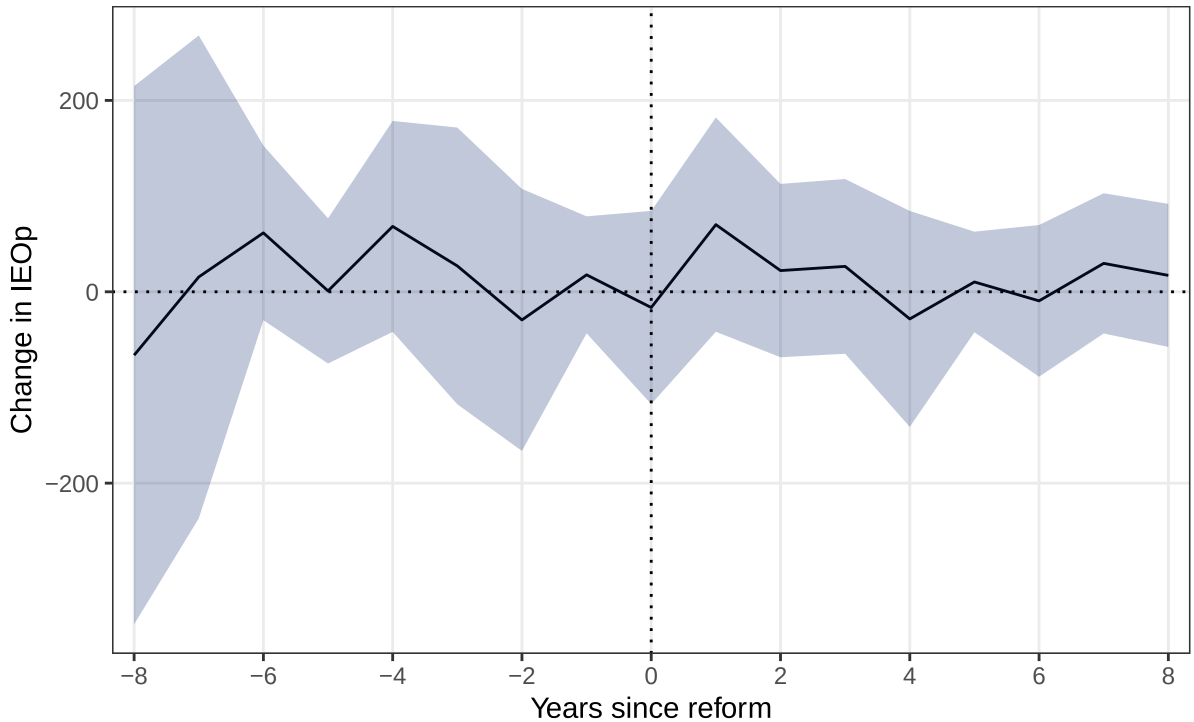
Note: The figure shows the estimates for the dynamic TWFE model. The solid line represents the coefficients, and the dashed lines represent the pointwise 95% confidence intervals using clustered standard errors at the state-subject level. The model is estimated on a pooled sample including both math and reading IEOp, and includes the following state-level control variables: the population shares of Whites, Blacks, and Hispanics; the share of unemployed individuals; the share of individuals with at least a four-year college degree; the poverty rate; and the mean individual income (in logarithms). The model also includes state-subject and year fixed-effects. The regression is unweighted. The figure excludes the estimates for β_{lower} and β_{upper}.

Figure 1.5 shows the event study plot for the estimates using the method by Callaway and Sant'Anna (2020). We use the never-treated units as control group and assume no anticipation ($\delta = 0$). The group-time average treatment effects are computed using an outcome regression, and pre-treatment ATTs are computed using varying base periods.[20] To aggregate the ATTs across lengths of exposure to the treatment, we balance the sample with respect to event time to ensure that the composition of groups remains comparable when event time changes. Specifically, we drop groups that are not exposed to a SFR for at least eight periods. The blue area represents the simultaneous 95% confidence bands using

[20]With varying base periods, pseudo-ATTs are computed for pre-treatment periods pretending that the treatment was implemented in that period instead of when it was actually implemented.

clustered bootstrapped standard errors at the state-subject level (which can be used to pre-test the parallel trends assumption as well as treatment effect estimates in post-treatment periods).[21] The results are consistent with the static and dynamic TWFE models: we can find no statistically significant effect of the SFRs on IEOp.

Figure 1.5: Group-time Average Treatment Effect Estimates



Note: The figure shows the event study plot for the group-time average treatment effects. The solid line represents the coefficients, and the blue area represents the simultaneous 95% confidence bands using clustered bootstrapped standard errors at the state-subject level. The model is estimated on a pooled sample including both math and reading IEOp, and includes the following time-invariant state-level control variables (1990 values): the population shares of Whites, Blacks, and Hispanics; the share of unemployed individuals; the share of individuals with at least a four-year college degree; the poverty rate; and the mean individual income (in logarithms). The never-treated units are the control group and there is no anticipation. The group-time average treatment effects are computed using an unweighted outcome regression, and pre-treatment ATTs are computed using varying base periods. To aggregate the ATTs across lengths of exposure to the treatment, the sample is balanced with respect to event time.

[21] Because the reform years and the NAEP waves are not consistent, and because NAEP is only available for every second year, we don't observe all relative periods for all states. For example, if we observe NAEP test scores in 1990 and 1992, and the SFR in a state was implemented in 1992, we don't observe $Y_{g-1} = Y_{1991}$ for that state. To compute $\widehat{ATT}(e) = \widehat{ATT}(-1)$, the algorithm by Callaway and Sant'Anna (2020) then selects the earliest available period prior to Y_{1991} (i.e. Y_{1990} in this example).

1.6.2 Robustness

In Figure 1.6, we reproduce Figure 1.5 but change key parameters. In panel **A**, we use the not-yet-treated units instead of the never-treated units as control group. In contrast to the latter, the conditional parallel trends assumption then restricts observed pre-treatment trends across groups. In panel **B**, we assume a one-year anticipation period ($\delta = 1$). This effectively shifts the reform year and changes the reference period from $g-1$ to $g-\delta-1$. In panel **C**, we use pointwise instead of simultaneous confidence bands. This yields narrower confidence bands, but suffers from multiple-testing problems. In panel **D**, we compute pre-treatment ATTs using a fixed base period ($ATT(g, g-1) = 0$ for all g). This changes the interpretation of pre-treatment ATTs, but leaves post-treatment ATTs unchanged. In panel **E**, we leave the sample unbalanced with respect to event time. This implies that the ATTs can be affected by changes in the sample across periods. Finally, in panel **F**, we use a weighted regression (weighted by the sum of NAEP student weights within a state) and leave the sample unbalanced with respect to event time. All results are similar to the baseline result.

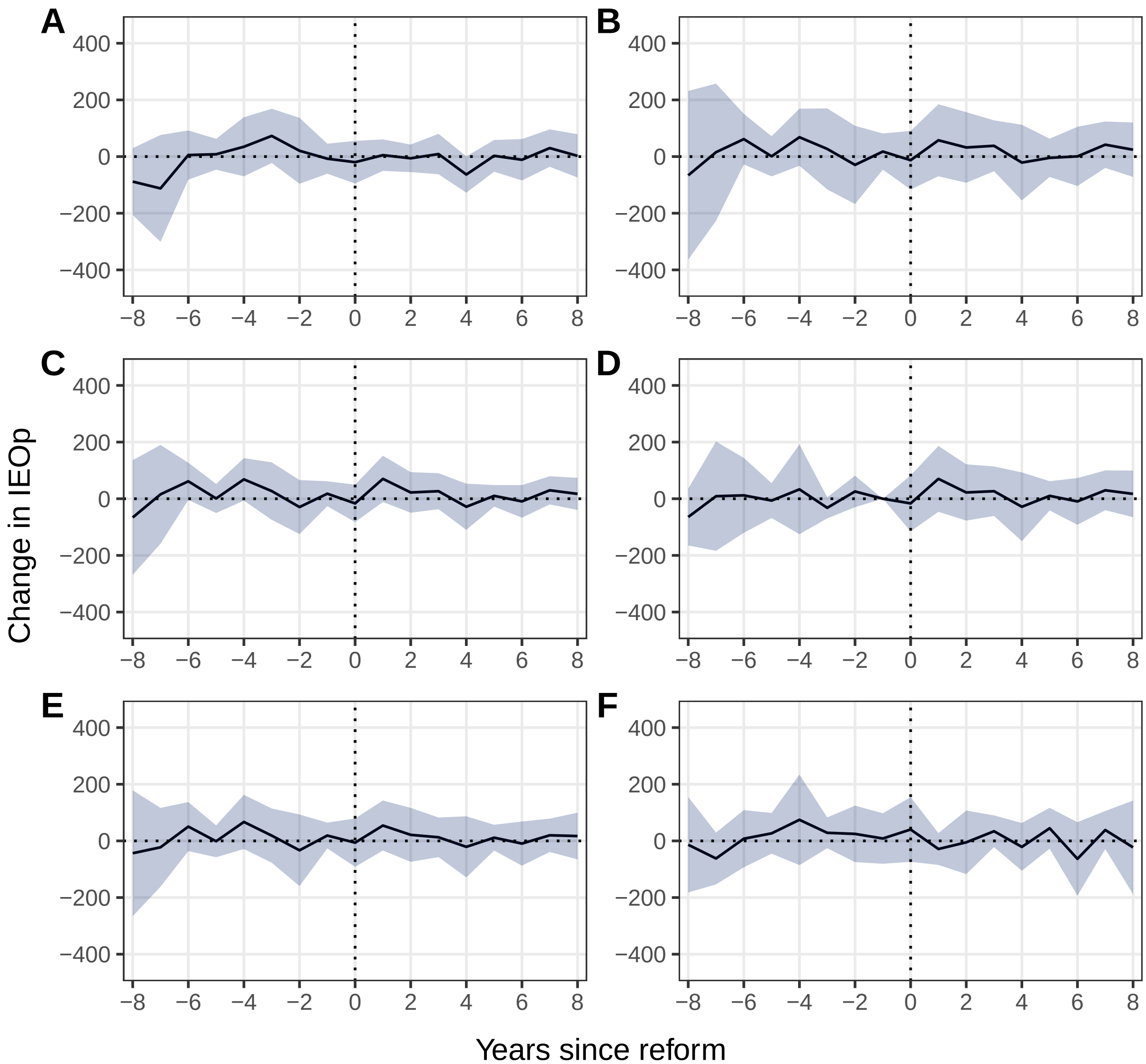
In Figure 1.7 panels **A** and **B**, we estimate the model on separate samples for math and reading IEOp. Again, the results remain largely unchanged and we find no effect of the SFRs on IEOp. In panel **C**, we exclude legislative reforms, i.e. we restrict the treatment events to court-ordered reforms only (see Table A.1 in the Appendix). Similarly, in panel **D**, we restrict the treatment events to reforms that implemented test-based school accountability systems. Buerger et al. (2021) show that these reforms were particularly effective in reducing the achievement gap between low- and high-income districts, presumably because they create incentives for school improvement. Yet again, we find no effect on IEOp. We also find no effect when we use IEOp measures that are computed using OLS or conditional inference trees instead of conditional inference forests (panels **E** and **F**, respectively).

The results also remain unchanged when we use different IEOp measures (see Figure A.9 in the Appendix).

1.6.3 Alternative Methods

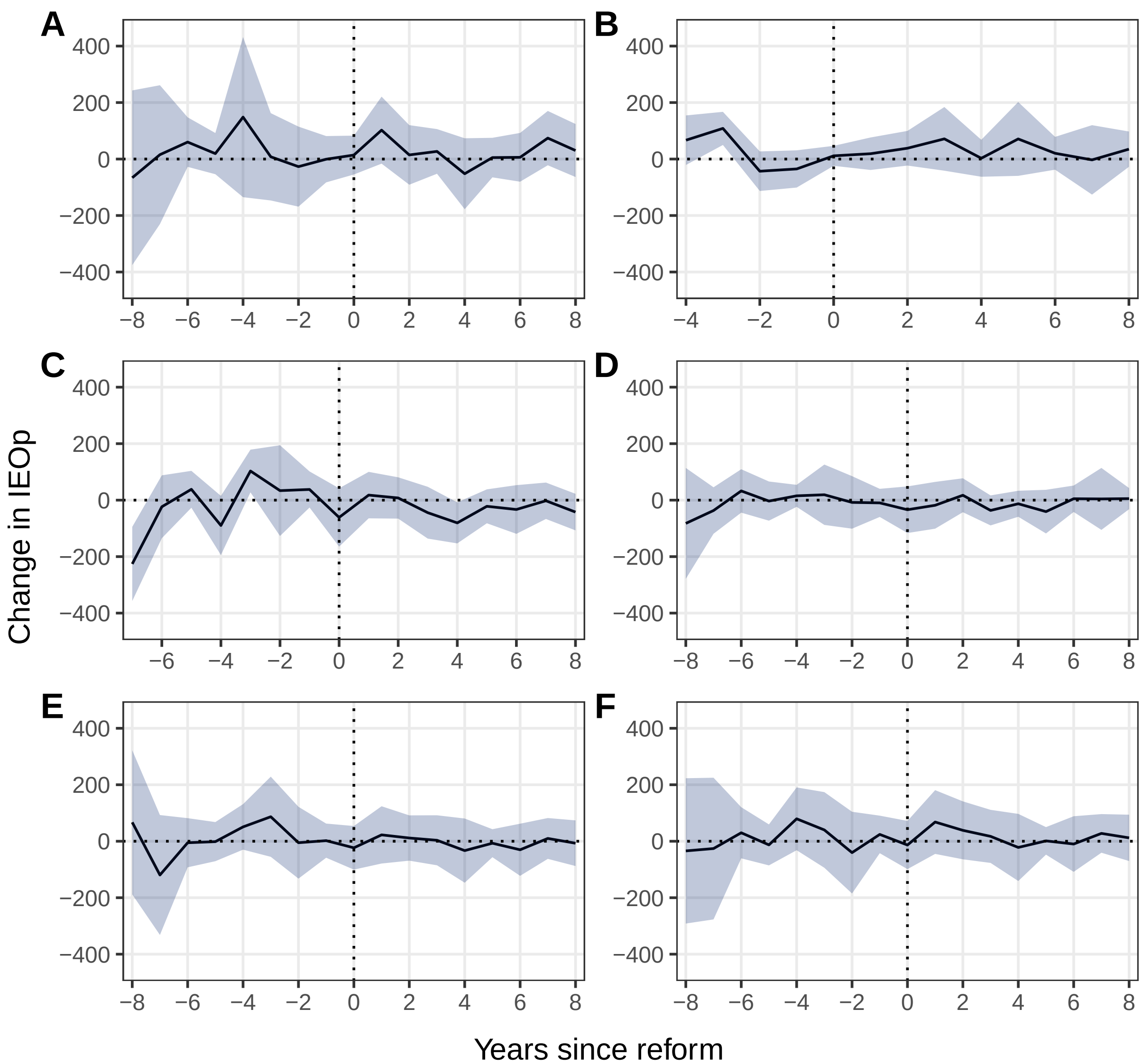
We additionally implement the estimators by Sun and Abraham (2020) and Roth and Sant'Anna (2021). Sun and Abraham (2020) propose as building blocks the cohort average treatment effects on the treated, given as the cohort-specific average difference in outcomes relative to never being treated. The approaches by Callaway and Sant'Anna (2020) and Sun and Abraham (2020) are similar in many respects, but implementation of the latter requires stronger assumption about treatment effect heterogeneity. Importantly, although both approaches allow for heterogeneity across time, Sun and Abraham (2020) do not allow for heterogeneity across groups. Because we expect treatment effect heterogeneity across

Figure 1.6: Robustness Checks



Note: The figure reproduces Figure 1.5 but with changing parameters: panel **A** uses the not-yet-treated units as control group; panel **B** assumes a one-year anticipation period ($\delta = 1$); panel **C** uses pointwise instead of simultaneous confidence bands; panel **D** computes pre-treatment ATTs using a fixed base period; panel **E** leaves the sample unbalanced with respect to event time; and panel **F** uses a weighted regression (weighted by the sum of NAEP student weights within a state) and leaves the sample unbalanced with respect to event time.

Figure 1.7: Robustness Checks (Continued)



Note: The figure reproduces Figure 1.5 but with different outcome measures. Panels **A** and **B** use math and reading IEOp only, respectively; panel **C** excludes legislative reforms; panel **D** restricts the treatment events to reforms that implemented test-based school accountability systems; and panels **E** and **F** use IEOp measured using OLS and conditional inference trees, respectively.

SFRs, we prefer Callaway and Sant'Anna (2020) to Sun and Abraham (2020).

Roth and Sant'Anna (2021) propose an analogous approach as Sun and Abraham (2020), but they make stronger treatment assignment assumptions. In particular, Roth and Sant'Anna (2021) assume that the treatment timing is (quasi-)randomly assigned, in the sense that any permutation of the treatment timing vector is equally likely. This implies that their building blocks are average treatment effects (ATE) rather than average treatment effects of the treated (ATT), as in Sun and Abraham (2020). Assuming random treatment timing allows them to use a more efficient estimator than Callaway and Sant'Anna (2020) and Sun and Abraham (2020). However, a drawback of Roth and Sant'Anna (2021) is that they do not allow for an unbalanced panel, which forces us to either drop reading IEOp or to restrict the time period to 2003–2015, i.e. when both math and reading IEOp are available. Because the latter option would mechanically introduce additional always-treated units, we estimate the ATE with math IEOp only.[22] In addition, implementation requires at least two units per treatment group. In our application, this means that for each state introducing a SFR in a given year, there needs to be at least one other state that introduces a SFR in the same year. In our application, this reduces the number of groups from 17 to just six, and the number of treated states from 24 to 13.[23]

The results are presented in section A.5 in the Appendix; they are similar to our baseline results and reinforce our conclusion.

1.6.4 Discussion

Our results are consistent, irrespective of the estimation method or IEOp measure: we find no evidence that school finance reforms affected inequality of educational opportunity. This is in line with Lafortune et al. (2018), who similarly find no effect of the reforms on statewide achievement gaps between high- and low-income students or between black and white students. Although they *do* find negative effects on achievement gaps between high- and low-income *districts*, because low-income students are not highly concentrated in low-income districts, these *students* do not necessarily benefit from the reforms.

This also highlights that the choice of the dependent variable is crucial and that one has to be careful how to interpret the results. In particular, although the SFRs during the adequacy era have reduced the achievement gaps between high- and low-income districts, this does not mean that the reforms were successful in leveling the playing field for students—which was the reforms' intended purpose. It instead appears that by targeting

[22] After dropping reading IEOp, there are still missing values left (see Figure 1.3). We impute them by either assigning the value in the subsequent period (if available) or the value in the previous period (if no value in the subsequent period is observed).

[23] There are 25 SFRs in total, but we exclude Kentucky in the analysis. This leaves us with 24 treated states.

low-income districts, the reforms failed to benefit disadvantaged students, and thus also failed to reduce inequality of educational opportunity.

Our study is related to Biasi (forthcoming), who studies the effect of equalizing revenues across school districts on intergenerational mobility. Using a simulated instruments approach, she finds positive effects of equalization on mobility, in particular for low-income students. Her measure of mobility is the national income rank of individuals given their parents' income quantile, i.e. linking outcomes measured during adulthood of two generations. This may suggest that one reason for our null finding is that the effects of the SFRs appear only later in life. Our study is also related to Card and Payne (2002), who find that the reforms during the equity era reduced the SAT gap between families from different socioeconomic backgrounds. Another explanation of our null finding may thus be that the reforms during the equity era already picked the low hanging fruits, and that there was little room left for improvements via revenue equalization during the adequacy era.

Note that we study the effect of school finance *reforms*, not the effect of school finances per se. Although our findings suggest that the SFRs during the adequacy era were unsuccessful in increasing equality of educational opportunity, this does not imply that the same is true for school finances in general. Our findings are well compatible with the view that increasing school finances can be productive for increasing equality of opportunity (although our study is silent about whether it is). To shed light on the effect of school finances on IEOp, it would seem natural to use the court-ordered SFRs as an instrument for school finances. Indeed, Jackson et al. (2015) study the effect of reform–induced changes in public school spending on long-run adult outcomes along these lines. However, we restrain from applying this approach because the effects are likely to be heterogeneous across reforms—potentially even with different signs, as Hoxby (2001) points out—which would violate the monotonicity assumption. Although Jackson et al. (2015) control for the type of a reform (e.g. foundation plans), there might still be heterogeneity within the same reform types.[24]

1.7 Conclusion

Since the 1970s, numerous states in the U.S. have implemented school finance reforms with the intention to either equalize resource or outcomes across districts. In this study, we investigate whether the SFRs since the late 1980s were successful in increasing equality of educational opportunity—the explicit goal of these reforms. We construct a measure of inequality of educational opportunity from NAEP data and exploit the exogenous timing of court-ordered reforms in an event-study framework to identify the effect of the reforms

[24]Biasi (forthcoming) proposes a simulated-instruments approach that doesn't suffer from this drawback, but her approach is not suitable to instrument school finances directly and she therefore focuses on instrumenting revenue equalization instead.

on opportunities. Although the previous literature has found effects of these reforms on the achievement gap between low- and high-income districts (Lafortune et al., 2018) or on educational attainment (Rothstein and Schanzenbach, 2021), we find no evidence that the SFRs have affected IEOp. This result is robust to an array of robustness checks. Our finding highlights that analyzing the effects of school finance reforms is not trivial, and that the choice of the dependent variable is crucial. In particular, although the SFRs have reduced the achievement gap across districts, there is no evidence that they reduced gaps across groups.

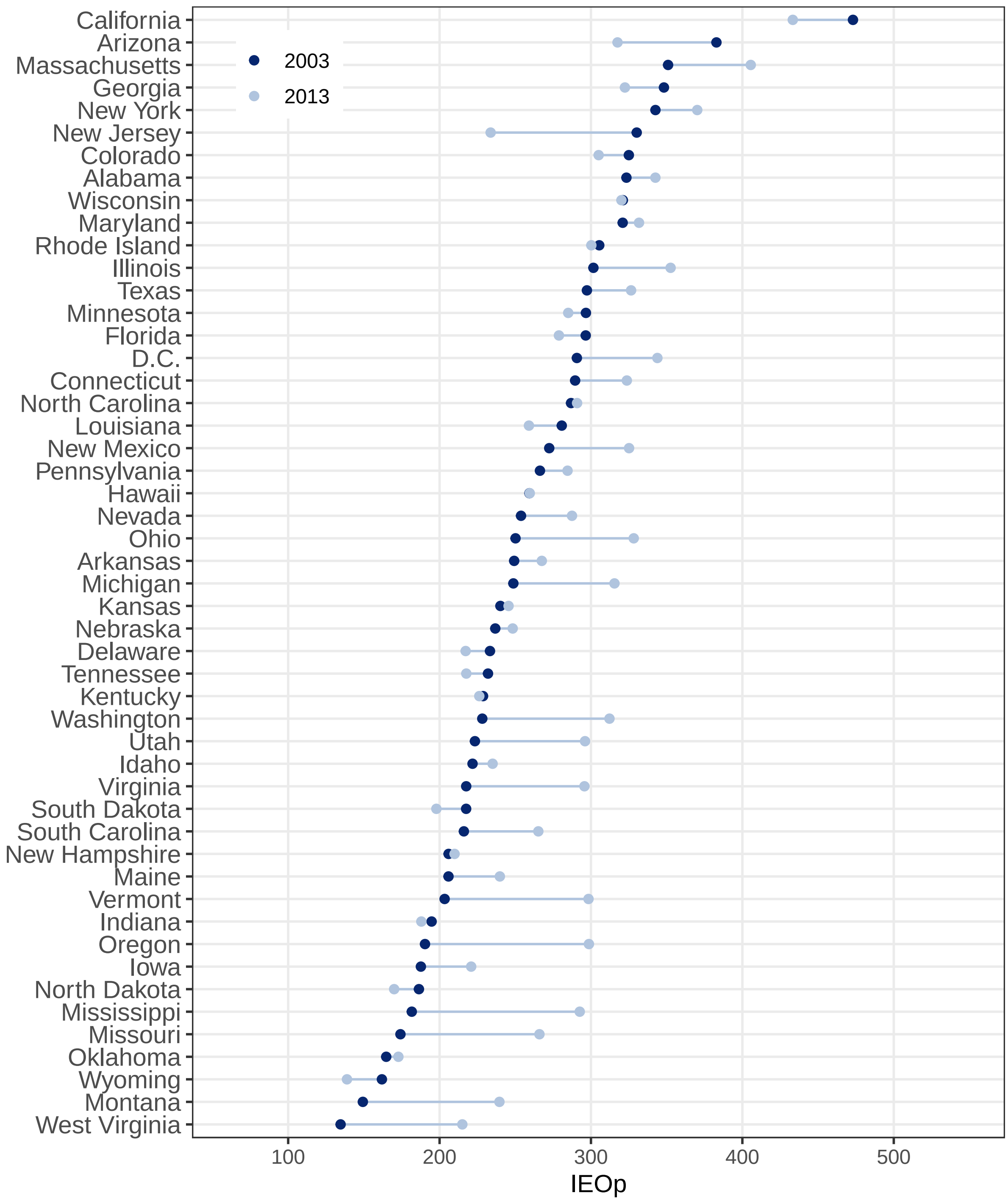
Note that our finding does not imply that the SFRs had no effect on other measures of equality of opportunity. It is well possible that the effects appear only later in students' lives, e.g. for the acquisition of income or wealth. Our measure of inequality of educational opportunity—constructed from achievement test scores when students are between 13 and 14 years old—may then be measured too early. Our finding does also not imply that SFRs are inherently unable to increase equality of educational opportunity. We only provide evidence that the court-ordered reforms since the late 1980s were unable to do so.

Our finding is important for policy-makers because it suggests that tackling inequality of educational opportunity via courts can be unproductive for leveling the playing field. At first glance, the reforms appear to have been successful because they were indeed able to equalize resources across districts. However, targeting *districts* is no guarantee that the policy reaches *individuals*. Given the economic costs of the court-ordered reforms, such as inefficient resource allocation, better-targeted policies may be preferable.

Appendix A

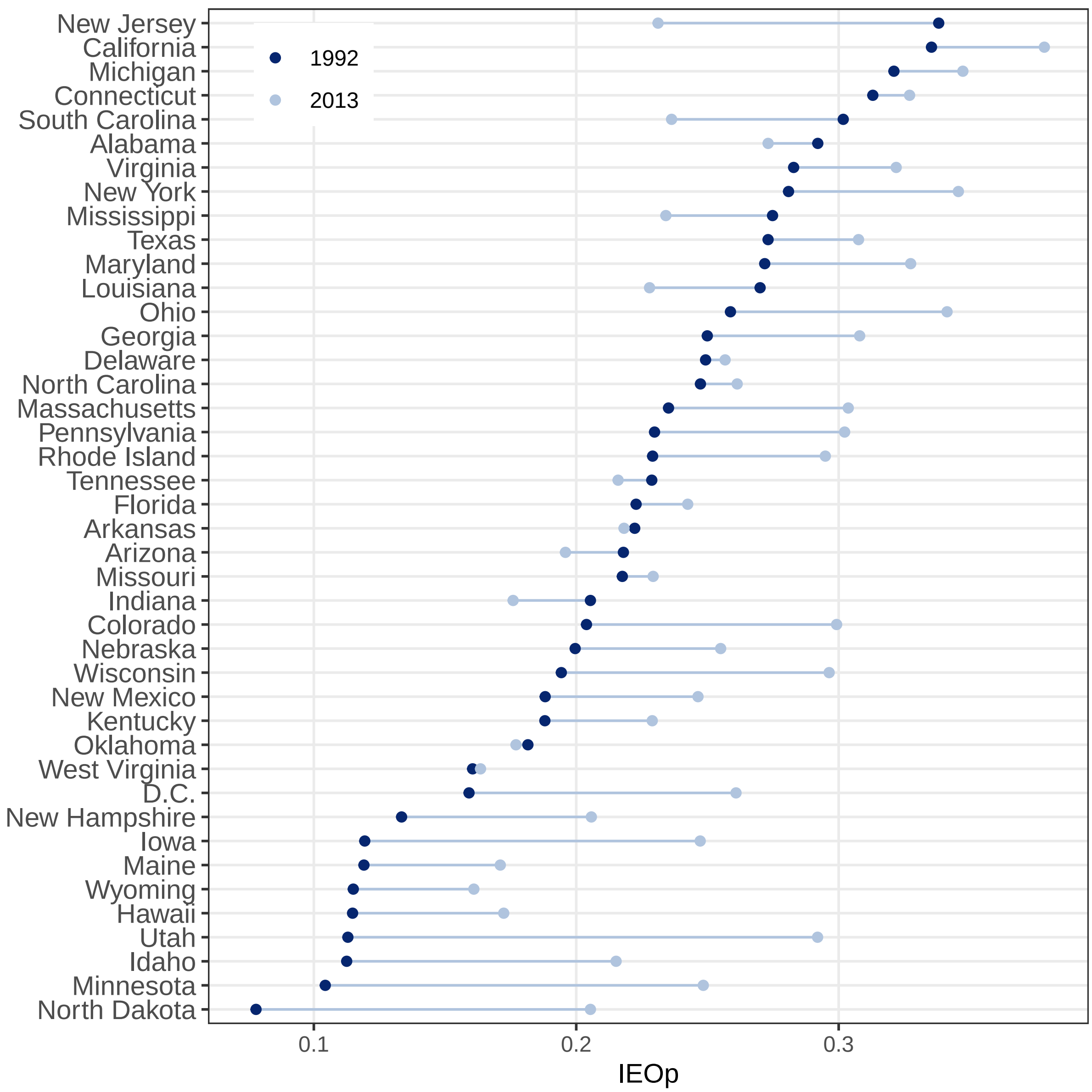
A.1 Data

Figure A.1: Inequality of Educational Opportunity: Reading



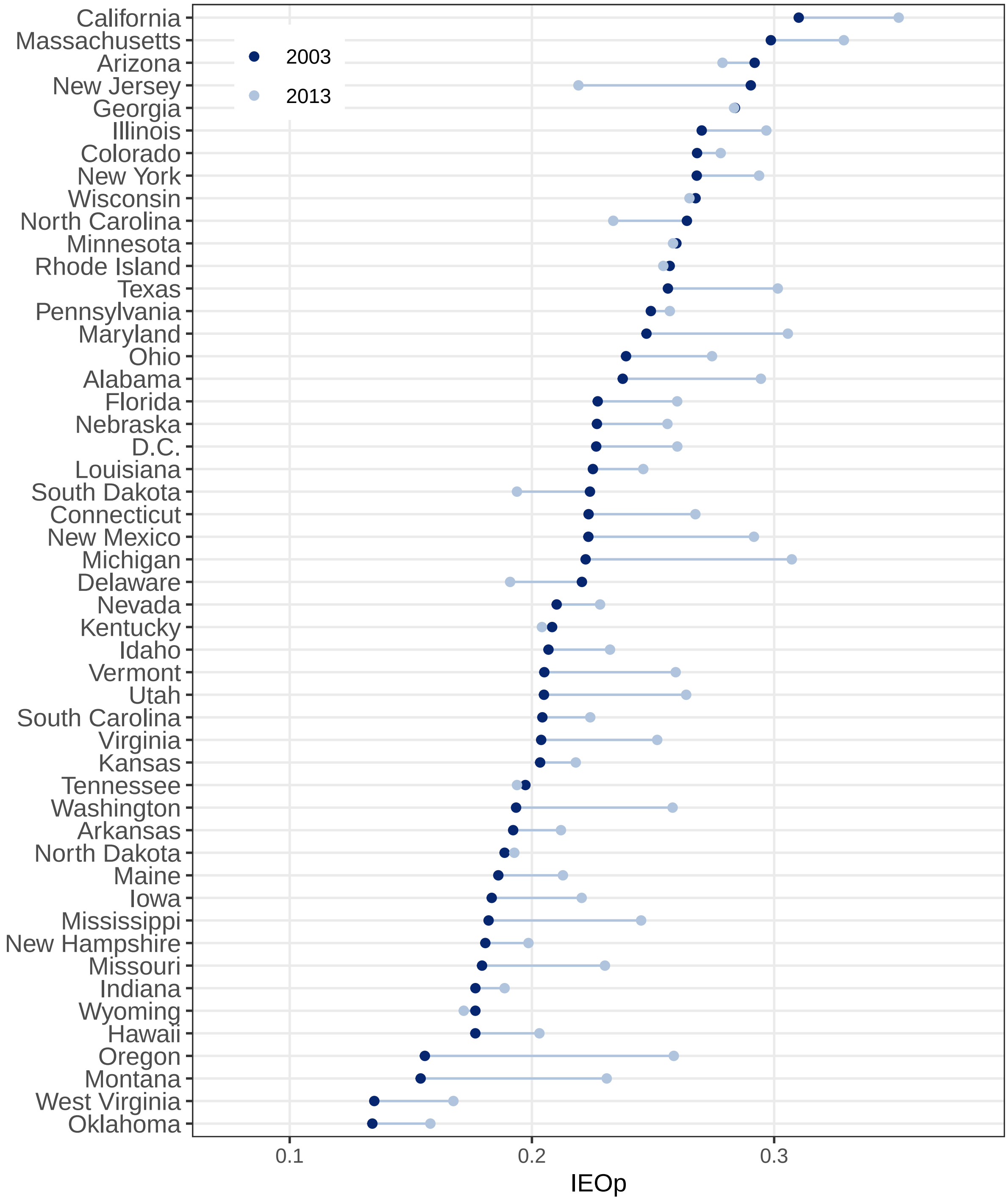
Note: The figure shows reading IEOp in 2003 and 2013, where IEOp is measured using conditional inference forests. The estimates refer to overall IEOp, given as the weighted variance of the counterfactual NAEP reading test score distribution. The included circumstance variables are gender, books at home, highest degree mother, highest degree father, ethnicity, and limited English proficiency. Data: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), Reading Assessment, Selected Years. Own calculations.

Figure A.2: Relative Inequality of Educational Opportunity: Math



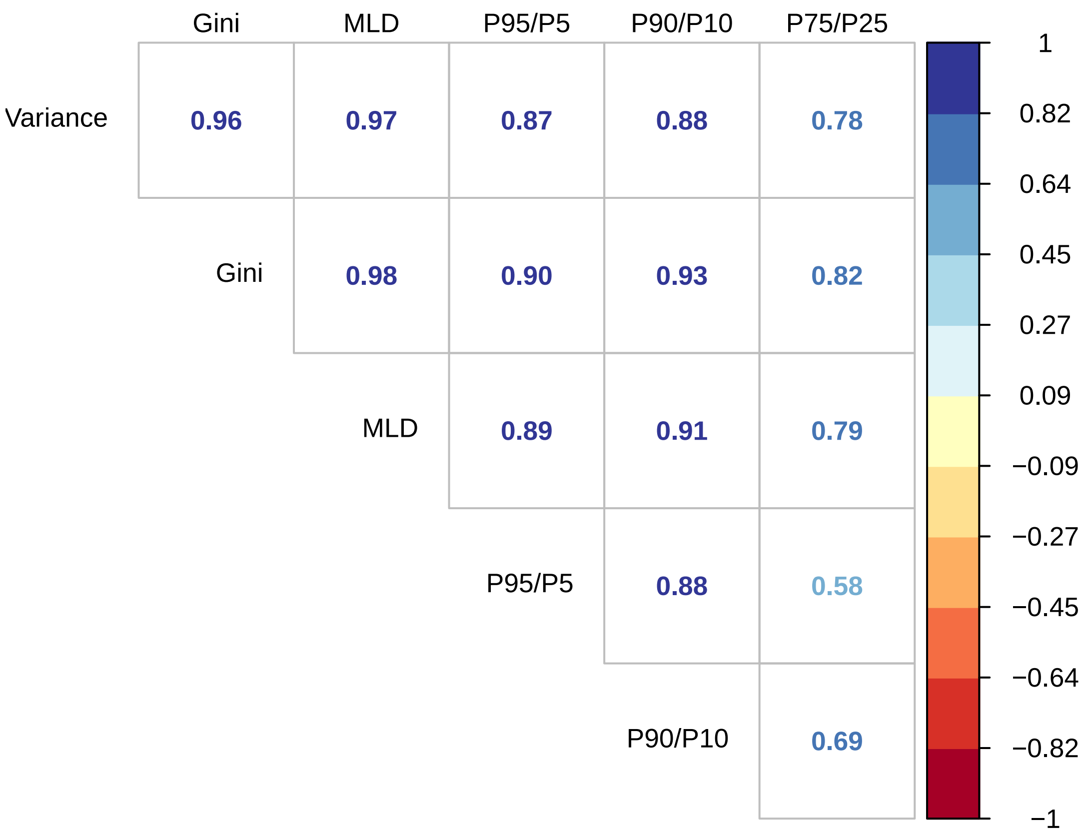
Note: The figure shows math IEOp in 1992 and 2013, where IEOp is measured using conditional inference forests. The estimates refer to relative IEOp, given as the ratio of the weighted variance of the counterfactual and the weighted variance of the original NAEP math test score distribution. The included circumstance variables are gender, books at home, highest degree mother, highest degree father, ethnicity, and limited English proficiency. Data: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), Mathematics Assessment, Selected Years. Own calculations.

Figure A.3: Relative Inequality of Educational Opportunity: Reading



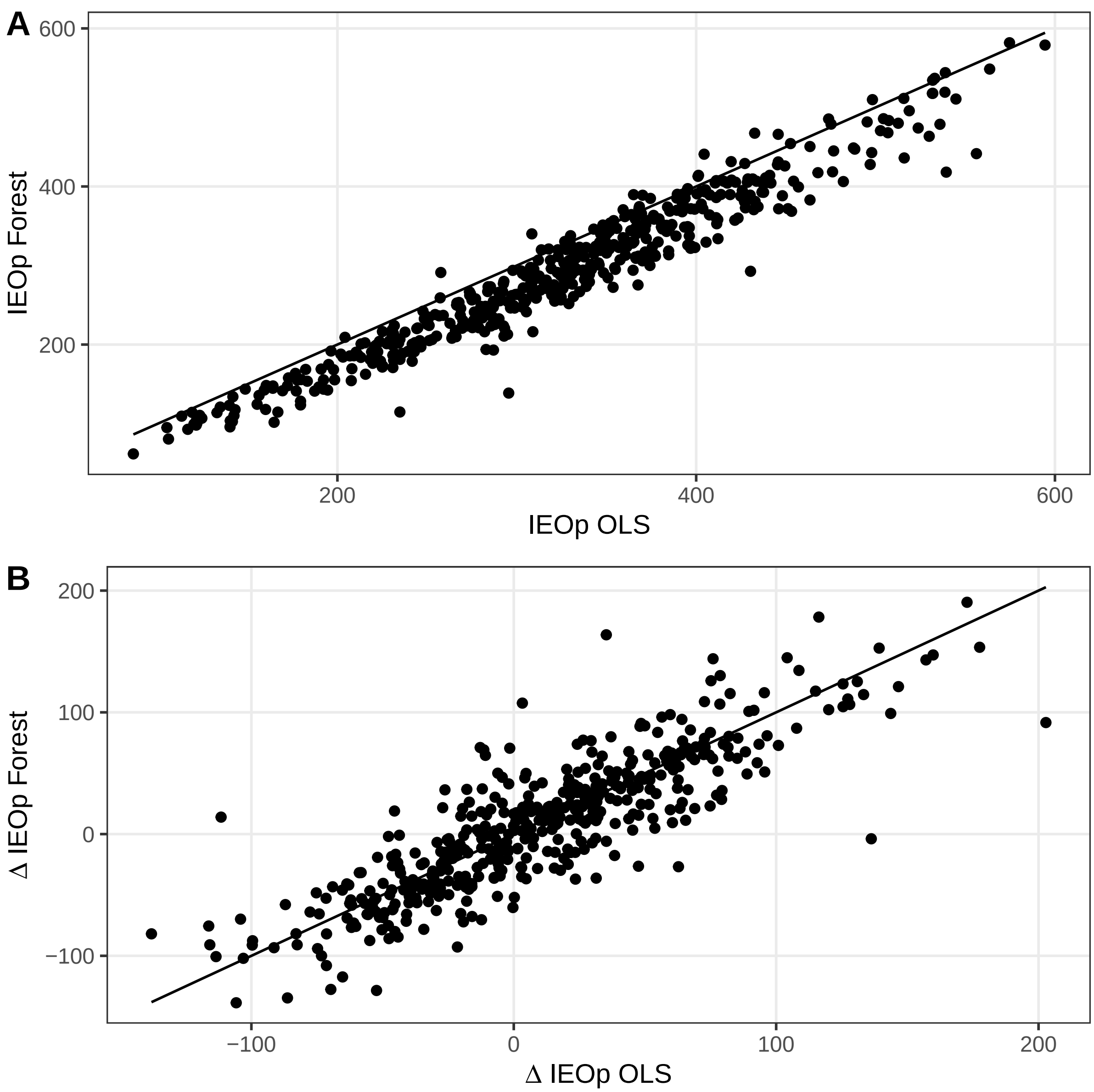
Note: The figure shows reading IEOp in 1992 and 2013, where IEOp is measured using conditional inference forests. The estimates refer to relative IEOp, given as the ratio of the weighted variance of the counterfactual and the weighted variance of the original NAEP reading test score distribution. The included circumstance variables are gender, books at home, highest degree mother, highest degree father, ethnicity, and limited English proficiency. Data: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), Reading Assessment, Selected Years. Own calculations.

Figure A.4: IEOp Correlation Matrix



Note: The figure shows the correlations between math IEOp given as the variance, Gini coefficient, mean log deviation (MLD), and three percentile ratios (P95/P5, P90/P10, and P75/P25) of the counterfactual NAEP test score distribution. IEOp is measured using conditional inference forests. The sample excludes Alaska, Kentucky, D.C. in 1996, New Jersey in 2015, Utah in 2015, and state-subject-year cells with less than 1,000 observations. Data: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), Mathematics Assessment, Selected Years. Own calculations.

Figure A.5: Conditional Inference Forest vs. OLS



Note: The figure plots math IEOP measured using conditional inference forests against math IEOp measured using OLS. The black line is the 45° line, i.e. points below this line indicate that IEOp is lower when using a conditional inference forest compared to OLS. Panel **A** shows math IEOp in levels, and Panel **B** shows the first differences. The sample excludes Alaska, Kentucky, D.C. in 1996, New Jersey in 2015, Utah in 2015, and state-subject-year cells with less than 1,000 observations. Data: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), Mathematics Assessment, Selected Years. Own calculations.

Table A.1: School Finance Reforms

State	Year	Event	Type
Alabama	1993	Alabama Coalition for Equity (ACE) v. Hunt; Harper v. Hunt	
Alaska	**1999**	**Kasayulie v. State of Alaska**	Court
Arizona	1994	Roosevelt v. Bishop	Court
	1997	Hull v. Albrecht	Court
	1998	**Hull v. Albrecht**	Court
	2007	Flores v. Arizona	
Arkansas	1994	Lake View v. Arkansas	Court
	1995	Approved Equitable School Finance Plan (Acts 917, 916, and 1194)	Bill
	2002	**Lake View v. Huckabee**	Court
	2005	Lake View v. Huckabee	Court
	2007	Various acts resulting from Master's Report findings	Bill
California	1998	Leroy F. Greene School Facilities Act of 1998	Bill
	2004	**Senate Bill 6, Senate Bill 550, Assembly Bill 1550, Assembly Bill 2727, and Assembly Bill 3001**	Bill
Colorado	**2000**	**Bill 181; Various other acts**	Bill
Connecticut	1995	Sheff v. O'Neill	
	2010	Coalition for justice in Education Funding, Inc. v. Rell	
Idaho	**1993**	**Idaho Schools for Equal Educational Opportunity v. Evans (ISEEO)**	Court
	1994	Senate Bill 1560	Bill
	1998	Idaho Schools for Equal Educational Opportunity v. State (ISEEO III)	
	2005	Idaho Schools for Equal Educational Opportunity v. Evans (ISEEO V)	Court
Indiana	**2011**	**HB 1001 (Pl229)**	Bill

Kansas	1992	The School District Finance and Quality Performance Act	Bill
	2005	**Montoy v. State; Montoy v. State funding increases**	Both
Kentucky	1989	Rose v. Council for Better Education, Inc.	Court
	1990	**Kentucky Education Reform Act (HB 940)**	Bill
Maryland	1996	Bradford v. Maryland State Board of Education	Court
	2002	**Bridge to Excellence in Public Schools Act (BTE) (Senate Bill 856)**	Bill
	2005	Bradford v. Maryland State Board of Education	
Massachusetts	1993	McDuffy v. Secretary of the Executive Office of Education; Massachusetts Education Reform Act	Both
Michigan	1997	Durant v. State of Michigan	
Missouri	**1993**	**Committee for Educational Equality v. State of Missouri; Outstanding Schools Act (S.B. 380)**	Both
	2005	Senate Bill 287	Bill
Montana	1993	House Bill 667	Bill
	2005	**Columbia Falls Elementary School v. State**	Court
	2007	M.C.A. §20-9-309	Bill
	2008	Montana Quality Education Coalition v. Montana	
New Hampshire	1993	Claremont New Hampshire v. Gregg	Court
	1997	Claremont School District v. Governor	Court
	1998	Opinion of the Justices—School Financing (Claremont III)	
	1999	Claremont v. Governor (Claremont III); RSA chapter 193-E	Both
	2000	Opinion of the Justices—School Financing (Claremont VI)	
	2002	Claremont School District v. Governor	Court
	2006	Londonderry School District v. New Hampshire	
	2008	**SB 539**	Bill
New Jersey	1990	The Quality Education Act; Abbot v. Burke	Both
	1991	Abbott v. Burke	
	1994	Abbott v. Burke	Court

	1996	Comprehensive Educational Improvement and Financing Act of 1996	Bill
	1997	Special Master's Report; Abbott v. Burke	Bill
	1998	**Abbott v. Burke**	Court
	2000	Abbott v. Burke	Court
	2008	The School Funding Reform Act of 2008	Bill
New Mexico	1998	Zuni School District v. State	
	1999	**Zuni School District v. State**	Court
	2001	Deficiencies Corrections Program; Public School Capital Outlay Act	Bill
New York	2003	Campaign for Fiscal Equity, Inc. v. State	Court
	2006	**Campaign for Fiscal Equity, Inc. v. State**	Court
	2007	Education Budget and Reform Act	Bill
North Carolina	**1997**	**Leandro v. State**	Court
	2004	Hoke County Board of Education v. State	Court
North Dakota	**2007**	**SB 2200**	Bill
Ohio	**1997**	**DeRolph v. Ohio**	Court
	2000	DeRolph v. Ohio; Increased school funding (see 93 Ohio St.3d 309)	Both
	2001	DeRolph v. Ohio	
	2002	DeRolph v. Ohio	Court
Oregon	2009	Pendleton School District 16R v. State	
South Carolina	2005	Abbeville County School District v. State	
Tennessee	1992	The Education Improvement Act	Bill
	1993	Tennessee Small School Systems v. McWherter	Court
	1995	**Tennessee Small School Systems v. McWherter**	Court
	2002	Tennessee Small School Systems v. McWherter	Court
Texas	1991	Edgewood Independent School District v. Kirby	Court
	1992	**Carrolton-Farmers Branch ISD v. Edgewood Independent School District**	Court
	1993	Senate Bill 7	Bill
	2004	West Orange-Cove ISD v. Nelson	
	2005	West Orange-Cove Consolidated ISD v. Neeley	

Vermont	1997	Brigham v. State	Court
	2003	**Revisions to Act 68; H.480**	Bill
Washington	1991	Seattle II	
	2007	Federal Way School District v. State	
	2010	**McCleary v. State**	Court
West Virginia	**1995**	**Tomblin v. Gainer**	Court
Wyoming	1995	Campbell County School District v. State	Court
	1997	The Wyoming Comprehensive Assessment System; The Education Resource Block Grant Model	Bill
	2001	**Campbell II; Recalibration of the MAP model**	Bill

Note: The table lists the school finance reforms in Lafortune et al. (2018). Each reform is either classified as court-ordered ("Court"), legislative ("Bill"), or both ("Both). Bold years indicate the single event per state selected by Lafortune et al. (2018).

A.2 Estimating Inequality of Opportunity from Regression Trees and Forests

A regression forest is based on the repeated computation of a regression tree. Intuitively, a conditional inference tree searches all available circumstance variables and tests their independence with the outcome variable (Hothorn et al., 2006). It then selects the variable for which independence is rejected with the highest statistical significance and splits the data in two parts along this variable. The algorithm then repeats this step with the two subgroups and selects new split variables. This process is repeated until the independence hypothesis cannot be rejected for any variable. In the end, each observation belongs to a *terminal node*—a group that is homogeneous in the expression of the circumstance variables. Terminal nodes are insofar analogous to the concept of a type in the framework Roemer (1998).

More precisely, a conditional inference tree applies the following algorithm:

1. Test the null hypothesis of independence for each input variable, and obtain a p-value associated with each test. Adjust the p-values for multiple hypothesis testing (Bonferroni correction).

2. Find the variable with the lowest adjusted p-value. If the adjusted p-value is larger than a specified significance level α, exit the algorithm. If it is lower or equal than α, continue and select this variable as the splitting variable.

3. Test the discrepancy between the subsamples for each possible binary partitioning based on the splitting variable, and obtain a p-value associated with each test. Split the sample by choosing the partitioning that yields the lowest p-value.

4. Repeat the algorithm for each of the resulting subsamples.

The counterfactual outcome can be obtained as the node-specific mean value of the outcome variable:

$$y_i^c = \hat{\mu}_{m(i)} = \frac{1}{N_m} \sum_{j \in g_m} y_j, \tag{A.1}$$

where $\sum_{j \in g_m} y_j$ is the sum of the outcomes of all individuals in terminal node g_m, and individual i belongs to that terminal node.

A conditional inference forest averages across B trees. To improve out-of-sample prediction performance, each tree draws a random sample of observations with the same sample size (with replacement) as well as P randomly selected input variables. Regression trees are sensitive to alternations in the data sample, and averaging the counterfactual outcomes computed on random samples of observations cushions the variance of their predictions.

Moreover, using a subset of the circumstance variables increases the likelihood that all observed circumstances with informational content will be used as splitting variable eventually.

We select B, P, and α according to their out-of-bag root mean squared error. The out-of-bag root mean squared error for α and P, $RMSE^{OBB}(\alpha, P)$, is given as:

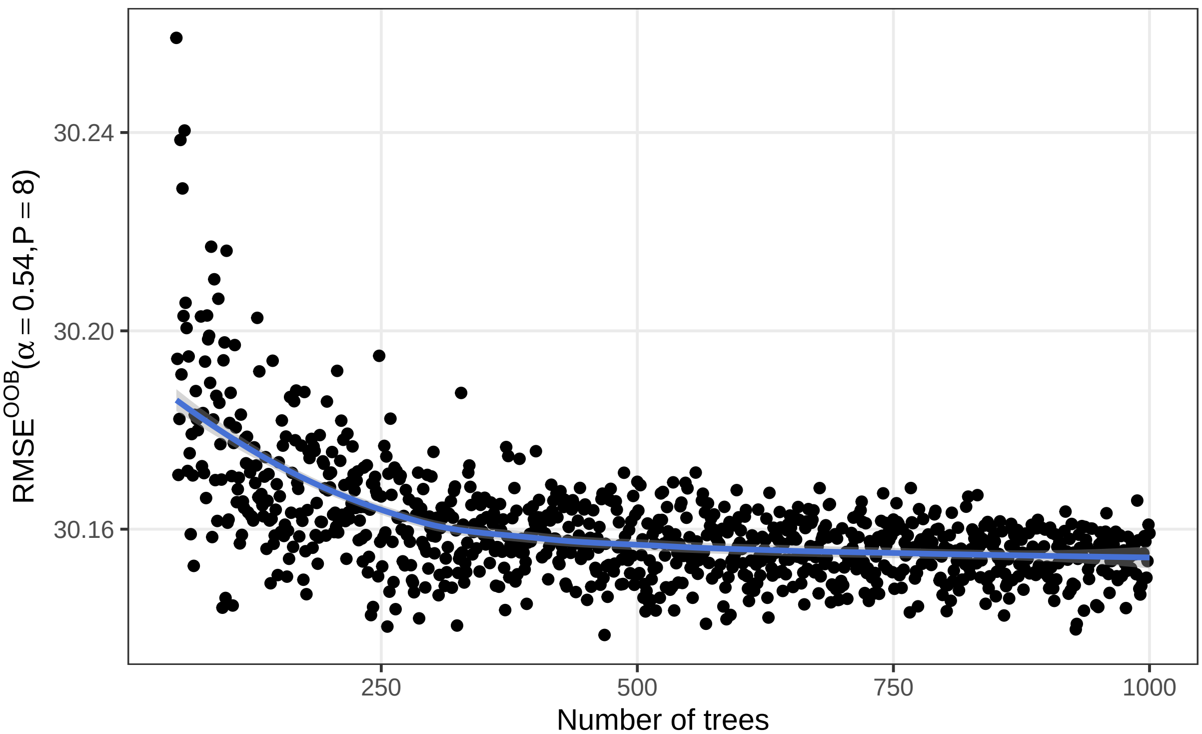
$$RMSE^{OBB}(\alpha, P) = \sqrt{\frac{1}{N}\sum_{i}^{N}(y_i - \widehat{f}^{OBB}(\Omega_i; \alpha, P))^2}, \tag{A.2}$$

where $\widehat{f}^{OBB}(\Omega_i; \alpha, P)$ is the average predicted value of observation i using each of the prediction functions estimated in subsamples in which i does not enter.

We fix $B = 400$ for all specifications because larger values yield no noticeable improvements in terms of reducing $RMSE^{OBB}$. As an illustrative example, Figure A.6 visualizes $RMSE^{OBB}$ as a function of B for math IEOp in California in 2015 (with $\alpha = 0.54$ and $P = 8$).

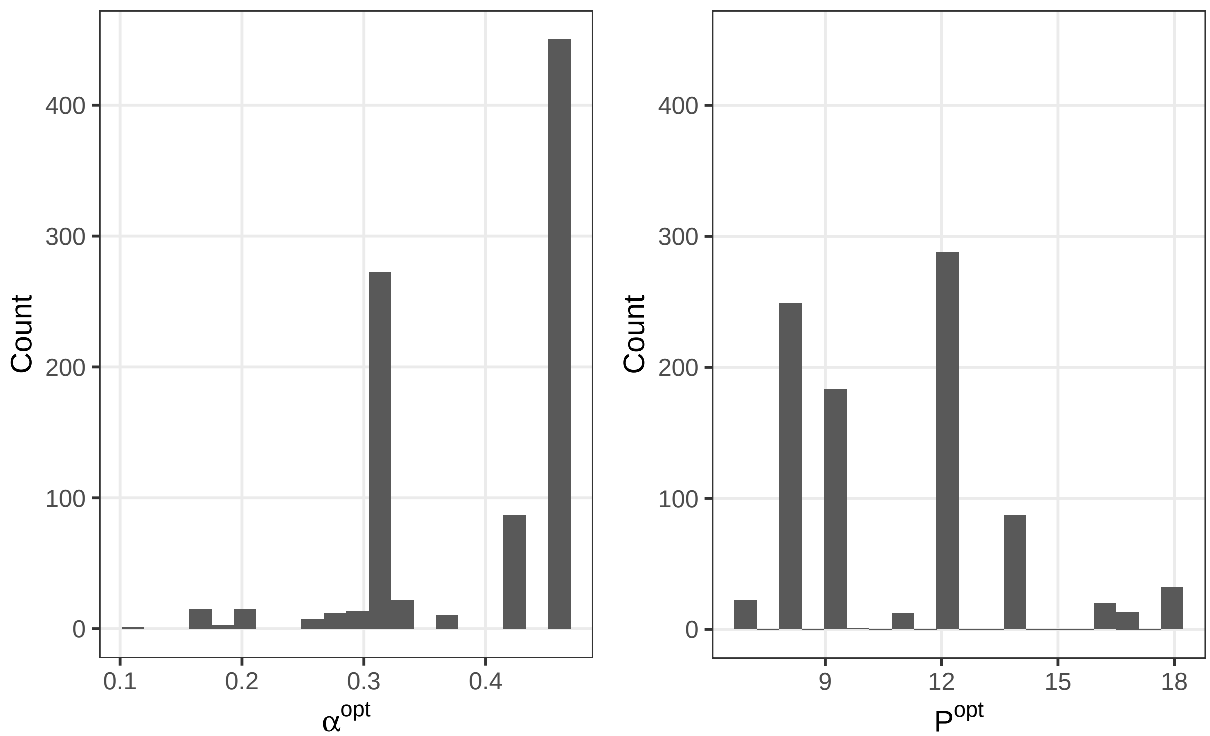
On the other hand, α and P are allowed to vary across state-subject-year cells. For each cell, we choose that α and P pair that minimizes $RMSE^{OBB}(\alpha, P)$. Figure A.7 shows the distributions of α^{opt} and P^{opt}, the optimal values that are chosen.

Figure A.6: Optimal Forest Size



Note: The figure shows $RMSE^{OBB}$ (y-axis) as a function of B (x-axis) for math NAEP test scores in California in 2015 (with $\alpha = 0.54$ and $P = 8$). Each point represents the root mean squared error for a particular B. The blue line is the fitted LOESS curve, and the grey area is the 95% confidence interval. Data: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), Mathematics Assessment, 2015. Own calculations.

Figure A.7: Optimal Parameters



Note: The figure shows the distributions of α^{opt} and P^{opt}, the optimal conditional inference forest parameters. For each state-subject-year cell, the optimal parameters minimize $RMSE^{OBB}(\alpha, P)$. The sample excludes Alaska, Utah in 2015, and state-subject-year cells with less than 1,000 observations in NAEP. Data: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), Mathematics and Reading Assessments, Selected Years. Own calculations.

A.3 Model Performance

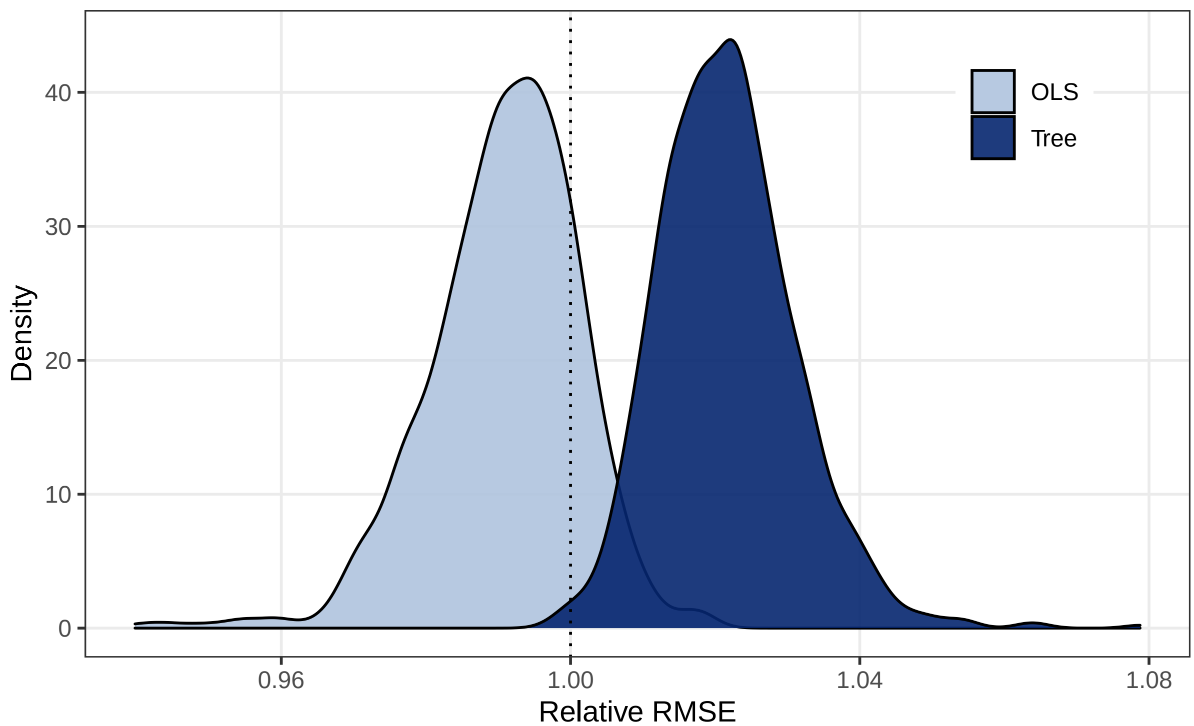
To assess the performance of the different approaches for computing IEOp, we compare their out-of-sample prediction accuracies. We follow the machine learning practice and, for each state-subject-year cell, split the data into a training and test set, with $i^{-H} \in \{1, \ldots, N^{-H}\}$ and $i^{H} \in \{1, \ldots, N^{H}\}$, respectively. For each cell with N observations, we randomly draw $N^{-H} = 2N/3$ observations for the training set, and $N^{H} = N/3$ observations for the test set. Each model is then estimated on the training set only, and prediction errors are calculated in the test set. Specifically, we first obtain the prediction function $\widehat{f}^{-H}$ by estimating the model using the training set. Second, we calculate the root mean squared error in the test sample for cell c and method m:

$$\text{RMSE}_{c,m} = \sqrt{\frac{1}{N_c^H} \sum_{i_c^H=1}^{N_c^H} (y_i - \widehat{f}_m^{-H}(\Omega_i))^2}, \tag{A.3}$$

where Ω_i denotes the circumstance vector of individual i. For each cell, we calculate relative RMSE by dividing $\text{RMSE}_{c,m}$ by the mean squared error of the conditional inference forests: $\text{RMSE}^{\text{rel}}_{c,m} = \text{RMSE}_{c,m}/\text{RMSE}_{c,m=forest}$. Values larger than one then denote a worse out-of-sample performance compared to forests, and vice versa.

Figure A.8 shows the distributions of $\text{RMSE}^{\text{rel}}_{c,m}$ for conditional inference trees and OLS. As expected, conditional inference forests clearly outperform trees. Yet, the performance of OLS is generally better. This suggests that our OLS model is already doing well at balancing the bias-variance trade-off, and that the benefit of using forest-based approaches is limited. However, the performance differences are small. In any case, the results of our study are virtually identical, irrespective of the approach.

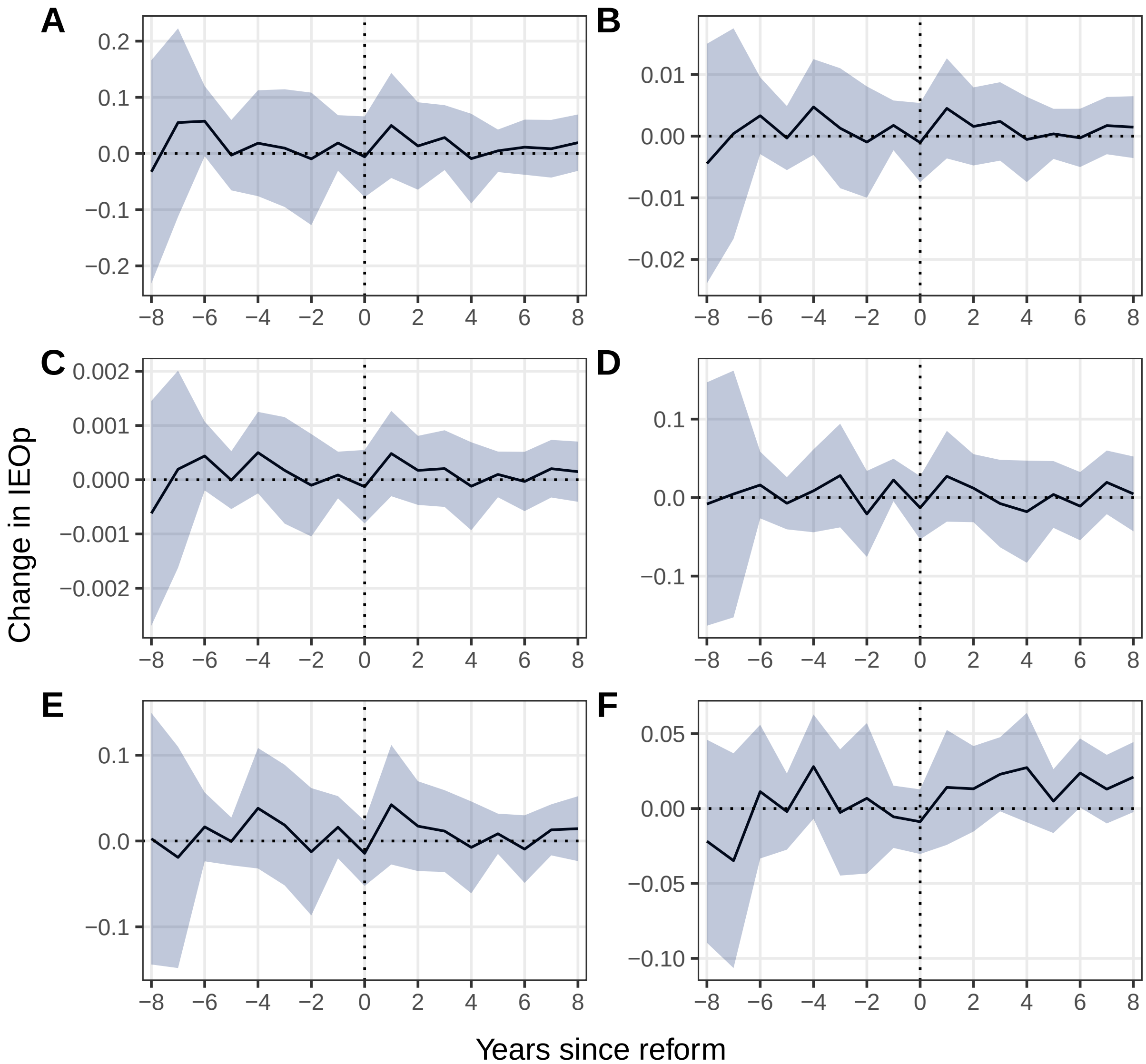
Figure A.8: Model Performance



Note: The figure shows the distribution of relative RMSEs for conditional inference trees and OLS, respectively, relative to conditional inference forests. Values larger than one denote a worse out-of-sample performance compared to forests, and vice versa. The sample excludes Alaska, Utah in 2015, and state-subject-year cells with less than 1,000 observations in NAEP. Data: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), Mathematics and Reading Assessments, Selected Years. Own calculations.

A.4 Further Robustness Checks

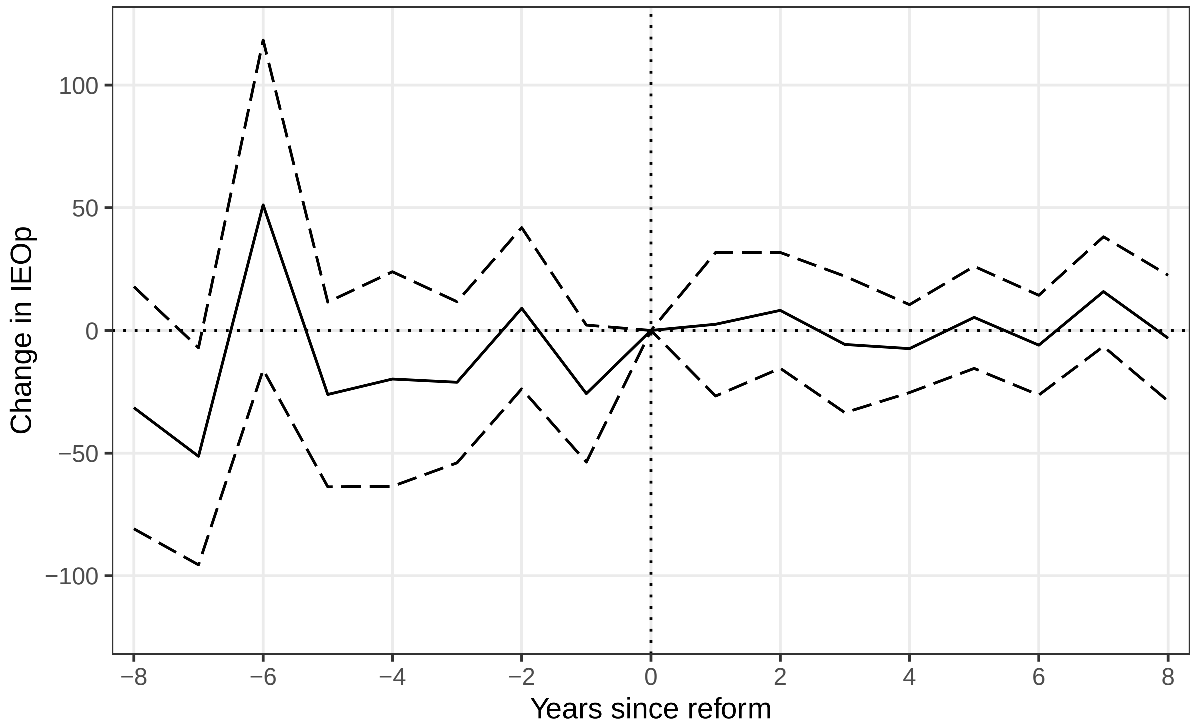
Figure A.9: Further Robustness Checks



Note: The figure reproduces Figure 1.5 but with different outcome measures. Panel **A** uses relative instead of absolute IEOp. The remaining panels apply different inequality measures $I()$ to the counterfactual NAEP score distribution: Gini index (panel **B**), mean log deviation (panel **C**), P95/P5 ratio (panel **D**), P90/P10 ratio (panel **E**), and P75/P25 ratio (panel **F**).

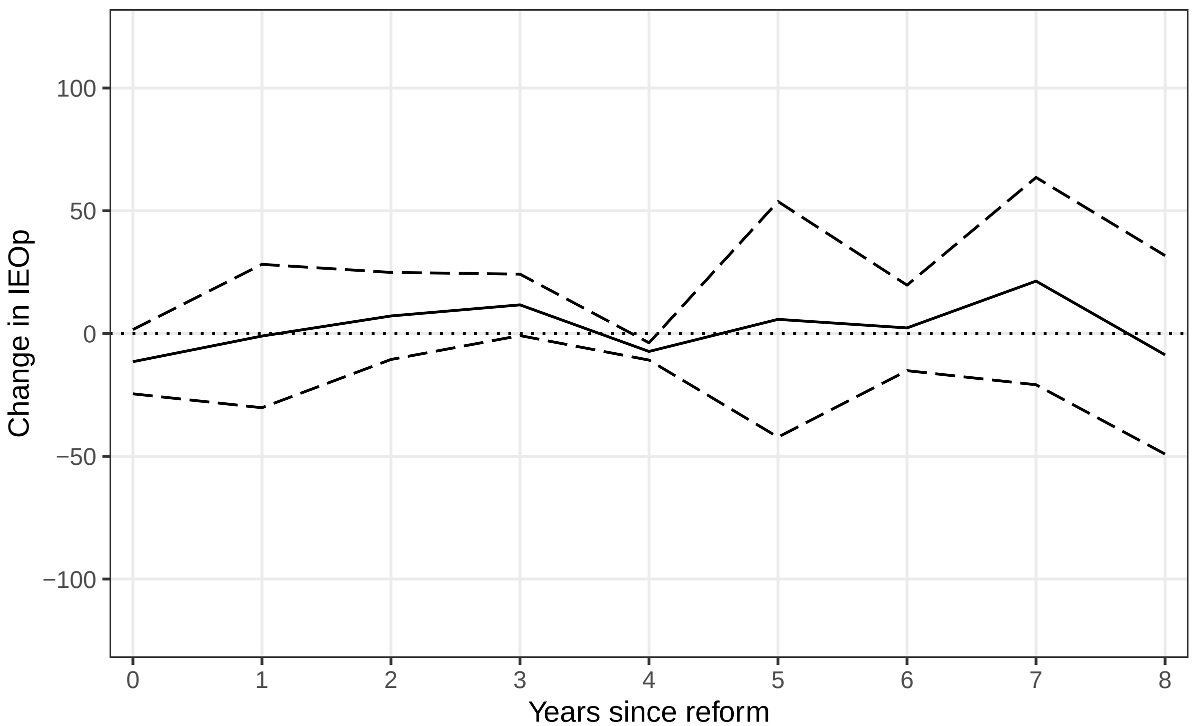
A.5 Alternative Estimators

Figure A.10: Sun and Abraham (2020): CATT Estimates



Note: The figure shows the event study plot for the cohort-specific average treatment effect on the treated, estimated with the method by Sun and Abraham (2020). The solid line represents the coefficients, and the dashed line represents the pointwise 95% confidence intervals using clustered standard errors at the state-subject level. The model is estimated on a pooled sample including both math and reading IEOp, and includes no control variables. The never-treated units are the control group. The lower and upper end of the time horizon are $k_{min} = -8$ and $k_{max} = 8$, respectively, and the lower and upper endpoint bins are $[-\infty, -9]$ and $[9, \infty]$, respectively. The figure excludes the estimates for β_{lower} and β_{upper}.

Figure A.11: Roth and Sant'Anna (2021): ATE Estimates



Note: The figure shows the event study plot for the average treatment effects, estimated with the method by Roth and Sant'Anna (2021). The solid line represents the coefficients, and the dashed line represents the pointwise 95% confidence intervals. The model is estimated on the math IEOp sample only, and includes no control variables. Missing values are imputed by either assigning them their value in the subsequent period (if available) or their value in the previous period (if no value in the subsequent period is observed).

Chapter 2

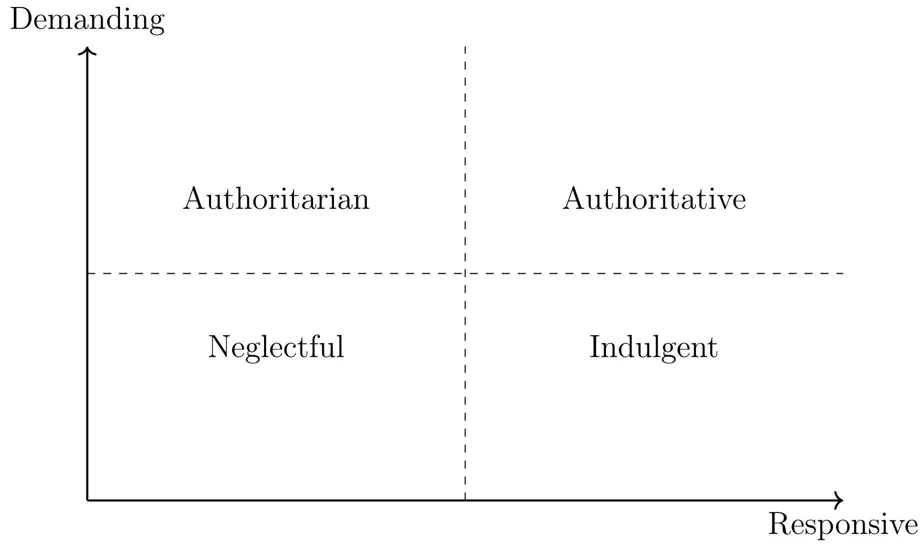
Parenting Styles and Child Skill Formation

2.1 Introduction

A large body of research shows that parenting styles are associated with a wide range of child outcomes, including subjective well-being, health, risky behavior, and school grades (Chan and Koo, 2011). To the extent that parenting styles are correlated with socioeconomic status, child-rearing practices insofar co-determine the strong link between parents' and children's outcomes. Indeed, Ermisch (2008) shows that differences in child-rearing can explain a large part of the income gradient in children's cognitive and non-cognitive development at age three. These skills, in turn, are crucial for children's success in life (Hanushek et al., forthcoming). However, although the association between parenting and child outcomes is well-documented, we know less about the relevance of parenting styles for the *dynamic process* of skill formation.

My paper addresses this gap and investigates whether parenting styles are relevant for the child skill formation process. I use data from the U.K. Millennium Cohort Study (MCS) to estimate child cognitive and non-cognitive skills production functions. The exceptionally rich data of the MCS allows me to construct measures of mothers' demandingness and responsiveness, the two parenting style dimensions highlighted by Baumrind (1991a) and Maccoby and Martin (1983). Demandingness describes parents' claims on the child in terms of maturity requirements, supervision, and discipline. Responsiveness refers to fostering individuality, self-regulation, and self-assertion of the child. The combination of the two dimensions yields the four best-known parenting styles: authoritarian (demanding but not responsive), authoritative (both demanding and responsive), indulgent (not demanding but responsive), and neglectful (neither demanding nor responsive). Figure 2.1 visualizes this typology.

Figure 2.1: Parenting Styles



Note: The figure shows the four parenting styles by Baumrind (1991a) and Maccoby and Martin (1983). Authoritarian parents are demanding but not responsive; authoritative parents are both demanding and responsive; indulgent parents are responsive but not demanding; neglectful parents are neither demanding nor responsive.

I use detailed information about parent-child interactions to measure demandingness and responsiveness by means of a principal component analysis. The factor that I interpret as demandingness largely loads on punitive parenting, i.e. the extent to which parents punish their child for misbehavior. The factor that I interpret as responsiveness loads on parent-child interactions that capture parental warmth and a positive mother-child relationship. I construct these measures for mothers at child ages 3, 5, 7, 11, and 14. Crucially, I obtain *time-varying* measures that allow me to estimate dynamic panel models.

The MCS assessed cognitive and non-cognitive skills at multiple stages during childhood. The former are measured with a battery of tests that capture different dimensions of cognitive ability. The latter are measured with the Strengths and Difficulties Questionnaire, a behavioral screening questionnaire designed to measure the psychological adjustment in children.

There are three main challenges in estimating the productivity of parenting styles (Todd and Wolpin, 2003). First, demandingness and responsiveness could be correlated with unobserved inputs, either past or present. This would be the case, for instance, if there were families who can rely on grandparents to raise their children (Deng and Tong, 2020). Low parental inputs are then simply due to the supply of grandparental inputs. The MCS contains information about inputs from various domains, including the extended family, the schooling environment, and the neighborhood. This allows me to control for inputs provided by grandparents, siblings, relatives, formal child care arrangements, and teachers.

To further rule out endogeneity issues due to unobserved inputs, I estimate a value-added model by controlling for the lagged outcome (Del Bono et al., 2016; Todd and Wolpin, 2007). That is, I regress cognitive or non-cognitive skills at child age a on skills at child age $a-1$. The lagged outcome serves as a sufficient statistic for unobserved inputs: if there still were unobserved inputs left, the coefficient of the lagged outcome would capture their effect on child skills.

Relatedly, unobserved innate ability may affect the learning rate of a child. The lagged outcome captures the past effects of innate ability on the skill formation process, but not the *contemporaneous* effect (Andrabi et al., 2011). I address this issue by estimating the model in first differences, i.e. regressing within-child changes in skills on changes in parenting styles.

The second challenge is measurement error in the skill measures. Measurement error in the lagged outcome leads to an attenuation bias, which may affect the estimates of my coefficients of interest (Andrabi et al., 2011). To address this issue, I use the double- or triple-lagged outcome as an instrument for the lagged outcome (Arellano and Bond, 1991).

The third challenge is feedback effects, i.e. that parents react to their child's observed outcomes and adjust their parenting style accordingly. If a child were to perform poorly in school, for instance, parents might become more involved in his or her life or become more demanding. To assess whether feedback effects are an issue, I employ dynamic panel estimators that use lagged inputs as instruments (Arellano and Bond, 1991; Blundell and Bond, 1998). The instruments can address feedback effects under the assumption that parents do not adjust their parenting style in anticipation of future child skill shocks (although adjustments to *past* shocks are allowed).

I find that parental demandingness is negatively associated with child development, both for cognitive and non-cognitive skills. Statistical significance is higher, and coefficient sizes are larger for non-cognitive than cognitive skills. Depending on the specification and age, a one-standard-deviation increase in the demandingness measure decreases cognitive skills by between 4% and 5% of a standard deviation. For comparison, this effect size is similar to increasing the class size in a school by one additional student (Fredriksson et al., 2013). On the other hand, parental responsiveness is positively associated with cognitive and non-cognitive skills. I also find a significant and positive interaction effect between demandingness and responsiveness for some specifications. This finding is consistent with the developmental psychology literature that highlights the benefits of authoritative parenting, i.e. high levels of parental control and high levels of parental warmth.

Using the dynamic panel estimators by Arellano and Bond (1991) and Blundell and Bond (1998)—difference GMM and system GMM, respectively—I find little evidence of feedback effects. This is crucial, as it suggests that my findings are not driven by reversed causality. My results are also robust to multiple robustness checks. In particular, they are essentially unchanged when using alternative skill measures: for cognitive skills, I use assessments that measure alternative skill domains (e.g. fluid ability instead of crystallized ability); for non-cognitive skills, I rely on teacher-assessed information instead of mother-assessed information.

When comparing the coefficients of parenting styles across child ages, I find that the negative association of demandingness with cognitive skills appears only during middle childhood, whereas for non-cognitive skills, the negative association is present throughout all developmental stages. In contrast, the positive association of responsiveness with skills is more pronounced during early childhood.

My paper contributes to a growing literature in economics on the importance of parents for child development (see Francesconi and Heckman, 2016, for an overview). A vast body of research on parental inputs focuses on monetary and time investments, but evidence on parenting styles is scarce. Del Bono et al. (2016) investigate the productivity of early and late parental time inputs and are closest to my study from an econometric point of view, but they treat parenting styles as a control rather than a variable of interest. Similarly, Dooley and Stewart (2007) and Khanam and Nghiem (2016) incorporate parenting styles in their analyses of the effect of family income on child outcomes, but rather because of a concern about omitted variable bias. On the other hand, Cobb-Clark et al. (2019), Deng and Tong (2020), Ermisch (2008), and Fiorini and Keane (2014) explicitly take parenting styles as an input in the skill production function into account.

The richness of the MCS allows me to go well beyond these previous analyses. First, neither Cobb-Clark et al. (2019), Deng and Tong (2020), nor Ermisch (2008) exploit panel data, which prohibits them from using time-varying parenting style measures. Yet, the panel structure is important to address unobserved innate ability and feedback effects. Although Fiorini and Keane (2014) *do* exploit panel data, they use only two waves with children between ages four and six. I instead observe both time-varying parenting style measures and child outcomes for five waves and child ages three to fourteen. This allows me to employ the Arellano-Bond and Blundell-Bond estimators, which require at least three time periods (and five periods to apply tests for serial correlation). The long time horizon also allows me to investigate whether the relevance of parenting styles varies across child ages.

Second, the MCS conducted multiple skill assessments during the early childhood, allowing me to take multiple ability domains into account. My cognitive skill measures also satisfy metric invariance, a necessary condition such that regression coefficients across waves can be meaningfully compared.

My findings have important policy implications. Because parenting matters for child development, equality of opportunity may be negatively affected by the increasing influence of parents on their children's lives (Doepke and Zilibotti, 2019). Targeting parenting practices with intervention programs may then be a suitable policy instrument to level the playing field. For example, programs such as "1-2-3 Magic" teach parents tactics to manage child behavior without arguing, yelling, or spanking—precisely the dimension of demandingness that I show to be detrimental for child development—and can be easily incorporated in home visits.[1]

The remainder of this paper is structured as follows. Section 2.2 gives a brief introduction to parenting styles, particularly to the typology used in developmental psychology. Section 2.3 discusses the theoretical framework, and section 2.4 discusses the empirical implementation. Section 2.5 describes the data, and the results are presented in section 2.6. Section 2.7 concludes.

2.2 Parenting Styles and Child Outcomes

Baumrind (1971, 1978, 1989) proposed the best known and most influential typology of parenting styles: authoritarian, authoritative, and permissive. *Authoritarian* parents are characterized by high levels of control and low levels of warmth. They demand obedience from their child, communicate through rules and orders, and employ harsh punishments. *Authoritative* parents similarly set rules and enforce boundaries, but they are also warm and responsive. They provide their child with autonomy and encourage independence. *Permissive* parents (either indulgent or neglecting) set few rules, and they are reluctant to enforce them (if there are any rules at all).

Maccoby and Martin (1983) highlight that this typology captures parenting styles as a function of two dimensions: demandingness and responsiveness (see also Darling and Steinberg, 1993; Spera, 2005). Demandingness describes parents' claims on the child in terms of maturity requirements, supervision, discipline, and confrontation if the child disobeys. Responsiveness refers to fostering individuality, self-regulation, and self-assertion of the child by being attuned and supportive to the child's needs and demands (Baumrind, 1991b). Baumrind's parenting styles emerge from the combination of the two dimensions: authoritarian parents are demanding but not responsive; authoritative parents are both demanding and responsive; indulgent parents are undemanding but responsive; and permissive parents are neither demanding nor responsive.

Early research already suggested that parenting styles matter for child outcomes. Baumrind (1967) compared preschool children from authoritative and non-authoritative households and found that the former were more mature, independent, and prosocial. Baumrind (1989) found similar effects for adolescents. Lamborn et al. (1991) and Steinberg et al. (1992) found positive effects of authoritative parenting for adolescents on school achievement, mental health, self-reliance, self-esteem, and antisocial behavior. More recent work, such as Chan and Koo (2011), also suggests that authoritative parenting is most beneficial for child outcomes. In economics, Ermisch (2008), Fiorini and Keane (2014), and Deng

[1] "1-2-3 Magic" divides the parenting responsibilities into three tasks: controlling negative behavior, encouraging good behavior, and strengthening the child-parent relationship.

and Tong (2020) similarly document the importance of parenting.

Although there is broad consensus that authoritative parenting is associated with the most favourable child outcomes—typically also thought to be a causal relationship, at least in the developmental psychology literature—there is less agreement as to why. The literature in psychology suggests that the effects of parenting are mediated by skill formation and the transmission of preferences. Durkin (1995), for example, suggests that authoritative parents provide their children with explanations for their actions, which fosters the transmission of values and goals.[2] These values and goals are, in turn, relevant to performing well in school. Moreover, the bi-directional communication typical of authoritative parents produces skills in interpersonal relations, which again is key for success in school and during adulthood. Steinberg (2001) similarly emphasizes skill formation. He argues that the combination of parental support and structure is beneficial for developing self-regulatory skills and that the verbal give-and-take fosters cognitive and social skills. Finally, Darling and Steinberg (1993) distinguish between three aspects to explain how parenting affects child outcomes: parental goals (the outcomes they want to achieve for their child), parenting practices (their child-rearing activities, similar to what the economics literature refers to as parental investment), and parenting style. In their framework, parenting styles can affect the productivity of parenting practices. That is, parenting styles moderate the relationship between parental investment and child outcomes. This conceptualization can take into account that parents with different parenting styles may not only differ in their investments, but also in the goals and values they have. The empirical relationship between parenting and child outcomes may thus result from particular parental goals (e.g. authoritative parents want their child to be successful), because they have higher investments, or because a particular parenting style increases the productivity of investments.

2.3 Framework

2.3.1 A Model of Child Skill Formation

I assume a model with T periods of childhood, with $t \in \{0, \ldots, T\}$. The childhood periods are divided into S stages of child development, with $s \in \{0, \ldots, S\}$ and $S \leq T$. During the developmental stages, cognitive (C) and non-cognitive skills (NC), $\theta_{k,t}$, with $k \in \{C, NC\}$, are produced. Examples of the former are IQ or crystallised ability. Examples of the latter are patience, self-control, temperament, or risk-aversion (Almlund et al., 2011). After childhood, adult outcomes are produced by the final skill levels, $\theta_{C,T+1}$ and $\theta_{NC,T+1}$.

Each individual is born with initial condition, $\theta_{k,0}$, that is influenced by family environments and genetic factors. The technology of skill production of skill k in period t

[2] Zumbuehl et al. (2021) provide recent evidence that parental involvement fosters the transmission of attitudes.

and developmental stage s depends on the stocks of skills in period t, $\theta_{k,t}$, parental inputs at t, $P_{k,t}$, innate ability, $\mu(G)$, shocks in period t, $\eta_{k,t}$, and the production function at developmental stage s:

$$\theta_{k,t+1} = f_{k,s}(\theta_{k,t}, P_{k,t}, \mu(G), \eta_{k,t}). \tag{2.1}$$

That is, the stock of skills at t, the child's innate ability, and inputs supplied by parents at t produce skills at period $t+1$. Innate ability is a function of the individual's genetic endowment, G. The inclusion of $\theta_{k,t}$ in Equation (2.1) yields what Cunha and Heckman (2007) refer to as self-productivity, i.e. that skills today produce skills tomorrow—skills beget skills.

Parental inputs consist of investments (both time and financial), $PI_{k,t}$, and parenting styles, $PS_{k,t}$: $P_{k,t} = [PI_{k,t}, PS_{k,t}]$. The subscript k indicates that parents may specifically target a skill of interest and tailor their inputs accordingly. Financial inputs include buying books for the child, whereas time inputs are e.g. reading a book to the child. Parenting styles, on the other hand, are the principles that parents follow in their child-rearing. They govern the parent-child relationship, e.g. how affectionate the father or mother is with his or her child. They also govern parent-child interactions: for every action of the child, the parenting style specifies the reaction of the parent. Crucially, they determine parental control. This includes, among others, behavioral and psychological control (Noack, 2011). Behavioral control refers to confronting a child who disobeys, monitoring him or her, or setting limits. Psychological control refers to appealing to guilt and expressing disappointment.[3]

Parental investment and parenting styles describe different concepts (Cobb-Clark et al., 2019; Deng and Tong, 2020). The former refer to resources that parents (consciously or not) spend on their children to increase their "quality" (or human capital). Parents may read to the child regularly or they may buy him or her colouring books. Because this input type is a finite resource, investments affect the parental time and financial budget. Assuming that households are optimizing agents, the supplied level of investment is a function of any factor that affects the budget as well (e.g. labour market participation). Parenting styles, on the other hand, may or may not affect the budget. It is not a finite resource, but rather *how* the investment is supplied.

I assume that parenting styles are time-varying but that there is a common component (or latent factor). That is, someone is e.g. an authoritative parent but can be more or less so as the child becomes older. Jones et al. (2018) present evidence for this assumption.

I also assume that in Equation (2.1), parenting styles directly affect skill formation, but that they additionally interact with parental investment. In particular, following the model

[3] In a recent study, Doepke and Zilibotti (2017) define permissive parenting as allowing the child to make free choices, i.e. low parental control. On the other hand, authoritative parents attempt to mold their children's preferences to induce choices that parents view as conducive to success in life, i.e. high psychological control. Authoritarian parents restrict children's choices, i.e. high behavioral control.

of Darling and Steinberg (1993), the productivity of investments is allowed to depend on parenting styles. The complementarity between parenting styles and investments, κ_{IS}, is given as:

$$\kappa_{IS} = \frac{\partial^2 f_{k,s}(.)}{\partial PI_{k,t} \partial PS_{k,t}}. \tag{2.2}$$

κ_{IS} can be positive or negative, depending on the particular parenting style and the investment type. For instance, reading to a child could be more productive when combined with a responsive parenting style (i.e. high parental warmth).

I further assume that different parenting style dimensions can interact with each other. Their complementarity, κ_{SS}, is given as:

$$\kappa_{SS} = \frac{\partial^2 f_{k,s}(.)}{\partial PS^D_{k,t} \partial PS^R_{k,t}}, \tag{2.3}$$

where $PS^D_{k,t}$ and $PS^R_{k,t}$ denote parental demandingness and responsiveness, respectively. $\kappa_{SS} > 0$ indicates that the combination of demandingness and responsiveness is particularly productive.

2.4 Empirical Implementation

2.4.1 Empirical Model

The technology of skill production in Equation (2.1) can be represented by the following regression analog (Todd and Wolpin, 2003):

$$\theta_{ia} = \boldsymbol{X}_{ia}\boldsymbol{\alpha}_1 + \boldsymbol{X}_{i,a-1}\boldsymbol{\alpha}_2 + \cdots + \boldsymbol{X}_{i1}\boldsymbol{\alpha}_a + \delta_a \mu_i + v_{ia}, \tag{2.4}$$

where θ_{ia} is a skill measure of child i at age a, $\boldsymbol{X}_{ia}$ is a vector of all inputs at age a that are relevant for skill formation, and $\boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_a}$ and δ_a are the coefficients (dropping the k subscript for simplicity).[4]

Equation (2.4) can't be empirically implemented unless the complete history of all relevant inputs is observed—which it never is. This is problematic because past inputs are likely to be correlated with contemporaneous inputs due to households optimizing their behavior. This implies that unobserved inputs (and unobserved innate ability, μ_i) may introduce endogeneity. Because of this, an alternative specification is usually implemented in practice. I opt for a value-added specification that includes the lagged outcome, $\theta_{i,a-1}$,

[4]Equation (2.4) requires the assumption that $f_{k,s}(.)$ in Equation (2.1) is constant over s, i.e. that the function is not age-varying.

on the right-hand-side:

$$\theta_{ia} = \boldsymbol{X}_{ia}\boldsymbol{\alpha}_1 + \gamma\theta_{i,a-1} + \zeta_{ia}. \tag{2.5}$$

The lagged outcome serves as a sufficient statistic for unobserved input histories and unobserved innate ability. That is, γ captures the effect of past inputs on θ_{ia}. Equation (2.5) can be derived from Equation (2.4) by subtracting $\gamma\theta_{i,a-1}$ from both sides of the latter. Rearranging the terms yields the following equation:

$$\begin{aligned}\theta_{ia} =& \boldsymbol{X}_{ia}\boldsymbol{\alpha}_1 + \gamma\theta_{i,a-1} + \boldsymbol{X}_{i,a-1}(\boldsymbol{\alpha}_2 - \gamma\boldsymbol{\alpha}_1) + \cdots + \boldsymbol{X}_{i1}(\boldsymbol{\alpha}_a - \gamma\boldsymbol{\alpha}_{a-1}) \\ &+ (\delta_a - \gamma\delta_{a-1})\mu_i + (\upsilon_{ia} - \gamma\upsilon_{i,a-1}).\end{aligned} \tag{2.6}$$

Equation (2.6) reduces to Equation (2.5) when the following requirements hold (Todd and Wolpin, 2007):

(i) For all l, $\boldsymbol{\alpha}_l = \gamma\boldsymbol{\alpha}_{l-1}$,

(ii) $\delta_a = \gamma\delta_{a-1}$, and

(iii) υ_{ia} is serially correlated at the rate γ.

Requirement (i) implies that input coefficients decline geometrically over time and that the rate of decline is the same for each input. Requirement (ii) further implies that the effect of innate ability declines geometrically at the same rate as the input effects. Requirement (iii) is necessary for $(\upsilon_{ia} - \gamma\upsilon_{i,a-1})$ to be an independently and identically distributed shock such that $\theta_{i,a-1}$ is uncorrelated with $\gamma\upsilon_{i,a-1}$. Because it is unclear whether these requirements hold in practice, I will explicitly control for lagged inputs and proxies for innate ability, and instrument the lagged outcome with the double-lagged outcome (Andrabi et al., 2011). This allows me to relax requirements (i)–(iii).

2.4.2 Cumulative Value-added Model

I implement Equation (2.5) with the following cumulative value-added model:

$$y_{ia} = \alpha_0 + \sum_{m=0}^{a-1} \boldsymbol{\beta}_{a-m}\boldsymbol{P}_{i,a-m} + \lambda y_{i,a-1} + \boldsymbol{\rho}\boldsymbol{W}_{ia} + \epsilon_{ia}, \tag{2.7}$$

with $\epsilon_{ia} = \mu_i + \varepsilon_{ia}$. y_{ia} is a skill measure of individual i at age a, $\boldsymbol{P}_{ia} = [\boldsymbol{PI}_{ia}, \boldsymbol{PS}_{ia}]$ is a vector of parental inputs—parental investment, $\boldsymbol{PI}_{ia}$, and parenting styles, $\boldsymbol{PS}_{ia}$—and $\boldsymbol{W}_{ia}$ is a vector of control variables (including inputs other than parental inputs). α_0,

$\boldsymbol{\beta}_{a-m}$, λ, and $\boldsymbol{\rho}$ are the coefficients. $\boldsymbol{\beta}_{a-m}$ are the parameter vectors of interest.[5]

$\boldsymbol{PI}_{ia}$ consists of measures of parental time investments such as playing with the child or helping him or her with homework. $\boldsymbol{PS}_{ia}$ consists of measures that capture different dimensions of parenting styles. In the benchmark specification, $\boldsymbol{PS}_{ia}$ are measures of parental demandingness and responsiveness.[6] I describe $\boldsymbol{PI}_{ia}$ and $\boldsymbol{PS}_{ia}$ in more detail in section 2.5.2.

To capture the intuition of Darling and Steinberg (1993) that parenting styles affect the productivity of parental investments, I additionally include the interaction between $\boldsymbol{PI}_{ia}$ and $\boldsymbol{PS}_{ia}$. If I find positive interaction terms, this will lend plausibility to the model of Darling and Steinberg (1993). I also include the interaction between parental demandingness and responsiveness.

λ represents the persistency of skills (or self-productivity in the language of Cunha and Heckman, 2007). Consistent estimation of the persistency coefficient is crucial such that Equation (2.6) reduces to Equation (2.5). Andrabi et al. (2011) point out that measurement error in the outcome variable may attenuate λ and also bias the coefficients of the input measures. I therefore instrument $y_{i,a-1}$ with the double-lagged outcome, $y_{i,a-2}$, in some specifications (see section 2.6.3).

2.5 Data

2.5.1 Millennium Cohort Study (MCS)

The MCS is an ongoing nationally representative longitudinal study of infants born in the U.K. between September 2000 and January 2002 (Plewis and Ketende, 2006). In the first wave, information on 18,818 infants from 18,533 families was collected. The sample was later augmented by 701 children born in the relevant time period who had been previously missed. The MCS followed the cohort members through their childhood and adolescence, collecting information from the cohort members directly, their resident parents, their older siblings, and their class teachers. Throughout seven waves so far, the survey involved home visits by interviewers at child ages nine months and years 3,

[5] Note that the subscripts a of the contemporaneous inputs and the outcome in Equation (2.7) are concurring. This is nonetheless consistent with Equation (2.1), i.e. that inputs and skill levels in period t produce skills in period $t+1$, because contemporaneous inputs (e.g. $\boldsymbol{PI}_{ia}$ or $\boldsymbol{PS}_{ia}$) proxy inputs over a longer time period, whereas y_{ia} measures skills exactly at t.

[6] There is no clear cut between parental investment and parenting style. Indeed, some parenting practices are characterized precisely by high investment, e.g. tiger moms or helicopter parents. In this sense, investment is not conceptually different, but rather a mediator of parenting styles. By controlling for investment, the coefficient estimates of the parenting style measures represents their effect on child skill formation net of the investment channel.

5, 7, 11, 14, and 17. Interviewers asked questions about, among others, socio-economic circumstances, demographics, parenting, the household environment, childcare, schooling, and child development. The interviews were conducted with the cohort members' parents in waves one to six and with the cohort members themselves in wave seven. Cohort members' cognitive and non-cognitive skills were assessed at child ages 3, 5, 7, 11, 14, and 17.

I restrict the sample to households where the natural mother was the main respondent in all waves and where she remains the resident mother throughout. This is important because some of the information about the child is obtained from the mother, and changes in the source of information could introduce non-random measurement error. This implies that my sample contains no single fathers but may (and does) contain single mothers. I further exclude twins because they might act as confounders. Parents tend to treat twins differently from singleton siblings (Bharadwaj et al., 2018), and being a twin may affect skill formation. I also only include children whose mothers were 18–45 years old at the child's birth. Again, teenage mothers or mothers significantly older than the average may act as confounders.

For the estimation sample, I additionally impose the requirement that the mother-child pair is observed for child ages 3, 5, 7, 11, and 14, and that there are no missing values for any of the control variables. This prevents that compositional changes drive the results. My final estimation sample includes 2,767 mother-child pairs.[7]

2.5.2 Parental Input Measures

Parenting styles I follow Baumrind (1991a) and Maccoby and Martin (1983) and construct measures that refer to parental demandingness and responsiveness. The MCS collects detailed information about parent-child interactions that I combine using a principal component analysis (PCA). PCA is a popular method for dimensionality reduction where each data point is projected onto only a few principal components.[8] This allows lowering the dimensionality while preserving as much variation as possible. For instance, the first principal component of a set of p variables, $\mathbf{x}_1, \ldots, \mathbf{x}_p$, is the linear combination of the variables $\mathbf{z}_1 = \phi_{11}\mathbf{x}_1 + \phi_{21}\mathbf{x}_2 + \cdots + \phi_{p1}\mathbf{x}_p$ that has the largest sample variance. $\phi_{11}, \ldots, \phi_{p1}$ are the eigenvectors of the first principal component, and $\mathbf{z}_1 = [z_{11}, \ldots, z_{n1}]$ are the scores for n observations.[9] The principal components are typically interpreted as a representation of an unobserved factor (parenting styles in my case). Instead of including all variables

[7]Estimating the model on a larger sample yields similar results.

[8]Cobb-Clark et al. (2019), Deng and Tong (2020), and Fiorini and Keane (2014) also use a PCA to construct their parenting style measures.

[9]The eigenvectors of the first principal component, i.e. the $\phi_{11}, \ldots, \phi_{p1}$ that maximize the sample variance, are obtained by an eigen decomposition. After the first principal component has been determined, the second (or third, etc.) principal component is the linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_p$ that maximizes the sample variance out of all linear combinations that is uncorrelated with $\mathbf{z}_1$.

individually in a regression, in a PCA regression, one only includes the PCA scores.

To capture demandingness, I include information about punitive behavior, rules-setting, and, for older children, independence. To capture responsiveness, I include information about parental warmth and a positive mother-child relationship. First, I use information from the Straus's Conflict Tactics Scales (CTS; Straus and Hamby, 1997). The CTS measures the tactics or behaviors used by parents when there is conflict or hostility toward a child. There are separate scales for reasoning, non-violent discipline, psychological aggression, and physical assault. In the MCS, the mother is asked about her discipline practices, i.e. how often she punishes her child in a particular way when he or she is naughty. Items include, among others, reasoning with the child (representing the *reasoning* scale in the CTS), ignoring the child (*non-violent discipline*), shouting at the child (*psychological aggression*), and smacking the child (*physical assault*). The items can be answered on a five-point Likert scale ranging from "never" to "daily". The CTS items are available for the second, third, fourth, and fifth wave of the MCS (i.e. child ages 3, 5, 7, and 11), although not all items are asked consistently.[10] In wave six, the discipline practices are assessed by the child directly instead of the mother. He or she is asked whether the parents ground him or her, tell him or her off, and punish him or her in some other way.

Second, I use the Child-Parent Relationship Scale (CPRS) in wave two (child age 3). The CPRS is a mother-assessed report of the child's relationship with the mother. The items include the mother's feelings and beliefs about her relationship with the child and the child's behaviour toward her. The CPRS in the MCS is a 15-item self-administered rating scale with responses on a five-point Likert scale.[11] Items include whether the mother shares an affectionate relationship with her child, whether the child seeks comfort from the mother, and whether the child spontaneously shares information with her. For waves three to six, there is less information available about the mother-parent relationship. The mother is either asked whether she shares an affectionate relationship with her child or how close they are, and whether she listens to her child. In wave six, the information is again assessed by the child instead of the mother.

Third, where available, I use information about whether the parents set rules about bedtime, whether the child has meals at regular times, and whether there are rules about watching TV or playing games on the computer. The first two items are answered on a four-point Likert scale, the questions about TV and computer are answered with yes or no. In wave four (child age 7), I additionally include whether the child has to do household chores, and in wave six (age 14) whether the parents are informed about the whereabouts of their child when he or she is not at home (a proxy for the level of independence).

I conduct a PCA separately for each wave and apply an oblique rotation to the factor loadings. Table B.2 in the Appendix shows an overview of the information I use in each

[10] I use a subset of all available items of the CTS. The items are chosen such that they are able to capture the demandingness dimension.

[11] Similar to the CTS, I only use a subset of the 15 available items of the CPRS.

wave. Table 2.1 shows the rotated loadings for waves two to six. To conduct the PCA, I transform the Likert scales into numeric scales by assigning a value of 1 to the lowest and a value of 5 (4) for the highest values of the five-point (four-point) Likert scales.

In the second wave (child age 3), the first principal component almost exclusively loads on the items from the CTS (i.e. disciplinary practices), whereas the second principal component almost exclusively loads on the items from the CPRS (i.e. the mother-child relationship). I interpret the former as a measure of demandingness and the latter as a measure of responsiveness. In the third wave (child age 5), the first principal component similarly loads on the CTS items, and I also interpret it as a measure of demandingness. The second principal component loads on the items about rules at home (not shown in Table 2.1), and the third on having a warm relationship with the child, reasoning with the child, and making sure he or she obeys. I use the third principal component as my measure of responsiveness. I use the first (second) principal component in wave four, the first (third) in wave five, and the third (second) in wave six as my measures for demandingness (responsiveness). In all cases, demandingness is similarly associated with the disciplinary practice and responsiveness with a positive relationship with the mother.

Note that wave six is an outlier to some extent. First, for consistency reasons, I do not use the first principal component because it heavily loads on the measures of independence (knowledge about whereabouts). Second, the information is assessed by the child, not by the mother. Although there is a subset of items where both the child and the mother were assessed in wave six, I only use child-assessed information for consistency. The factors obtained from wave six are thus less comparable to the factors from waves two to five. Child- and mother-assessed information is nonetheless correlated: the correlation between the items measuring a positive relationship reported by the child and the mother is 0.34, and the correlation between the items measuring knowledge about whereabouts is 0.38.[12]

Table 2.1: Parenting Styles Factor Loadings

	Demandingness	Responsiveness
Age 3		
Warm relationship with child	−0.00	0.49
Child seeks comfort	0.07	0.53
Child values relationship	−0.01	0.61
Child beams with pride when praised	−0.01	0.64
Child shares information	0.03	0.64
In tune with child	−0.10	0.60
Child shares feelings	−0.00	0.60
Ignores child	0.52	−0.03
Smacks child	0.54	−0.06
Shouts at child	0.71	−0.02

[12]In both cases, I compare the mean responses of the child with the mean responses of the mother.

Sends child to room	0.62	0.06
Takes away treats from child	0.63	0.06
Tells child off	0.73	0.05
Bribes child	0.45	−0.04
Rules about bed times	0.03	0.20
Rules about eating	−0.03	0.17
Age 5		
Warm relationship with child	−0.20	0.69
Ignores child	0.50	−0.14
Smacks child	0.50	−0.28
Shouts at child	0.70	−0.04
Sends child to room	0.66	−0.00
Takes away treats from child	0.64	0.08
Tells child off	0.75	0.19
Bribes child	0.42	−0.02
Reasons with child	0.51	0.46
Makes sure child obeys	0.09	0.62
Rules about bed times	0.07	0.02
Rules about eating	0.01	0.01
Age 7		
How close to child	−0.17	0.57
Listens to child	−0.15	0.68
Warm relationship with child	−0.04	0.69
Ignores child	0.50	−0.07
Smacks child	0.47	−0.15
Shouts at child	0.70	−0.07
Sends child to room	0.69	0.11
Takes away treats from child	0.68	0.15
Tells child off	0.77	0.06
Bribes child	0.42	−0.14
Reasons with child	0.58	0.16
Rules about bed times	0.01	0.28
Rules about TV times	0.11	0.39
Rules about TV hours	0.06	0.35
Child has chores	0.01	0.28
Age 11		
How close to child	−0.14	0.73
Talks to child	0.13	0.80
Sends child to room	0.84	−0.09
Takes away treats from child	0.86	−0.04
Reasons with child	0.78	0.11

Rules about bed times	−0.10	0.03
Rules about TV times	0.08	−0.06
Rules about TV shows	0.04	0.03
Age 14		
How close to mother	−0.10	0.82
How close to father	−0.11	0.82
Keeps child indoor	0.81	0.02
Tells child off	−0.02	−0.16
Punish child in other way	0.56	−0.14
Parents know where after 9pm	0.10	0.13
Parents know where overnight	0.21	0.17
Parents know where when out	−0.10	−0.13
Parents know with whom when out	−0.11	−0.17
Parents know what when out	−0.09	−0.16

Note: The table shows the rotated factor loadings for waves two to six (child ages 3, 5, 7, 11, and 14, respectively). For child ages 3, 5, 7, and 11, demandingness refers to the first principal component; for child age 14, it refers to the third principal component. For child ages 3, 7, and 14, responsiveness refers to the second principal component; for child ages 5 and 11, it refers to the third principal component. Data: Millennium Cohort Study. Own calculations.

Wave six aside, the factor loadings are largely stable across waves (see e.g. sending the child to the room or shouting at him or her). This is important because it ensures that the variation in the parenting style measures across child ages is driven by differences in parenting behavior and not by differences in the factor loadings.

Also, note that rules-setting does not load on the demandingness factor at any child age; it instead loads on the third principal component in most waves (not shown in Table 2.1). My measure of demandingness thus predominantly captures the punitive parenting aspect of demandingness and not the firm-rules aspect.[13]

To check whether the demandingness and responsiveness factors indeed measure parenting styles, I compare them with self-assessed information from wave two (child age 3). Parents were asked to choose one of the following options that best describes their own parenting style: i.) "doing my best for the children", ii.) "firm rules and discipline", iii.) "firm discipline plus lots of fun", iv.) "lots of fun", and v.) "have not really thought about it". Table B.5 in the Appendix reports the child-age-specific mean values of demandingness and responsiveness. "Firm rules and discipline", i.e. authoritarian parenting, is indeed associated with the highest values of demandingness and low values of responsiveness. "Firm discipline plus lots of fun", presumably authoritative parenting, is associated with above-average demandingness and high values of responsiveness. "Lots of fun", i.e. indulgent parenting, is associated with the lowest values of demandingness. Self-assessed parenting styles from wave two and demandingness/responsiveness in wave six are largely

[13]My empirical results are virtually unchanged when I include the third principal component in my analysis (not presented).

uncorrelated, likely because the former is assessed by the parents and the latter by the children.

Table B.6 in the Appendix also reports correlations between parenting styles of a parent at child age a and $a-1$. I interpret this correlation as a measure of "persistency" of demandingness and responsiveness from one child age to the next. Demandingness is highly persistent across child ages 3 and 11, with correlation coefficients between 0.55 and 0.65. Persistency of responsiveness is lower but still sizeable, ranging from 0.17 to 0.32 across ages 3 and 11. Persistency of demandingness and responsiveness from age 11 to 14 is lower in both cases (0.21 and 0.10, respectively). Again, this is likely because parenting styles are measured using information obtained from the child at age 14 but from the mother at ages 3 to 11.

Parental investment I follow Del Bono et al. (2016) and construct an index of parental time investment from self-assessed information about the type and frequency of parent-child activities. From wave three onward, I additionally include information about parental school involvement: whether parents are attending the parents' evenings, whether they arrange special meetings with the teacher, how much effort they put in getting the child into their school of choice, and how much they participate in school activities. The exact parent-child activities and school involvement measures vary across the waves. Table B.3 in the Appendix shows an overview of the information I use in each wave.

I again combine the parent-child activities using a PCA. The rotated factor loadings are shown in Table B.4 in the Appendix. The first three principal components load primarily on variables that can be interpreted as recreational time investments, educational time investments, and parental school involvement, respectively.[14] The first principal component, which I interpret as recreational time investment, includes playing games with the child, painting with him or her, going to the park, and singing songs together. The second principal component, educational time investment, includes helping the child learn reading, writing, and math. The third component loads on the school involvement variables.[15] Note that, as is the case for the parenting style measures, wave six is an outlier because the information is assessed by the child instead of the mother. Here it is the first principal component that I interpret as parental school involvement and the second as educational time investment. There is no principal component that can be interpreted as recreational time investment in wave six.

Table 2.2 shows descriptive statistics for the parental input measures, computed on a

[14]The distinction between educational and recreational investments is also suggested by Del Bono et al. (2016).

[15]I only use the educational time investment factor in my main regressions for consistency reasons, because recreational time investments and school involvement are not available for all waves. Adding them in the regressions does not change the results.

pooled sample across all child ages 3, 5, 7, 11, and 14.[16] All measures are age-standardized to mean zero and standard deviation one on the full sample.

Table 2.2: Descriptive Statistics for Parental Input Measures

	Mean	S.D.	Min.	Max.	N
Parenting Style					
Demandingness	0.02	0.90	−2.17	2.29	13,835
Responsiveness	0.10	0.84	−2.97	1.58	13,835
Parental Investment					
Education	0.09	0.88	−2.72	1.75	13,835
Recreation	0.08	0.87	−2.47	1.97	10,951
Involvement	−0.03	0.79	−1.13	3.91	10,912

Note: The table shows descriptive statistics for the parental input measures (mean, standard deviation, minimum and maximum values, and number of observations; all unweighted). All measures are age-standardized to mean zero and standard deviation one (on the full sample). The values are computed on the pooled estimation sample across child ages 3, 5, 7, 11, and 14. Recreation is not observed at child age 14, and involvement is not observed at child age 3. Data: Millennium Cohort Study. Own calculations.

2.5.3 Skill Measures

Cognitive skills The MCS administered multiple achievement and ability assessments throughout the cohort members' lives. Cognitive skill assessments were part of the waves corresponding to child ages 3, 5, 7, 11, and 14. Table 2.3 shows an overview of which tests were administered at what wave and what cognitive concept they capture (see also Moulton et al., 2020, for a discussion). A description of the skill assessments can be found in section B.1.1 in the Appendix. The ability concepts in Table 2.3 are based on the Cattell-Horn-Carroll (CHC) model of cognitive ability and refer to crystallised ability (Gc), fluid ability (Gf), reading and writing (Grw), visual processing (Gv), and quantitative knowledge (Gq). The two best-known concepts are crystallised and fluid ability. Crystallised ability refers to the skill or knowledge base that an individual has acquired, e.g. the fundamental meaning of words. Fluid intelligence is the ability to solve novel problems without relying on previously acquired knowledge (Schneider and McGrew, 2018).

In an ideal setting, the skill measures are comparable both within and across waves. Because there are at least three measures in each age group required to test for measure-

[16]Because of outliers in the parental inputs measures, I remove the respective top and bottom 1% of the pooled distributions from the sample.

Table 2.3: Cognitive Skill Measures

	Age 3	Age 5	Age 7	Age 11	Age 14
Naming Vocabulary	Gc	Gc			
Pattern Construction		Gv	Gv		
Picture Similarities		Gf			
Word Reading			Gc/Grw		
Progress in Maths			Gq		
Verbal Similarities				Gc	
Vocabulary Test					Gc

Note: The table shows the cognitive skill measures in the MCS and their ability concepts based on the Cattell-Horn-Carroll model of cognitive ability: Gc (crystallised ability), Gv (visual processing), Gf (fluid ability), Grw (reading and writing), Gq (quantitative knowledge).

ment invariance—a necessary condition for comparability across waves—and the MCS does not satisfy this requirement, strict invariance can't be tested, and thus can't be assumed (McElroy et al., 2021). This implies that mean differences between the waves of the skill measures are not necessarily meaningful. However, for my application, *metric* invariance, a weaker condition, is sufficient. If metric invariance holds, regression coefficients across waves can be meaningfully compared. To achieve this, I choose measures in the same domain across waves and then to do sensitivity analysis with alternative measures.

In my benchmark regressions, I combine the following measures: Naming Vocabulary (ages 3 and 5), Word Reading (age 7), Verbal Similarities (age 11), and Vocabulary Test (age 14). That is, I combine measures for crystallised ability. For the sensitivity analysis, I use Picture Similarities (age 5), Pattern Construction (ages 5 and 7), and Progress in Maths (age 7).

Non-cognitive skills I use the Strengths and Difficulties Questionnaire (SDQ) as measure of non-cognitive skills, a well-established and validated instrument in psychology. The SDQ is a behavioral screening questionnaire designed to quantify the psychological adjustment in children aged three to sixteen (Goodman, 1997) and is highly correlated with alternative measures of non-cognitive skills. In the economics literature, the SDQ is interpreted as a measure of positive child development (see e.g. Attanasio et al., 2020) and has been used in a series of recent studies (e.g. Cornelissen and Dustmann, 2019; Jensen et al., 2017). Non-cognitive skills, as measured by the SDQ, have been found to be an important predictor of adult life satisfaction (Layard et al., 2014) and labour market outcomes (Clark and Lepinteur, 2019).

There are 25 SDQ items that are divided between five scales with five items each. The five scales are i.) hyperactivity/inattention, ii.) conduct problems, iii.) emotional symptoms, iv.) peer problems, and v.) pro-social behavior. Table B.1 in the Appendix presents the items of the five scales. Each of the items of each scale can be marked "not true", "somewhat true", or "certainly true". The score for each of the scale is generated by adding up the scores for their respective five items; each score thus ranges between zero and ten. I reverse-code the scales such that higher scores indicate better child behavior.

The MCS administered the SDQ primarily to the cohort members' parents, who were asked to complete the questionnaire on behalf of their child. For my main results, I use the sum of the hyperactivity/inattention, conduct problems, emotional symptoms, and peer problems scales (i.e. excluding the pro-social behavior scale).[17] In two waves, ages 7 and 11, the SDQ was additionally administered to the cohort members' teacher and at age 17 to the cohort members themselves. I make use of the teacher-assessed questionnaires in the robustness analyses.

Table 2.4 shows descriptive statistics for the skill measures. All measures are age-standardized to mean zero and standard deviation one on the full sample (i.e. before dropping observations with missing values). The means above zero indicate that higher-skilled children are less likely to have missing values in the input measures or control variables.[18]

2.5.4 Other Inputs in the Production Function

At child ages 3, 5, and 7, the MCS includes information about who looks after the child and how many hours per week. I control for the time children are taken care of by grandparents, siblings, other relatives, nurseries, and childminders.[19]

At child age 7, the MCS administered a teacher questionnaire to class teachers for a subset of children that includes information about teacher quality and classroom characteristics. With respect to the former, I include information about the class teacher's highest educational degree, his or her teaching experience in years, the time since he or she received the teaching qualification, and the time he or she is already teaching at the child's school. These characteristics have been shown to be predictive for teacher value-added (Clotfelter et al., 2010; Jacob et al., 2018; Papay and Kraft, 2015; Rockoff, 2004). Information about the classroom includes class size, the share of children with special educational needs, the share of children with English as a second language, the number of children excluded from class, and whether the child shares his or her class with a child that disturbs the lessons.

[17]Excluding the pro-social behavior scale is common practice and the default in the MCS total SDQ score.

[18]Age-standardizing to mean zero and standard deviation one on the estimation sample instead of the full sample yields similar results.

[19]Ideally I would also be able to control for grandparents' parenting style or the style of the child care arrangement (e.g. whether parents hire a strict nanny or childminder). Yet, this information is not available in the MCS. Inputs in the form of parenting styles are captured insofar as they are correlated with included time inputs.

Table 2.4: Descriptive Statistics for Skill Measures

	Mean	S.D.	Min.	Max.	N
Cognitive Skills					
Age 3	0.27	0.84	−3.51	3.77	2,767
Age 5	0.33	0.83	−3.60	3.91	2,767
Age 7	0.23	0.90	−3.12	3.74	2,767
Age 11	0.19	0.84	−5.94	3.41	2,767
Age 14	0.16	0.99	−2.69	4.55	2,767
Non-cognitive Skills					
Age 3	0.23	0.84	−3.08	1.81	2,767
Age 5	0.25	0.81	−4.10	1.47	2,767
Age 7	0.24	0.82	−5.02	1.37	2,767
Age 11	0.20	0.84	−4.79	1.31	2,767
Age 14	0.23	0.85	−3.97	1.37	2,767

Note: The table shows descriptive statistics for the skill measures (mean, standard deviation, minimum and maximum values, and number of observations; all unweighted). All measures are age-standardized to mean zero and standard deviation one (on the full sample). The values are computed on the estimation sample. Cognitive skills are measured by the following assessments: Naming Vocabulary (ages 3 and 5), Word Reading (age 7), Verbal Similarities (age 11), and Vocabulary Test (age 14). Non-cognitive skills are measured by the SDQ at all child ages. Data: Millennium Cohort Study. Own calculations.

Finally, I control for the Index of Multiple Deprivation (IMD), a measure of relative levels of deprivation at the Lower Super Output Area level. The consistent sub-scales across the U.K. are income, employment, health, and education. The IMD ranks every area from most to least deprived in the respective country for every sub-scale. At child ages 3, 5, and 7, the MCS contains the deciles in the country-specific IMD distribution of the area a respondent lived in at the time of the interview.

Table B.9 in the Appendix presents descriptive statistics for the input variables (see section B.1 in the Appendix for a detailed description of how the variables are coded). For the time-varying variables—household and neighborhood inputs—the values are computed on a pooled sample across child ages 3, 5, and 7.

2.5.5 Further Control Variables

As child-level controls, I include the child's age at the time of the assessment (in months), gender, ethnicity, birth order, number of siblings, and birth weight. I also include a dummy variable for whether the child was born preterm. As mother-level controls, I include the mother's age at the birth of the child (in years), her age when she left full-time education (in years), household income, employment status, marital status, cognitive skills (Word

Activity assessment), mental health (Kessler scale), and personality traits (extroversion and neuroticism). I additionally include dummy variables for whether the natural or a non-natural father is present in the household, and I control for the government region where the child lived in at the time of the interview. A further description of the control variables can be found in section B.1 in the Appendix. Table 2.5 shows descriptive statistics for the control variables.

Table B.10 in the Appendix shows associations between a subset of control variables and parenting styles. Responsiveness is positively correlated with the mother's educational attainment and parental income; it is also positively correlated with parental investment. On the other hand, demandingness is negatively correlated with the mother's age at birth and the child's birth rank.

2.6 Results

2.6.1 Benchmark Estimates

Table 2.6 reports the main results for the cognitive and non-cognitive production function, respectively. Columns (1) and (2) are the benchmark estimates; columns (3) and (4) are extensions and are discussed in the subsequent sections 2.6.2 and 2.6.3. In each column, I pool child ages 7, 11, and 14, and regress the cognitive/non-cognitive skill measure y_a on parental demandingness PS_a^D, responsiveness PS_a^R, their interaction $PS_a^D \times PS_a^R$, and the one-period lagged cognitive/non-cognitive skill measure y_{a-1}.[20,21] I also include parental time investment (educational activities, see section 2.5.2) and its interactions with demandingness and responsiveness, respectively.[22] For demandingness, responsiveness, and parental investment, I further include their first and second lag.

In all columns, I control for the age of the child at the time of the assessment (in months) and the number of siblings, and I include the following time-varying mother-level controls: equivalized household income, whether the mother is employed, dummy variables for whether the natural or a non-natural father is present in the household, a dummy variable for whether the mother is married with the father figure in the household (if present), and mother's mental health. In columns (1), (2), and (4), I also control for the following time-invariant child-level controls: gender, ethnicity, birth order, birth weight, and a dummy variable for whether the child was born preterm. Likewise, I control for the following time-invariant mother-level controls: age at the birth of the child, her age when

[20]For child age 7, the one-period lagged skill measure refers to age 5; for age 11, the one-period lagged skill measure refers to age 7; for age 14, the one-period lagged skill measure refers to age 11.

[21]Table B.11 in the Appendix shows more detailed regression results. Tables B.12 and B.13 in the Appendix also show the results for each child age separately.

[22]I exclude recreational activities and school involvement because they are not observed at all child ages.

Table 2.5: Descriptive Statistics for Control Variables

	Mean	S.D.	Min.	Max.	N
	Time-invariant Controls				
Child Controls					
Female	0.52	0.50	0.00	1.00	2,767
Ethnicity (% of sample)					
White	0.94	0.24	0.00	1.00	2,767
Black	0.01	0.11	0.00	1.00	2,767
Indian, Pakistani, Bangladeshi	0.02	0.14	0.00	1.00	2,767
Other	0.03	0.17	0.00	1.00	2,767
Birth rank	1.83	0.91	1.00	7.00	2,767
Birth weight (pounds)	3.42	0.57	0.74	5.87	2,767
Preterm birth	0.06	0.24	0.00	1.00	2,767
Mother Controls					
Age at birth (years)	29.97	5.25	18.00	45.00	2,767
Age left full-time education (years)	18.51	2.57	14.00	25.00	2,767
Cognitive skills	0.23	0.87	−2.34	1.96	2,767
Extroversion	−0.07	0.94	−3.35	2.23	2,767
Neuroticism	−0.18	0.98	−2.65	3.71	2,767
	Time-varying Controls				
Child Controls					
Age at assessment (months)	96.70	46.79	34.00	180.00	13,835
Number of siblings	1.38	0.98	0.00	8.00	13,835
Mother Controls					
Two-parent household (nat. parents)	0.80	0.40	0.00	1.00	13,835
Two-parent household (nat. mother)	0.05	0.22	0.00	1.00	13,835
Married	0.68	0.47	0.00	1.00	13,835
Equivalized weekly household income	391.72	181.96	17.04	1298.52	13,835
Employed	0.76	0.43	0.00	1.00	13,835
Mental health	−0.14	0.86	−1.02	5.93	13,835

Note: The table shows descriptive statistics for the control variables (mean, standard deviation, minimum and maximum values, and number of observations; all unweighted). The values for the time-varying controls are computed on the pooled estimation sample across child ages 3, 5, 7, 11, and 14. The measures for cognitive skills, extroversion, neuroticism, and mental health are standardized to mean zero and standard deviation one (on the full sample prior to removing observations with missing values). A further description of the control variables can be found in section B.1 in the Appendix. Data: Millennium Cohort Study. Own calculations.

Table 2.6: Main Results

	OLS		IV	
	(1)	(2)	(3)	(4)
		Cognitive Skills		
PS_a^D	-0.04^{***}	-0.04^{**}	-0.03^{*}	-0.03^{**}
	(0.01)	(0.02)	(0.01)	(0.02)
PS_a^R	0.02^{**}	0.05^{***}	0.01	0.03^{*}
	(0.01)	(0.02)	(0.01)	(0.01)
$PS_a^D \times PS_a^R$	0.00	-0.01	-0.01	0.01
	(0.01)	(0.02)	(0.01)	(0.01)
y_{a-1}	0.19^{***}	0.19^{***}	0.01	0.97^{***}
	(0.01)	(0.02)	(0.02)	(0.07)
First stage			-0.68^{***}	0.22^{***}
F-statistic			222.16	10.67
Adj. R^2	0.17	0.19		
Num. obs.	8,301	4,005	8,301	8,301
		Non-cognitive Skills		
PS_a^D	-0.10^{***}	-0.11^{***}	-0.09^{***}	-0.09^{***}
	(0.01)	(0.01)	(0.01)	(0.01)
PS_a^R	0.05^{***}	0.05^{***}	0.04^{***}	0.04^{***}
	(0.01)	(0.01)	(0.01)	(0.01)
$PS_a^D \times PS_a^R$	0.02^{**}	0.02	0.02^{**}	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
y_{a-1}	0.56^{***}	0.53^{***}	0.20^{***}	0.81^{***}
	(0.01)	(0.02)	(0.03)	(0.02)
First stage			-0.36^{***}	0.49^{***}
F-statistic			114.09	40.15
Adj. R^2	0.17	0.19		
Num. obs.	8,301	4,005	8,301	8,301
Time-var. controls	x	x	x	x
Time-invar. controls	x	x		x
School inputs		x		

Note: The table shows the estimates for the cognitive (upper panel) and non-cognitive (lower panel) production function (robust standard errors in parentheses). PS_a^D and PS_a^R denote contemporaneous parental demandingness and responsiveness, respectively. Columns (1), (2), and (4) include all child and mother controls; column (3) includes time-varying controls only. Column (2) additionally controls for school inputs. Column (3) presents the estimates for the model in first difference, and where Δy_{a-1} is instrumented with y_{a-2}. The first stage coefficient and F-statistic refer to the regression of Δy_{a-1} on y_{a-2}. In column (4), the lagged outcome is instrumented with the double-lagged outcome, and the first stage coefficient and F-statistic refer to the regressions of y_{a-1} on y_{a-2}. The models are estimated on a pooled sample across child ages 7, 11, and 14, and include year fixed-effects. Data: Millennium Cohort Study. Own calculations.

she left full-time education, and measures for the mother's cognitive skills and personality traits.[23] Columns (1), (2), and (4) further control for household and neighborhood inputs.[24] In all columns, I include controls for the government region where the child lived in at the time of the interview as well as time fixed-effects.

Column (1) in the upper panel of Table 2.6 reports the estimates of the value-added specification. I find that parental demandingness is negatively associated with cognitive skills: a one-standard-deviation increase in the demandingness measure reduces child skills by 4% of a standard deviation. On the other hand, parental responsiveness is positively associated: a one-standard-deviation increase in the responsiveness measure increases skills by 2% of a standard deviation. The interaction term doesn't reach statistical significance at the conventional levels. The coefficients of the lagged outcome indicates that 19% of child skills persist from one period to the other.

In column (2), I additionally control for school inputs. Because school inputs are available only for a subset of children, the sample size is more than halved.[25] The results remain largely unchanged.

The lower panel of Table 2.6 reports the benchmark estimates for the non-cognitive production function. Columns (1) and (2) are in line with the results for cognitive skills: parental demandingness is negatively associated with non-cognitive skills, whereas responsiveness is positively associated. Yet, there are several key differences compared to the results for cognitive skills. First, the coefficients are more than twice as large. Second, the outcome persistence is two to three times larger. Third, the interaction between demandingness and responsiveness reaches statistical significance in column (1). This suggests that responsiveness is particularly productive for the formation of non-cognitive skills when combined with demandingness—that is, authoritative parenting.

I find no evidence that the productivity of parental time investments depends on a particular parenting style in columns (1) and (2). The interactions between parenting styles and investment are insignificant for both the cognitive and non-cognitive production function (not shown in Table 2.6). That is, I find no empirical evidence for the framework of Darling and Steinberg (1993), who suggest that parenting affects the productivity of investments. Moreover, excluding the educational activities measure leaves the coefficients of demandingness and responsiveness virtually unchanged (not shown). This suggests that parenting styles operate above and beyond their influence on investments.

[23]The mother's cognitive skills are measured in wave six (child age 14), and the personality traits in wave four (child age 7). This implies that the timing when the outcome variable and the mother's skills and personality traits are measured does not always coincide.

[24]Household and neighborhood inputs are measured at child ages 3, 5, and 7. This implies that for regressions where the outcome refers to child ages above 7, the contemporaneous inputs are not included but only their lags.

[25]School inputs are only measured at child age 7. For child ages 11 and 14, they enter the regression only as lagged variables (first and second lag, respectively).

In sum, columns (1) and (2) suggest that parental demandingness is detrimental for skill development, whereas responsiveness is beneficial. Considering that my measure of demandingness mostly reflects punitive parenting and my measure of responsiveness reflects parental warmth, my results are in line with the previous literature: authoritarian parenting, i.e. punitive and low in warmth, is typically found to hem child development (e.g. Cuartas, 2022). Authoritative parenting, which in contrast to authoritarian parenting is also high in warmth, is highlighted as the most beneficial parenting style in most studies (e.g. Rothenberg et al., 2021).[26]

The effect sizes are economically significant. To put the estimates into perspective, Fredriksson et al. (2013), exploiting variation in the class size created by a maximum class size rule, find that increasing the class size by one student reduces cognitive skills at age 13 by between 3.2% and 6.3% of a standard deviation.[27] This is strikingly similar to the 4% of a standard deviation reduction I find for demandingness in column (1) of Table 2.6. With respect to my own estimates, the coefficient of demandingness of the cognitive production function is more than twice as large as the (positive) effect of postponing the mother's age when she left full-time education by one year.

To give an intuition what a one-standard-deviation increase in demandingness or responsiveness entails, I regress the frequency of how often the mother tells her child off (how often she expresses affection) on demandingness (responsiveness). The results are presented in Table B.8 in the Appendix. Increasing demandingness (responsiveness) by one standard deviation is associated with 1.16 (1.48) additional days per week where the mother tells her child off (expresses her affection).

2.6.2 Unobserved Ability

Unobserved innate ability of a child is included in the error term in Equation (2.7). Although the past effects of innate ability on the skill formation process are captured by the lagged outcome, the contemporaneous effect is not. This is problematic if parents adjust their inputs according to their child's ability. It is also problematic if high-ability children have higher skill growth, e.g. because they learn faster. This implies that $\mathrm{Cov}(y_{i,a-1}, \epsilon_{ia}) > 0$, and λ will be biased upwards because $\epsilon_{ia} = \mu_i + \varepsilon_{ia}$ (Andrabi et al., 2011).

To account for this possibility, I estimate the model in Equation (2.7) by first-differencing

[26] Fiorini and Keane (2014) find that *both* parental discipline and warmth are beneficial for child development. Yet, their measure of parental discipline is not directly comparable with my measure of demandingness. Their measure largely captures effective, not punitive, discipline, i.e. how successful parents can impose disciplinary measures. In fact, their factor loads negatively on punishing the child. In this sense, their positive and my negative coefficients are not at odds.

[27] See Table III in Fredriksson et al. (2013).

such that time-invariant variables—including innate ability—drop out of the equation:

$$y_{ia} - y_{i,a-1} = \sum_{m=0}^{a-2} \boldsymbol{\beta}_{a-m}(\boldsymbol{P}_{i,a-m} - \boldsymbol{P}_{i,a-m-1}) + \lambda(y_{i,a-1} - y_{i,a-2}) \\ + \boldsymbol{\rho}(\boldsymbol{W}_{ia} - \boldsymbol{W}_{i,a-1}) + (\varepsilon_{ia} - \varepsilon_{i,a-1}). \tag{2.8}$$

First-differencing accounts for the issue of unobserved ability under the assumption that the effect of innate ability on skill formation is constant across child ages. Note that $(y_{i,a-1} - y_{i,a-2})$ will be correlated with the error term by construction because the latter includes $\varepsilon_{i,a-1}$. I therefore follow Arellano and Bond (1991) and instrument $(y_{i,a-1} - y_{i,a-2})$ with $y_{i,a-2}$. The double-lagged outcome is uncorrelated with $(\varepsilon_{ia} - \varepsilon_{i,a-1})$ but correlated with $(y_{i,a-1} - y_{i,a-2})$ under the assumption that parents do not adjust their inputs according to the child's expected future skill shocks.

Column (3) of Table 2.6 reports the estimates for the model in first differences. The results largely support the benchmark results, although the estimates for the cognitive production function lose statistical significance. This may be due to the lagged cognitive skill measure being an ill-suited instrument—which the insignificant coefficient of the lagged outcome suggests (a coefficient of zero would imply that there is no skill persistency, which is implausible).[28]

2.6.3 Measurement Error

Andrabi et al. (2011) and Del Bono et al. (2016) show that the persistence parameter λ in Equation 2.7 can be attenuated by measurement error in the outcome variable. This matters for my application because the bias in the persistence parameter can lead to a bias in the input coefficients that can go in either direction. To check whether this is an issue, I use the double-lagged outcome to instrument the lagged outcome (as proposed by e.g. Del Bono et al., 2016).

Column (4) of Table 2.6 reports the instrumental variable estimates for the cognitive and non-cognitive production function. Two points are worth highlighting: first, the persistence parameter increases considerably when the lagged outcome is instrumented, both for the cognitive and non-cognitive production functions. This is consistent with the intuition that instrumenting the lagged outcome corrects for measurement error, thus reducing the downward bias. Second, although the IV estimates are less statistically significant compared to the OLS estimates, the results are stable and the coefficients change only little. I conclude from this exercise that, although measurement error might be present, it doesn't alter my findings and conclusion.

[28] Note that the first stage coefficient is negative because there is a mechanical negative association between y_{a-2} and $\Delta y_{a-1} = (y_{a-1} - y_{a-2})$.

2.6.4 Feedback Effects

One explanation of the results is that parents react to child outcomes and adjust their parenting style accordingly. The negative (positive) coefficient of demandingness (responsiveness) may then occur because negative skill shocks induce parents to become more (less) demanding (responsive), and not because parenting styles have a causal impact on skill formation. To address such feedback effects, I employ dynamic panel estimators that use lagged inputs as instruments (Arellano and Bond, 1991; Blundell and Bond, 1998). These estimators make the assumption that inputs are predetermined but not strictly exogenous. An input is predetermined if it is uncorrelated with contemporaneous and future error terms, but is potentially correlated with past error terms. This implies that parents are allowed to adjust their parenting styles in reaction to past child skill shocks (but not to contemporaneous or future skill shocks).

First, the Arellano-Bond estimator takes the model in first differences—see Equation (2.8)—and uses lagged inputs and outcomes to instrument the first-differences. Specifically, $\Delta y_{i,a-1}$ is instrumented with $y_{i,a-2}$, and $\Delta \boldsymbol{P}_{i,a}$ with $\boldsymbol{P}_{i,a-1}$. Identification of λ and $\boldsymbol{\beta}_a$ is achieved by imposing moment conditions derived from the assumption that inputs are predetermined (Arellano and Bond, 1991). The Arellano-Bond estimator is also known as "difference GMM".

Second, the Blundell-Bond estimator extends the Arellano-Bond estimator and estimates a system of equations, one for Equation (2.7), and another for Equation (2.8). In addition to Arellano and Bond (1991), Blundell and Bond (1998) additionally instrument the inputs and lagged outcome in levels with the inputs and lagged outcome in first differences: $y_{i,a-1}$ with $\Delta y_{i,a-1}$, and $\boldsymbol{P}_{i,a}$ with $\Delta \boldsymbol{P}_{i,a}$. This is valid because additional moment conditions can be derived under the assumption that there is a constant correlation between the time-varying inputs and innate ability, and that time-invariant inputs and controls are uncorrelated with innate ability.

The Blundell-Bond estimator, also known as "system GMM", is more efficient than difference GMM because of these additional moment conditions. However, the constant-correlation assumption is rather strong. In addition, as Andrabi et al. (2011) points out, the Blundell-Bond estimator requires that innate ability does not influence skill growth, which arises for instance when imperfect persistence cancels the benefit of faster learning. In light of these issues, I prefer difference GMM to system GMM, but nonetheless show estimates for the latter because it is less susceptible to weak instruments.

Table 2.7 reports the estimates using the Arellano-Bond and Blundell-Bond estimators, both for the cognitive (top panel) and non-cognitive (bottom panel) production function. In column (1), I reproduce the results from column (3) of Table 2.6 as a reference (with the exception that I do not control for the interactions of the parental inputs because they complicate instrumentation in the remaining columns; I also initially exclude lagged first-differenced inputs, but include them in later specifications). Column (2) shows the

Arellano-Bond benchmark estimates, where I instrument $\Delta y_{i,a-1}$ with $y_{i,a-2}$ and $\Delta \boldsymbol{P}_{i,a}$ with $\boldsymbol{P}_{i,a-1}$. Under the assumption that parents react to past positive skill shocks by decreasing (increasing) their demandingness (responsiveness), feedback effects bias the estimate of λ upwards and the estimated coefficients of parental inputs towards zero. Correcting for feedback effects would thus lead to a reduction in the estimate of λ, and a more negative (more positive) estimated coefficient of demandingness (responsiveness). Comparing column (1) with (2) shows little indication of feedback effects: although the estimates for the cognitive production function lose significance, the stable coefficients for the non-cognitive production function are reassuring.

In column (3), I instrument $\Delta y_{i,a-1}$ with $y_{i,a-3}$ instead of $y_{i,a-2}$ to account for measurement error.[29] If true skill y^*_{ia} is measured with error u_{ia} such that $y_{ia} = y^*_{ia} + u_{ia}$, Equation (2.8) becomes $\Delta y_{ia} = \sum_{m=0}^{a-2} \boldsymbol{\beta}_{a-m} \Delta \boldsymbol{P}_{i,a-m} + \lambda \Delta y_{i,a-1} + \boldsymbol{\rho} \Delta \boldsymbol{W}_{ia} + [\Delta \varepsilon_{ia} + \Delta u_{ia} - \lambda \Delta u_{i,a-1}]$ (Andrabi et al., 2011). $y_{i,a-2}$ will then be correlated with $\Delta u_{i,a-1} = (u_{i,a-1} - u_{i,a-2})$ by construction and downward-bias the estimate of λ. Comparing column (2) with (3), the persistency parameter indeed increases considerably for the cognitive production, yet not for the non-cognitive production function (in contrast, λ is lower when the instrument is $y_{i,a-3}$ instead of $y_{i,a-2}$). In any case, the coefficients of demandingness and responsiveness change little.

In column (4), I additionally include $\Delta \boldsymbol{P}_{i,a-1}$ in the difference equation. In this specification, I use $\boldsymbol{P}_{i,a-2}$ and $\boldsymbol{P}_{i,a-3}$ as instruments instead of $\boldsymbol{P}_{i,a-1}$. This is because when inputs are assumed to be predetermined (and not exogenous), $\boldsymbol{P}_{i,a-1}$ is an invalid instrument for $\Delta \boldsymbol{P}_{i,a-1}$. The estimates in column (4) are considerably different than the estimates in columns (1) to (3). For the cognitive production function, the coefficient of responsiveness is significantly larger. For the non-cognitive production function, the coefficients of both demandingness and responsiveness are insignificant. This may be because the instruments are weak in this specification.

For columns (2) to (4), I present p-values for the m2 test, a test for second-order serial correlation in the idiosyncratic error term ε_{ia} in Equation (2.7). Serial correlation would invalidate the instruments because $\boldsymbol{P}_{i,a-1}$ and $\boldsymbol{P}_{i,a-2}$ are potentially correlated with past and contemporaneous error terms, and may then also be correlate with future error terms. Reassuringly, in column (4) of the upper panel and in columns (2) to (4) of the bottom panel, I fail to reject the null hypothesis of no serial correlation.

I also present p-values for the Sargan test of the overidentifying restrictions. In all but one GMM specification, I reject the null hypothesis that the overidentifying restrictions are valid. This may be problematic, yet, as Andrabi et al. (2011, p. 47) point out, it is not entirely unexpected because different instruments identify different local average treatment effects.

[29]To keep the number of observations constant, I nonetheless instrument $\Delta y_{i,a-1}$ with $y_{i,a-2}$ in the earliest time period.

Table 2.7: GMM Estimates

	IV	Difference GMM			System GMM		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				Cognitive Skills			
PS_a^D	-0.03^{*}	0.01	0.01	-0.06	-0.00	0.01	0.04
	(0.02)	(0.02)	(0.02)	(0.11)	(0.02)	(0.02)	(0.05)
PS_a^R	0.02^{*}	0.00	0.00	0.34^{***}	-0.01	-0.01	0.31^{***}
	(0.01)	(0.02)	(0.02)	(0.13)	(0.01)	(0.01)	(0.08)
y_{a-1}	0.01	0.01	0.23^{***}	0.08^{**}	0.04^{**}	0.05^{***}	0.10^{***}
	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
Num. obs.	8,301	8,301	8,301	8,301	8,301	8,301	8,301
m2 (p-value)		0.00	0.00	0.75	0.00	0.00	0.00
Sargan (p-value)	0.00	0.00	0.00	0.51	0.00	0.00	0.00
				Non-cognitive Skills			
PS_a^D	-0.10^{***}	-0.11^{***}	-0.10^{***}	-0.08	-0.15^{***}	-0.14^{***}	0.01
	(0.01)	(0.02)	(0.02)	(0.05)	(0.01)	(0.01)	(0.04)
PS_a^R	0.04^{***}	0.03^{**}	0.03^{**}	-0.08	0.04^{***}	0.04^{***}	-0.02
	(0.01)	(0.01)	(0.01)	(0.06)	(0.01)	(0.01)	(0.06)
y_{a-1}	0.19^{***}	0.18^{***}	0.13^{***}	0.18^{***}	0.35^{***}	0.35^{***}	0.37^{***}
	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
Num. obs.	8,301	8,301	8,301	8,301	8,301	8,301	8,301
m2 (p-value)		0.66	0.80	0.19	0.03	0.04	0.00
Sargan (p-value)	0.16	0.00	0.00	0.00	0.00	0.00	0.00

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Note: The table shows the estimates for the cognitive and non-cognitive production function (robust standard errors in parentheses). PS_a^D and PS_a^R denote parental demandingness and responsiveness, respectively. Each column includes the full set of time-varying child and mother controls. Each model is estimated in first-differences. In column (1), $\Delta y_{i,a-1}$ is instrumented with $y_{i,a-2}$. Columns (2)–(4) are estimated with the Arellano-Bond estimator, and columns (5)–(7) with the Blundell-Bond estimator. In columns (2) and (5), $\Delta y_{i,a-1}$ is instrumented with $y_{i,a-2}$, and $\Delta \boldsymbol{P}_{i,a}$ with $\boldsymbol{P}_{i,a-1}$. In columns (3) and (6), $\Delta y_{i,a-1}$ is instrumented with $y_{i,a-3}$, and $\Delta \boldsymbol{P}_{i,a}$ with $\boldsymbol{P}_{i,a-1}$. Columns (4) and (7) additionally include $\Delta \boldsymbol{P}_{i,a-1}$, and $\Delta \boldsymbol{P}_{i,a}$ and $\Delta \boldsymbol{P}_{i,a-1}$ are instrumented with $\boldsymbol{P}_{i,a-2}$ and $\boldsymbol{P}_{i,a-3}$. The Blundell-Bond estimator additionally estimates a levels equation in columns (5)–(7). In column (5), $y_{i,a-1}$ is instrumented with $\Delta y_{i,a-1}$ and $\boldsymbol{P}_{i,a}$ with $\Delta \boldsymbol{P}_{i,a}$. In column (6), $y_{i,a-1}$ is instrumented with $\Delta y_{i,a-2}$ and $\boldsymbol{P}_{i,a}$ with $\Delta \boldsymbol{P}_{i,a}$. In column (7), $y_{i,a-1}$ is instrumented with $\Delta y_{i,a-1}$ and $\boldsymbol{P}_{i,a}$ and $\boldsymbol{P}_{i,a-1}$ with $\Delta \boldsymbol{P}_{i,a-2}$ and $\Delta \boldsymbol{P}_{i,a-3}$. Data: Millennium Cohort Study. Own calculations.

Columns (5) to (7) show the results for the Blundell-Bond estimator. Column (5) corresponds to column (2) with the addition that I instrument $y_{i,a-1}$ with $\Delta y_{i,a-1}$ and $\boldsymbol{P}_{i,a}$ with $\Delta \boldsymbol{P}_{i,a}$ in a levels equation. Similarly, column (6) corresponds to column (3), but I instrument $y_{i,a-1}$ with $\Delta y_{i,a-2}$ in the levels equation. Finally, column (7) corresponds to column (4), but is estimated with the Blundell-Bond instead of the Arellano-Bond estimator. In all specifications, the persistency parameter increases considerably, likely because the increased efficiency is better able to account for the attenuation bias. For the non-cognitive production function, the coefficient of demandingness becomes slightly more negative. However, note that the m2 test rejects the null hypothesis of no serial correlation in all specifications at least at the 5%-level.

In sum, Table 2.7 provides little evidence of feedback effects. A few caveats apply, however. First, the dynamic panel estimators only allow for parents adjusting their parenting style in reaction to *past* child skill shocks. Yet, it is possible that parents also anticipate future shocks, which would invalidate the instruments. More problematic, parents are also not allowed to react to *contemporaneous* shocks. Still, my results indicate that parents do not react to past shocks, which supports the assumption that parents also do not respond to contemporaneous shocks. Second, the Sargan test rejects the null hypothesis that the overidentifying restrictions are valid in most cases. This may be problematic, but as mentioned, not unexpected in my application.

2.6.5 Effects across Child Ages

To see whether the effect of parenting styles on skills is heterogeneous across child ages, I estimate the model separately for each child age. Compared to the baseline model in Table 2.6, I make the following adjustments: first, I additionally estimate the model with cognitive and non-cognitive skills at child ages 3 and 5 on the left-hand-side, respectively.[30] Second, I exclude child age 14. The coefficients for child age 14 are less comparable because the parenting style measures are based on child-assessed instead of mother-assessed information. Third, I construct a demandingness measure that is more comparable across child ages. When I employ a PCA on the punitive parenting information only, I find that the first three principal components can be interpreted as non-restrictive control (loading on e.g. ignoring or bribing the child), restrictive control (loading on e.g. sending the child to the room or taking away treats), and corporal punishment (loading on e.g. smacking the child).[31] Crucially, not all of these factors are observed in all waves, which implies that the benchmark demandingness measure has a different interpretation across waves. I choose restrictive control as my measure of demandingness.[32]

[30]For the regression with skills at child age 3, I don't control for the lagged outcome or inputs because I don't observe either of them prior to age 3. At child age 5, I include the lagged outcome and the first lag of the input variables.

[31]The factor loadings are shown in Table B.7 in the Appendix.

[32]The results are similar for nonrestrictive control and corporal punishment.

Figure 2.2 shows the point estimates and the 95% confidence intervals for the cognitive and non-cognitive production functions, respectively. For cognitive skills, the negative association with demandingness appears only during middle childhood, but for non-cognitive skills, the negative association is present throughout all developmental stages. For parental responsiveness, the opposite is the case: the positive association of responsiveness with skills is more pronounced during early childhood (for cognitive skills, the coefficient of responsiveness is statistically significant only at child age 3).[33] Figure 2.2 suggests not only that parenting styles are relevant for all ages, but also that the relative importance of demandingness and responsiveness changes across child ages.

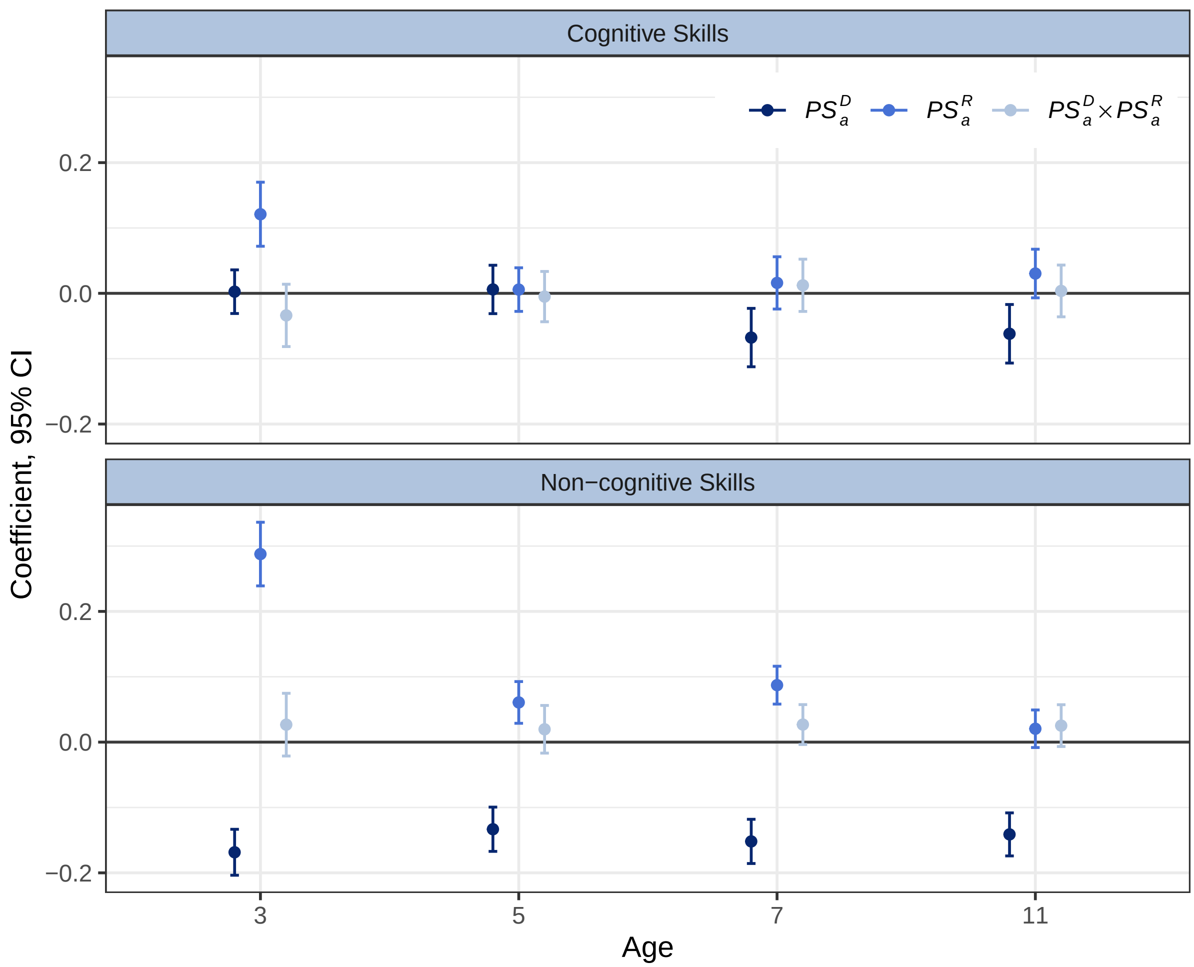
2.6.6 Robustness Analyses

The MCS administered multiple cognitive skill measures at child ages 3, 5, and 7 (see Table 2.3). To test whether the results are sensitive to the choice of my baseline measures, I estimate the production functions using alternative measures. The top panel of Table 2.8 reports the estimates where I substitute the right-hand-side lagged outcome with an alternative outcome. Columns (1) and (4) are the benchmark results, corresponding to column (5) in Table 2.6, but with a smaller sample size. In column (2), I use as lagged outcome the Picture Similarities (age 7), the Pattern Construction (age 11), and the Verbal Similarities (age 14). In column (3), I use as lagged outcome the Pattern Construction (age 7), the Progress in Maths (age 11), and the Verbal Similarities (age 14). In column (5), I use the Strengths and Difficulties Questionnaire that was administered to the cohort member's school teacher. Column (5) is estimated on a pooled sample with child ages 11 and 14 only because the teacher-administered SDQ is only available at child ages 7 and 11. I find that the benchmark results are robust to the choice of cognitive skill measure on the right-hand-side.

The bottom panel of Table 2.8 reports the estimates where I substitute the left-hand-side outcome with an alternative outcome. Columns (1) and (4) again correspond to column (5) in Table 2.6. In column (2), I use the Pattern Construction, and in column (3) I use the Progress in Maths. Columns (1) to (3) are estimated on child age 7 outcomes only because there are no alternative skill measures available for later waves. In column (5), I use the teacher-administered SDQ. Column (5) is estimated on a pooled sample with child ages 7 and 11. The results are again similar to the benchmark estimates, although less statistically significant.

[33] The coefficient at child age 3 is significantly different from the coefficients at child age 5, 7, and 11 (both for the cognitive and non-cognitive production function).

Figure 2.2: Heterogeneity in Productivity across Child Ages



Note: The table shows the point estimates and 95% confidence intervals for the cognitive (top) and non-cognitive (bottom) production function. Demandingness refers to restrictive control for all child ages. All regressions control for the full set of child and mother controls. Child ages 7 and 11 additionally control for the first and second lagged parenting styles, and age 5 controls for the first lagged parenting style. Ages 5, 7, and 11 additionally control for the lagged outcome. Data: Millennium Cohort Study. Own calculations.

Table 2.8: Robustness Analysis

	Cognitive Skills			Non-cognitive Skills	
	(1)	(2)	(3)	(4)	(5)
			Right-hand Side		
PS_a^D	-0.04^{***}	-0.03^{***}	-0.03^{***}	-0.08^{***}	-0.09^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
PS_a^R	0.02^{**}	0.03^{**}	0.03^{**}	0.05^{***}	0.07^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$PS_a^D \times PS_a^R$	0.00	-0.00	-0.00	0.02^{*}	0.04^{**}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
y_{a-1}	0.19^{***}			0.55^{***}	
	(0.01)			(0.02)	
y'_{a-1}		0.16^{***}			0.25^{***}
		(0.01)			(0.02)
y''_{a-1}			0.20^{***}		
			(0.01)		
Adj. R^2	0.17	0.16	0.18	0.49	0.34
Num. obs.	8,263	8,263	8,263	3,808	3,808
			Left-hand Side		
PS_a^D	-0.05^{**}	-0.02	-0.03	-0.19^{***}	-0.12^{***}
	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)
PS_a^R	0.01	0.02	-0.03^{*}	0.03^{***}	0.01
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)
$PS_a^D \times PS_a^R$	-0.01	0.02	0.00	0.03^{**}	0.01
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)
y_{a-1}	0.22^{***}	0.16^{***}	0.29^{***}	0.54^{***}	0.20^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Adj. R^2	0.24	0.10	0.16	0.51	0.17
Num. obs.	2,755	2,755	2,755	3,808	3,808

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Note: The table shows the estimates for the cognitive and non-cognitive production function (robust standard errors in parentheses). PS_a^D and PS_a^R denote contemporaneous parental demandingness and responsiveness, respectively. Each column includes the full set of child and mother controls, including household and neighborhood inputs as well as lagged inputs. Columns (1) and (4) refer to the benchmark estimates. Upper panel: In column (2), the lagged outcome is the Picture Similarities at age 7, the Pattern Construction at age 11, and the Verbal Similarities at age 14; in column (3), the lagged outcome is the Pattern Construction at age 7, the Progress in Maths at age 11, and the Verbal Similarities at age 14; and in column (5), the lagged outcome is the teacher-administered SDQ. Columns (1) to (3) are estimated on a pooled sample across child ages 7, 11, and 14, and include year fixed-effects; columns (4) and (5) are estimated on a pooled sample across child ages 11 and 14, and include year fixed-effects. Lower panel: In column (2), the outcome is the Pattern Construction; in column (3), the outcome is the Progress in Maths; and in column (5), the outcome is the teacher-administered SDQ. Columns (1) to (3) are estimated on child age 7 outcomes only; columns (4) and (5) are estimated on a pooled sample across child ages 7 and 11, and include year fixed-effects. Data: Millennium Cohort Study. Own calculations.

2.7 Conclusion

In this paper, I investigate the relevance of parenting styles for the formation of cognitive and non-cognitive skills. I use two measures of parenting styles that are common in developmental psychology: parental demandingness and responsiveness. I obtain these measures from the U.K. Millennium Cohort Study based on detailed information about parent-child interactions. Crucially, I propose *time-varying* measures that allow me to estimate dynamic panel models and to address the major sources of endogeneity: unobserved inputs, heterogeneity in innate ability, measurement error, and feedback effects.

I find that both demandingness and responsiveness are relevant for child development. The former is negatively associated with child skills, whereas the latter is positively associated. I also find a positive interaction between demandingness and responsiveness. This confirms the finding of psychologists that authoritative parenting—high demandingness and high responsiveness—is associated with the best child outcomes. Finally, my results suggest that parenting styles are more important for non-cognitive skills than cognitive skills.

A few caveats apply. First, I obtain my parenting style measures based on parent-child interactions assessed by the mother. However, child-assessed information or information assessed by a trained observer may be more reliable. Using the waves where I can additionally observe child-assessed information, I reassuringly nonetheless find that child- and parent-assessed information is correlated. Second, I investigate only the productivity of the mother's parenting style, not the father's. This is because the information I use to construct my measures of demandingness and responsiveness is assessed by the main interview respondent only, which is always the natural mother by construction in this study. In any case, in households with two parents, mothers and fathers typically have similar parenting styles, such that the former acts as a proxy for the latter (Steinberg, 2001). Third, my measure of demandingness largely captures punitive parenting. Although this is an important component of being a demanding parent, it by no means captures the full spectrum of it. My results thus have to be seen in this light: the negative association between demandingness and child skills means that punitive parenting is detrimental, but not necessarily other components of being demanding. Finally, my results may be affected by feedback effects (parents adjusting their parenting style according to child outcomes). Yet, using lagged inputs as instruments, I cannot find evidence for feedback effects.

My results help understanding how parents affect the lives of their children. I show that it is not only the resources that parents invest that matter, e.g. the time they spend with their child, but also *how* they spend these resources. In light of the changes in parenting styles over the last decades, my findings are important for policy-makers and researchers concerned with intergenerational mobility and equality of opportunity. Parenting styles may be an important intervention target for leveling the playing field for children.

Appendix B

B.1 Data Appendix

B.1.1 Skill Measures

The paragraphs below describes the cognitive and non-cognitive skill measures in the Millennium Cohort Study. I age-standardize each measure to mean zero and standard deviation one (separately for each wave).

Naming Vocabulary (waves two and three) The Naming Vocabulary assesses the spoken vocabulary of young children. The child is shown coloured pictures of objects and asked to name them. The assessment measures expressive language ability and requires the child to recall words from long-term memory.

Pattern Construction (waves three and four) The Pattern Construction measures spatial awareness, dexterity, coordination, and traits like perseverance and determination. The child is asked to construct a design by putting together flat squares or solid cubes with black and yellow patterns on each side. The score is based on accuracy and speed.

Picture Similarities (wave three) The Picture Similarities measures problem solving abilities. The child is shown a row of four pictures on a page and asked to place a card with a fifth picture under the picture most similar to it.

Word Reading (wave four) The Word Reading assesses English reading ability. The child is asked to read aloud a series of words presented on a card to the interviewer. The words are organised into 9 blocks of 10 words in ascending order of difficulty.

Progress in Maths (wave four) The Progress in Maths assesses mathematical skills and knowledge. The child is asked questions covering topics such as numbers, shapes, measurement and data handling. The test is read aloud to the child and they are asked to complete a series of calculations in a paper-and-pencil exercise.

Verbal Similarities (wave five) The Verbal Similarities assesses verbal reasoning and verbal knowledge. The child is asked to say how three words that the interviewer reads out are similar or go together. The assessment duration depends on the performance of the child.

Vocabulary Test (wave six) The Vocabulary Test measures the understanding of the meaning of words. The child is presented with a list of target words, each with five other words next to them. From the five other words, the child has to select the one with the same meaning as the target word.

Strengths and Difficulties Questionnaire In each wave, cohort members' parents were asked to complete the Strengths and Difficulties Questionnaire on behalf of their child. In waves four and five, the SDQ was additionally administered to the cohort members' teacher. There are 25 SDQ items that are divided between five scales with five items each. The five scales are i.) hyperactivity/inattention, ii.) conduct problems, iii.) emotional symptoms, iv.) peer problems, and v.) pro-social behavior. Table B.1 shows the items of the five scales. Each of the five items of each scale can me marked "not true", "somewhat true", or "certainly true". The score for each of the scales is computed by adding up the scores for their respective five items. Each score thus ranges between zero and ten. The scales are reverse-coded such that higher scores indicate better child behavior. The finale measure is the sum of the hyperactivity/inattention, conduct problems, emotional symptoms, and peer problems scales (i.e. excluding the pro-social behavior scale).

Table B.1: Strengths and Difficulties Questionnaire

Hyperactivity/Inattention
• Restless, overactive, cannot stay still for long • Constantly fidgeting or squirming • Easily distracted, concentration wanders • Thinks things out before acting • Sees tasks through to the end, good attention span
Conduct Problems
• Often has temper tantrums or hot tempers • Generally obedient, usually does what adults request • Often fights with other children or bullies them • Often lies or cheats • Steals from home, school or elsewhere
Emotional Symptoms
• Often complains of headaches, stomach-ache or sickness • Many worries, often seems worried • Often unhappy, down-hearted or tearful • Nervous or clingy in new situations, easily loses confidence • Many fears, easily scared
Peer Problems
• Rather solitary, tends to play alone • Has at least one good friend • Generally liked by other children • Picked on or bullied by other children • Gets on better with adults than with other children
Pro-social Behavior
• Considerate of other people's feelings • Shares readily with other children (treats, toys, pencils, etc.) • Helpful if someone is hurt, upset or feeling ill • Kind to younger children • Often volunteers to help others (parents, teachers, other children)

Note: The table shows the items of the five sub-scales of the Strengths and Difficulties Questionnaire.

B.1.2 Parental Inputs

Table B.2: Parenting Style Items

	Lowest Value	Highest Value
Age 3		
I share an affectionate, warm relationship with *Jack*	Def. not (1)	Definitely (5)
Jack will seek comfort from me	Def. not (1)	Definitely (5)
Jack values his/her relationship with me	Def. not (1)	Definitely (5)
Jack spontaneously shares information about him/herself	Def. not (1)	Definitely (5)
It is easy to be in tune with what *Jack* is feeling	Def. not (1)	Definitely (5)
Jack openly shares his/her feelings/experiences with me	Def. not (1)	Definitely (5)
When I praise *Jack*, he/she beams with pride	Def. not (1)	Definitely (5)
How often do you do the following when *Jack* is naughty:		
Ignore him/her	Never (1)	Daily (5)
Shout at him/her	Never (1)	Daily (5)
Take away treats	Never (1)	Daily (5)
Tell him/her off	Never (1)	Daily (5)
Bribe him/her (e.g. with sweets or a treat)	Never (1)	Daily (5)
Smack him/her	Never (1)	Daily (5)
Send to his/her bedroom/naughty chair, etc.	Never (1)	Daily (5)
Does *Jack* go to bed at regular times?	Never (1)	Always (4)
Does *Jack* have meals at regular times?	Never (1)	Always (4)
Age 5		
Overall, how close would you say you are to *Jack*?	Not close (1)	Extr. close (4)
How often do you do the following when *Jack* is naughty:		
Ignore him/her	Never (1)	Daily (5)
Try to reason with him/her	Never (1)	Daily (5)
Shout at him/her	Never (1)	Daily (5)
Take away treats	Never (1)	Daily (5)
Tell him/her off	Never (1)	Daily (5)
Bribe him/her (e.g. with sweets or a treat)	Never (1)	Daily (5)
Smack him/her	Never (1)	Daily (5)
Send to his/her bedroom/naughty chair, etc.	Never (1)	Daily (5)
How often do you make sure that *Jack* does a request?	Never (1)	All the time (5)
Does *Jack* go to bed at a regular time (weekdays)?	Never (1)	Always (4)
Does *Jack* have meals at regular times?	Never (1)	Always (4)
Age 7		
Overall, how close would you say you are to *Jack*?	Not close (1)	Extr. close (4)
How often do you:		
Enjoy listening to/doing things with *Jack*?	Never (1)	Always (5)
Express affection by hugging/kissing/holding *Jack*?	Never (1)	Always (5)
How often do you do the following when *Jack* is naughty:		

Ignore him/her	Never (1)	Daily (5)
Try to reason with him/her	Never (1)	Daily (5)
Shout at him/her	Never (1)	Daily (5)
Send to his/her bedroom/naughty chair, etc.	Never (1)	Daily (5)
Take away treats	Never (1)	Daily (5)
Tell him/her off	Never (1)	Daily (5)
Bribe him/her (e.g. with sweets or a treat)	Never (1)	Daily (5)
Smack him/her	Never (1)	Daily (5)
Does *Jack* go to bed at a regular time (weekdays)?	Never (1)	Always (4)
Do you have rules about:		
How early or late *Jack* may watch TV?	No (0)	Yes (1)
How many hours *Jack* may watch TV?	No (0)	Yes (1)
How often is *Jack* involved in household activities/chores?	Not at all (1)	Every day (6)
Age 11		
Overall, how close would you say you are to *Jack*?	Not close (1)	Extr. close (4)
How often do you do the following when *Jack* misbehaves:		
Send to his/her bedroom/ground him/her, etc.	Never (1)	Daily (5)
Take away treats/remove privileges	Never (1)	Daily (5)
Try to reason with him/her	Never (1)	Daily (5)
Does *Jack* go to bed at a regular time (weekdays)?	Never (1)	Always (4)
Do you have rules about		
How early or late *Jack* may watch TV etc.?	No (0)	Yes (1)
What kind of TV *Jack* may watch etc.?	No (0)	Yes (1)
How often is *Jack* involved in household activities/chores?	Not at all (1)	Every day (6)
Age 14		
Overall, how close would you say you are:		
To your mother?	Not close (1)	Extr. close (4)
To your father?	Not close (1)	Extr. close (4)
If you do something that you shouldn't, do your parents:		
Ground you/stop you from seeing friends?	No (0)	Yes (1)
Tell you off or shout at you?	No (0)	Yes (1)
Punish you in some other way?	No (0)	Yes (1)
Without your parents knowing where you were, did you:		
Stay out after 9 pm at night in the past month?	Never (1)	10 times (4)
Stay away over night in the past year?	Never (1)	Yes (3)
When you go out, how often do your parents know:		
Where you are going?	Never (1)	Always (4)
Who you are going out with?	Never (1)	Always (4)
What you are doing?	Never (1)	Always (4)

Note: The table shows the variables that enter the principal component analysis to construct the parenting style measures. The second and third column present the lowest and highest value of the variable, respectively (numerical value in parenthesis).

Table B.3: Parental Investment Items

	Lowest Value	Highest Value
Age 3		
How often:		
Does anyone else at home read to *Jack*?	Not at all (1)	Every day (6)
Do you read to *Jack*?	Not at all (1)	Every day (6)
Does someone at home take *Jack* to the library?	Never (0)	1x/week (4)
Does *Jack* paint or draw at home?	Never (0)	7x/week (7)
How often does someone at home:		
Help *Jack* to learn the ABC or the alphabet?	Never (0)	7x/week (7)
Try to teach *Jack* numbers or counting?	Never (0)	7x/week (7)
Try to teach *Jack* any songs/poems/nursery rhymes?	Never (0)	7x/week (7)
Has *Jack* eaten with a family member in the past week?	No (0)	Yes (1)
Does anyone help *Jack* learn a sport/dance/activity?	No (0)	Yes (1)
Age 5		
How often do you:		
Read to *Jack*?	Not at all (1)	Every day (6)
Tell stories to *Jack* not from a book?	Not at all (1)	Every day (6)
Make music/dance with *Jack*?	Not at all (1)	Every day (6)
Draw, paint or make things with *Jack*?	Not at all (1)	Every day (6)
Play sports/outdoors games/indoors games with *Jack*?	Not at all (1)	Every day (6)
Play with toys/games indoors with *Jack*?	Not at all (1)	Every day (6)
Take *Jack* to the park or to an outdoor playground?	Not at all (1)	Every day (6)
How often does someone at home:		
Help *Jack* with reading?	Never (0)	Every day (6)
Help *Jack* with writing?	Never (0)	Every day (6)
Help *Jack* with numbers, counting and adding up?	Never (0)	Every day (6)
How often:		
Has *Jack* been to a library (past year)?	Never (1)	Every day (7)
Do you take part in physical activities with *Jack*?	Never (1)	Every day (7)
Does your family do things together for an evening?	Never (1)	Every day (7)
Age 7		
How often do you:		
Read with or to *Jack*?	Not at all (1)	Every day (6)
Tell stories to *Jack* not from a book?	Not at all (1)	Every day (6)
Play music/nursery rhymes/dance etc. with *Jack*?	Not at all (1)	Every day (6)
Draw, paint or make things with *Jack*?	Not at all (1)	Every day (6)
Play sports or physically active games with *Jack*?	Not at all (1)	Every day (6)
Take part in physical activities with *Jack*?	Not at all (1)	Every day (6)
Play with toys or games indoors with *Jack*?	Not at all (1)	Every day (6)
Take *Jack* to the park or to an outdoor playground?	Not at all (1)	Every day (6)
How often does someone at home:		
Help *Jack* with reading?	Never (0)	Every day (6)

Help *Jack* with writing and spelling?	Never (0)	Every day (6)
Help *Jack* with maths?	Never (0)	Every day (6)
Age 11		
How often does anyone at home:		
Help *Jack* with his/her homework?	Never (1)	Always (4)
Make sure *Jack* has finished homework?	Never (1)	Always (4)
How often do you:		
Play sports or physically active games with *Jack*?	Not at all (1)	Every day (6)
Play indoor games with *Jack*?	Not at all (1)	Every day (6)
Age 14		
How often:		
Does anyone make sure you do your homework?	Never (1)	Always (4)
Do you talk to *Jack* about important things to him/her?	Not at all (1)	Every day (6)

Note: The table shows the variables that enter the principal component analysis to construct the parental investment measures. The second and third column present the lowest and highest value of the variable, respectively (numerical value in parenthesis).

Table B.4: Parental Investment Factor Loadings

	Education	Recreation	Involvement
Age 3			
Reading to the child	0.68	0.13	
Anyone else reading to the child	0.68	0.06	
Visiting the library with the child	0.67	−0.14	
Teaching the child sports	0.20	0.21	
Teaching the child learn the alphabet	−0.09	0.70	
Teaching the child counting	−0.05	0.81	
Teaching the child songs/poems/rhymes	0.13	0.68	
Painting with the child	−0.02	0.53	
Eating with the child as a family	0.21	−0.02	
Age 5			
Reading to the child	0.36	0.36	0.00
Telling stories to the child	0.09	0.50	−0.01
Doing musical activities with the child	0.00	0.56	−0.01
Painting with the child	0.06	0.64	−0.03
Playing physically active games with the child	−0.11	0.73	0.03
Playing indoor games with the child	0.03	0.68	0.03
Taking the child to park or playground	−0.12	0.55	0.01
Helping the child reading	0.77	−0.04	−0.05
Helping the child writing	0.74	0.07	−0.08
Helping the child with maths	0.71	0.07	−0.04
Visiting the library with the child	0.13	0.20	0.04
Family playing physically active games	−0.07	0.53	0.01
Family playing indoor games with the child	0.05	0.33	−0.03
Attending parent evenings	0.19	−0.03	0.05
Parent-initiated meeting with teacher	0.03	0.02	0.77
Teacher-initiated meeting with teacher	−0.04	0.01	0.76
Age 7			
Reading to the child	0.12	0.47	0.09
Telling stories to the child	0.10	0.50	−0.07
Doing musical activities with the child	−0.00	0.54	−0.04
Painting with the child	0.08	0.66	−0.11
Playing physically active games with the child	−0.06	0.72	0.00
Playing indoor games with the child	−0.00	0.72	−0.06
Taking the child to park or playground	0.00	0.47	0.02
Helping the child reading	0.80	0.04	0.07
Helping the child writing	0.86	0.02	0.03
Helping the child with maths	0.81	0.04	−0.00

Family playing physically active games	−0.06	0.57	0.07
Attending parent evenings	−0.00	0.08	0.25
Steps taken to get child into school	−0.01	−0.01	0.27
Participation in school activities	−0.11	0.28	0.25
Parent-initiated meeting with teacher	0.08	−0.06	0.75
Teacher-initiated meeting with teacher	0.06	−0.13	0.69
Age 11			
Helping the child with homework	0.81	−0.05	0.01
Ensuring that the homework is complete	0.79	0.03	−0.01
Playing physically active games with the child	−0.04	0.86	0.01
Playing indoor games with the child	0.01	0.85	−0.01
Steps taken to get child into school	−0.10	−0.13	0.91
Attending parent evenings	0.07	0.10	0.45
Age 14			
Ensuring that the homework is complete	0.67		−0.14
Talking about important things with the child	0.67		0.05
Attending parent evenings	0.50		0.06
Parent-initiated meeting with teacher	0.08		0.80
Teacher-initiated meeting with teacher	−0.08		0.80

Note: The table shows the rotated factor loadings for waves two to six (child ages 3, 5, 7, 11, and 14, respectively). For child ages 3, 5, 7, 11, and 14, education refers to the second principal component. For child ages 3, 5, 7, and 11, recreation refers to the first principal component. For child ages 5, 7, and 11, involvement refers to the third principal component; for child age 14, it refers to the first principal component. Data: Millennium Cohort Study. Own calculations.

B.1.3 Other Inputs

Household inputs At waves two, three, and four, the MCS includes information about who looks after the child and for how many hours per week. For each wave, I compute three variables by aggregating the hours of the following items: childcare services (childminder, workplace / college nursery / creche, private / independent day nursery / creche, local authority nursery, nursery school, nursery or reception class in a primary or infants school, special day school or nursery or unit for children with SEN, playgroup, combined centre / family centre); grandparents (grandparent at home, care in grandparent's home); others (other relative including non-resident parent at home, care in other relative's home including non-resident parent, non-relative including nannies and au pairs at home non-relative elsewhere). For each variable, I winsorize the top 1% of the respective age-specific distribution.

Neighborhood inputs At waves two, three, and four, the MCS includes the deciles in the country-specific Index of Multiple Deprivation (IMD) distribution of the area a respondent lived in at the time of the interview. The IMD is a measure of relative levels of deprivation at the Lower Super Output Area level. I use the sub-scales for income, employment, health, and education. For each sub-scale, the IMD ranks every area from most to least deprived in the respective country (England, Wales, Scotland, and Northern Ireland).

School inputs At wave four, the MCS administered a teacher questionnaire to class teachers that includes information about teacher quality and classroom characteristics. With respect to the former, I include information about the class teacher's highest educational degree (Bachelor degree, diploma, or Master degree), his or her teaching experience in years, the time since he or she received the teaching qualification (computed as the the survey year minus the year when he or she received the qualification), and the time he or she is already teaching at the child's school. Information about the classroom includes class size, the share of children with special educational needs, the share of children with English as a second language, the number of children excluded from class, and whether the child shares his or her class with a child that disturbs the lessons (yes or no).

B.1.4 Control Variables

Child controls I obtain information about the child's gender, birth weight, gestation period, birth rank, and ethnicity from wave one. A birth is considered preterm if the gestation period is below 37 weeks. I group the child's ethnicity as follows: White; Black (Black Caribbean, Black African, other Black); Indian, Pakistani, Bangladeshi; other Asian (Chinese, other Asian); other (mixed, other ethnic group). I obtain information about the child's age at the skill assessments and the number of siblings from waves two to six. The child's ages are transformed to months.

Parental controls I obtain information about the mother's age at the child's birth from wave one. I impute missing values by using information about the mother's age at the interviews. If there are discrepancies across waves for the imputed values for a particular mother, I use the mode value. Information about mother's age when she left full-time education comes from waves one, two, three, and six. I assign the largest value across these waves, and winsorize values below 12 and above 25 years.

Information about the number of parent figures in the household, marital status, employment status, household income, mental health, and the government region comes from waves two to six. I assign a household to be either a two-parent household with two natural parents, a two-parent household with the natural mother and a partner, or a single-mother household. I assign a mother to be either employed or not at the time when the respective interview was conducted. Weekly household incomes are equivalized using the modified OECD scales and adjusted for inflation using the consumer price index published by the Office for National Statistics. The mother's mental health was assessed using the Kessler scales, a measure of psychological distress (Kessler et al., 2003). In the MCS, the mother was asked six questions about depressive and anxiety symptoms that she has experienced in the last 30 days: about how often did she (i) feel so depressed that nothing could cheer her up, (ii) feel hopeless, (iii) feel restless or fidgety, (iv) feel that everything was an effort, (v) feel worthless, and (vi) feel nervous. The government regions where the child lived in at the time of the interview are North East, North West, Yorkshire and Humberside, East Midlands, West Midlands, East of London, London, South East, South West, Wales, Scotland, and Northern Ireland.

I obtain information about the mother's personality traits from wave four. Respondents were asked 15 questions that were used to construct a measure of extroversion and neuroticism, respectively, two of the Big Five personality traits (see Almlund et al., 2011, for a discussion of the Big Five personality traits).

Wave six includes a Word Activity assessment that was administered to the cohort member's resident parents. Respondents were presented with a list of target words, each with five other words next to them, and asked which of those words meant the same as the target word. I use the Word Activity assessment as a proxy for cognitive skills measuring knowledge of vocabulary.

B.2 Descriptive Statistics

B.2.1 Parenting Styles

Table B.5: Self-assessed Parenting Styles

	Demandingness	Responsiveness
Age 3		
Doing my best for the children	−0.04	0.05
Firm discipline plus lots of fun	0.05	0.18
Firm rules and discipline	0.22	−0.02
Have not really thought about it	0.07	−0.22
Lots of fun	−0.20	0.09
Age 5		
Doing my best for the children	0.01	−0.02
Firm discipline plus lots of fun	0.05	0.13
Firm rules and discipline	0.04	−0.24
Have not really thought about it	0.11	−0.20
Lots of fun	−0.21	0.06
Age 7		
Doing my best for the children	−0.01	−0.02
Firm discipline plus lots of fun	0.03	0.15
Firm rules and discipline	0.09	−0.12
Have not really thought about it	0.12	−0.31
Lots of fun	−0.16	0.05
Age 11		
Doing my best for the children	−0.04	0.04
Firm discipline plus lots of fun	0.01	0.08
Firm rules and discipline	0.13	−0.06
Have not really thought about it	0.06	−0.08
Lots of fun	−0.11	0.10
Age 14		
Doing my best for the children	−0.06	−0.04
Firm discipline plus lots of fun	−0.03	−0.06
Firm rules and discipline	−0.00	−0.06
Have not really thought about it	0.02	−0.14
Lots of fun	−0.15	−0.07

Note: The table shows the mean values of mothers' demandingness and responsiveness by self-assessed parenting style. Self-assessed parenting styles are obtained in wave two (child age 3). Demandingness and responsiveness are standardized (mean zero and standard deviation one). Data: Millennium Cohort Study. Own calculations.

Table B.6: Persistency of Parenting Styles

Child age (a)	$cor(PS^D_a, PS^D_{a-1})$	$cor(PS^R_a, PS^R_{a-1})$
Age 3	-	-
Age 5	0.56	0.17
Age 7	0.65	0.32
Age 11	0.55	0.32
Age 14	0.21	0.10

Note: The table shows the correlations between parental demandingness (PS^D_a) and responsiveness (PS^R_a), respectively, of a parent at child age a and age $a-1$. $a-1$ refers to the child age of the MCS wave prior to child age a. The values are computed on the full sample. Data: Millennium Cohort Study. Own calculations.

Table B.7: Punitive Parenting Factor Loadings

	Nonrestrictive	Restrictive	Corporal
Age 3			
Ignores child	0.61	0.25	−0.04
Smacks child	−0.14	−0.05	0.85
Shouts at child	0.14	0.02	0.77
Sends child to room	−0.07	0.84	0.04
Takes away treats from child	0.04	0.80	0.01
Tells child off	0.18	0.27	0.53
Bribes child	0.88	−0.18	0.06
Age 5			
Ignores child	0.34	0.26	0.08
Smacks child	−0.14	0.00	0.91
Shouts at child	0.31	0.11	0.61
Sends child to room	−0.09	0.84	0.09
Takes away treats from child	0.03	0.82	−0.05
Tells child off	0.42	0.36	0.28
Bribes child	0.76	−0.24	0.14
Reasons with child	0.75	0.17	−0.21
Age 7			
Ignores child	0.39	0.20	0.11
Smacks child	−0.13	−0.04	0.93
Shouts at child	0.27	0.14	0.61
Sends child to room	−0.07	0.84	0.08
Takes away treats from child	0.02	0.85	−0.04
Tells child off	0.35	0.41	0.28
Bribes child	0.83	−0.25	0.06
Reasons with child	0.68	0.23	−0.15
Age 11			
Sends child to room	−0.03	0.93	
Takes away treats from child	0.04	0.89	
Reasons with child	1.00	0.00	

Note: The table shows the rotated factor loadings for waves two to five (child ages 3, 5, 7, and 11, respectively). For child age 3, nonrestrictive refers to the third principal component; for child age 5, it refers to the first principal component; and for child ages 7 and 11, it refers to the second principal component. For child ages 3 and 5, restrictive refers to the second principal component; for child ages 7 and 11, it refers to the first principal component. For child age 3, corporal refers to the first principal component; for child ages 5 and 7, it refers to the third principal component. Data: Millennium Cohort Study. Own calculations.

Table B.8: Parenting Styles and Parental Behavior

	Telling child off	Expressing affection
PS_a^D	1.16^{***}	
	(0.01)	
PS_a^R		1.48^{***}
		(0.02)
Adj. R^2	0.40	0.43
Num. obs.	12,407	12,407

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Note: The table show the estimates from a regression of the number of days per week the mother tells her child off (first column) and expresses her affection (second column) on parental demandingness and responsiveness, respectively. The categorical values are transformed to numerical values as follows: "never" = 0, "rarely" = 0.25, "sometimes (about once a month)" = 0.5, "often (about once a week or more)" = 2, "daily" = 7. The model is estimated on a sample from wave four only (child age 7). Data: Millennium Cohort Study. Own calculations.

B.2.2 Inputs and Control Variables

Table B.9: Descriptive Statistics for Other Input Measures

	Mean	S.D.	Min.	Max.	N
	Time-varying Inputs				
Household Inputs					
Childcare services	1.78	5.67	0.00	45.00	8,301
Grandparents	2.34	5.23	0.00	35.00	8,301
Others	1.61	6.02	0.00	48.00	8,301
Neighborhood Inputs					
Overall	5.93	2.83	1.00	10.00	8,301
Income	5.88	2.84	1.00	10.00	8,301
Employment	5.94	2.78	1.00	10.00	8,301
Health	5.95	2.79	1.00	10.00	8,301
Education	5.72	2.91	1.00	10.00	8,301
	Time-invariant Inputs				
School Inputs: Teacher Characteristics					
Years since teacher qualification	15.61	11.58	0.00	44.00	1,335
Teaching experience (years)	13.92	9.88	1.00	40.00	1,335
Tenure at school (years)	8.80	7.47	1.00	36.00	1,335
Highest Degree					
Bachelor	0.48	0.50	0.00	1.00	1,335
Diploma	0.13	0.34	0.00	1.00	1,335
Master	0.38	0.49	0.00	1.00	1,335
School Inputs: Class Characteristics					
Class size	25.74	4.98	1.00	37.00	1,335
# Students w/ SEN statement	1.53	2.41	0.00	17.00	1,335
# Students excluded from school	0.06	0.44	0.00	12.00	1,335
# Students w/ English as second language	1.69	3.73	0.00	30.00	1,335
Student preventing others from learning (y/n)	0.31	0.46	0.00	1.00	1,335
Regular support (number of supporters)	1.77	0.95	0.00	6.00	1,335

Note: The table shows descriptive statistics for the input measures (mean, standard deviation, minimum and maximum values, and number of observations; all unweighted). The values for the time-varying variables are computed on the pooled estimation sample across child ages 3, 5, and 7. The time-invariant variables are measured at child age 7. The household input measures are winsorized at the top 1% of the respective age-specific distribution. Data: Millennium Cohort Study. Own calculations.

Table B.10: Parenting Styles by Observables

	Demandingness	Responsiveness	N
Age left Full-time Education			
Before 16	−0.03	−0.07	2,921
Between 16 and 20	−0.02	0.02	36,188
After 20	0.02	0.12	10,174
Parental Income			
First decile	−0.03	−0.10	9,857
Second decile	−0.01	−0.04	9,857
Third decile	0.02	0.05	9,857
Fourth decile	−0.03	0.11	9,856
Fifth decile	−0.01	0.17	9,856
Age at Birth			
18–21 years	0.12	−0.10	5,745
22–27 years	0.07	0.01	12,591
28–33 years	−0.02	0.07	19,904
34–39 years	−0.15	0.07	9,923
40+ years	−0.32	0.06	1,120
Birth Rank			
First born	0.07	0.06	20,392
Second born	−0.01	0.05	18,023
Third+ born	−0.17	−0.02	10,868
Parental Investment			
First decile	−0.01	−0.16	9,857
Second decile	0.02	−0.05	9,857
Third decile	−0.01	0.07	9,857
Fourth decile	−0.00	0.12	9,856
Fifth decile	−0.05	0.21	9,856

Note: The table shows mean values for parental demandingness and responsiveness, respectively, by a selection of observables. The values are computed on the pooled sample across child ages 3, 5, 7, 11, and 14. Data: Millennium Cohort Study. Own calculations.

B.3 Results

B.3.1 Results

Table B.11: Benchmark Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
			Cognitive Skills			
PS_a^D	-0.05^{***}	-0.04^{***}	-0.04^{***}	-0.04^{***}	-0.04^{***}	-0.04^{**}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
PS_a^R	0.03^{***}	0.03^{**}	0.02^{**}	0.02^{**}	0.02^{**}	0.05^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
$PS_a^D \times PS_a^R$	0.00	0.00	0.00	0.00	0.00	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
y_{a-1}	0.21^{***}	0.21^{***}	0.20^{***}	0.20^{***}	0.19^{***}	0.19^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Adj. R^2	0.16	0.16	0.16	0.16	0.17	0.19
Num. obs.	8,301	8,301	8,301	8,301	8,301	4,005
			Non-cognitive Skills			
PS_a^D	-0.11^{***}	-0.11^{***}	-0.10^{***}	-0.10^{***}	-0.10^{***}	-0.11^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
PS_a^R	0.05^{***}	0.05^{***}	0.05^{***}	0.05^{***}	0.05^{***}	0.05^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$PS_a^D \times PS_a^R$	0.02^{**}	0.02^{**}	0.02^{**}	0.02^{**}	0.02^{**}	0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
y_{a-1}	0.58^{***}	0.58^{***}	0.57^{***}	0.57^{***}	0.56^{***}	0.53^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
PS_{a-1}		x	x	x	x	x
PS_{a-2}			x	x	x	x
Househ. inputs				x	x	x
Neighb. inputs					x	x
School inputs						x
Adj. R^2	0.50	0.50	0.50	0.50	0.50	0.50
Num. obs.	8,301	8,301	8,301	8,301	8,301	4,005

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Note: The table shows the estimates for the cognitive (upper panel) and non-cognitive (lower panel) production function (robust standard errors in parentheses). PS_a^D and PS_a^R denote contemporaneous parental demandingness and responsiveness, respectively. Each column includes the full set of child and mother controls. Column (2) additionally controls for the lagged inputs, and columns (3) to (6) control for the double-lagged inputs. Columns (4), (5), and (6) also control for household inputs, neighborhood inputs, and school inputs, respectively. The models are estimated on a pooled sample across child ages 7, 11, and 14, and include year fixed-effects. Data: Millennium Cohort Study. Own calculations.

Table B.12: Benchmark Estimates: Cognitive Skills

	(1)	(2)	(3)	(4)	(5)	(6)
Age 3	Outcome: Naming Vocabulary					
PS_a^D	0.02			0.02	0.02	
	(0.02)			(0.02)	(0.02)	
PS_a^R	0.12***			0.12***	0.12***	
	(0.03)			(0.03)	(0.03)	
$PS_a^D \times PS_a^R$	−0.01			−0.00	−0.00	
	(0.03)			(0.03)	(0.03)	
y_{a-1}						
Adj. R^2	0.16			0.16	0.17	
Num. obs.	2,767			2,767	2,767	
Age 5	Outcome: Naming Vocabulary					
PS_a^D	−0.02	−0.01		−0.01	−0.00	
	(0.02)	(0.02)		(0.02)	(0.02)	
PS_a^R	0.00	0.00		0.00	0.01	
	(0.02)	(0.02)		(0.02)	(0.02)	
$PS_a^D \times PS_a^R$	−0.01	−0.01		−0.01	−0.00	
	(0.02)	(0.02)		(0.02)	(0.02)	
y_{a-1}	0.35***	0.35***		0.35***	0.34***	
	(0.02)	(0.02)		(0.02)	(0.02)	
Adj. R^2	0.27	0.28		0.28	0.28	
Num. obs.	2,767	2,767		2,767	2,767	
Age 7	Outcome: Word Reading					
PS_a^D	−0.06***	−0.04*	−0.05*	−0.05*	−0.05**	−0.06*
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
PS_a^R	0.03	0.02	0.01	0.01	0.01	0.03
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
$PS_a^D \times PS_a^R$	−0.01	−0.01	−0.01	−0.01	−0.01	−0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
y_{a-1}	0.23***	0.23***	0.22***	0.23***	0.22***	0.25***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Adj. R^2	0.23	0.23	0.23	0.23	0.24	0.26
Num. obs.	2,767	2,767	2,767	2,767	2,767	1,335

Age 11	Outcome: Verbal Similarities					
PS_a^D	-0.04^{**}	-0.04	-0.03	-0.03	-0.04	-0.03
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)
PS_a^R	0.05^{***}	0.04^{**}	0.03^{*}	0.03^{*}	0.03^{*}	0.08^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
$PS_a^D \times PS_a^R$	-0.00	-0.00	-0.00	-0.00	0.00	0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
y_{a-1}	0.16^{***}	0.16^{***}	0.16^{***}	0.16^{***}	0.16^{***}	0.18^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Adj. R^2	0.15	0.15	0.15	0.16	0.16	0.18
Num. obs.	2,767	2,767	2,767	2,767	2,767	1,335

Age 14	Outcome: Vocabulary Test					
PS_a^D	-0.04^{**}	-0.03^{*}	-0.04^{*}	-0.04^{*}	-0.04^{**}	-0.04
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
PS_a^R	0.03	0.02	0.02	0.02	0.02	0.05
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
$PS_a^D \times PS_a^R$	0.00	0.00	0.00	0.00	0.00	-0.02
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
y_{a-1}	0.29^{***}	0.28^{***}	0.27^{***}	0.27^{***}	0.27^{***}	0.28^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)
PS_{a-1}		x	x	x	x	x
PS_{a-2}			x	x	x	x
Househ. inp.				x	x	x
Neighb. inp.					x	x
School inp.						x
Adj. R^2	0.18	0.19	0.19	0.19	0.20	0.22
Num. obs.	2,767	2,767	2,767	2,767	2,767	1,335

Note: The table shows the estimates for the cognitive production functions (robust standard errors in parentheses). PS_a^D and PS_a^R denote contemporaneous parental demandingness and responsiveness, respectively. Each column includes the full set of child and mother controls. Column (2) additionally controls for the lagged inputs, and columns (3) to (6) control for the double-lagged inputs. Columns (4), (5), and (6) also control for household inputs, neighborhood inputs, and school inputs, respectively. Data: Millennium Cohort Study. Own calculations.

Table B.13: Benchmark Estimates: Non-cognitive Skills

	(1)	(2)	(3)	(4)	(5)	(6)
Age 3	Outcome: Strengths and Difficulties Questionnaire					
PS_a^D	-0.26^{***}			-0.26^{***}	-0.27^{***}	
	(0.02)			(0.02)	(0.02)	
PS_a^R	0.26^{***}			0.26^{***}	0.25^{***}	
	(0.02)			(0.02)	(0.02)	
$PS_a^D \times PS_a^R$	0.01			0.01	0.01	
	(0.03)			(0.03)	(0.03)	
y_{a-1}						
Adj. R^2	0.30			0.30	0.31	
Num. obs.	2,767			2,767	2,767	
Age 5	Outcome: Strengths and Difficulties Questionnaire					
PS_a^D	-0.19^{***}	-0.22^{***}		-0.22^{***}	-0.22^{***}	
	(0.02)	(0.02)		(0.02)	(0.02)	
PS_a^R	0.05^{***}	0.04^{***}		0.04^{***}	0.04^{**}	
	(0.02)	(0.02)		(0.02)	(0.02)	
$PS_a^D \times PS_a^R$	0.02	0.01		0.02	0.01	
	(0.02)	(0.02)		(0.02)	(0.02)	
y_{a-1}	0.40^{***}	0.41^{***}		0.41^{***}	0.40^{***}	
	(0.02)	(0.02)		(0.02)	(0.02)	
Adj. R^2	0.42	0.43		0.43	0.43	
Num. obs.	2,767	2,767		2,767	2,767	
Age 7	Outcome: Strengths and Difficulties Questionnaire					
PS_a^D	-0.14^{***}	-0.19^{***}	-0.20^{***}	-0.20^{***}	-0.21^{***}	-0.19^{***}
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
PS_a^R	0.05^{***}	0.06^{***}	0.06^{***}	0.06^{***}	0.07^{***}	0.06^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
$PS_a^D \times PS_a^R$	0.02	0.02	0.02	0.02	0.02	0.02
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
y_{a-1}	0.54^{***}	0.56^{***}	0.56^{***}	0.56^{***}	0.55^{***}	0.53^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Adj. R^2	0.52	0.53	0.53	0.53	0.53	0.52
Num. obs.	2,767	2,767	2,767	2,767	2,767	1,335

Age 11	Outcome: Strengths and Difficulties Questionnaire					
PS_a^D	-0.18^{***}	-0.19^{***}	-0.18^{***}	-0.18^{***}	-0.17^{***}	-0.21^{***}
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
PS_a^R	0.03^{**}	0.03^{**}	0.03^{**}	0.03^{**}	0.03^{*}	0.03
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
$PS_a^D \times PS_a^R$	0.03^{*}	0.03	0.03^{*}	0.03^{*}	0.04^{**}	0.06^{**}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
y_{a-1}	0.55^{***}	0.54^{***}	0.54^{***}	0.54^{***}	0.54^{***}	0.55^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Adj. R^2	0.51	0.51	0.51	0.51	0.51	0.52
Num. obs.	2,767	2,767	2,767	2,767	2,767	1,335

Age 14	Outcome: Strengths and Difficulties Questionnaire					
PS_a^D	-0.03^{**}	-0.02^{*}	-0.02^{*}	-0.02^{*}	-0.02^{*}	-0.03^{*}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
PS_a^R	0.06^{***}	0.06^{***}	0.06^{***}	0.06^{***}	0.06^{***}	0.07^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
$PS_a^D \times PS_a^R$	0.02	0.02	0.02	0.02	0.01	-0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
y_{a-1}	0.63^{***}	0.60^{***}	0.60^{***}	0.60^{***}	0.60^{***}	0.60^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
PS_{a-1}		x	x	x	x	x
PS_{a-2}			x	x	x	x
Househ. inp.				x	x	x
Neighb. inp.					x	x
School inp.						x
Adj. R^2	0.51	0.51	0.51	0.51	0.51	0.50
Num. obs.	2,767	2,767	2,767	2,767	2,767	1,335

Note: The table shows the estimates for the non-cognitive production functions (robust standard errors in parentheses). PS_a^D and PS_a^R denote contemporaneous parental demandingness and responsiveness, respectively. Each column includes the full set of child and mother controls. Column (2) additionally controls for the lagged inputs, and columns (3) to (6) control for the double-lagged inputs. Columns (4), (5), and (6) also control for household inputs, neighborhood inputs, and school inputs, respectively. Data: Millennium Cohort Study. Own calculations.

Chapter 3

Genetic Endowments, Educational Outcomes, and the Mediating Influence of School Investments*

3.1 Introduction

Education is a core determinant of life outcomes, both for individuals and societies at large (Acemoglu and Autor, 2011; Hanushek and Woessmann, 2008; Krueger and Lindahl, 2001). Hence, improving equity and efficiency in education systems is a central policy goal in modern societies. To achieve such improvements, it is important to understand the role of genetic endowments for educational attainment: on the one hand, genetic endowments are strong predictors of education; in heritability studies they account for 40% of the variation in years of education (Branigan et al., 2013; Lee et al., 2018). On the other hand, the importance of genetic endowments varies with social environments like families, neighborhoods, and schools (Cesarini and Visscher, 2017; Koellinger and Harden, 2018). Therefore, the link between genetic endowments and life outcomes may be modified by policy interventions. This observation raises important questions: can school reforms moderate the link between genetic endowments and educational outcomes? If yes, which domains of school environments are particularly effective in doing so? Answers to these questions are of utmost importance to address equity and efficiency concerns in the production of educational attainment. Despite this importance, current evidence is scant.

In this paper, we study the interaction of genetic endowments and school environments in the production of educational attainment. We focus on two dimensions of school environments that have been studied extensively in the literature on education economics:

*This chapter is based on joint work with Benjamin Arold and Paul Hufe.

teacher quality and class size (Angrist and Lavy, 1999; Angrist et al., 2019; Chetty et al., 2014a,b; Fredriksson et al., 2013; Jackson, 2019; Leuven and Løkken, 2020; Rivkin et al., 2005; Rockoff, 2004). Furthermore, these dimensions can be directly influenced by policymakers, but their reform applies to all children and does not presuppose any form of genetic screening—a practice many of us would be uncomfortable with (Martschenko et al., 2018).

We use data from the National Longitudinal Study of Adolescent to Adult Health (Add Health) to study the interaction of genetic endowments and school environments in a between-family design. Add Health is a 5-wave panel study that follows a representative sample of U.S. high school students from 1994/95 until the present day. To the best of our knowledge, Add Health is the only (publicly available) data set that offers detailed information on schooling environments from both survey and administrative sources for a genotyped sample of reasonable size.

To measure genetic endowments, we leverage recent advances in molecular biology and use a polygenic score for educational attainment ($\mathrm{PGS}^{\mathrm{EA}}$, Dudbridge, 2013; Lee et al., 2018). $\mathrm{PGS}^{\mathrm{EA}}$ is an individual measure for the genetic propensity to attain education.[1] The score is fixed at conception and cannot be modified by environmental interventions thereafter. Therefore, $\mathrm{PGS}^{\mathrm{EA}}$ confers important advantages over traditional proxies for "innate ability", such as student test scores and IQ tests (Brinch and Galloway, 2012; Hanushek and Woessmann, 2008, 2012; Heckman et al., 2010). To measure the quality of school environments, we use information from headmaster surveys and administrative data sources such as the Common Core of Data, and conduct a principal component analysis on the following school-level characteristics: teacher experience, teacher turnover, teacher education, teacher diversity as well as class sizes and student-teacher ratios. From this analysis, we extract two factors that are indicative for the quality of teachers ($\mathrm{I}_{\mathrm{Qual}}$) and the quantity of teachers relative to the number of students ($\mathrm{I}_{\mathrm{Quant}}$), respectively.

Clean causal identification of gene-environment interactions is challenging. In this study, we rely on a between-family comparison in which we control for an extensive set of pre-determined family background characteristics. We discuss the associated identification assumptions in detail and provide tests for their satisfaction. First, while genetic endowments are fixed at conception, they are correlated with other family characteristics that co-determine educational attainment. Therefore, our parameters of interest may be confounded by *genetic nurture effects*. In response, we show that the relevant point estimates from the between-family design replicate in a smaller sibling sample that allows us to control for genetic nurture by including family fixed effects. Second, school characteristics may be correlated with other family characteristics that co-determine educational

[1] In addition, $\mathrm{PGS}^{\mathrm{EA}}$ has been shown to be highly predictive for a number of life outcomes that are closely related to educational attainment. These outcomes include earnings, wealth, and (non-)cognitive skills (Barth et al., 2020; Buser et al., 2021a; Demange et al., 2021; Houmark et al., 2020; Lee et al., 2018; Muslimova et al., 2020; Papageorge and Thom, 2020).

attainment. Therefore, our parameters of interest may be confounded by *selection effects*. In response, we show that our main findings are robust to bounding exercises à la Cinelli and Hazlett (2020) and Oster (2019). Lastly, gene-environment interactions can only be identified if genetic endowments and the environmental variable of interest are distributed independently of each other. In response, we show that we cannot reject the equality of $\mathrm{PGS}^{\mathrm{EA}}$ distributions in various school environments. In summary, although the between-family design does not allow to cleanly identify causal effects, all our empirical tests point towards the satisfaction of the relevant identification requirements. In addition, our results withstand a series of empirical tests for competing mechanisms that we discuss in detail below.

Our results can be summarized as follows. First, genetic endowments and teacher quality are highly predictive for years of education: a one-standard-deviation increase in $\mathrm{PGS}^{\mathrm{EA}}$ (teacher quality) increases educational attainment by ≈ 0.37 (0.22) years. These increases can be compared to a sample average of 14.81 years and correspond to 16.44% (9.8%) of a standard deviation. Second, genetic endowments and teacher quality act as substitutes in the production of educational attainment: a one-standard-deviation increase in teacher quality reduces the positive association of educational attainment with $\mathrm{PGS}^{\mathrm{EA}}$ by $\approx 19\%$. This result implies that improvements in the quality of teachers may reduce the genetic gradient in educational attainment. Furthermore, it suggests that teacher quality may countervail the effects of family socio-economic status—an environmental characteristic that tends to magnify the genetic gradient in educational attainment (Papageorge and Thom, 2020; Ronda et al., forthcoming).[2] Third, in contrast to teacher quality, teacher quantity is not associated with educational attainment—a null result that does not vary across the $\mathrm{PGS}^{\mathrm{EA}}$ distribution.

We perform a series of robustness checks to evaluate whether our results are conflated by competing mechanisms. We begin by showing that our measures for teacher quality and quantity do not pick up the effects of other school characteristics that may correlate with student outcomes. These characteristics comprise school peer characteristics, school-level policies such as sanctions for academic misconduct, and overall school value-added. Next, we demonstrate that our results are not driven by gene-environment interactions that reflect family instead of school environments. To that end, we run a fully interacted model controlling for all possible interactions between $\mathrm{PGS}^{\mathrm{EA}}$, $\mathrm{I}_{\mathrm{Qual}}$, $\mathrm{I}_{\mathrm{Quant}}$, and a broad set of parental background characteristics (Keller, 2014). In addition, we show that there is no differential association between $\mathrm{PGS}^{\mathrm{EA}}$ and parental time investments depending on school quality.

We also analyze the mechanisms that underpin the substitutability of genetic endowments and teacher quality. Educational attainment summarizes information from various educational stages, where each stage requires a different mix of skills (Cunha et al., 2006, 2010). Therefore, we repeat our analysis by replacing total educational attainment with

[2]See also our replication of their findings in section 3.5.

binary variables for whether or not respondents achieved a given educational degree. We find that the substitutability of genetic endowments and teacher quality is largest at the stage of college education. In contrast, we find little substitutability for the probability of graduating from high school, and none for the probability of obtaining post-graduate degrees. These results provide a notable contrast to Papageorge and Thom (2020), who find a growing complementarity of parental background characteristics and genetic endowments as individuals progress through the educational system. To uncover which type of skills drives our results, we analyze the associations of PGS^{EA} and teacher quality with a set of intermediate outcomes including subjective and objective health, cognitive skills, economic preferences, and personality measures. We find that substitutabilities of genetic endowments and teacher quality with respect to subjective health, verbal intelligence, risk-aversion, and patience underpin our main result.

Our study contributes to three strands of literature. First, we contribute to the literature on gene-environment interactions. Existing evidence shows that the association between socio-economic outcomes and genetic endowments varies with the socio-economic status of parents (Houmark et al., 2020; Papageorge and Thom, 2020; Ronda et al., forthcoming). Evidence on gene-environment interactions regarding school environments is more scant. Barcellos et al. (2021) use a compulsory schooling reform to show that returns to schooling are lower for genetically advantaged students. However, they focus on the length of education and not the quality of school environments. Trejo et al. (2018) show a stronger genetic gradient in schools with better educated parents. However, the composition of schools is difficult to control in the presence of endogenous sorting. Therefore, we focus on margins that can be directly targeted by policymakers: the quality and quantity of teachers.

Second, we contribute to the literature on teacher quality. The positive effects of teacher quality on short- and long-term outcomes of students are well-documented (Chetty et al., 2014a,b; Jackson, 2019; Rivkin et al., 2005; Rockoff, 2004). However, the literature is far less conclusive about the equalizing effect of teacher quality across student subgroups. For example, Aaronson et al. (2007) find that low-achieving students benefit more from high-quality teachers. In contrast, Chetty et al. (2014b) show that students from minority and low-income backgrounds benefit less. While existing studies have evaluated heterogeneities along dimensions that conflate genetic and social factors, we are able to measure the genetic predisposition for educational success as fixed at conception. We show that investments in the quality of teachers cushion the genetic gradient in educational attainment.

Third, we contribute to the literature on class size. Here, the average effects on students outcomes are subject to academic debate. On the one hand, experimental studies on class size reductions tend to find positive effects on student achievement (Chetty et al., 2011; Krueger, 1999). On the other hand, quasi-experimental analyses exploiting maximum class-size rules tend to find mixed results even if they analyse similar settings (Angrist and Lavy, 1999; Angrist et al., 2019; Fredriksson et al., 2013; Leuven and Løkken, 2020).

Similarly, the equalizing effects of class size reductions are contested. For example, Krueger (1999) shows that class size reductions are more beneficial to students from minority and low-income background. In contrast, Fredriksson et al. (2013) document that wage increases following class size reductions are more pronounced for students from high income backgrounds. Our study is the first to evaluate heterogeneities along the genetic dimension. We show that teacher quantity is not associated with gains in educational attainment, irrespective of genetic endowment.

Our results are policy relevant. First, we show that higher-quality teachers are conducive to educational attainment in the lower tail of the PGS^{EA} distribution but do not compromise achievement in the upper tail. This finding suggests that policymakers do not face an equity-efficiency trade-off when investing into the quality of teachers. Second, in contrast to teacher quality, we find no effect of teacher quantity on the educational outcomes of students, irrespective of their genetic endowments. This finding suggests that policymakers who are willing to address the equity and efficiency concerns related to genetic endowments do not face a trade-off between investments into teacher quality and teacher quantity. This last finding is economically relevant, as salaries and employee benefits of teachers are by far the largest cost factor in the U.S. school system, accounting for about half of the expenditures in US public primary and secondary schools (Figure C.1).

The remainder of this paper is structured as follows. In section 3.2, we provide a primer on the measurement of genetic endowments. In section 3.3, we detail our empirical strategy. After introducing our data sources in section 3.4, we present results in section 3.5. Section 3.6 concludes the paper.

3.2 Measuring Genetic Endowment

The "First Law of Behavior Genetics" states that all human behavioral traits are heritable (Turkheimer, 2000). That is, genetic endowments explain the expression of any trait at least to some extent. The empirical challenge is to identify the specific sequences in the genome that are related to the traits of interest.[3] Recent advances in molecular genetics have enabled a novel method of genetic discovery: genome-wide association studies (GWAS). GWAS exploit the most common type of genetic variation between humans, so-called single-nucleotide polymorphisms (SNP). SNPs occur when a single nucleotide—the basic building block of DNA molecules—differs at a specific position in the genome. Humans have between four and five million SNPs. GWAS estimate separate linear regressions

[3]Human genetic information is stored in 23 chromosome pairs that consist of deoxyribonucleic acid (DNA) molecules. These chromosomes, in turn, contain 20,000 to 25,000 genes—specific DNA sequences that provide instructions for building particular proteins. About 99% of the sequences are identical across humans.

that relate a SNP of individual i at genome location j to an outcome of interest y:

$$y_i = \psi_j^y SNP_{ij} + \delta C_i + \varepsilon_i. \tag{3.1}$$

$SNP_{ij} \in \{0, 1, 2\}$ is a count variable and indicates the number of minor allele that individual i possesses at location j. Minor alleles are the less frequent genetic variation within a population. As humans inherit one of each chromosome from each parent, they possess either zero, one, or two minor alleles at each location j. C_i is a vector of control variables to filter out spurious correlations due to non-biological differences across population groups. A particular SNP coefficient ψ_j^y is referred to as genome-wide significant if the null hypothesis of non-association is rejected at a level of $p < 5 \times 10^{-8}$ (Chanock et al., 2007). The p-value is deliberately low to adjust for multiple hypothesis testing.

The association of any single SNP with y is minuscule, but jointly they can explain a substantial share of the observed outcome differences between individuals (Lee et al., 2018). In particular, the estimated SNP coefficients can be used to construct polygenic scores (PGS). A PGS is a single quantitative measure of an individual's genetic propensity toward an outcome relative to the population. Formally, individual i's PGS for outcome y, PGS_i^y, is constructed by linearly aggregating all SNP_{ij} using ψ_j^y as weighting factors:

$$PGS_i^y = \sum_j \widehat{\psi}_j^y SNP_{ij}, \tag{3.2}$$

where $\widehat{\psi}_j^y$ is the estimated SNP coefficient from Equation (3.1). To avoid overfitting, Equation (3.1) is estimated in a discovery sample, whereas the PGS is constructed in a hold-out sample (Wray et al., 2014).

The predictive power of a PGS is broadly determined by two factors: the heritability of the outcome, which serves as an upper bound of the variance the PGS can explain; and the size of the discovery sample (Dudbridge, 2013). All else equal, the more heritable the outcome, or the larger the discovery sample to estimate the aggregation weights $\widehat{\psi}_j^y$, the higher the predictive accuracy of the PGS. For example, the heritability of educational attainment is around 40% (Branigan et al., 2013). The PGS for educational attainment constructed by Lee et al. (2018) is based on information from 1.1 million individuals and explains 12.7% of the variance in educational attainment.

The interpretation of PGS is non-trivial. First, PGS are not purely measures of biological influence. In particular, GWAS coefficients may capture environmental factors such as population stratification across geographic regions (Abdellaoui et al., 2019). To address this concern, we follow standard practice and always control for the first 20 principal components of the genetic data in our empirical analysis.[4] Second, the explanatory power of

[4]The first principal components of the full matrix of genetic data capture most of the geographical variation in allele frequencies (see Mills et al., 2020, chapter 9.4, for a discussion). Therefore, they control for the geographic correlation between allele frequencies and socio-economic status.

PGS is contingent on the context of its application. If a PGS is applied in one context, whereas the underlying GWAS was estimated in a very different context, the predictive power of the PGS will be attenuated. In our context, this concern is limited: we apply PGS to a sample from the United States, whereas the underlying GWAS predominantly draws on samples from other industrialized countries with comparable education systems. Third, PGS are noisy measures of genetic endowments. Due to current GWAS sample sizes, they do not capture all genetic variation relevant for the outcome of interest. As a direct consequence, alternative PGS are still predictive for educational attainment over and above PGS^{EA}. However, in Appendix Table C.4, we show that PGS^{EA} is significantly more predictive than any plausible alternative PGS. Therefore, it is the best among other noisy measures for genetic endowments.

PGS are now available for a wide variety of outcomes. These include, for example, body mass index and height (Yengo et al., 2018), attention deficit hyperactivity disorder (Demontis et al., 2019), major depressive disorder (Howard et al., 2019), intelligence (Savage et al., 2018), smoking (Liu et al., 2019), and sleep duration (Jansen et al., 2019). For our analysis, we rely on the PGS for educational attainment by Lee et al. (2018).

3.3 Empirical Strategy

3.3.1 Empirical Model

Consider a model in which skills θ of child i at age a are determined by prior skill levels θ_{ia-1}, parental investments I_{ia}^{P}, school investments I_{ia}^{S}, and genetic endowments G_i.[5] There are three phases of skill accumulation:

$$\theta_{ia} = \begin{cases} f_a(G_i) & \text{, for child age } a = 0, \\ f_a(\theta_{ia-1}, I_{ia}^{P}, G_i) & \text{, for child age } a = 1, ..., 5, \\ f_a(\theta_{ia-1}, I_{ia}^{P}, I_{ia}^{S}, G_i) & \text{, for child age } a = 6, ..., A. \end{cases} \tag{3.3}$$

That is, skills at conception are determined by genetic endowments only. For child ages $a = 1, ..., 5$, i.e. prior to attending school, parents are the only source of investments into skills in this model. Parental investments include monetary investments, such as buying toys or books, but also time investments, such as playing with or talking to the child. For $a = 6, ..., A$, schools are an additional source of investments into skills. School-based investments include instruction by teachers or interactions with peers.

[5]For the sake of simplicity, we abstract from other actors in the child development process.

Furthermore, assume completed education Y to be a function of individual skills accumulated by the end of childhood at age $a = A$:

$$Y_i = g(\theta_{iA}). \tag{3.4}$$

Recursively substituting Equations (3.3) and (3.4) across child ages $a = 1, ..., A$, we obtain a model in which educational attainment is determined by initial genetic endowments, the history of family inputs, and the history of schooling inputs:

$$Y_i = h(I^P_{iA}, ..., I^P_{i1}, I^S_{iA}, ..., I^S_{i6}, G_i). \tag{3.5}$$

We are interested in the complementarity of schooling inputs and genetic endowments at a particular child age a:

$$\kappa = \frac{\partial^2 h(\overline{I^P_{ia}}, \overline{I^P_{ia-1}}, ..., \overline{I^P_{i1}}, I^S_{ia}, \overline{I^S_{ia-1}}, ..., \overline{I^S_{i6}}, G_i)}{\partial I^S_{ia} \partial G_i}. \tag{3.6}$$

If $\kappa < 0$, genetic endowments and school investments at age a are *substitutes* in the production of educational attainment. School investments are then less productive for individuals with high genetic endowments. Reversely, if $\kappa > 0$, genetic endowments and school investments at age a are *complements* in the production of educational attainment. School investments are then more productive for individuals with high genetic endowments.

In this study, we focus on school investments during high school ($14 \leq a \leq 18$). We estimate the complementarity parameter κ from a linear regression model with an interaction term:

$$Y_i = \alpha G_i + \beta I^S_{ia} + \kappa(G_i \times I^S_{ia}) + \mathbf{X}_i(a)\gamma + \epsilon_i, \tag{3.7}$$

where $\mathbf{X}_i(a)$ denotes a vector of control variables to condition on the history of family and schooling inputs until age $a = 14$.

3.3.2 Identification

Unbiased estimation of κ is based on the following set of requirements: (i) the effect of G_i is identified, (ii) the effect of I^S_{ia} is identified, and (iii) G_i and I^S_{ia} are assigned independently from each other (Almond and Mazumder, 2013; Johnson and Jackson, 2019; Nicoletti and Rabe, 2014). In the following, we will discuss each of these requirements, potential threats to their satisfaction, and how we address them in the context of this paper.

(i) Absence of genetic nurture effects. Genetic endowments are fixed at conception, yet they are not exogenous to family characteristics that co-determine educational attainment. During meiosis, genetic endowments of children are randomly drawn from the genetic pool of their biological parents.[6] As a consequence, G_i is a function of maternal and paternal genetic endowments. These parental endowments, however, may also correlate with parental investments $I^P_{i1}, ..., I^P_{ia}$. Hence, in estimating Equation (3.7), α and κ may be confounded by *genetic nurture effects* (Kong et al., 2018). Genetic nurture can be controlled either by estimating a sibling fixed effects model that relies on within-family variation in G_i only (Houmark et al., 2020; Kweon et al., 2020; Selzam et al., 2019); in a non-transmitted genes design, where one includes both maternal and paternal genetic endowments in control vector $\mathbf{X}_i(a)$; or in an adoption design, where offspring are biologically unrelated to their parents (see Demange et al., 2020, for a detailed comparison of all three approaches). All approaches, however, are very data demanding. For example, the sibling design requires the availability of both a large set of siblings and individual measurements of G_i. Therefore, it can only be applied in a very limited set of existing data sets.

In this study, we estimate a between-family model in which we use an extensive set of pre-determined family background characteristics to control for genetic nurture effects. This approach is standard in the literature and intends to approximate requirement (i) while maximizing statistical power to detect the sought-after gene-environment interaction (Domingue et al., 2020). Reassuringly, controlling for $\mathbf{X}_i(a)$ in our between-family model, we obtain a point estimate of α that replicates the corresponding estimate from a sibling fixed effects model on a subsample of our data ($N = 525$).

(ii) Absence of selection effects. Parents choose schools based on school characteristics. Therefore, the latter may not be exogenous to family characteristics that co-determine educational attainment (Altonji et al., 2005; Beuermann et al., 2018). As a consequence, I^S_{ia} is a function of observed and unobserved family and child characteristics that may correlate with parental investments $I^P_{i1}, ..., I^P_{ia}$. Hence, in estimating Equation (3.7), β and κ may be confounded by *selection effects* (Altonji et al., 2005; Altonji and Mansfield, 2018; Biasi, forthcoming). Selection into schools can be controlled in (quasi-)experimental settings, e.g. using variation based on admission lotteries (Angrist et al., 2016; Cullen et al., 2006), or the geographic design of catchment areas (Laliberté, 2021). Existing data sets that avail such variation, however, do not contain sequenced DNA data that are required to measure G_i at the individual level.

In this study, we use an extensive set of pre-determined family background characteristics to control for selection into schools based on observables. To assess potential confounding through selection based on unobservables, we calculate *bias-adjusted treatment effects*

[6] In this process, chromosomes of fathers and mothers are re-combined to produce genetically distinct offspring. Therefore, singleton children of the same parents are never genetically identical to their siblings. Furthermore, conditional on the parents' genome, the offspring's set of genes is randomly distributed.

along the lines of Cinelli and Hazlett (2020) and Oster (2019). Both procedures assume that changes in the coefficients of interest due to the introduction of observables are informative for the extent of confounding due to unobservables. Reassuringly, applying these correction methods, our results for β and κ remain qualitatively unaltered in comparison to our benchmark estimates.

(iii) Independent assignment of genetic endowments and school environments. Requirements (i) and (ii) must be combined such that G_i and I_{ia}^S are distributed independently of each other. Strong correlation between G_i and I_{ia}^S implies little overlap in the distributions of G_i at different levels of I_{ia}^S, and vice versa. As a consequence, two empirical challenges arise. First, there may not be sufficient variation to identify $\alpha, \beta,$ and κ separately from each other. Second, κ would be identified from the tails in the respective distributions. One then would always compare individuals with similar G_i that score unusually high or low in their school-quality specific distribution of genetic endowments, and vice versa. Among others, these concerns would be addressed in a setting that avails (quasi-)experimental variation in I_{ia}^S at the level of siblings from the same biological parents. However, as highlighted in our previous discussion, such a setting is unlikely to be found in existing data sources.

To verify the satisfaction of requirement (iii), we present empirical evidence that G_i and I_{ia}^S are indeed distributed independently of each other. This conclusion holds both unconditionally and controlling for $\mathbf{X}_i(a)$.

In summary: in an ideal setting, one would estimate the complementarity parameter κ by combining a sibling fixed effects model with experimental variation in school characteristics among children of the same biological parents. To date, there is no single data set that simultaneously avails genetic data at the individual level, a large set of siblings, and quasi-experimental variation in school assignment. Therefore, we approximate the conditions of such an ideal setting with the best data available to us. Our estimates of $\alpha, \beta,$ and κ do not have a strict causal interpretation. However, we demonstrate their robustness to a large battery of potential confounders including school peer effects, school sanction policies, parental time investments, and potential non-linearities of genetic effects by other individual characteristics. Furthermore, we show that our baseline estimates of genetic effects are consistent with the estimates from a sibling fixed effects model. Erring on the side of caution, we nevertheless speak of *associations* instead *causal effects* in the remainder of the paper.

3.4 Data

We use data from the National Longitudinal Study of Adolescent to Adult Health (Add Health), a 5-wave panel study that focuses on the determinants of health-related behaviors and health outcomes. Add Health is a nationally representative sample of adolescents enrolled in grades 7–12 in 1994/95. Initial information (wave 1, $N = 20,745$) was collected from a stratified sample of 80 high schools across the U.S. as well as their associated feeder schools. In addition to in-depth interviews with adolescents, questionnaires were administered to school representatives, parents, and roughly 90,000 students of the sampled schools. Follow-up in-home questionnaires were collected in 1996 (wave 2, $N = 14,738$), 2001/02 (wave 3, $N = 15,179$), and 2008/09 (wave 4, $N = 15,701$). In the most recent wave (wave 5, 2016/18, $N = 12,300$), Add Health respondents are between 33 and 43 years old.

In the following, we describe our main variables of interest. Detailed descriptions of all variables used in our analysis are disclosed in Appendix section C.3.

Outcomes. We measure educational attainment Y_i by total years of education. In each wave, respondents were asked about their highest level of education at the time of the interview. For each individual, we use the most recent information and transform education levels into years of education following the mapping suggested by Domingue et al. (2015).[7]

To analyze the mechanisms behind our headline results, we additionally use a series of measures for academic degrees, health, and (non-)cognitive skills. First, academic degrees allow us to investigate at which educational stage our results emerge. We focus on whether respondents finished high school, obtained a college degree, or obtained a post-graduate degree. Second, measures for health and (non-)cognitive skills serve as proxy variables for θ_{iA} and allow us to analyze the dimensions of skill development that drive the main findings on educational attainment. We proxy health by quality-adjusted life years (QALY) that we derive from self-assessed health measures as well as a summary index of diagnosed health conditions. We proxy cognitive skills by the Picture Vocabulary Test (PVT), a test for receptive hearing vocabulary that is a widely-used proxy for verbal ability and scholastic aptitude. We proxy non-cognitive skills by self-reported measures of general risk aversion and patience (Falk et al., 2018) as well as self-reported information on the Big Five personality traits (Almlund et al., 2011).

[7]Numeric values in parentheses: eighth grade or less (8), some high school (10), high school graduate (12), GED (12), some vocational/technical training (13), some community college (14), some college (14), completed vocational/technical training (14), associate or junior college degree (14), completed college (16), some graduate school (17), completed a master's degree (18), some post-baccalaureate professional education (18), some graduate training beyond a master's degree (19), completed post-baccalaureate professional education (19), completed a doctoral degree (20).

Genetic endowments. Add Health obtained saliva samples from consenting participants in wave 4. After quality control procedures, genotyped data is available for 9,974 individuals and 609,130 SNPs. Add Health uses this data to calculate different PGS using summary statistics from existing GWAS. We use a PGS for educational attainment, denoted by PGS^{EA}, that is based on the GWAS of Lee et al. (2018).[8]

Lee et al. (2018) perform a meta-analysis of 71 quality-controlled cohort-level GWAS. Their meta-analysis produced association statistics for around 10 million SNPs, of which 1,271 reached genome-wide significance. Genes near these genome-wide significant SNPs are relevant for the central nervous system, and many of them encode proteins that carry out neurophysiological functions such as neurotransmitter secretion or synaptic plasticity. They are relevant for brain-development processes prior to and after birth.

PGS^{EA} is highly predictive for educational attainment and has been widely used in existing studies. Lee et al. (2018) suggest that PGS^{EA} is a better predictor for years of education than household income. Including the score in a regression of years of education on a set of controls yields an incremental R^2 of 0.127 in the Add Health sample. Among others, PGS^{EA} has been used to study the formation of early childhood skills (Belsky et al., 2016), educational attainment (Domingue et al., 2015; Houmark et al., 2020), earnings (Papageorge and Thom, 2020), wealth accumulation (Barth et al., 2020), and social mobility (Belsky et al., 2018).

We standardize PGS^{EA} on our analysis sample to have a mean of zero ($\mu = 0$) and a standard deviation of one ($\sigma = 1$).

School investments. In wave 1 and 2, Add Health administered detailed questionnaires to headmasters of Add Health schools. The schools are also linked to administrative data from the Common Core of Data (CCD) and the Private School Survey (PSS). We use these sources to construct indicators for I^S_{ia} using a principal component analysis that includes the following school-level information: (i) average class size, (ii) average student-teacher ratio, (iii) share of teachers with a master degree, (iv) share of new teachers in the current school year, (v) share of teachers with school-specific tenure of more than five years, and Herfindahl indices measuring teacher diversity with respect to (vi) race and (vii) Hispanic background.

Many of these characteristics have been shown to predict teacher value-added. For example, Hanushek et al. (2016) and Ronfeldt et al. (2013) show that a high teacher turnover, which we proxy by the share of new teachers, harms the quality of instruction and student achievement. Papay and Kraft (2015) and Rockoff (2004) show that teaching

[8]Lee et al. (2018) construct PGS^{EA} for two prediction cohorts, Add Health and the Health and Retirement Study (HRS). PGS^{EA} is based on results from the meta-analysis that excluded these two cohorts from the discovery sample. PGS^{EA} was generated from HapMap3 SNPs using the software LDpred—a Bayesian method that weights each SNP by the posterior mean of its conditional effect given other SNPs.

experience, which we proxy by the share of teachers with more than five years of tenure, correlates with teacher performance.[9] Finally, Clotfelter et al. (2010) and Jacob et al. (2018) show that academic credentials, which we proxy by the share of teachers with a master degree, are positively associated with teacher effectiveness.

Figure 3.1 shows the rotated loadings on the first two principal components. The first component loads almost exclusively on average class size and average student-teacher ratio. Hence, we interpret this component as an indicator for the "quantity" of teachers, denoted by $\mathrm{I}_{\mathrm{Quant}}$. The second component loads positively on the percentage of teachers with a master degree and the share of teachers with a tenure of more than five years; it loads negatively on the share of new teachers in the current school year. We interpret this component as an indicator for the "quality" of teachers, denoted by $\mathrm{I}_{\mathrm{Qual}}$. Both factors are coded such that higher values indicate higher school investments, i.e. higher teacher "quantity" investments (smaller classes) and higher teacher "quality" investments (better teachers), respectively. The calculated factors are orthogonal to each other by construction and standardized to $\mu = 0$ and $\sigma = 1$.[10]

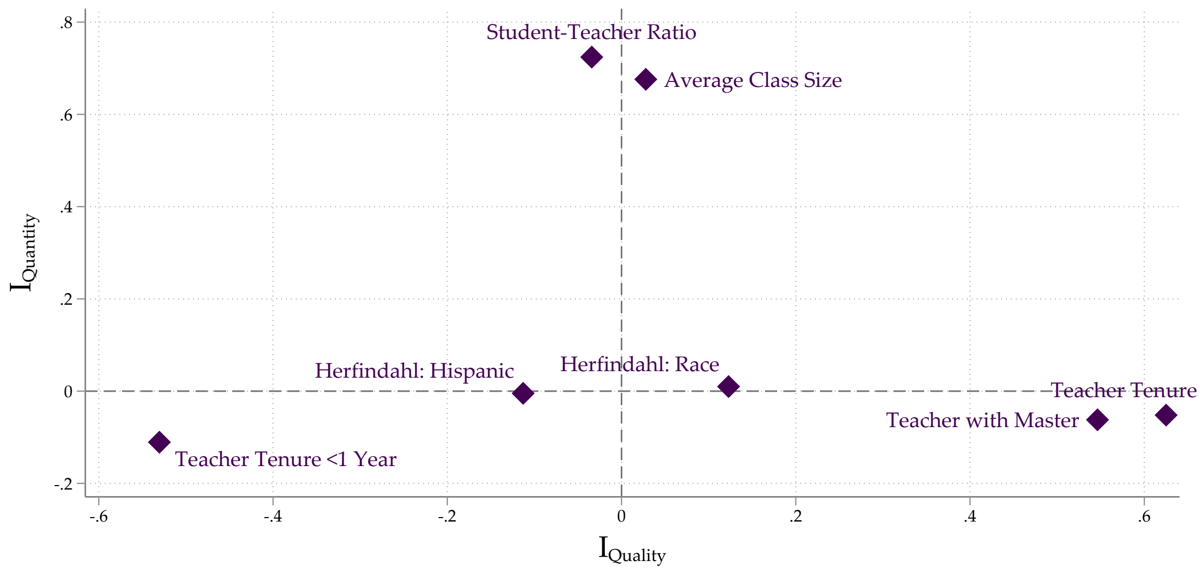
Control variables. Add Health provides extensive information about the environments that respondents were exposed to during childhood. We approximate the identification prerequisites discussed in section 3.3 by choosing a vector of pre-determined variables $\mathbf{X}_i(a)$ to control for genetic nurture effects and selection into schools. Specifically, we control for family background characteristics by including maternal and paternal education (in years), the family's religious affiliation (Christian/Non-Christian), and maternal age at birth (in years). Furthermore, we include the mean and standard deviation of potential wages for both mother and father across child ages 0–14.[11] At the level of children, we control for age in months, sex as well as their interaction. We follow standard practice in the literature and account for population stratification in genetic endowments by including the first 20 principal components of the full matrix of genetic data. Lastly, all estimations include a vector of state fixed effects.

[9] These teachers have taught for at least five years in their life and hence do not suffer from a lack of basic teaching experience. Since we measure tenure at the current school, the measure combines information about teaching experience with information about teacher turnover.

[10] Intuitively, one may expect a negative correlation between teacher quality and quantity: conditional on a given budget, a school administrator may prefer to invest in teacher quality at the expense of average class sizes or vice versa. However, this is not what we observe in the data. If quality and quantity were substitutes, we would expect loadings on the two principal components to pull into diametrically opposed directions. To the contrary, we find that the variables capturing the quality and quantity dimensions are orthogonal to each other and almost exclusively load on one principal component only.

[11] Note that Add Health contains information on actual income. However, actual income may be a bad control as it reflects parental responses to both G_i and I^S_{ia}. Therefore, we follow the procedure of Shenhav (2021) and combine the 1970 Census and the March Current Population Survey (1975–2000) to construct potential wages for gender/education/census region/race/ethnicity cells and match these potential wages to parents at each child age $a = 1, ..., 14$.

Figure 3.1: Rotated Loadings on Factors for School Characteristics



Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This figure shows the rotated factor loadings on I_{Qual} and I_{Quant}. The principal component analysis is conducted using the following school-level information: (i) average class size, (ii) average student-teacher ratio, (iii) share of teachers with a Master degree, (iv) share of new teachers in the current school year, (v) share of teachers with school-specific tenure of more than five years, and Herfindahl indices to measure teacher diversity with respect to (vi) race and (vii) Hispanic background.

Note that we focus on pre-determined variables—variables that are fixed prior to the period of observation—to avoid smearing through "bad controls" (Angrist and Pischke, 2009). However, in robustness analyses, we expand the vector of controls by potentially endogenous parental time investments and family income. Our results remain unaffected.

Analysis sample. We apply the following sample selection criteria. First, we restrict our sample to genotyped respondents of European descent.[12] This is common practice in the literature because GWAS are predominantly conducted on this ancestry group. As a consequence, there is a lack of statistical power to account for population stratification between ancestry groups and estimates of genetic influence would be biased without this restriction (Martin et al., 2017; Ware et al., 2017).

Second, we retain the subsample of individuals who (i) visited an Add Health high school or an associated feeder school in wave 1, and (ii) for whom the high school exit record indicates that they had graduated from the same school. These sample selection criteria strike a balance between sample size and the matching accuracy of individuals

[12] Ancestry groups in Add Health are identified by principal component analysis on all unrelated members of the full Add Health genotyped sample.

with our measures for schooling environments. For example, imposing criterion (i), we assume that individuals indeed transition from feeder schools to designated Add Health schools. Thereby, we increase our sample but may erroneously assign information on I^S_{ia} to individuals transitioning to high schools out of the Add Health universe. Reversely, imposing criterion (ii), we exclude individuals that may have moved to other high schools throughout grades 9–12. Thereby, we reduce our sample size but minimize the risk of erroneously assigning information on I^S_{ia} to movers. We note that neither strengthening (i) by excluding individuals from feeder schools, nor relaxing (ii) by assuming that individuals remain at the same school across grades 9–12 overturns our main conclusions (Appendix Table C.5).

Third, we drop all observations with missing information in Y_i, G_i, I^S_{ia}, and $\mathbf{X}_i(a)$ through list-wise deletion.

Applying these restrictions, we obtain a sample of $3,081$ individuals from 77 high schools across the U.S. for which we provide summary statistics in Table 3.1. 55% are female, and the average age measured at wave 1 equals $\approx$ 16 years (194 months). The average educational attainment in our sample is 14.8 years, which exceeds the average educational attainment in the parental generation by $\approx$ 1.1 years. Almost all individuals graduate from high school, which is not surprising given our sample restriction to individuals of European descent who stayed at the same high school in grades 9–12. The college completion rate equals $\approx$ 50%.

To assess the sample representativity, we compare our analysis sample to the 1974–1983 birth cohorts of Non-Hispanic Whites in the American Community Survey (ACS) and the Current Population Survey (CPS) (Appendix Table C.1). This comparison shows a slight over-representation of females and children from young mothers in our sample. Otherwise, our sample is by-and-large comparable to the corresponding groups in ACS and CPS. In robustness analyses, we re-weight our analysis sample to match ACS and CPS with respect to gender composition, educational attainment of parents, and the age of mothers at birth. Our results remain unaffected (Appendix Table C.5).

Table 3.1: Summary Statistics

	N=3,081; Siblings=525; High Schools=77			
	Mean	SD	Min	Max
Educational Attainment				
Years Education	14.81	2.25	8.00	20.00
High School Degree	0.97	0.18	0.00	1.00
2-year College Degree	0.53	0.50	0.00	1.00
4-year College Degree	0.42	0.49	0.00	1.00
Post-Graduate Degree	0.15	0.36	0.00	1.00
Variables of Interest				
PGS^{EA}	0.00	1.00	-4.18	3.35
I_{Qual}	0.00	1.00	-3.41	1.91
I_{Quant}	0.00	1.00	-3.25	3.21
Child Background Characteristics				
Female	0.55	0.50	0.00	1.00
Age in Months (Wave 1)	193.64	19.76	144.00	256.00
Maternal Age at Birth	25.49	4.83	16.00	44.33
Christian	0.82	0.38	0.00	1.00
Education Mother (in Years)	13.63	2.50	8.00	19.00
Education Father (in Years)	13.67	2.68	8.00	19.00
Potential Wage/Hour Mother	12.61	1.38	9.45	14.27
Potential Wage/Hour Father	15.48	1.31	11.14	17.11

Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This table shows summary statistics for the core analysis sample. The sample is restricted to genotyped individuals of (i) European descent, (ii) who visited an Add Health high school or an associated feeder school in wave 1, and (iii) who graduated from the same school. Observations with missing information in any of the displayed variables are dropped by list-wise deletion.

3.5 Results

We present our results in four steps. In section 3.5.1, we discuss the association of educational attainment, genetic endowments, and school investments in light of the identification requirements discussed in section 3.3. In section 3.5.2, we present our estimates for the complementarity parameter κ. After a robustness analysis in section 3.5.3, we conclude with an analysis of mechanisms in section 3.5.4.

3.5.1 The Association of Educational Attainment with Genetic Endowments and School Investments

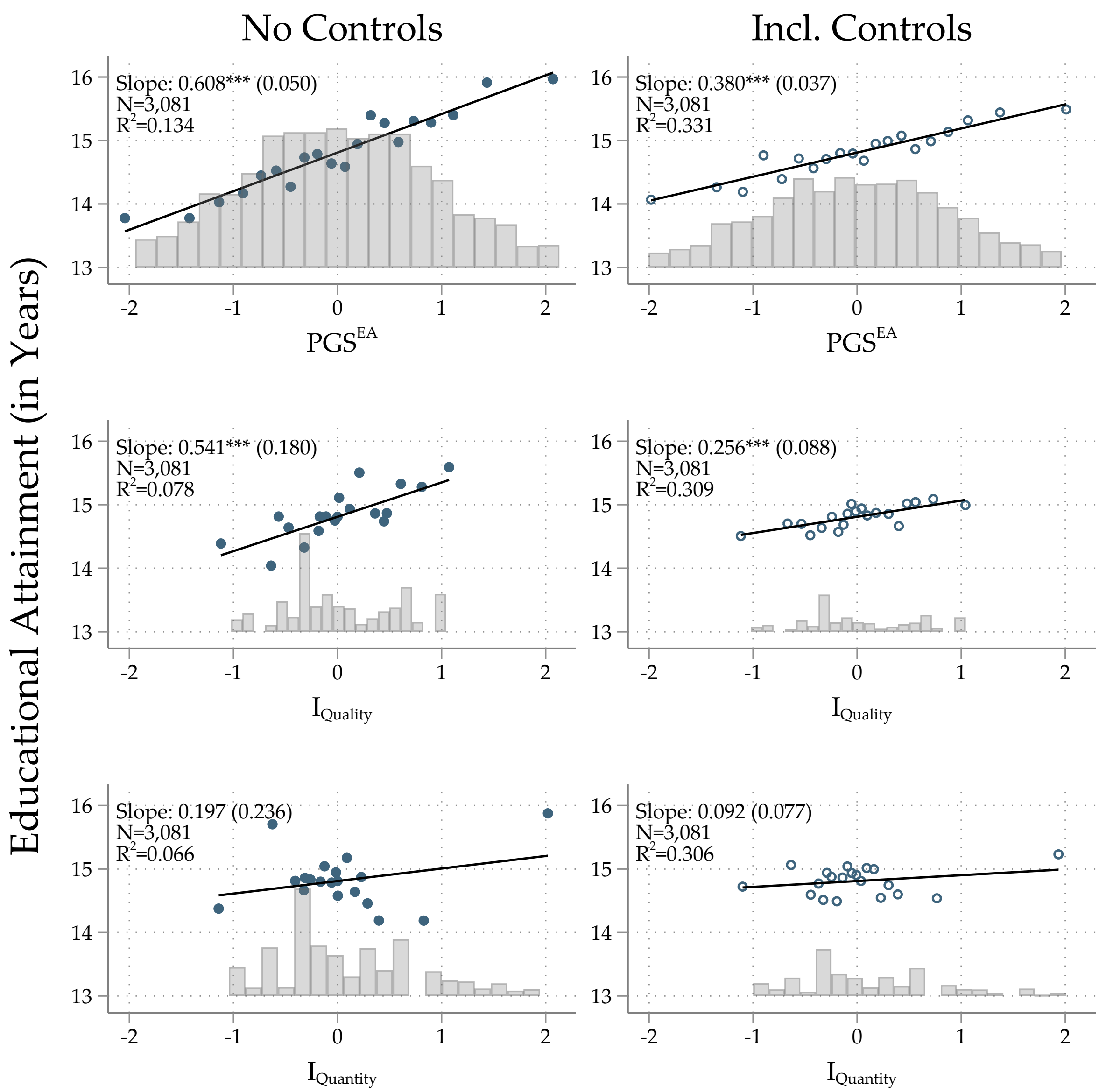
Figure 3.2 visualizes the association of educational attainment with our measures for genetic endowments G_i and school investments I^S_{ia}. In the left column, we show raw correlations that do not account for the control variables $\mathbf{X}_i(a)$. In the right column, we show associations conditional on $\mathbf{X}_i(a)$.

First, $\mathrm{PGS}^{\mathrm{EA}}$ is highly predictive of educational attainment. Without controls, a one-standard-deviation (1 SD) increase in $\mathrm{PGS}^{\mathrm{EA}}$ is associated with an increase in educational attainment of 0.608 years. This association does not have a causal interpretation as it may be confounded by genetic nurture effects. When we control for pre-determined child and family characteristics, a 1 SD increase in $\mathrm{PGS}^{\mathrm{EA}}$ is associated with an increase in educational attainment of 0.380 years. Sibling studies show that genetic nurture effects usually account for 40–50% of the raw association between $\mathrm{PGS}^{\mathrm{EA}}$ and educational attainment (Kweon et al., 2020; Muslimova et al., 2020; Ronda et al., forthcoming; Selzam et al., 2019). In our case, the association decreases by 38% when we control for child and family background characteristics. This result suggests that $\mathbf{X}_i(a)$ is indeed able to account for genetic nurture effects as confounding factors. This conclusion is further bolstered by a comparison of our between-family model with a sibling fixed effects model that we estimate on a subsample of our data ($N = 525$). In the within-family comparison, which allows us to perfectly control for genetic nurture effects, we obtain a point estimate of 0.458 that is significant at the 1%-level (Appendix Table C.2). This point estimate is very close to the result of the between-family comparison controlling for $\mathbf{X}_i(a)$, and lends further credence to our research design.

Second, $\mathrm{I}_{\mathrm{Qual}}$ is highly predictive of educational attainment. Without controls, a 1 SD increase in $\mathrm{I}_{\mathrm{Qual}}$ is associated with an increase in educational attainment of 0.541 years. This association does not have a causal interpretation as it may be confounded by selection effects. When we control for pre-determined child and family characteristics, a 1 SD increase in $\mathrm{I}_{\mathrm{Qual}}$ is associated with an increase in educational attainment of 0.256 years. This 53% decrease reflects positive selection into schools based on "teacher quality"—a pattern that has been thoroughly documented in existing literature for the U.S. (Biasi, forthcoming). Nevertheless, even when accounting for selection, the association of $\mathrm{I}_{\mathrm{Qual}}$ and educational attainment remains strong and positive. This result confirms prior literature, which has repeatedly demonstrated positive effects of teacher quality on students' educational success (Chetty et al., 2014a; Hanushek and Rivkin, 2010).

Third, $\mathrm{I}_{\mathrm{Quant}}$ is not significantly associated with educational attainment. The weakly positive correlation is imprecisely estimated and does not attain statistical significance at conventional levels. Furthermore, this result does not change when accounting for selection effects through the introduction of control vector $\mathbf{X}_i(a)$. This finding is in line with prior literature, which has not been able to establish consistent effects of teacher

Figure 3.2: Association of Educational Attainment with PGS^{EA}, I_{Qual}, and I_{Quant}

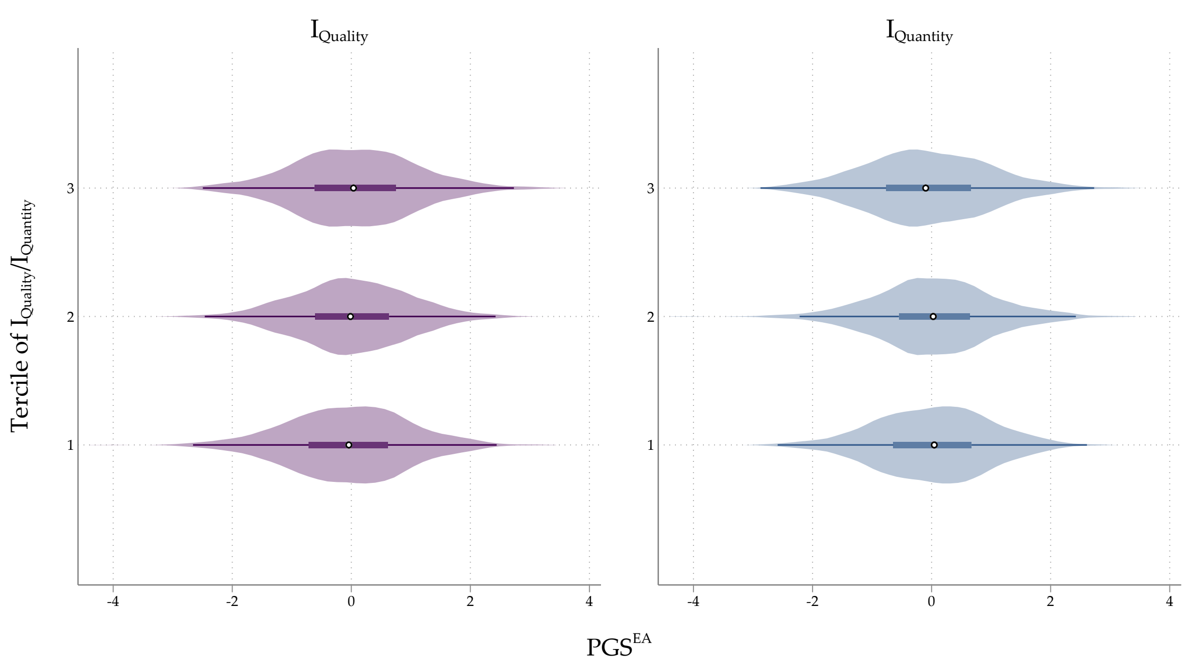


Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This figure visualizes the correlation of completed years of education with PGS^{EA}, I_{Qual}, and I_{Quant}, respectively. We bin scatterplots using 20 quantiles of the variable of interest. Gray bars indicate density distributions of the (residualized) variable of interest. Black lines are fitted from linear regressions of educational attainment on the variable of interest. In the left-column, we control for state fixed effects only. In the right column, we introduce the full set of control variables. *Child Controls*: Gender times birth cohort dummies, 20 principal components of the full matrix of genetic data. *Family Controls*: Age of mother at birth, years of education of both mother and father, average potential wages of both mother and father, the standard deviation of potential wages of both mother and father, a dummy for Christian religion, state fixed effects. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the school level.

quantity on students' educational success (Angrist et al., 2019; Fredriksson et al., 2013; Leuven and Løkken, 2020). However, this average association may mask heterogeneity across students with different genetic endowments—a hypothesis that we will test in the following subsection.

Next to genetic nurture effects and selection effects, a high correlation between G_i and I_{ia}^S would pose another threat to identification of the gene-environment interaction. Figure 3.3 suggests that this threat is not operational in our setting. In this figure, we plot the unconditional PGS^{EA} distribution by tercile of I_{Qual} and I_{Quant}, respectively. Visual inspection suggests that PGS^{EA} distributions are almost congruent to each other within each tercile of the two indicators. In Appendix Table C.3, we present formal statistical tests for this observation. In particular, we residualize PGS^{EA}, I_{Qual}, and I_{Quant} using control vector $\mathbf{X}_i(a)$. We then perform two-sample Kolmogorov-Smirnov tests for the equality of PGS^{EA} distributions within the terciles of I_{Qual} and I_{Quant}, respectively. We do not reject the null hypothesis of equal distributions for any of the comparisons at conventional levels of statistical significance. Hence, we conclude that PGS^{EA}, I_{Qual}, and I_{Quant} are indeed assigned independently of each other.

Figure 3.3: PGS^{EA} Distribution by I_{Qual} and I_{Quant}



Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This figure shows unconditional PGS^{EA} distributions by terciles of both I_{Qual} and I_{Quant}. The central point indicates the median, the bar indicates the interquartile range. The density represents the estimated Epanechnikov kernel density.

3.5.2 The Interplay of Genetic Endowments and School Investments in the Production of Educational Attainment

Table 3.2 shows our baseline estimates for the interaction of genetic endowments and school investments. In all regressions, we include the vector $\mathbf{X}_i(a)$ to control for genetic nurture and selection into schools.

Table 3.2: Association of PGS^{EA} and School Environments with Years of Education

	Baseline			Oster (2019)
Outcome: Years of Education	(1)	(2)	(3)	(4)
PGS^{EA}	0.374*** (0.033)	0.376*** (0.037)	0.371*** (0.033)	0.202*** (0.044)
I_{Qual}	0.227*** (0.081)	–	0.222*** (0.083)	0.047 (0.078)
$\text{PGS}^{\text{EA}} \times \text{I}_{\text{Qual}}$	-0.073** (0.033)	–	-0.072** (0.033)	-0.082** (0.035)
I_{Quant}	–	0.064 (0.068)	0.062 (0.058)	-0.012 (0.066)
$\text{PGS}^{\text{EA}} \times \text{I}_{\text{Quant}}$	–	0.036 (0.035)	0.026 (0.031)	-0.031 (0.040)
Child Controls	✓	✓	✓	✓
Family Controls	✓	✓	✓	✓
N	$3,081$	$3,081$	$3,081$	$3,081$
R^2	0.335	0.332	0.335	–
R^2_{max}	–	–	–	0.436
Outcome Mean	14.810	14.810	14.810	14.810
Outcome SD	2.250	2.250	2.250	2.250

Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This table shows the joint association of PGS^{EA}, I_{Qual}, and I_{Quant} with completed years of education. The first panel establishes our baseline estimates. The second panel displays bias-adjusted treatment effects following the procedure of Oster (2019). We impose R^2_{max} by multiplying R^2 from column (3) with 1.3. *Child Controls*: Gender times birth cohort dummies, 20 principal components of the full matrix of genetic data. *Family Controls*: Age of mother at birth, years of education of both mother and father, average potential wages of both mother and father, the standard deviation of potential wages of both mother and father, a dummy for Christian religion, state fixed effects. All non-binary variables are standardized on the estimation sample to have $\mu = 0$, $\sigma = 1$. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the school level. Standard errors of bias adjusted treatment effects are bootstrapped based on 200 draws.

In column (1), we focus on the teacher quality indicator, $\mathrm{I}_{\mathrm{Qual}}$. The point estimates for $\mathrm{PGS}^{\mathrm{EA}}$ and $\mathrm{I}_{\mathrm{Qual}}$ replicate the findings from Figure 3.2 and indicate a strong and positive association of $\mathrm{PGS}^{\mathrm{EA}}$ and $\mathrm{I}_{\mathrm{Qual}}$ with educational attainment.[13] A 1 SD increase in $\mathrm{PGS}^{\mathrm{EA}}$ ($\mathrm{I}_{\mathrm{Qual}}$) increases educational attainment by ≈ 0.37 (≈ 0.23) years.

$\mathrm{PGS}^{\mathrm{EA}} \times \mathrm{I}_{\mathrm{Qual}}$ is our estimate for the complementarity parameter κ. The negative coefficient of the interaction term indicates that genetic endowments and teacher quality act as *substitutes* in the production of educational attainment. A 1 SD increase in teacher quality reduces the positive association of educational attainment with $\mathrm{PGS}^{\mathrm{EA}}$ by $\approx 19\%$ ($= 0.07/0.37$). This result provides a notable contrast to existing literature investigating the gene-environment interaction between $\mathrm{PGS}^{\mathrm{EA}}$ and parental socio-economic status, which tend to act as *complements* in the production of educational attainment (Papageorge and Thom, 2020; Ronda et al., forthcoming).

In column (2), we focus on $\mathrm{I}_{\mathrm{Quant}}$. The point estimate for $\mathrm{I}_{\mathrm{Quant}}$ is again statistically indistinguishable from zero. The estimate for $\mathrm{PGS}^{\mathrm{EA}} \times \mathrm{I}_{\mathrm{Quant}}$ indicates that this null result is not driven by heterogeneity along the $\mathrm{PGS}^{\mathrm{EA}}$ distribution. Our estimate for the complementarity parameter κ is small and not statistically different from zero.

In column (3), we estimate both complementarity parameters in the same model and show that our results remain virtually unchanged. This stability is expected since $\mathrm{I}_{\mathrm{Qual}}$ and $\mathrm{I}_{\mathrm{Quant}}$ are distributed independently of each other by construction.

In column (4), we assess the potential for confounding due to unobserved differences across individuals. In spite of the rich control set $\mathbf{X}_i(a)$, our results may still reflect genetic nurture effects and selection effects due to unobservables. We follow Oster (2019) and calculate bias-adjusted treatment effects to account for this issue. The procedure assumes that changes in the coefficients of interest due to the introduction of $\mathbf{X}_i(a)$ are informative for the extent of confounding due to unobservables. The estimator requires two key inputs. The first input is $\mathrm{R}^2_{\mathrm{max}}$—the R^2 from a hypothetical regression of educational attainment on our variables of interest as well as observed and unobserved controls. The second input is δ—a measure for the relative degree of confounding through observed and unobserved controls. We follow the suggestion of Oster (2019) and specify $\mathrm{R}^2_{\mathrm{max}}$ as 1.3 times the empirical R^2 from column (3), and $\delta = 1$. Intuitively, $\delta = 1$ assumes that observed and unobserved confounders are equally related to the treatment.[14] The results remain qualitatively unaltered in comparison to column (3), yet the point estimates of

[13] In comparison to Figure 3.2, there are minor changes in coefficients due to the correlation of $\mathrm{PGS}^{\mathrm{EA}}$ and $\mathrm{I}_{\mathrm{Qual}}$. This correlation, however, is small and does not threaten the identification of the gene-environment interaction—see our discussion in section 3.5.1.

[14] Cinelli and Hazlett (2020) question this interpretation of δ as it is a function of (i) the association of the *treatment variable* with observed and unobserved confounders and (ii) the association of the *outcome variable* with observed and unobserved confounders. Therefore, Cinelli and Hazlett (2020) propose a bounding procedure based on parameter k_D that varies with (i) but not with (ii). Implementing their alternative procedure, our main conclusions remain unaffected—see our discussion in the following.

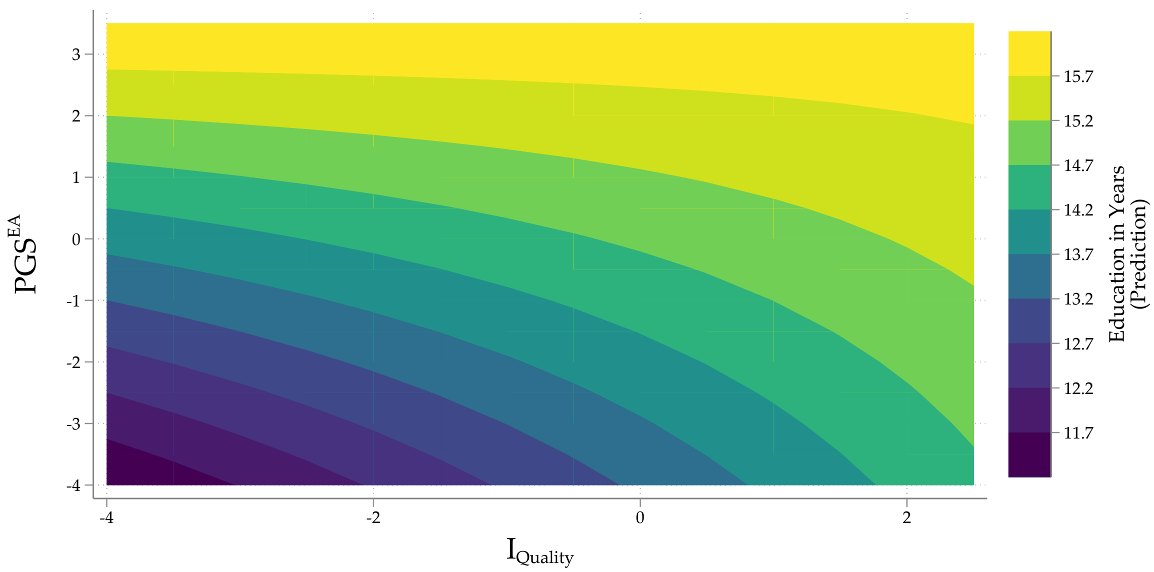
PGS^{EA} and I_{Qual} drop significantly. For example, under the maintained assumptions of R^2_{max} and $\delta = 1$, the point estimate of PGS^{EA} drops by almost half in comparison to our baseline estimate. However, note that it is also $\approx 50\%$ lower than in a sibling fixed effect estimation (Appendix Table C.2). The latter controls perfectly for genetic nurture effects. Hence, in view of our rich controls for family socio-economic background, the assumption of $\delta = 1$ is likely too conservative.[15]

In Appendix Figure C.2, we provide further sensitivity analyses with respect to the influence of unobserved confounding variables. Following the procedure of Cinelli and Hazlett (2020), we show that our results for PGS^{EA} and its interaction with I_{Qual} remain statistically significant at the 5%-level even if the set of unobserved confounding variables were more than ten times as strong as paternal education. Similarly, our results for I_{Qual} would remain statistically significant at the 5%-level if the set of unobserved confounding variables were more than five times as strong as paternal education. In view of the strong correlation of parental education with genetic endowments, its decisive role for school choices, and its strong predictive power for educational outcomes of children, these results bestow further confidence into the fact that our results are not just a reflection of genetic nurture effects and selection into schools by family background.

In principle, the negative gene-environment interaction between PGS^{EA} and I_{Qual} could be driven by low PGS^{EA} students gaining from higher-quality teachers, or high PGS^{EA} students losing from higher-quality teachers. In Figure 3.4, we provide evidence for the former, but not the latter. In this figure, we show years of education as predicted from the estimates in column (3) of Table 3.2. Moving horizontally from left to right at a given PGS^{EA} level, we see that predicted education increases strongly in the lower parts of the PGS^{EA} distribution. To the contrary, in the upper parts of the PGS^{EA} distribution, predicted education remains virtually unchanged, regardless of the quality of teachers at a given school. This pattern is encouraging as it suggests that investments into teacher quality mitigate inequity in educational outcomes without compromising the attainment of genetically advantaged students.

[15] Another popular way of reporting the results from sensitivity analyses à la Oster (2019) is to calculate the level of δ required to make coefficients equal to zero. We report these levels and associated bootstrapped standard errors for PGS^{EA}, I_{Qual}, and their interaction in the following: PGS^{EA} (1.989 [0.358]), I_{Qual} (1.263 [0.503]), $PGS^{EA} \times I_{Qual}$ (-7.731 [-37.409]). Note that $PGS^{EA} \times I_{Qual}$ is very insensitive to the inclusion of controls. Therefore, standard errors are large and the corresponding point estimate for δ cannot be reliably calculated.

Figure 3.4: Association of PGS^{EA} with Years of Education by I_{Qual}



Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This figure shows a prediction of completed years of education by PGS^{EA} and I_{Qual} cell. Predictions are calculated from the model estimated in column (3) of Table 3.2.

3.5.3 Robustness Analysis

We probe the robustness of our results in two steps. First, we investigate whether I_{Qual} and I_{Quant} pick up the effect of other school characteristics that may correlate with student outcomes. Second, we test whether our estimates of the complementarity parameter κ are confounded by interactions between genetic endowments and family environments.

Other school characteristics. First, in Figure 3.2 we document positive sorting into schools based on I_{Qual}. As a consequence, students in schools with high-quality teachers may additionally be exposed to a more favorable composition of their peer group. A broad literature has documented that skill formation is influenced by school peers (Bietenbeck, 2019; Isphording and Zölitz, 2020; Sacerdote, 2014).[16] Hence, our results for I_{Qual} may reflect both the quality of teachers and peer group composition. To test this hypothesis, we make use of Add Health's in-school questionnaire that elicits background information from a total of 90,000 students in the sampled schools. Based on this questionnaire, we calculate proxy indicators for the quality of the student's peers. In particular, we calculate (i) average years of paternal education, (ii) the share of single parent families, and (iii) student's average self-assessment with respect to the likelihood of attaining a college

[16] Sotoudeh et al. (2019) show genetic endowments of peers are also associated with individual outcomes.

degree.[17] Then, we include these indicators as well as their interaction with PGS^{EA} into our estimation model.

Table 3.3 displays the results. Column (1) replicates our baseline estimates. In columns (2)–(4), we sequentially introduce the peer quality indicators as well as their interaction with PGS^{EA}. Each proxy for the quality of peers is highly predictive of educational attainment. For example, a 1 SD increase in average paternal education of peers is associated with a 0.26 increase in years of education. Importantly, however, for all considered peer quality indicators, our conclusions with respect to I_{Qual}, I_{Quant}, and their interaction with genetic endowments remain unaffected.

Second, I_{Qual} and I_{Quant} may be correlated with school rules and sanction policies. Existing literature suggests that school rules may promote educational attainment (Bacher-Hicks et al., 2019). For example, the success of charter schools has been attributed to strict "no excuses" policies (Angrist et al., 2013). Hence, our results for I_{Qual} may reflect both the quality of teachers and the effect of school rules. To test this hypothesis, we exploit information from headmaster questionnaires and conduct a principal component analysis on various school policies.[18] We extract three components that reflect the school's strictness regarding (i) drug use, (ii) social misconduct, and (iii) academic misconduct.

In columns (5)–(7) of Table 3.3, we sequentially introduce the strictness indicators as well as their interaction with PGS^{EA}. Neither of the indicators is predictive of educational attainment, nor is there an interaction with genetic endowments. Our conclusions with respect to I_{Qual}, I_{Quant}, and their interaction with genetic endowments remain unaffected.

Third, there may be unobservable school characteristics that drive the relationship between I_{Qual}, I_{Quant}, and educational attainment. To address this concern, we use transcript records across grades 9–12 of roughly 12,000 Add Health respondents to calculate cohort-specific measures of school value-added in GPAs for Science, Math, and English. In the extant literature, value-added measures are mostly calculated with respect to test scores that are unaffected by evaluation biases of teachers. To the contrary, GPAs capture student progress in cognitive and behavioral outcomes as well as teacher perceptions (Jackson, 2019). In spite of these intricacies, GPAs are highly predictive of long-term student outcomes (Borghans et al., 2016; Kirkebøen, 2021). Therefore, GPA-based value-added measures provide a good way to capture the quality of schooling environments beyond the measures reported headmaster surveys and administrative data. Specifically, we fol-

[17] To avoid mechanical relationships between individual characteristics and peer group composition, we calculate leave-one-out (jackknife) indicators. A detailed description of these variables is disclosed in Appendix section C.3.

[18] In wave 1, headmasters were asked about the school's policy in the following domains of behavior: cheating, fighting with or injuring another student, alcohol or drug possession, drinking alcohol or using illegal drugs, smoking, verbally or physically abusing a teacher, and stealing school property. Possible policies are (i) no policy, (ii) verbal warning, (iii) minor action, (iv) in-school suspension, (v) out-of-school suspension, and (vi) expulsion. A detailed description of these variables is disclosed in Appendix section C.3.

Table 3.3: Robustness to Additional School Characteristics

Outcome: Years of Edu.	Baseline (1)	+ School Peer Characteristics: Edu. Father (2)	Single Parents (3)	College Aspir. (4)	+ School Sanction Policies: Drugs (5)	Social (6)	Acad. (7)	+ School VA (8)
PGS^{EA}	0.371*** (0.033)	0.361*** (0.035)	0.370*** (0.035)	0.360*** (0.034)	0.370*** (0.035)	0.368*** (0.035)	0.368*** (0.035)	0.367*** (0.035)
I_{Qual}	0.222*** (0.083)	0.150** (0.072)	0.221*** (0.082)	0.211*** (0.073)	0.209** (0.091)	0.195** (0.088)	0.231** (0.093)	0.171** (0.085)
PGS^{EA} × I_{Qual}	-0.072** (0.033)	-0.077** (0.036)	-0.073** (0.033)	-0.076** (0.034)	-0.072** (0.035)	-0.073** (0.035)	-0.069** (0.035)	-0.071** (0.035)
I_{Quant}	0.062 (0.058)	-0.033 (0.059)	-0.011 (0.056)	-0.017 (0.056)	0.055 (0.063)	0.038 (0.057)	0.030 (0.061)	0.014 (0.068)
PGS^{EA} × I_{Quant}	0.026 (0.031)	0.017 (0.028)	0.030 (0.032)	0.024 (0.030)	0.027 (0.032)	0.026 (0.032)	0.022 (0.032)	0.030 (0.034)
School Char.	–	0.261*** (0.055)	-0.201*** (0.049)	0.212*** (0.045)	0.003 (0.052)	-0.108 (0.076)	0.071 (0.072)	0.109* (0.062)
PGS^{EA} × School Char.	–	-0.044 (0.042)	0.031 (0.039)	-0.046 (0.035)	0.002 (0.041)	0.024 (0.030)	0.029 (0.039)	-0.016 (0.031)
Child Contr.	✓	✓	✓	✓	✓	✓	✓	✓
Family Contr.	✓	✓	✓	✓	✓	✓	✓	✓
N	3,081	2,965	2,965	2,965	2,999	2,999	2,999	2,773
R^2	0.335	0.344	0.343	0.343	0.338	0.339	0.338	0.315

Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This table shows the joint association of PGS^{EA}, I_{Qual}, and I_{Quant} with completed years of education. We control for additional school characteristics and their interaction with PGS^{EA}. The relevant school characteristics are indicated in the column header. *Child Controls*: Gender times birth cohort dummies, 20 principal components of the full matrix of genetic data. *Family Controls*: Age of mother at birth, years of education of both mother and father, average potential wages of both mother and father, the standard deviation of potential wages of both mother and father, a dummy for Christian religion, state fixed effects. All non-binary variables are standardized on the estimation sample to have $\mu = 0$, $\sigma = 1$. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the school level.

low the indirect calculation procedures suggested in Chetty et al. (2014a) and Jackson et al. (2020): we residualize subject-specific GPAs from lagged GPAs in English, Math, and Science, lagged and contemporaneous measures of tracks in these subjects as well as a rich set of individual background characteristics. In turn, we sum residuals to calculate school-times-cohort fixed effects. To avoid mechanical relationships between individual outcomes and cohort-specific school effects, we calculate leave-cohort-out predictions while giving greater weight to neighboring cohorts. We calculate these measures separately for each subject but summarize the school-specific information by extracting the first principal component from the three value-added measures (see Appendix section C.3 for details).

In column (8) of Table 3.3, we introduce school value-added as well as its interaction with PGS^{EA} as additional controls. While school value-added is indeed predictive of educational attainment, there is no effect heterogeneity across the PGS^{EA} distribution. Furthermore, the associations of I_{Qual}, I_{Quant}, and PGS^{EA} with educational attainment remain unaffected. Hence, we find no evidence that our relationships of interest are confounded by unobservable school characteristics.

Family environments. In our baseline analysis, we control for a rich set of parental background characteristics to control for genetic nurture effects and selection into schools. However, even if we were able to perfectly control for these confounding factors, the complementarity parameter for genetic endowments and school investments may still be confounded by interactions between genetic endowments and family socio-economic status (Domingue et al., 2020; Keller, 2014). To test this hypothesis, we enrich our estimation model by interacting genetic endowments and school environments with the entire control vector $\mathbf{X}_i(a)$. Therefore, we allow for the possibility that family socio-economic status interacts with both genetic endowments and school investments.

Table 3.4 displays the results. Column (1) replicates our baseline estimates. Column (2) displays the enriched estimation model. In spite of a slight decrease in precision, our conclusions with respect to I_{Qual}, I_{Quant}, and their interaction with PGS^{EA} remain unaffected.

In addition, we assess whether I_{Qual}, I_{Quant}, and their interaction with genetic endowments predict parental investments.[19] A positive association of PGS^{EA} and I_{Qual} with parental investments may suggest that $\mathbf{X}_i(a)$ does not fully account for genetic nurture effects and selection into schools. Therefore, we collect information on a series of activities that the child has conducted with her mother or father in the last four weeks.[20] Following Anderson (2008) and Kling et al. (2007), we standardize each response dimension to $\mu = 0$ and $\sigma = 1$ and sum them linearly by parent to obtain aggregate indexes of time investment. We then use the resulting indexes for parental time investment as the outcome of interest, respectively.

The results are shown in columns (3)–(4) of Table 3.4. PGS^{EA} and parental investments are indeed positively associated. However, this association does not necessarily

[19]Note that parental investments may reflect responses to PGS^{EA}, I_{Qual}, and I_{Quant}. Therefore, we analyze them as separate outcomes instead of including them in $\mathbf{X}_i(a)$.

[20]These activities include shopping, playing sports, church attendance, talking about dates, going to movies and similar events, talking about personal problems, having an argument, talking about school work, working together on school work, and talking about other things at school. See Appendix section C.3 for details.

Table 3.4: Robustness to Family Environments

	Outcome: Years of Education		Outcome: Parental Investment		Outcome: Years of Education
	Baseline	Full Interaction	Mother	Father	Endogenous Controls
	(1)	(2)	(3)	(4)	(5)
PGS^{EA}	0.371*** (0.033)	0.414*** (0.086)	0.044** (0.017)	0.058*** (0.017)	0.339*** (0.099)
I_{Qual}	0.222*** (0.083)	0.198* (0.116)	0.004 (0.043)	0.034 (0.042)	0.219* (0.126)
$PGS^{EA} \times I_{Qual}$	-0.072** (0.033)	-0.092*** (0.035)	-0.013 (0.014)	-0.009 (0.016)	-0.098** (0.040)
I_{Quant}	0.062 (0.058)	0.064 (0.093)	0.015 (0.040)	-0.013 (0.036)	0.061 (0.112)
$PGS^{EA} \times I_{Quant}$	0.026 (0.031)	0.008 (0.032)	-0.019 (0.018)	-0.031* (0.016)	0.014 (0.037)
Child Controls	✓	✓	✓	✓	✓
Family Controls	✓	✓	✓	✓	✓
Full Interaction	×	✓	×	×	✓
Endogenous Controls	×	×	×	×	✓
N	3,081	3,081	3,081	2,541	2,125
R^2	0.335	0.354	0.101	0.078	0.379

Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. The first panel of this table shows the joint association of PGS^{EA}, I_{Qual} and I_{Quant} with completed years of education. In column (2) we control for all possible interactions between PGS^{EA}, I_{Qual} and I_{Quant} and the control variables. The second panel of this table shows the joint association of PGS^{EA}, I_{Qual} and I_{Quant} with an index of parental time investments. The third panel of this table shows the joint association of PGS^{EA}, I_{Qual} and I_{Quant} with completed years of education while accounting for endogenous control variables. Endogenous control variables include the index for maternal time investments, the index for paternal time investments, and log family income. *Child Controls*: Gender times birth cohort dummies, 20 principal components of the full matrix of genetic data. *Family Controls*: Age of mother at birth, years of education of both mother and father, average potential wages of both mother and father, the standard deviation of potential wages of both mother and father, a dummy for Christian religion, state fixed effects. All non-binary variables are standardized on the estimation sample to have $\mu = 0$, $\sigma = 1$. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the school level.

imply the existence of genetic nurture effects. Instead they could also reflect *evocative gene-environment correlations*, i.e. that children select into environments depending on their genetic endowments G_i (Smith-Woolley et al., 2018). If we were to follow this interpretation, the observed association is not indicative for a third factor confounding the relationship of interest, but rather speaks to a particular mechanism of how genetic endowments influence child outcomes. Importantly, there is no differential association between

PGS^{EA} and parental time investments depending on school quality.

Lastly, in column (5) of Table 3.4, we again estimate a fully interacted model, but also incorporate controls for family environments that are potentially endogenous to PGS^{EA} and schooling environments. In particular, we include the indexes for maternal and paternal time investments as well as the log of annual family income. Despite a decrease in sample size and the associated loss in precision, our results remain unaffected.

Overall, these results bolster confidence that our estimates for the complementarity parameter κ are not confounded by interactions between genetic endowments and family environments.

3.5.4 Mechanisms

In this section, we analyze mechanisms that underpin the substitutability of genetic endowments and teacher quality. We abstract from I_{Quant} in view of its robust non-association with educational outcomes (see sections 3.5.1–3.5.3).

Educational degrees. Total years of education summarizes information from various educational stages, where each stage requires a different mix of skills θ_i (Cunha et al., 2006, 2010). Therefore, we repeat our analysis by replacing total years of education with binary variables for whether respondents achieved (i) at least a high school degree or GED, (ii) completed a 2-year college degree, (iii) completed a 4-year college degree, and (iv) completed a post-graduate degree.

In Figure 3.5, we display the resulting point estimates for the complementarity parameter κ and the associated 95% confidence bands. The series in circles indicates that the compensating effect of teacher quality has a U-shaped pattern throughout the educational life-cycle. There is a small reduction of the probability to drop out of high school, followed by larger substitutability with respect to 2-year and 4-year college degrees. The substitutability of high-quality teachers and genetic endowments levels off at the post-graduate level. This pattern is consistent with the following interpretation. High school graduation is a relatively "inclusive" educational outcome that is accessible for most, including low PGS^{EA} students in low-quality schooling environments. Evidence to this effect is provided by a high school graduation rate of 97% in our sample (Table 3.1). To the contrary, post-graduate education is a relatively "exclusive" educational outcome that is more accessible for students with a advantageous genetic endowments *and* who experienced conducive environments. In both cases, there is limited scope for high-quality teachers to make a difference for low PGS^{EA} students. College education, however, takes a middle ground between these two polar outcomes and therefore offers scope for disadvantageous genetic endowments to be offset by conducive school environments, and vice versa.[21]

The series in triangles indicates analogous complementarity parameters for genetic endowments and a summary index of family socio-economic status (SES).[22] Consistent with Buser et al. (2021a) and Papageorge and Thom (2020), the complementarity between genetic endowments and family SES increases across the educational life-cycle of individuals. The differential complementarity patterns of school investment and family SES point to the complexity of the skill production function, where endowments and different investments interact in distinct and time-variant ways across the life-cycle of individuals.

Skill formation. In section 3.3, we formulated educational attainment Y_i as a function of child skills θ_i at the end of childhood. Skills that influence educational attainment are multidimensional and comprise a broad set of health indicators and (non-)cognitive skills (Almlund et al., 2011; Heckman and Mosso, 2014). Furthermore, an emerging literature provides evidence for each of these skill dimensions being partially shaped by genetic influence (Buser et al., 2021a; Demange et al., 2020, 2021).

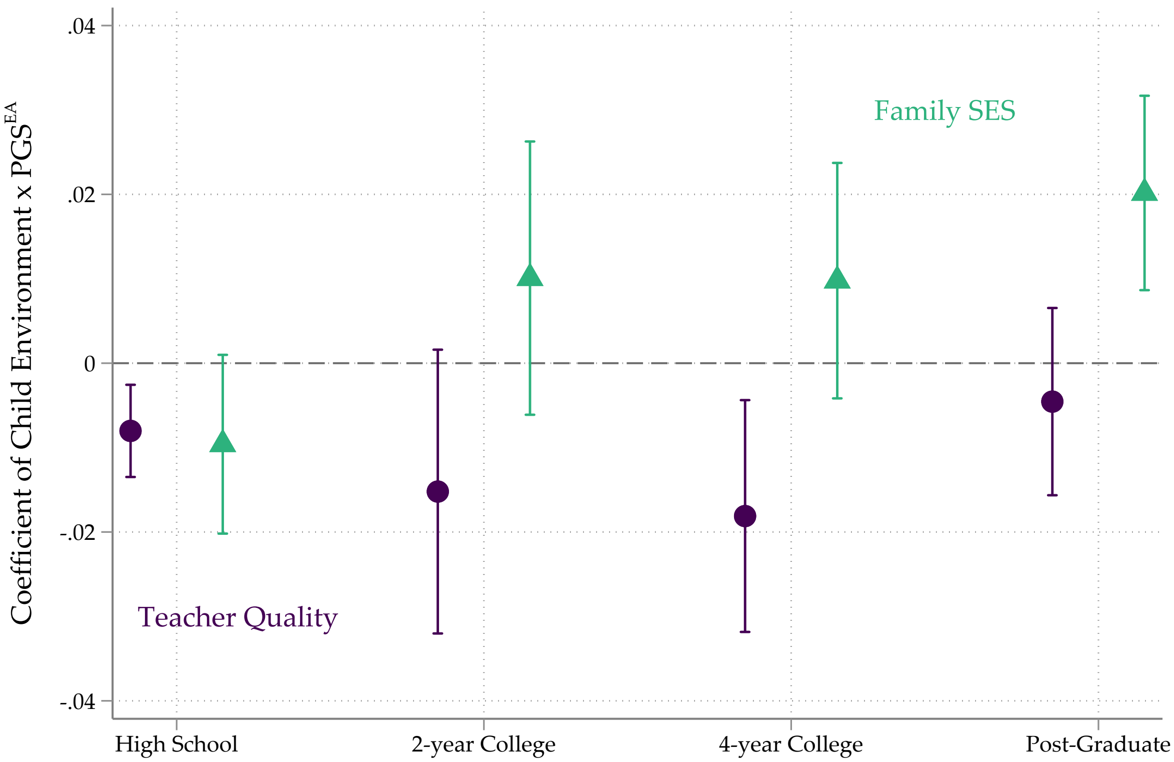
We evaluate these potential channels by analyzing the associations of $\mathrm{PGS}^{\mathrm{EA}}$ and $\mathrm{I}_{\mathrm{Qual}}$ with a set of intermediate outcomes. In terms of health outcomes, we focus on subjective health as measured by quality-adjusted life years (QALY) and objective health as measured by an index that comprises information about whether the respondent is obese, afflicted by stage one hypertension, or has high cholesterol. In terms of cognitive skills, we use the Picture Vocabulary Test (PVT) as a measure for verbal intelligence. Lastly, we focus on personality and preferences as two distinct conceptualizations of non-cognitive skills (Becker et al., 2012; Humphries and Kosse, 2017). In particular, we focus on risk aversion, patience, and the Big Five personality traits. All measures are collected in waves 3 and 4 of Add Health, and hence after respondents have left high school but potentially before they have concluded their highest level of education (see Appendix section C.3 for details).

Health, cognitive skills, risk aversion, and patience have been shown to be strong predictors of educational attainment (Burks et al., 2015; Castillo et al., 2018a,b; Jackson, 2009). Furthermore, openness and emotional stability—the inverse of neuroticism—are positively associated with educational attainment (Becker et al., 2012; Buser et al., 2021b). Based on this evidence, we expect positive associations of both $\mathrm{PGS}^{\mathrm{EA}}$ and $\mathrm{I}_{\mathrm{Qual}}$ with each of these intermediate outcomes. The sign of the gene-environment interaction is a priori un-

[21] Gene-environment interactions on binary outcomes may be misinterpreted when estimated in a linear probability model. As a remedy, Domingue et al. (2020) recommend to present results for the underlying continuous variable, i.e. years of education in the case of this paper. Reassuringly, our results for years of education are in line with our results for educational degrees.

[22] In particular, we use the "social origins score" from Belsky et al. (2018), measured at wave 1. Results for alternative measures of family SES, such as family income or potential wages of either parent, are similar.

Figure 3.5: Association of PGS^{EA} and School/Family Environments with Degree Attainment



Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This figure shows point estimates and 95% confidence bands of interaction associations between PGS^{EA} and school/family environments during childhood with completed education degrees. Estimates follow the specification of Equation (3.7). *Child Controls*: Gender times birth cohort dummies, 20 principal components of the full matrix of genetic data. *Family Controls*: Age of mother at birth, years of education of both mother and father, average potential wages of both mother and father, the standard deviation of potential wages of both mother and father, a dummy for Christian religion, state fixed effects. Standard errors are clustered at the school level.

clear. However, in view of the substitutability of PGS^{EA} and I_{Qual} in the production of educational attainment, we expect similar substitutability patterns for a subset of these intermediate outcomes as well.

Table 3.5 summarizes the results. In column (1)–(2) of Panel (a), we focus on health outcomes. In line with expectations, our results show a positive association of PGS^{EA} with both subjective and objective health. A 1 SD increase in PGS^{EA} increases subjective (objective) health by 0.069 SD (0.043 SD). Furthermore, the negative coefficient on the interaction of PGS^{EA} and I_{Qual} suggests that this increase is particularly pronounced for low-PGS^{EA} students: a 1 SD increase in teacher quality reduces the positive association of subjective health with the PGS^{EA} by $\approx 41\%$ $(= 0.028/0.069)$.

In column (3) of Panel (a), we focus on the PVT as a measure of cognitive skills. In

Table 3.5: Association of PGS^{EA} and School Environments with Skill Measures

Panel (a)	Health		Cognitive	Preferences	
	Subjective (1)	Objective (2)	PVT (3)	Risk (4)	Patience (5)
PGS^{EA}	0.069*** (0.017)	0.043** (0.018)	0.181*** (0.017)	0.038** (0.015)	0.074*** (0.017)
I_{Qual}	0.021 (0.042)	0.032 (0.036)	0.102*** (0.039)	0.047 (0.030)	0.045 (0.038)
$PGS^{EA} \times I_{Qual}$	-0.028** (0.014)	-0.000 (0.020)	-0.034* (0.018)	-0.046*** (0.015)	-0.044*** (0.013)
Child Controls	✓	✓	✓	✓	✓
Family Controls	✓	✓	✓	✓	✓
N	3,081	3,081	3,001	3,077	3,077
R^2	0.078	0.054	0.207	0.112	0.096

Panel (b)	Personality				
	Open-ness (1)	Conscient-iousness (2)	Extra-version (3)	Agree-ableness (4)	Neuro-ticism (5)
PGS^{EA}	0.073*** (0.017)	-0.017 (0.017)	-0.006 (0.019)	0.038* (0.020)	-0.084*** (0.019)
I_{Qual}	0.038 (0.033)	-0.031 (0.036)	-0.043 (0.030)	0.057 (0.037)	-0.018 (0.033)
$PGS^{EA} \times I_{Qual}$	0.012 (0.013)	-0.007 (0.015)	-0.001 (0.023)	-0.007 (0.019)	0.023 (0.019)
Child Controls	✓	✓	✓	✓	✓
Family Controls	✓	✓	✓	✓	✓
N	3,059	3,079	3,075	3,077	3,077
R^2	0.084	0.041	0.031	0.133	0.092

Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This table shows the joint association of PGS^{EA}, I_{Qual}, and I_{Quant} with health, cognitive skills, preferences, and personality. *Child Controls*: Gender times birth cohort dummies, 20 principal components of the full matrix of genetic data. *Family Controls*: Age of mother at birth, years of education of both mother and father, average potential wages of both mother and father, the standard deviation of potential wages of both mother and father, a dummy for Christian religion, state fixed effects. All non-binary variables are standardized on the estimation sample to have $\mu = 0$, $\sigma = 1$. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the school level.

line with expectations, our results show positive associations of both PGS^{EA} and I_{Qual} with the PVT. A 1 SD increase in PGS^{EA} (I_{Qual}) is associated with a 0.181 SD (0.102 SD) increase in the PVT. Furthermore, both factors are substitutes for each other. A 1 SD increase in teacher quality reduces the positive association of PVT and PGS^{EA} by $\approx$ 19% (= 0.034/0.181).

In columns (4)–(5) of Panel (a), we focus on economic preferences. In line with expectation, we find strong positive associations of PGS^{EA} with both risk aversion and patience. A 1 SD increase in PGS^{EA} is associated with a 0.038 SD (0.074 SD) increase in risk aversion (patience). Furthermore, PGS^{EA} and I_{Qual} are substitutes for each other. A 1 SD increase in I_{Qual} reduces the positive associations of risk aversion and patience with the PGS^{EA} by $\approx$ 124% (= 0.046/0.037) and $\approx$ 59% (= 0.044/0.074), respectively.

In Panel (b), we focus on personality traits. In line with expectation, we find a positive association of PGS^{EA} with openness and a negative association of PGS^{EA} with neuroticism. However, I_{Qual} is not predictive of any of the Big Five dimensions. Furthermore, we do not find evidence for complementarity of PGS^{EA} and I_{Qual} in the production of personality traits.

To summarize: we find negative gene-environment interactions of genetic endowments and teacher quality in the production of subjective health, cognitive skills, risk aversion, and patience. Given their predictive power for educational attainment, these intermediate outcomes are plausible channels to explain the substitutability of genetic endowments and teacher quality in the production of educational attainment.

3.6 Conclusion

The question of how natural endowments and environmental factors determine life outcomes has a long history of inquiry in philosophy and science (Darwin, 1859; Descartes, 1641; Lamarck, 1838; Locke, 1690). The assumption that life outcomes are the result of genetic and environmental factors initially led to the so-called "nature versus nurture" debate. However, current research has moved beyond this simplistic dichotomy and recognizes that individual life outcomes are the result of a complex interplay between nature and nurture. Importantly, this insight illustrates that the importance of genetic endowments for life outcomes is not immutable. Instead, it opens an avenue for policy interventions that shape the relevant environment.

In this paper, we contribute to this research agenda by studying the interplay of genetic endowments and schooling environments in the production of educational outcomes. Making use of recent advances in molecular genetics, we link an individual-level index of genetic predispositions for educational success with measures of our environmental factors of interest, namely teacher "quality" and "quantity" during high school. In turn, we can investigate directly whether the importance of genetic endowments varies with the quality of their high school environments.

Our findings suggest that school investments have the potential to cushion the genetic gradient in educational attainment. However, this conclusion depends on the particular type of investment. On the one hand, increases in "teacher quality" offset genetic disadvantages. On the other hand, we do not find any substitutability with respect to "teacher quantity." Our findings furthermore suggest that increased gains in educational attainment for students of low genetic endowments are mediated by gains in subjective health, cognitive skills and risk aversion, and patience.

Genes are important co-determinants of many life outcomes. However, although they are fixed at conception, their importance can be mediated by suitable policy intervention. In the case of education, increasing the quality of teachers in high schools may provide an important step to level the playing field for all students, regardless of their draw in the genetic lottery.

Appendix C

C.1 Supplementary Tables

Table C.1: Sample Representativeness

	Population (Cohorts 1974-1983)		Analysis Sample	
	All	Non-Hispanic White	Unweighted	Re-Weighted
Gender				
Male	0.498	0.503	0.453	0.503
Female	0.502	0.497	0.547	0.497
Education Mother				
$\leq$ High School	0.536	0.489	0.494	0.489
$>$ High School; $<$ College	0.281	0.302	0.217	0.301
$\geq$ College	0.183	0.209	0.289	0.210
Education Father				
$\leq$ High School	0.472	0.425	0.491	0.425
$>$ High School; $<$ College	0.255	0.271	0.196	0.271
$\geq$ College	0.273	0.304	0.312	0.303
Age Mother at Birth				
$<$ 25 Years	0.353	0.330	0.485	0.330
$\geq$ 25 Years	0.647	0.670	0.515	0.670
Parental Income				
$<$ \$50,000	0.557	0.491	0.531	0.516
$\geq$ \$50,000; $<$ \$100,000	0.352	0.403	0.390	0.401
$\geq$ \$100,000	0.091	0.106	0.079	0.083
Education Respondent				
$\leq$ High School	0.301	0.225	0.181	0.173
$>$ High School; $<$ College	0.327	0.344	0.399	0.402
$\geq$ College	0.372	0.431	0.419	0.425

Data: National Longitudinal Study of Adolescent to Adult Health, American Community Survey (ACS), Current Population Survey (CPS).
Note: Own calculations. This table shows summary statistics of the core analysis sample in comparison to other population samples. It shows respondents' characteristics for the following samples: (i) the U.S. population from birth cohorts 1974–1983, (ii) the Non-Hispanic White U.S. population from birth cohorts 1974–1983, (iii) the core estimation sample, and (iv) the core estimation sample re-weighted to match (ii) with respect to *Gender*, *Education Mother*, *Education Father*, and *Age Mother at Birth*. Population data on *Gender* and *Education Respondent* from IPUMS ACS 2019 (Ruggles et al., 2020). Population data on *Education Mother*, *Education Father*, *Age Mother at Birth*, and *Parental Income* from IPUMS CPS 1994 (Flood et al., 2020).

Table C.2: Association of PGS^{EA} and Years of Education: Comparing Between-Family and Within-Family Models

Outcome: Years of Education	Between-Family (1)	Within-Family (2)
PGS^{EA}	0.374^{***} (0.037)	0.458^{***} (0.160)
I_{Qual}	0.226^{***} (0.083)	–
I_{Quant}	0.063 (0.057)	–
Child Controls	✓	✓
Family Controls	✓	✓
Sibling Fixed Effect	×	✓
N	3,081	525
R^2	0.334	0.785
Outcome Mean	14.810	14.928
Outcome SD	2.250	2.262

Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This table shows the joint association of PGS^{EA}, I_{Qual}, and I_{Quant} with completed years of education. The first panel establishes our baseline estimates based on a between-family model. The second panel displays results from a family fixed effect model using within-family variation only. *Child Controls*: Gender times birth cohort dummies, 20 principal components of the full matrix of genetic data. *Family Controls*: Age of mother at birth, years of education of both mother and father, average potential wages of both mother and father, the standard deviation of potential wages of both mother and father, a dummy for Christian religion, state fixed effects. All non-binary variables are standardized on the estimation sample to have $\mu = 0$, $\sigma = 1$. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the school level.

Table C.3: Tests for Equality of PGS^{EA} Distributions

	Terciles of I_{Qual} / I_{Quant}		
	1	2	3
Panel (a): I_{Qual}			
1	–	–	–
2	0.66	–	–
3	0.25	0.77	–
Panel (b): I_{Quant}			
1	–	–	–
2	0.77	–	–
3	1.00	0.70	–

Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This table shows the results of pairwise Kolmogorov-Smirnov tests for the PGS^{EA} distributions within different terciles of I_{Qual} and I_{Quant}, respectively. Results are summarized by the p-value for the null hypothesis that the two PGS^{EA} distributions are equal within the corresponding terciles of I_{Qual} and I_{Quant}.

Table C.4: Alternative Polygenic Scores

Outcome: Years of Education	Baseline (1)	+ Controls for Other Polygenic Scores: Body Mass Index (2)	ADHD (3)	Depressive Symptoms (4)	Intelligence (5)	Ever Smoker (6)	Sleep Duration (7)
PGS^{EA}	0.371*** (0.033)	0.357*** (0.035)	0.346*** (0.032)	0.372*** (0.034)	0.358*** (0.039)	0.347*** (0.038)	0.374*** (0.033)
I_{Qual}	0.222*** (0.083)	0.227*** (0.082)	0.223*** (0.081)	0.226*** (0.081)	0.228*** (0.081)	0.225*** (0.080)	0.227*** (0.081)
$PGS^{EA} \times I_{Qual}$	-0.072** (0.033)	-0.084** (0.036)	-0.074** (0.033)	-0.073** (0.034)	-0.079** (0.037)	-0.080** (0.037)	-0.073** (0.033)
Other PGS	–	-0.070** (0.032)	-0.130*** (0.034)	-0.024 (0.032)	0.029 (0.040)	-0.128*** (0.041)	-0.004 (0.032)
Other PGS $\times I_{Qual}$	–	-0.037 (0.037)	0.019 (0.031)	0.007 (0.030)	0.011 (0.040)	-0.019 (0.039)	0.003 (0.034)
Child Controls	✓	✓	✓	✓	✓	✓	✓
Family Controls	✓	✓	✓	✓	✓	✓	✓
N	3,081	3,081	3,081	3,081	3,081	3,081	3,081
R^2	0.335	0.336	0.338	0.335	0.335	0.338	0.335

Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This table shows the joint association of PGS^{EA}, I_{Qual}, and I_{Quant} with completed years of education. We control for other PGS and their interaction with I_{Qual} and I_{Quant}. The relevant PGS are indicated in the column header. *Child Controls*: Gender times birth cohort dummies, 20 principal components of the full matrix of genetic data. *Family Controls*: Age of mother at birth, years of education of both mother and father, average potential wages of both mother and father, the standard deviation of potential wages of both mother and father, a dummy for Christian religion, state fixed effects. All non-binary variables are standardized on the estimation sample to have $\mu = 0$, $\sigma = 1$. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the school level.

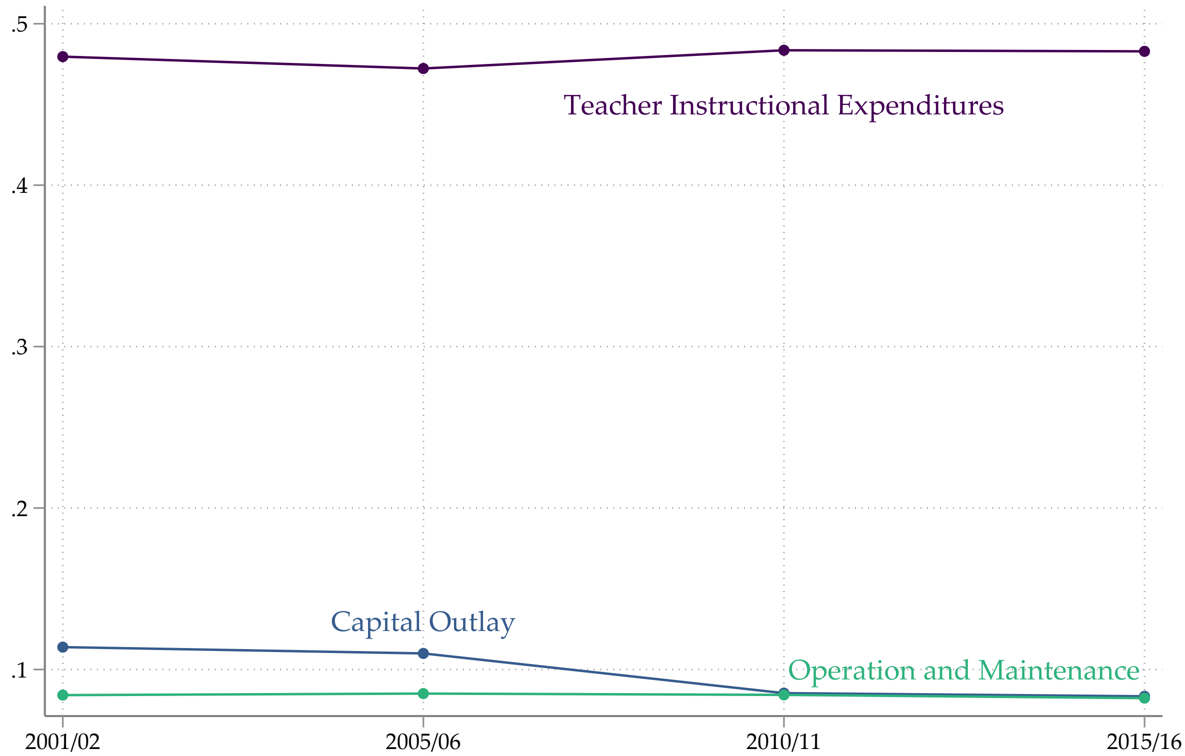
Table C.5: Robustness to Sample Composition

Outcome: Years of Education	Baseline (1)	Alternative Sample Composition: Re-Weighted (2)	Excl. all (Potential) Movers (3)	Inc. all (Potential) Movers (4)
PGS^{EA}	0.371***	0.359***	0.360***	0.381***
	(0.033)	(0.035)	(0.040)	(0.028)
I_{Qual}	0.222***	0.198**	0.154	0.217***
	(0.083)	(0.084)	(0.099)	(0.078)
$PGS^{EA} \times I_{Qual}$	-0.072**	-0.073**	-0.068*	-0.050*
	(0.033)	(0.034)	(0.039)	(0.029)
I_{Quant}	0.062	0.049	0.067	0.050
	(0.058)	(0.062)	(0.069)	(0.075)
$PGS^{EA} \times I_{Quant}$	0.026	0.037	0.013	0.037
	(0.031)	(0.034)	(0.035)	(0.023)
Child Controls	✓	✓	✓	✓
Family Controls	✓	✓	✓	✓
N	3,081	3,027	2,526	4,185
R^2	0.335	0.315	0.328	0.319

Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This table shows the joint association of PGS^{EA}, I_{Qual}, and I_{Quant} with completed years of education. In column (2), we re-weight our analysis sample to match ACS and CPS with respect to gender composition, educational attainment of parents, and the age of mothers at birth—see also Appendix Table C.1. In column (3), we exclude respondents that visit feeder schools in wave 1 and for whom we do not have information on subsequent high schools. In column (4), we include respondents that are in Add Health high schools in wave 1 and for whom we do not have information on subsequent high schools. *Child Controls*: Gender times birth cohort dummies, 20 principal components of the full matrix of genetic data. *Family Controls*: Age of mother at birth, years of education of both mother and father, average potential wages of both mother and father, the standard deviation of potential wages of both mother and father, a dummy for Christian religion, state fixed effects. All non-binary variables are standardized on the estimation sample to have $\mu = 0$, $\sigma = 1$. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the school level.

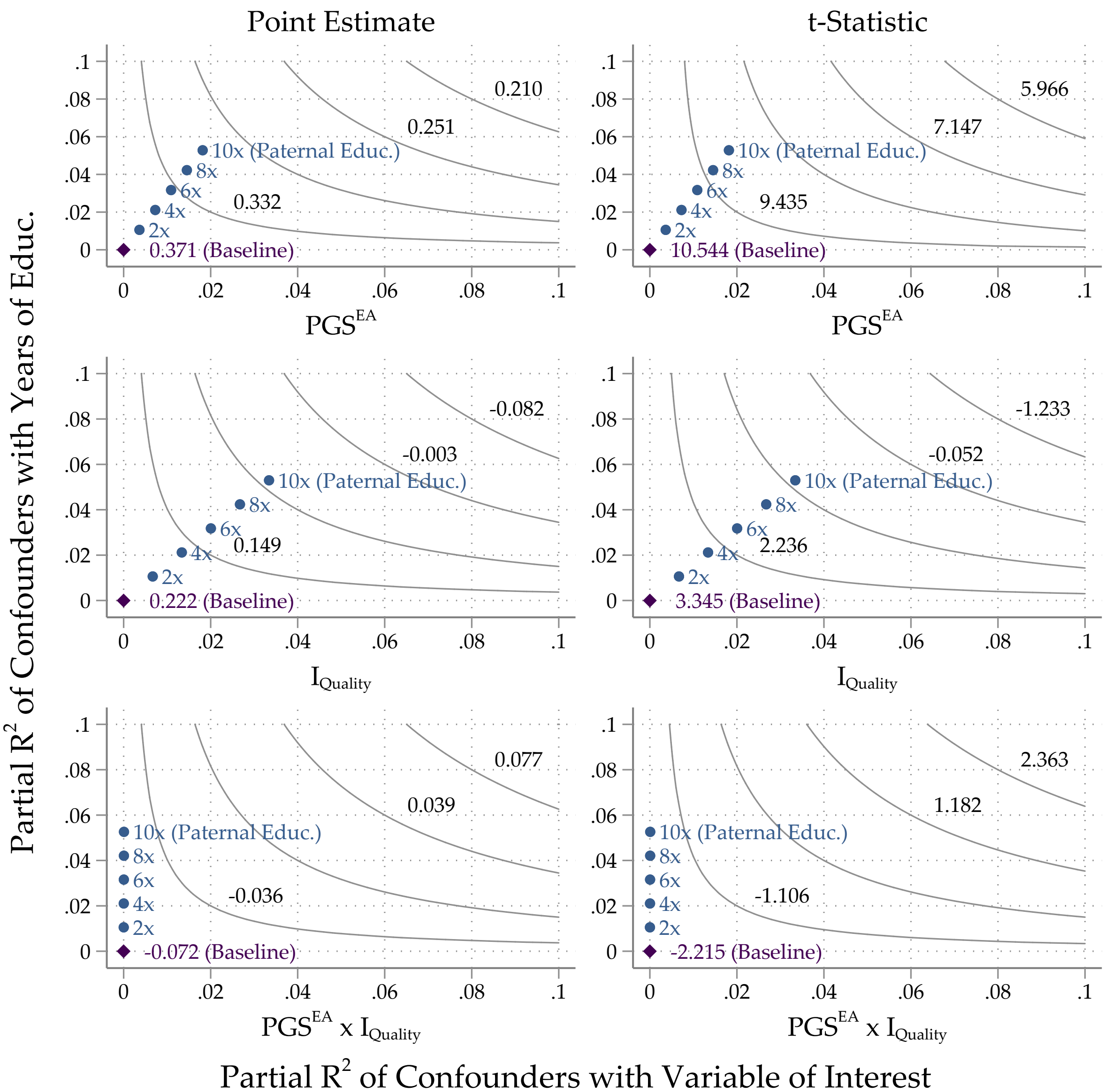
C.2 Supplementary Figures

Figure C.1: Top 3 School Expenditure Categories (in % of Total)



Data: Common Core of Data (CCD), National Public Education Financial Survey.
Note: Own calculations. This figure shows per-pupil expenditures shares in public elementary and secondary schools in the U.S. Teacher Instructional Expenditures includes teachers' salaries and employee benefits. Capital Outlay includes expenditures for property and for buildings and alterations completed by school district staff or contractors. Operation and Maintenance includes expenditures for the supervision of operations and maintenance, the operation of buildings, the care and upkeep of grounds and equipment, vehicle operations (other than student transportation) and maintenance, and security.

Figure C.2: Sensitivity To Unobserved Confounders



Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This figure shows the sensitivity of the point estimates for PGS^{EA}, I_{Qual}, and their interaction to unobserved confounding variables. Following the procedure of Cinelli and Hazlett (2020), we calculate the bias-adjusted treatment effect of $\text{PGS}^{\text{EA}}$and I_{Qual}, and their interaction under different assumptions about the partial R^2 of confounding variables with the variables of interest and the partial R^2 of confounding variables with years of education. Each contour line shows point estimates (left-hand panel) and t-statistics (right-hand panel) for different combinations of the two partial R^2. Each circle shows resulting values for different multiples of confounders as strong as parental education. Diamonds show baseline estimates from Table 3.2. Standard errors are clustered at the school level.

C.3 Data Appendix

C.3.1 Outcome Variables

Educational Attainment. We measure educational attainment by total *years of education*. In each wave, respondents were asked about their highest level of education at the time of the interview. For each respondent, we use the most recent information and transform education levels into years of education following the mapping suggested by Domingue et al. (2015). Numeric values in parentheses: eighth grade or less (8), some high school (10), high school graduate (12), GED (12), some vocational/technical training (13), some community college (14), some college (14), completed vocational/technical training (14), associate or junior college degree (14), completed college (16), some graduate school (17), completed a master's degree (18), some postbaccalaureate professional education (18), some graduate training beyond a master's degree (19), completed post-baccalaureate professional education (19), completed a doctoral degree (20).

We use the most recent available information to construct the following measures for educational degrees: *High School* (including GED), *2-year College*, *4-year College*, and *Post-Graduate*. Two-year college degrees include associate and junior college degrees as well as vocational and technical training after high school. Four-year college degrees include bachelor's degrees. Post-graduate degrees include master's degrees, doctoral degrees, and post-baccalaureate professional degrees. If available, information is taken from wave 5; otherwise we take it from waves 4 or 3, respectively. We only include respondents for which we observe educational degrees when they are at least 27 years old at the time of observation. We assume an ordinal ranking of degrees (high school $<$ 2-year college $<$ 4-year college $<$ post-graduate) and assign the possession of a lower-ranked degree if a respondent obtained a higher-ranked degree. For example, we assume that a respondent has finished high school if he or she has obtained a college degree, even if we don't have explicit information about high school graduation status.

Health. We proxy *subjective health* by quality-adjusted life years (QALY) that we derive from self-assessed health (SAH) measures. We use information from waves 3 and 4, where participants were asked "in general, how is your health?". We convert their (categorical) responses into a continuous measure using a mapping proposed by Van Doorslaer and Jones (2003). Using information about objective health—the Health Utility Index Mark III—Van Doorslaer and Jones (2003) scale the intervals of the SAH categories. This approach yields "quality weights" for health between 0 and 1. The value for each health status category is as follows (quality weights in parentheses): "excellent" (0.9833), "very good" (0.9311), "good" (0.841), "fair" (0.707), and "poor" (0.401).[1] We average resulting QALY measures across waves 3 and 4.

[1] See Table 4 in Van Doorslaer and Jones (2003).

We construct an index of *objective health* based on information from wave 4. Specifically, we sum the standardized values about whether a respondent (i) is obese, (ii) has stage one hypertension, and (iii) has high cholesterol (as indicated by the respondent). Each item was answered with either "yes" (= 1) or "no" (= 0). We reverse-code our measure of objective health such that higher values indicate better health.

Cognitive Skills. The *Picture Vocabulary Test* (PVT) is a test for receptive hearing vocabulary and is a widely-used proxy for verbal ability and scholastic aptitude. To administer the PVT, an examiner presents a series of pictures to the respondent. There are four pictures per page, and the examiner speaks a word describing one of the pictures. The respondent then has to indicate the picture that the word describes. In our analysis, we use age-adjusted PVT percentile ranks from wave 3 (Harris, 2020).

Preferences. We construct two measures of preferences: *risk aversion* and *patience*. In waves 3 and 4, participants were asked (i) whether they like to take risks, and (ii) whether they live their life without much thought for the future. Questions were answered on a five-point Likert scale ranging from "strongly agree" to "strongly disagree." We reverse-code both measures and use averages from waves 3 and 4 in our analysis.

Personality. The Big Five personality traits are openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism (Almlund et al., 2011). We use information from wave 4 to construct personality measures. Participants were asked a set of questions that each relate to one of the five personality traits. Questions were answered on a five-point Likert scale ranging from "strongly agree" to "strongly disagree." We use averages of the following questions in our analysis. *Openness*: (i) "I have a vivid imagination," (ii) "I have difficulty understanding abstract ideas" (reverse-coded), (iii) "I am not interested in abstract ideas" (reverse-coded), (iv) "I do not have a good imagination" (reverse-coded). *Conscientiousness*: (i) "I get chores done right away," (ii) "I like order," (iii) "I often forget to put things back in their proper place" (reverse-coded), (iv) "I make a mess of things" (reverse-coded). *Extraversion*: (i) "I am the life of the party," (ii) "I talk to a lot of different people at parties," (iii) "I don't talk a lot" (reverse-coded), (iv) "I keep in the background" (reverse-coded). Agreeableness: (i) "I sympathize with others' feelings," (ii) "I feel others' emotions," (iii) "I am not interested in other people's problems" (reverse-coded), (iv) "I am not really interested in others" (reverse-coded). *Neuroticism*: (i) "I have frequent mood swings," (ii) "I get upset easily," (iii) "I am relaxed most of the time" (reverse-coded), (iv) "I seldom feel blue" (reverse-coded).

Parental Investment. To measure *parental time investments*, we use information on a series of activities that children have done with their mother or father in the last four weeks. Specifically, the child is asked whether he or she has (i) gone shopping, (ii) played a sport, (iii) gone to a religious service or church-related event, (iv) talked about someone he or she is dating, or a party he or she went to, (v) gone to a movie, play, museum, concert, or sports event, (vi) had a talk about a personal problem he or she was having, (vii) had a serious argument about him or her behavior, (viii) talked about his or her school work or grades, (ix) worked on a project for school, (x) talked about other things he or she is doing in school. Questions were answered with "yes" (= 1) or "no" (= 0). We standardize each response to have mean zero and standard deviation one and then sum by parent (Anderson, 2008; Kling et al., 2007).

C.3.2 Variables of Interest

Polygenic Scores. Add Health obtained saliva samples from consenting participants in wave 4. After quality control procedures, genotyped data is available for 9,974 individuals and 609,130 SNPs. Add Health uses this data and calculates a set of different PGS using summary statistics from existing GWAS. Our baseline measure PGS^{EA} is based on statistics from Lee et al. (2018). In our analysis, we also use the PGS for body mass index (*BMI*) (Yengo et al., 2018), attention deficit hyperactivity disorder (*ADHD*) (Demontis et al., 2019), *depressive symptoms* (Howard et al., 2019), *intelligence* (Savage et al., 2018), *smoking* (Liu et al., 2019), and *sleep duration* (Jansen et al., 2019). All polygenic scores are standardized to $\mu = 0$ and $\sigma = 1$ on the full sample of genotyped Add Health respondents.

School Characteristics. In wave 1 and 2, Add Health administered questionnaires to headmasters of Add Health schools. We use this information to construct indicators for high school investments using a principal components analysis that includes the following school-level information: (i) average class size, (ii) share of teachers with a master degree, (iii) share of new teachers in the current school year, (iv) share of teachers with school-specific tenure of more than five years, and Herfindahl indices to measure teacher diversity with respect to (v) race and (vi) Hispanic background.[2] We also include school-level information about the average student-teacher ratio (number of full-time students per full-time equivalent teachers) in 1995/96, taken from the Common Core of Data (CCD) and the Private School Survey (PSS). We apply a factor rotation for interpretability reasons (oblique oblimin rotation of the Kaiser normalized matrix with $\gamma = 0$; see Gorsuch, 1983). The first component loads almost exclusively on average class size and average student-

[2]Herfindahl indices are calculated by first squaring the share of each component and then summing up resulting values (i.e. $H = \sum_{i=1}^{N} a_i^2$, where a_i is the share of component i, and N is the total number of components). For the Herfindahl index for race, we include the schools' share of full-time classroom teachers that are (i) White, (ii) Black or African American, (iii) American Indian or Native American, (iv) and Asian or Pacific Islander. For the Herfindahl index for Hispanic background, we include the schools' share of full-time classroom teachers that are (i) Hispanic or of Spanish origin, and (ii) neither Hispanic nor of Spanish origin.

Table C.6: Summary Statistics (Outcomes)

	Obs.	Mean	SD	Min	Max
Educational Attainment					
Years Education	3,081	14.81	2.25	8.00	20.00
High School Degree	3,081	0.97	0.18	0.00	1.00
2-year College Degree	3,081	0.53	0.50	0.00	1.00
4-year College Degree	3,081	0.42	0.49	0.00	1.00
Post-Graduate Degree	3,081	0.15	0.36	0.00	1.00
Health					
Subjective	3,081	0.91	0.07	0.40	0.98
Objective	3,081	0.03	1.94	-6.46	1.62
Cognitive Skills					
Picture Vocabulary Test	3,001	59.94	25.94	0.00	100.00
Preferences					
Risk Aversion	3,077	2.83	0.86	1.00	5.00
Patience	3,077	3.93	0.72	1.00	5.00
Personality					
Openness	3,059	3.63	0.63	1.00	5.00
Conscientousness	3,079	3.65	0.70	1.25	5.00
Extraversion	3,075	3.33	0.77	1.00	5.00
Agreeableness	3,077	3.87	0.58	1.00	5.00
Neuroticism	3,077	2.56	0.70	1.00	5.00
Parental Time Investments					
Mother	3,081	0.53	4.34	-8.51	14.89
Father	2,541	0.32	4.28	-6.47	16.74

Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This table shows summary statistics for outcome variables in our core analysis sample. The sample is restricted to genotyped individuals of (i) European descent, (ii) who visited an Add Health high school or an associated feeder school in wave 1, and (iii) who graduated from the same school. Observations with missing information in any of the displayed variables are dropped by list-wise deletion.

teacher ratio. Hence, we interpret this component, I_{Quant}, as an indicator for the "quantity" of teachers. The second component primarily loads positively on the percentage of teachers with a master degree and the share of teachers with a tenure of more than five years; it loads negatively on the share of new teachers in the current school year. We interpret this component, I_{Qual}, as an indicator for the "quality" of teachers. Both factors are coded such that higher values indicate higher school investments, i.e. higher teacher "quantity" investments (smaller classes) and higher teacher "quality" investments (better teachers), respectively. The calculated factors are orthogonal to each other by construction. They are standardized to $\mu = 0$ and $\sigma = 1$ on the full sample of Add Health high schools.[3]

Family Socio-Economic Status. We use the *social origins factor score* constructed by Belsky et al. (2018). Their measure uses information about parental education, parental occupation, household income, and household receipt of public assistance in wave 1. The score is standardized to $\mu = 0$ and $\sigma = 1$ on the full sample of Add Health respondents in wave 1.

Table C.7: Summary Statistics (Variables of Interest)

	Obs.	Mean	SD	Min	Max
Polygenic Scores					
PGS^{EA}	3,081	0.05	1.00	-4.13	3.39
BMI	3,081	-0.02	1.01	-3.42	3.56
ADHD	3,081	-0.05	1.00	-3.82	3.48
Depressive Symptoms	3,081	-0.02	1.01	-3.79	3.55
Intelligence	3,081	0.02	0.99	-3.30	4.06
Ever Smoker	3,081	-0.04	1.00	-4.25	4.25
Sleep Duration	3,081	0.02	0.99	-3.74	2.99
School Characteristics					
I_{Qual}	3,081	0.07	1.17	-3.90	2.30
I_{Quant}	3,081	-0.03	1.02	-3.34	3.25
Family SES					
Social Origins Factor Score	3,024	0.37	1.12	-4.40	3.51

Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This table shows summary statistics for variables of interest in our core analysis sample. The sample is restricted to genotyped individuals of (i) European descent, (ii) who visited an Add Health high school or an associated feeder school in wave 1, and (iii) who graduated from the same school. Observations with missing information in any of the displayed variables are dropped by list-wise deletion.

[3]Note that in an oblique rotation, factors may be slightly correlated.

C.3.3 Control Variables

Child Characteristics. The child's *gender* (female or male, as indicated by the interviewer) is taken from the in-home questionnaire in wave 1.

We calculate the child's *age* (in months) at each wave by subtracting the child's birth date from the date of interview. Because birth dates have minor inconsistencies across waves, we take averages across waves 1 to 4.

We use the first 20 *principal components* of full matrix of the genetic data. The components are obtained from a principal components analysis on the matrix of SNPs in Add Health (see Braudt and Harris, 2020, for a discussion). The principal components are standardized to $\mu = 0$ and $\sigma = 1$ on the full sample of genotyped Add Health respondents.

Family Socio-Economic Status. We use information from wave 1 to construct measures of *parents' education*. We transform parents' highest degree into years of education following the mapping suggested by Domingue et al. (2015). Numeric values in parentheses: never went to school (0), eighth grade or less (8), some high school (10), completed vocational/technical training instead of high school (10), went to school but level unknown (12), respondent doesn't know (12), high school graduate (12), GED (12), completed vocational/technical training after high school (14), some college (14), completed college (16), professional training beyond a master's degree (19). Where available, mothers' and fathers' education refers to the resident parent. If this information is not available, we use the biological parents' education instead.

Information about *mother's age at birth* (in years) is obtained from wave 1 if available, and wave 2 otherwise. To calculate age at birth, we take information about mother's age (as indicated by the child) and subtract the age of the child at the respective wave.

Information about religion (*Christian* or not) is obtained from wave 1 (as indicated by the child).

We calculate *potential wages* for population group g in time period t according to the following formula (Shenhav, 2021):

$$\hat{w}_{gt} = \sum_j \frac{E_{jg,1970}}{E_{g,1970}} \times \sum_o \frac{E_{ojg,1970}}{E_{jg,1970}} \left(\pi_{ojt,-r}\right) \times w_{ojt,-r},$$

where $\frac{E_{jg,1970}}{E_{g,1970}}$ describes the group-specific employment share of industry j in 1970, $\frac{E_{ojg,1970}}{E_{jg,1970}}$ describes the group- and industry-specific employment share of occupation o in 1970, $\pi_{ojt,-r}$ describes the leave-region-out industry-specific employment growth in occupation o for year t relative to 1970 (scaled by the overall employment growth in occupation o for year t relative to 1970), and $w_{ojt,-s}$ describes the leave-region-out average hourly wage paid in year t

for each occupation/industry/region cell. We define groups g by individuals that are homogeneous in gender (male, female), educational attainment (< High School, High School, > High School), and ethnicity (Non-Hispanic White, Hispanic, Non-Hispanic Black). We define regions r by census regions (North-East, Midwest, South, West). Employment shares in 1970 are taken from the 1970 decennial census. Employment shares and wages in periods t are taken from the March Supplements of the Current Population Survey (CPS) over the time period 1975-2000. We match time series of $\hat{w}_{gt}$ to the parents of respondents in Add Health based on information about g. Then we calculate (i) mean potential wages across respondent ages 0–14, and (ii) the standard deviation in potential wages across respondent ages 0–14.

School Characteristics. We use information about school *peer characteristics* from the in-school questionnaire in wave 1. Specifically, for each school, we calculate average years of education of students' fathers, the share of single parents, and the average subjective likelihood of students to attend college. We transform the father's highest degree into years of education following the mapping suggested by Domingue et al. (2015). Numeric values in parentheses: never went to school (0), eighth grade or less (8), some high school (10), went to school but level unknown (12), respondent doesn't know (12), high school graduate (12), GED (12), completed vocational/technical training after high school (14), some college (14), completed college (16), professional training beyond a four-year college (19). For college aspiration, students indicate how likely it is that they will graduate from college. Responses range from "no chance" (= 0) to "it will happen" (= 8). We define a student to have college aspiration if his or her response is above "about 50-50" (= 4), and to have no college aspiration otherwise. To prevent mechanical correlation between school peer characteristics and respondent characteristics, we calculate averages and shares while excluding individual respondents (leave-one-out).

We use information from the school administrator questionnaire in wave 1 to construct measures of *sanction policies* by means of a principal components analysis. School administrators were asked what happens to a student who is caught in their school (i) cheating, (ii) fighting with another student, (iii) injuring another student, (iv) possessing alcohol, (v) possessing an illegal drug, (vi) possessing a weapon, (vii) drinking alcohol at school, (viii) using an illegal drug at school, (ix) smoking at school, (x) verbally abusing a teacher, (xi) physically injuring a teacher, and (xii) stealing school property. Responses are "minor action", "in-school suspension", "out-of-school suspension", and "expulsion." Administrators were asked about sanctions in response to both first and second occurrences. We apply a factor rotation for interpretability reasons (oblique oblimin rotation of the Kaiser normalized matrix with $\gamma = 0$; see Gorsuch, 1983). The first three components load on variables reflecting the school's strictness regarding (i) drug use, (ii) social misconduct, and (iii) academic misconduct. The calculated factors are orthogonal to each other by construction. They are standardized to $\mu = 0$ and $\sigma = 1$ on the full sample of Add Health high schools.[4]

We calculate *value-added measures* with respect to GPAs in subject s for cohort c visiting high school j following a two-step procedure (Chetty et al., 2014a):

$$\text{GPA}^s_{igjc} = \beta^s Z_{igjc} + \text{VA}^s_{jc} + \epsilon^s_{igjc},$$
$$\widehat{\text{VA}}^s_{jc} = \frac{1}{N}\sum_{i \in jc}^{N}(\text{VA}^s_{jc} + \widehat{\epsilon}^s_{igjc}).$$

Z_{igjc} contains grade fixed effects δ_g, lagged GPAs from grade levels $g-1$ for English, Math and Science as well as current and lagged grade- and subject-specific indicators for academic tracks in English, Math and Science (3 levels per grade times subject cell). To avoid mechanical relationships, we predict $\widetilde{\text{VA}}^s_{jc}$ excluding data from cohort c and choosing a weighting vector $\phi^s = [\phi^s_{c-5}, ..., \phi^s_{c+5}]$ that minimizes the out-of-sample mean-squared error. Hence, $\widetilde{\text{VA}}^s_{jc}$ is our best prediction based on other cohorts of how much school j will increase GPAs in subject s in one year of high school relative to the improvements of similar students at other schools. We calculate $\widetilde{\text{VA}}^s_{jc}$ for English, Math and Science. In turn, we run a principal component analysis and use the first principal component as the aggregate measure of school value-added. The principal component is standardized to $\mu = 0$ and $\sigma = 1$ on the full sample of high schools with available transcript data on Add Health respondents.

Table C.8: Summary Statistics (Controls)

	Obs.	Mean	SD	Min	Max
Child Characteristics					
Female	3,081	0.55	0.50	0.00	1.00
Age in Months (Wave 1)	3,081	193.64	19.76	144.00	256.00
Principal Component 1	3,081	0.00	0.01	-0.14	0.10
Principal Component 2	3,081	-0.00	0.01	-0.37	0.07
Principal Component 3	3,081	0.00	0.01	-0.10	0.02
Principal Component 4	3,081	0.00	0.01	-0.09	0.65
Principal Component 5	3,081	-0.00	0.01	-0.07	0.18
Principal Component 6	3,081	-0.00	0.01	-0.14	0.19
Principal Component 7	3,081	-0.00	0.01	-0.13	0.33
Principal Component 8	3,081	-0.00	0.01	-0.37	0.08
Principal Component 9	3,081	0.00	0.01	-0.06	0.07
Principal Component 10	3,081	-0.00	0.01	-0.58	0.26
Principal Component 11	3,081	0.00	0.01	-0.25	0.37
Principal Component 12	3,081	0.00	0.01	-0.39	0.18

[4]Note that in an oblique rotation, factors may be slightly correlated.

Principal Component 13	3,081	-0.00	0.01	-0.35	0.18
Principal Component 14	3,081	-0.00	0.01	-0.12	0.23
Principal Component 15	3,081	0.00	0.01	-0.28	0.23
Principal Component 16	3,081	0.00	0.02	-0.15	0.66
Principal Component 17	3,081	-0.00	0.01	-0.50	0.24
Principal Component 18	3,081	-0.00	0.01	-0.29	0.20
Principal Component 19	3,081	0.00	0.01	-0.26	0.46
Principal Component 20	3,081	-0.00	0.01	-0.18	0.27
Family SES					
Education Mother (in Years)	3,081	13.63	2.50	8.00	19.00
Education Father (in Years)	3,081	13.67	2.68	8.00	19.00
Maternal Age at Birth	3,081	25.49	4.83	16.00	44.33
Christian	3,081	0.82	0.38	0.00	1.00
Potential Wage/Hour Mother (Mean)	3,081	12.61	1.38	9.45	14.27
Potential Wage/Hour Father (Mean)	3,081	15.48	1.31	11.14	17.11
Potential Wage/Hour Mother (SD)	3,081	0.36	0.11	0.12	0.51
Potential Wage/Hour Father (SD)	3,081	0.40	0.08	0.20	0.65
School Characteristics					
Peer Characteristics (Educ. Father)	2,965	13.57	1.05	10.90	17.84
Peer Characteristics (Single Parents)	2,965	0.24	0.08	0.00	0.60
Peer Characteristics (College Aspir.)	2,965	0.76	0.08	0.44	1.00
Sanction Policies (Drugs)	2,999	0.15	1.87	-5.71	9.06
Sanction Policies (Social)	2,999	0.25	1.61	-6.30	5.00
Sanction Policies (Acad.)	2,999	0.04	1.22	-3.41	2.38
Value-Added (GPA)	2,773	0.21	1.55	-4.18	4.41

Data: National Longitudinal Study of Adolescent to Adult Health.
Note: Own calculations. This table shows summary statistics for control variables in our core analysis sample. The sample is restricted to genotyped individuals of (i) European descent, (ii) who visited an Add Health high school or an associated feeder school in wave 1, and (iii) who graduated from the same school. Observations with missing information in any of the displayed variables are dropped by list-wise deletion.

Bibliography

Aaronson, D., L. Barrow, and W. Sander (2007). "Teachers and Student Achievement in the Chicago Public High Schools". *Journal of Labor Economics* 25 (1), pp. 95–135.

Abdellaoui, A., D. Hugh-Jones, L. Yengo, K. Kemper, M. Nivard, L. Veul, Y. Holtz, B. Zietsch, T. Frayling, N. Wray, J. Yang, K. Verweij, and P. M. Visscher (2019). "Genetic Correlates of Social Stratification in Great Britain". *Nature Human Behaviour* (3), pp. 1332–1342.

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

Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011). "Personality Psychology and Economics". *Handbook of the Economics of Education*. Ed. by E. A. Hanushek, S. Machin, and L. Woessmann. Vol. 4. Elsevier, pp. 1–181.

Almond, D. and B. Mazumder (2013). "Fetal Origins and Parental Responses". *Annual Review of Economics* 5 (1), pp. 37–56.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005). "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools". *Journal of Political Economy* 113 (1), pp. 151–184.

Altonji, J. G. and R. K. Mansfield (2018). "Estimating Group Effects Using Averages of Observables to Control for Sorting on Unobservables: School and Neighborhood Effects". *American Economic Review* 108 (10), pp. 2902–2946.

Anderson, M. L. (2008). "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects". *Journal of the American Statistical Association* 103 (484), pp. 1481–1495.

Andrabi, T., J. Das, A. I. Khwaja, and T. Zajonc (2011). "Do Value-added Estimates Add Value? Accounting for Learning Dynamics". *American Economic Journal: Applied Economics* 3 (3), pp. 29–54.

Angrist, J. D., S. R. Cohodes, S. M. Dynarski, P. A. Pathak, and C. R. Walters (2016). "Stand and Deliver: Effects of Boston's Charter High Schools on College Preparation, Entry, and Choice". *Journal of Labor Economics* 34 (2), pp. 275–318.

Angrist, J. D. and V. Lavy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement". *Quarterly Journal of Economics* 114 (2), pp. 533–575.

Angrist, J. D., V. Lavy, J. Leder-Luis, and A. Shany (2019). “Maimonides’ Rule Redux”. *American Economic Review: Insights* 1 (3), pp. 309–324.

Angrist, J. D., P. A. Pathak, and C. R. Walters (2013). “Explaining Charter School Effectiveness”. *American Economic Journal: Applied Economics* 5 (4), pp. 1–27.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton: Princeton University Press.

Arellano, M. and S. R. Bond (1991). “Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations”. *Review of Economic Studies* 58 (2), pp. 277–297.

Attanasio, O., R. Blundell, G. Conti, and G. Mason (2020). “Inequality in Socio-emotional Skills: A Cross-cohort Comparison”. *Journal of Public Economics* 191, p. 104171.

Bacher-Hicks, A., S. B. Billings, and D. J. Deming (2019). “The School to Prison Pipeline: Long-run Impacts of School Suspensions on Adult Crime”. *NBER Working Paper* 26257.

Barcellos, S. H., L. Carvalho, and P. Turley (2021). “The Effect of Education on the Relationship between Genetics, Early-life Disadvantages, and Later-life SES”. *NBER Working Paper* 28750.

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

Baumrind, D. (1967). “Child Care Practices Anteceding Three Patterns of Preschool Behavior”. *Genetic Psychology Monographs* 75 (1), pp. 43–88.

— (1971). “Current Patterns of Parental Authority”. *Developmental Psychology Monograph* 4 (1, Pt. 2), pp. 1–103.

— (1978). “Parental Disciplinary Patterns and Social Competence in Children”. *Youth & Society* 9 (3), pp. 239–276.

— (1989). “Rearing Competent Children”. *Child Development Today and Tomorrw*. Ed. by W. Damon. Jossey-Bass/Wiley, pp. 349–378.

— (1991a). “Parenting Styles and Adolescent Development”. *The Encyclopedia on Adolescence*. Ed. by R. M. Lerner, A. C. Petersen, and J. Brooks-Gunn. New York: Garland Publishing, pp. 746–758.

— (1991b). “The Influence of Parenting Style on Adolescent Competence and Substance Use”. *Journal of Early Adolescence* 11 (1), pp. 56–95.

Becker, A., T. Deckers, T. Dohmen, A. Falk, and F. Kosse (2012). “The Relationship Between Economic Preferences and Psychological Personality Measures”. *Annual Review of Economics* 4 (1), pp. 453–478.

Belsky, D. W., B. W. Domingue, R. Wedow, L. Arseneault, J. D. Boardman, A. Caspi, D. Conley, J. M. Fletcher, J. Freese, P. Herd, T. E. Moffitt, R. Poulton, K. Sicinski, J. Wertz, and K. M. Harris (2018). “Genetic Analysis of Social-class Mobility in Five Longitudinal Studies”. *Proceedings of the National Academy of Sciences* 115 (31), E7275–E7284.

Belsky, D. W., T. E. Moffitt, D. L. Corcoran, B. W. Domingue, H. Harrington, S. Hogan, R. Houts, S. Ramrakha, K. Sugden, B. S. Williams, R. Poulton, and A. Caspi (2016). “The Genetics of Success: How Single-nucleotide Polymorphisms Associated with Educational

Attainment Relate to Life-course Development". *Psychological Science* 27 (7), pp. 957–972.

Beuermann, D., C. K. Jackson, L. Navarro-Sola, and F. Pardo (2018). "What is a Good School, and Can Parents Tell? Evidence on the Multidimensionality of School Output". *NBER Working Paper* 25342.

Bharadwaj, P., J. P. Eberhard, and C. A. Neilson (2018). "Health at Birth, Parental Investments, and Academic Outcomes". *Journal of Labor Economics* 36 (2), pp. 349–394.

Biasi, B. (forthcoming). "School Finance Equalization Increases Intergenerational Mobility". *Journal of Labor Economics*.

Bietenbeck, J. (2019). "The Long-term Impacts of Low-achieving Childhood Peers: Evidence from Project STAR". *Journal of the European Economic Association* 18 (1), pp. 392–426.

Blundell, R. and S. R. Bond (1998). "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models". *Journal of Econometrics* 87 (1), pp. 115–143.

Borghans, L., B. H. H. Golsteyn, J. Heckman, and J. E. Humphries (2016). "What Grades and Achievement Tests Measure". *Proceedings of the National Academy of Sciences* 113 (47), pp. 13354–13359.

Bourguignon, F., F. H. G. Ferreira, and M. Menéndez (2007). "Inequality of Opportunity in Brazil". *Review of Income and Wealth* 53 (4), pp. 585–618.

Branigan, A. R., K. J. McCallum, and J. Freese (2013). "Variation in the Heritability of Educational Attainment: An International Meta-analysis". *Social Forces* 92 (1), pp. 109–140.

Braudt, D. and K. M. Harris (2020). *Polygenic Scores (PGSs) in the National Longitudinal Study of Adolescent to Adult Health (Add Health)—Release 2*. Carolina Population Center.

Brinch, C. N. and T. A. Galloway (2012). "Schooling in Adolescence Raises IQ Scores". *Proceedings of the National Academy of Sciences* 109 (2), pp. 425–430.

Brunori, P., P. Hufe, and D. Mahler (2021). "The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees and Forests". *IZA Working Paper* 14689.

Buerger, C., S. H. Lee, and J. D. Singleton (2021). "Test-based Accountability and the Effectiveness of School Finance Reforms". *AEA Papers and Proceedings* 111, pp. 455–459.

Burks, S. V., C. Lewis, P. A. Kivi, A. Wiener, J. E. Anderson, L. Götte, C. G. DeYoung, and A. Rustichini (2015). "Cognitive Skills, Personality, and Economic Preferences in Collegiate Success". *Journal of Economic Behavior & Organization* 115, pp. 30–44.

Buser, T., R. Ahlskog, M. Johannesson, P. D. Koellinger, and S. Oskarsson (2021a). "Using Genes to Explore the Effects of Cognitive and Non-cognitive Skills on Education and Labor Market Outcomes". *Tinbergen Institute Discussion Paper* 2021-088.

Buser, T., M. Niederle, and H. Oosterbeek (2021b). "Can Competitiveness Predict Education and Labor Market Outcomes? Evidence from Incentivized Choice and Survey Measures". *NBER Working Paper* 28916.

Callaway, B. and P. H. C. Sant'Anna (2020). "Difference-in-differences with Multiple Time Periods". *Journal of Econometrics* 225 (2), pp. 200–230.

Camarero Garcia, S. (2022). "Inequality of Educational Opportunities and the Role of Learning Intensity". *Labour Economics* 74, p. 102084.

Card, D. and A. Payne (2002). "School Finance Reform, the Distribution of School Spending, and the Distribution of Student Test Scores". *Journal of Public Economics* 83 (1), pp. 49–82.

Castillo, M., J. L. Jordan, and R. Petrie (2018a). "Discount Rates of Children and High School Graduation". *Economic Journal* 129 (619), pp. 1153–1181.

— (2018b). "Children's Rationality, Risk Attitudes and Field Behavior". *European Economic Review* 102, pp. 62–81.

Cesarini, D. and P. M. Visscher (2017). "Genetics and Educational Attainment". *npj Science of Learning* 2 (4).

Chan, T. W. and A. Koo (2011). "Parenting Style and Youth Outcomes in the UK". *European Sociological Review* 27 (3), pp. 385–399.

Chanock, S. J., T. Manolio, M. Boehnke, E. Boerwinkle, D. J. Hunter, G. Thomas, J. N. Hirschhorn, G. R. Abecasis, D. Altshuler, J. E. Bailey-Wilson, et al. (2007). "Replicating Genotype-phenotype Associations". *Nature* 447, pp. 655–660.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011). "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR". *Quarterly Journal of Economics* 126 (4), pp. 1593–1660.

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

— (2014b). "Measuring the Impacts of Teachers II: Teacher Value-added and Student Outcomes in Adulthood". *American Economic Review* 104 (9), pp. 2633–79.

Cinelli, C. and C. Hazlett (2020). "Making Sense of Sensitivity: Extending Omitted Variable Bias". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82 (1), pp. 39–67.

Clark, A. E. and A. Lepinteur (2019). "The Causes and Consequences of Early-adult Unemployment: Evidence from Cohort Data". *Journal of Economic Behavior and Organization* 166, pp. 107–124.

Clotfelter, C. T., H. F. Ladd, and J. L. Vigdor (2010). "Teacher Credentials and Student Achievement in High School: A Cross-subject Analysis with Student Fixed Effects". *Journal of Human Resources* 45 (3), pp. 655–681.

Cobb-Clark, D. A., N. Salamanca, and A. Zhu (2019). "Parenting Style as an Investment in Human Development". *Journal of Population Economics* 32 (4), pp. 1315–1352.

Coleman, J. S., E. Campbell, C. Hobson, J. McPartland, A. Mood, F. Weinfeld, and R. York (1966). *Equality of Educational Opportunity.* U.S. Department of Health, Education & Welfare.

Coons, J. E., W. H. Clune, and S. D. Sugarman (1970). *Private Wealth and Public Education.* Cambridge, MA: Belknap Press.

Cornelissen, T. and C. Dustmann (2019). "Early School Exposure, Test Scores, and Noncognitive Outcomes". *American Economic Journal: Microeconomics* 11 (2), pp. 35–63.

Cuartas, J. (2022). “The Effect of Spanking on Early Social-emotional Skills”. *Child Development* 93, pp. 180–193.

Cullen, J. B., B. A. Jacob, and S. Levitt (2006). “The Effect of School Choice on Participants: Evidence from Randomized Lotteries”. *Econometrica* 74 (5), pp. 1191–1230.

Cunha, F. and J. Heckman (2007). “The Technology of Skill Formation”. *American Economic Review* 97 (2), pp. 31–47.

Cunha, F., J. Heckman, L. Lochner, and D. V. Masterov (2006). “Interpreting the Evidence on Life Cycle Skill Formation”. Ed. by E. A. Hanushek and F. Welch. Vol. 1. Handbook of the Economics of Education. Amsterdam: Elsevier. Chap. 12, pp. 697–812.

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

Darling, N. and L. Steinberg (1993). “Parenting Style as Context: An Integrative Model”. *Psychological Bulletin* 113 (3), pp. 487–496.

Darwin, C. R. (1859). *On the Origin of Species*. London: John Murray.

De Chaisemartin, C. and X. D’Haultfoeuille (2021). “Two-way Fixed Effects and Differences-in-differences with Heterogeneous Treatment Effects: A Survey”. *NBER Working Paper* 29691.

De Chaisemartin, C. and X. D’Haultfœuille (2020). “Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects”. *American Economic Review* 110 (9), pp. 2964–2996.

Del Bono, E., M. Francesconi, Y. Kelly, and A. Sacker (2016). “Early Maternal Time Investment and Early Child Outcomes”. *Economic Journal* 126 (596), F96–F135.

Demange, P. A., J. J. Hottenga, A. Abdellaoui, M. Malanchini, B. W. Domingue, E. L. de Zeeuw, K. Rimfeld, T. Eley, D. I. Boomsma, E. van Bergen, G. Breen, M. G. Nivard, and R. Cheesman (2020). “Parental Influences on Offspring Education: Indirect Genetic Effects of Non-cognitive Skills”. *bioRxiv* 2020.09.15.296236.

Demange, P. A., M. Malanchini, T. T. Mallard, P. Biroli, S. R. Cox, A. D. Grotzinger, E. M. Tucker-Drob, A. Abdellaoui, L. Arseneault, A. Caspi, D. Corcoran, B. W. Domingue, C. Mitchell, E. van Bergen, D. I. Boomsma, K. M. Harris, H. F. Ip, T. E. Moffitt, R. Poulton, J. Prinz, K. Sugden, J. Wertz, B. Williams, E. L. de Zeeuw, D. W. Belsky, K. P. Harden, and M. G. Nivard (2021). “Investigating the Genetic Architecture of Non-cognitive Skills Using GWAS-by-subtraction”. *Nature Genetics* 53, pp. 35–44.

Demontis, D., R. K. Walters, J. Martin, M. Mattheisen, T. D. Als, E. Agerbo, G. Baldursson, R. Belliveau, J. Bybjerg-Grauholm, M. Bækvad-Hansen, et al. (2019). “Discovery of the First Genome-wide Significant Risk Loci for Attention Deficit/Hyperactivity Disorder”. *Nature Genetics* 51 (1), pp. 63–75.

Deng, L. and T. Tong (2020). “Parenting Style and the Development of Noncognitive Ability in Children”. *China Economic Review* 62, p. 101477.

Descartes, R. (1641). “Meditations on First Philosophy”. *Descartes: Selected Philosophical Writings*. Ed. by J. Cottingham, D. Murdoch, R. Stoothoff, and A. Kenny. Cambridge: Cambridge University Press, pp. 73–122.

Doepke, M. and F. Zilibotti (2017). “Parenting with Style: Altruism and Paternalism in Intergenerational Preference Transmission”. *Econometrica* 85 (5), pp. 1331–1371.

Doepke, M. and F. Zilibotti (2019). *Love, Money, and Parenting: How Economics Explains the Way We Raise Our Kids*. Princeton University Press.

Domingue, B. W., D. W. Belsky, D. Conley, K. M. Harris, and J. Boardman (2015). "Polygenic Influence on Educational Attainment: New Evidence From the National Longitudinal Study of Adolescent to Adult Health". *AERA Open* 1 (3), pp. 1–13.

Domingue, B. W., S. Trejo, E. Armstrong-Carter, and E. M. Tucker-Drob (2020). "Interactions between Polygenic Scores and Environments: Methodological and Conceptual Challenges". *Sociological Science* 7 (19), pp. 465–486.

Dooley, M. and J. Stewart (2007). "Family Income, Parenting Styles and Child Behavioural-emotional Outcomes". *Health Economics* 16 (2), pp. 145–162.

Dudbridge, F. (2013). "Power and Predictive Accuracy of Polygenic Risk Scores". *PLOS Genetics* 9 (3), e1003348.

Durkin, K. (1995). *Developmental Social Psychology: From Infancy to Old Age*. Blackwell Publishing.

Ermisch, J. (2008). "Origins of Social Immobility and Inequality: Parenting and Early Child Development". *National Institute Economic Review* 205, pp. 62–71.

Falk, A., A. Becker, T. J. Dohmen, B. Enke, D. B. Huffman, and U. Sunde (2018). "Global Evidence on Economic Preferences". *Quarterly Journal of Economics* 133 (4), pp. 1645–1692.

Ferreira, F. H. G. and J. Gignoux (2011). "The Measurement of Inequality of Opportunity: Theory and an Application to Latin America". *Review of Income and Wealth* 57 (4), pp. 622–657.

Fiorini, M. and M. P. Keane (2014). "How the Allocation of Children's Time Affects Cognitive and Noncognitive Development". *Journal of Labor Economics* 32 (4), pp. 787–836.

Fleurbaey, M. and V. Peragine (2013). "Ex Ante Versus Ex Post Equality of Opportunity". *Economica* 80 (317), pp. 118–130.

Flood, S., M. King, R. Rodgers, S. Ruggles, and J. R. Warren (2020). "Integrated Public Use Microdata Series, Current Population Survey: Version 8.0 [dataset]". *Minneapolis, MN: IPUMS*.

Francesconi, M. and J. Heckman (2016). "Child Development and Parental Investment: Introduction". *Economic Journal* 126 (596), F1–F27.

Fredriksson, P., B. Öckert, and H. Oosterbeek (2013). "Long-term Effects of Class Size". *Quarterly Journal of Economics* 128 (1), pp. 249–285.

Goldin, C. and L. Katz (2010). *The Race between Education and Technology*. Harvard University Press.

Goodman, R. (1997). "The Strengths and Difficulties Questionnaire: A Research Note". *The Journal of Child Psychology and Psychiatry* 38 (5), pp. 581–586.

Goodman-Bacon, A. (2021). "Difference-in-differences with Variation in Treatment Timing". *Journal of Econometrics* 225 (2), pp. 254–277.

Gorsuch, R. (1983). *Factor Analysis*. 2nd ed. Hillsdale, NJ: Psychology Press.

Hanushek, E. A., L. Kinne, P. Lergetporer, and L. Woessmann (forthcoming). "Patience, Risk-taking, and Human Capital Investment across Countries". *Economic Journal*.

Hanushek, E. A., P. E. Peterson, L. M. Talpey, and L. Woessmann (2019). "The Unwavering SES Achievement Gap: Trends in U.S. Student Performance". *NBER Working Paper* 25648.

Hanushek, E. A. and S. G. Rivkin (2010). "Generalizations about Using Value-added Measures of Teacher Quality". *American Economic Review* 100 (2), pp. 267–271.

Hanushek, E. A., S. G. Rivkin, and J. C. Schiman (2016). "Dynamic Effects of Teacher Turnover on the Quality of Instruction". *Economics of Education Review* 55, pp. 132–148.

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

— (2012). "Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation". *Journal of Economic Growth* 17 (4), pp. 267–321.

Harris, K. M. (2020). "National Longitudinal Study of Adolescent to Adult Health, Public Use Peabody Picture Vocabulary Test (PVT) Score Data, Wave III [dataset]". *National Longitudinal Study of Adolescent to Adult Health*.

Heckman, J., S. H. Moon, R. Pinto, P. Savelyev, and A. Yavitz (2010). "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program". *Quantitative Economics* 1 (1), pp. 1–46.

Heckman, J. and S. Mosso (2014). "The Economics of Human Development and Social Mobility". *Annual Review of Economics* 6 (1), pp. 689–733.

Hendren, N. and B. Sprung-Keyser (2020). "A Unified Welfare Analysis of Government Policies". *The Quarterly Journal of Economics* 135 (3), pp. 1209–1318.

Horowitz, H. W. (1965). "Unseparate but Unequal: The Emerging Fourteenth Amendment Issue in Public School Education". *UCLA Law Review* 13, pp. 1147–1172.

Hothorn, T., K. Hornik, and A. Zeileis (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework". *Journal of Computational and Graphical Statistics* 15 (3), pp. 651–674.

Houmark, M. A., V. Ronda, and M. Rosholm (2020). "The Nurture of Nature and the Nature of Nurture: How Genes and Investments Interact in the Formation of Skills". *IZA Discussion Paper Series* 13780.

Howard, D. M., M. J. Adams, T.-K. Clarke, J. D. Hafferty, J. Gibson, M. Shirali, J. R. Coleman, S. P. Hagenaars, J. Ward, E. M. Wigmore, et al. (2019). "Genome-wide Meta-analysis of Depression Identifies 102 Independent Variants and Highlights the Importance of the Prefrontal Brain Regions". *Nature Neuroscience* 22 (3), pp. 343–352.

Hoxby, C. M. (2001). "All School Finance Equalizations are not Created Equal". *The Quarterly Journal of Economics* 116 (4), pp. 1189–1231.

Hufe, P., R. Kanbur, and A. Peichl (forthcoming). "Measuring Unfair Inequality : Reconciling Equality of Opportunity and Freedom from Poverty". *Review of Economic Studies*.

Hufe, P. and A. Peichl (2019). "Beyond Equal Rights: Equality of Opportunity in Political Participation". *Review of Income and Wealth* 66 (3), pp. 477–511.

Humphries, J. E. and F. Kosse (2017). "On the Interpretation of Non-cognitive Skills—What is Being Measured and Why it Matters". *Journal of Economic Behavior & Organization* 136, pp. 174–185.

Isphording, I. E. and U. Zölitz (2020). “The Value of a Peer”. *University of Zurich, Department of Economics, Working Paper* 342.

Jackson, C. K. (2019). “What Do Test Scores Miss? The Importance of Teacher Effects on Non-test Score Outcomes”. *Journal of Political Economy* 126 (5), pp. 2072–2107.

Jackson, C. K., R. C. Johnson, and C. Persico (2015). “The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms”. *Quarterly Journal of Economics* 131 (1), pp. 157–218.

Jackson, C. K. and C. Mackevicius (2021). “The Distribution of School Spending Impacts”. *NBER Working Paper* 28517.

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

Jackson, M. I. (2009). “Understanding Links between Adolescent Health and Educational Attainment”. *Demography* 46 (4), pp. 671–694.

Jacob, B. A., J. E. Rockoff, E. S. Taylor, B. Lindy, and R. Rosen (2018). “Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools”. *Journal of Public Economics* 166, pp. 81–97.

Jansen, P., K. Watanabe, S. Stringer, N. Skene, J. Bryois, A. Hammerschlag, C. de Leeuw, J. Benjamins, A. Muñoz-Manchado, M. Nagel, et al. (2019). “Genome-wide Analysis of Insomnia in 1,331,010 Individuals Identifies New Risk Loci and Functional Pathways”. *Nature Genetics* 51 (3), pp. 394–403.

Jensen, B., P. Jensen, and A. W. Rasmussen (2017). “Does Professional Development of Preschool Teachers Improve Children’s Socio-emotional Outcomes?” *Labour Economics* 45, pp. 26–39.

Johnson, R. C. and C. K. Jackson (2019). “Reducing Inequality through Dynamic Complementarity: Evidence from Head Start and Public School Spending”. *American Economic Journal: Economic Policy* 11 (4), pp. 310–49.

Jones, J. D., R. C. Fraley, K. B. Ehrlich, J. A. Stern, C. W. Lejuez, P. R. Shaver, and J. Cassidy (2018). “Stability of Attachment Style in Adolescence: An Empirical Test of Alternative Developmental Processes”. *Child Development* 89 (3), pp. 871–880.

Keller, M. (2014). “Gene × Environment Interaction Studies have not Properly Controlled for Potential Confounders: The Problem and the (Simple) Solution”. *Biological Psychiatry* 75 (1), pp. 18–24.

Kessler, R. C., P. R. Barker, L. J. Colpe, J. F. Epstein, J. C. Gfroerer, E. Hiripi, M. J. Howes, S.-L. T. Normand, R. W. Manderscheid, E. E. Walters, and A. M. Zaslavsky (2003). “Screening for Serious Mental Illness in the General Population”. *Archives of General Psychiatry* 60 (2), pp. 184–189.

Khanam, R. and S. Nghiem (2016). “Family Income and Child Cognitive and Noncognitive Development in Australia: Does Money Matter?” *Demography* 53 (3), pp. 597–621.

Kirkebøen, L. (2021). “School Value-added and Long-term Student Outcomes”. *Statistics Norway Discussion Papers* 970.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007). “Experimental Analysis of Neighborhood Effects”. *Econometrica* 75 (1), pp. 83–119.

Koellinger, P. D. and K. P. Harden (2018). “Using Nature to Understand Nurture”. *Science* 359 (6374), pp. 386–387.

Kong, A., G. Thorleifsson, M. L. Frigge, B. J. Vilhjalmsson, A. I. Young, T. E. Thorgeirsson, S. Benonisdottir, A. Oddsson, B. V. Halldorsson, G. Masson, D. F. Gudbjartsson, A. Helgason, G. Bjornsdottir, U. Thorsteinsdottir, and K. Stefansson (2018). “The Nature of Nurture: Effects of Parental Genotypes”. *Science* 359 (6374), pp. 424–428.

Krueger, A. B. (1999). “Experimental Estimates of Education Production Functions”. *Quarterly Journal of Economics* 114 (2), pp. 497–532.

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

Kweon, H., C. A. Burik, R. K. Linnér, R. de Vlaming, A. Okbay, D. Martschenko, K. P. Harden, T. A. DiPrete, and P. D. Koellinger (2020). “Genetic Fortune: Winning or Losing Education, Income, and Health”. *Tinbergen Institute Discussion Papers* 20-053.

Lafortune, J., J. Rothstein, and D. W. Schanzenbach (2018). “School Finance Reform and the Distribution of Student Achievement”. *American Economic Journal: Applied Economics* 10 (2), pp. 1–26.

Laliberté, J.-W. (2021). “Long-term Contextual Effects in Education: Schools and Neighborhoods”. *American Economic Journal: Economic Policy* 13 (2), pp. 336–377.

Lamarck, J.-B. P. A. d. M. de (1838). *Histoire Naturelle des Animaux sans Vertèbres.* Paris: Verdière.

Lamborn, S. D., N. S. Mounts, and L. Steinberg (1991). “Patterns of Competence and Adjustment among Adolescents from Authoritative, Authoritarian, Indulgent, and Neglectful Families”. *Child Development* 62 (5), pp. 1049–1065.

Layard, R., A. E. Clark, F. Cornaglia, N. Powdthavee, and J. Vernoit (2014). “What Predicts a Successful Life? A Life-course Model of Well-being”. *Economic Journal* 124 (580), F720–F738.

Lee, J. J. 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), pp. 1112–1121.

Leuven, E. and S. A. Løkken (2020). “Long-term Impacts of Class Size in Compulsory School”. *Journal of Human Resources* 55 (1), pp. 309–348.

Liu, M., Y. Jiang, R. Wedow, Y. Li, D. Brazel, F. Chen, G. Datta, J. Davila-Velderrain, D. McGuire, C. Tian, et al. (2019). “Association Studies of up to 1.2 Million Individuals Yield New Insights into the Genetic Etiology of Tobacco and Alcohol Use”. *Nature Genetics* 51 (2), pp. 237–244.

Locke, J. (1690). *An Essay Concerning Human Understanding.* Hamilton: McMaster University Archive for the History of Economic Thought.

Maccoby, E. and J. Martin (1983). “Socialization in the Context of the Family: Parent-child Interaction”. *Handbook of Child Psychology.* Ed. by M. Hetherington. Vol. 4. New York: Wiley, pp. 1–101.

Marcus, M. and P. H. C. Sant’Anna (2021). “The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics”. *Journal of the Association of Environmental and Resource Economists* 8 (2), pp. 235–275.

Martin, A. R., C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale, S. Gravel, M. J. Daly, C. D. Bustamante, and E. E. Kenny (2017). “Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations”. *American Journal of Human Genetics* 100 (4), pp. 635–649.

Martschenko, D., S. Trejo, and B. W. Domingue (2018). “Genetics-infused Education Research: What Can and Can’t Knowledge of Genetics Bring to Education?” *SocArXiv*.

McElroy, E., M. Richards, E. Fitzsimons, G. Conti, G. B. Ploubidis, A. Sullivan, and V. Moulton (2021). *Feasibility of Retrospectively Harmonising Cognitive Measures in Five British Birth Cohort Studies*. CLOSER.

Mills, M. C., N. Barban, and F. C. Tropf (2020). *An Introduction to Statistical Genetic Data Analysis*. Cambridge, MA: MIT Press.

Moulton, V., E. McElroy, M. Richards, E. Fitzsimons, K. Northstone, G. Conti, G. B. Ploubidis, A. Sullivan, and D. O’Neill (2020). *A Guide to the Cognitive Measures in Five British Birth Cohort Studies*. CLOSER.

Murray, S. E., W. N. Evans, and R. M. Schwab (1998). “Education-finance Reform and the Distribution of Education Resources”. *American Economic Review* 88 (4), pp. 789–812.

Muslimova, D., H. van Kippersluis, C. A. Rietveld, S. von Hinke, and S. F. W. Meddens (2020). “Dynamic Complementarity in Skill Production: Evidence from Genetic Endowments and Birth Order”. *Tinbergen Institute Discussion Paper* 2020-082/V.

Nicoletti, C. and B. Rabe (2014). “School Inputs and Skills: Complementarity and Self-productivity”. *IZA Working Paper Series* 8693.

Noack, P. (2011). “Socialization”. *Encyclopedia of Adolescence*. Ed. by B. Brown and M. Prinstein. San Diego: Academic Press, pp. 352–359.

Oster, E. (2019). “Unobservable Selection and Coefficient Stability: Theory and Evidence”. *Journal of Business & Economic Statistics* 37 (2), pp. 187–204.

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

Papay, J. P. and M. A. Kraft (2015). “Productivity Returns to Experience in the Teacher Labor Market: Methodological Challenges and New Evidence on Long-term Career Improvement”. *Journal of Public Economics* 130, pp. 105–119.

Plewis, I. and S. Ketende (2006). *Millennium Cohort Study Technical Report on Response Centre for Longitudinal Studies*. Centre for Longitudinal Studies.

Plomin, R. (2018). *Blueprint: How DNA Makes Us Who We Are*. MIT Press.

Ramos, X. and D. Van de Gaer (Dec. 2016). “Approaches to Inequality of Opportunity: Principles, Measures and Evidence”. *Journal of Economic Surveys* 30 (5), pp. 855–883.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). “Teachers, Schools, and Academic Achievement”. *Econometrica* 73 (2), pp. 417–458.

Rockoff, J. E. (2004). “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data”. *American Economic Review* 94 (2), pp. 247–252.

Roemer, J. E. (1998). *Equality of Opportunity*. Cambridge, MA: Harvard University Press.

Ronda, V., E. Agerbo, D. Bleses, P. B. Mortensen, A. Børglum, D. M. Hougaard, O. Mors, M. Nordentoft, T. Werge, and M. Rosholm (forthcoming). "Family Disadvantage, Gender and the Returns to Genetic Human Capital". *Scandinavian Journal of Economics*.

Ronfeldt, M., S. Loeb, and J. Wyckoff (2013). "How Teacher Turnover Harms Student Achievement". *American Educational Research Journal* 50 (1), pp. 4–36.

Roth, J. and P. H. C. Sant'Anna (2021). "Efficient Estimation for Staggered Rollout Designs". *arXiv* 2102.01291.

Rothenberg, A., J. Lansford, M. Bornstein, L. M. Uribe Tirado, S. Yotanyamaneewong, L. P. Alampay, S. Al-Hassan, D. Bacchini, L. Chang, K. Deater-Deckard, et al. (2021). "Cross-cultural Associations of Four Parenting Behaviors with Child Flourishing: Examining Cultural Specificity and Commonality in Cultural Normativeness and Intergenerational Transmission Processes". *Child Development* 92 (6), e1138–e1153.

Rothstein, J. and D. W. Schanzenbach (2021). "Does Money Still Matter? Attainment and Earnings Effects of Post-1990 School Finance Reforms". *NBER Working Paper* 29177.

Ruggles, S., S. Flood, R. Goeken, J. Grover, E. Meyer, J. Pacas, and M. Sobek (2020). "Integrated Public Use Microdata Series USA: Version 10.0 [dataset]". *Minneapolis, MN: IPUMS*.

Sacerdote, B. (2014). "Experimental and Quasi-experimental Analysis of Peer Effects: Two Steps Forward?" *Annual Review of Economics* 6 (1), pp. 253–272.

Savage, J. E. et al. (2018). "Genome-wide Association Meta-analysis in 269,867 Individuals Identifies New Genetic and Functional Links to Intelligence". *Nature Genetics* 50 (7), pp. 912–919.

Schneider, J. and K. McGrew (2018). "The Cattell–Horn–Carroll Theory of Cognitive Abilities". *Contemporary Intellectual Assessment: Theories, Tests, and Issues*. Ed. by D. Flanagan and E. McDonough. The Guilford Press, pp. 73–163.

Selzam, S., S. J. Ritchie, J.-B. Pingault, C. A. Reynolds, P. F. O'Reilly, and R. Plomin (2019). "Comparing Within- and Between-family Polygenic Score Prediction". *American Journal of Human Genetics* 105 (2), pp. 351–363.

Shenhav, N. (2021). "Lowering Standards to Wed? Spouse Quality, Marriage, and Labor Market Responses to the Gender Wage Gap". *Review of Economics and Statistics* 103 (2), pp. 265–279.

Smith-Woolley, E., J.-B. Pingault, S. Selzam, K. Rimfeld, E. Krapohl, S. von Stumm, K. Asbury, P. S. Dale, T. Young, R. Allen, Y. Kovas, and R. Plomin (2018). "Differences in Exam Performance between Pupils Attending Selective and Non-selective Schools Mirror the Genetic Differences between Them". *npj Science of Learning* 3 (1), pp. 1–7.

Sotoudeh, R., K. M. Harris, and D. Conley (2019). "Effects of the Peer Metagenomic Environment on Smoking Behavior". *Proceedings of the National Academy of Sciences* 116 (33), pp. 16302–16307.

Spera, C. (2005). "A Review of the Relationship among Parenting Practices, Parenting Styles, and Adolescent School Achievement". *Educational Psychology Review* 17 (2), pp. 125–146.

Steinberg, L. (2001). "We Know Some Things: Parent-adolescent Relationships in Retrospect and Prospect". *Journal of Research on Adolescence* 11 (1), pp. 1–19.

Steinberg, L., S. D. Lamborn, and S. M. Dornbusch (1992). "Impact of Parenting Practices on Adolescent Achievement: Authoritative Parenting, School Involvement, and Encouragement to Succeed". *Child Development* 63 (5), pp. 1266–1281.

Straus, M. and S. Hamby (1997). *Measuring Physical & Psychological Maltreatment of Children with the Conflict Tactics Scales*. ERIC.

Sun, L. and S. Abraham (2020). "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects". *Journal of Econometrics* 225 (2), pp. 175–199.

Todd, P. E. and K. I. Wolpin (2003). "On the Specification and Estimation of the Production Function for Cognitive Achievement". *Economic Journal* 113 (485), F3–F33.

— (2007). "The Production of Cognitive Achievement in Children: Home, School and Racial Test Score Gaps". *Journal of Human Capital* 1 (1), pp. 91–136.

Trejo, S., D. W. Belsky, J. Boardman, J. Freese, K. M. Harris, P. Herd, K. Sicinski, and B. W. Domingue (2018). "Schools as Moderators of Genetic Associations with Life Course Attainments: Evidence from the WLS and Add Health". *Sociological Science* 5, pp. 513–540.

Tucker-Drob, E. M. and T. C. Bates (2016). "Large Cross-national Differences in Gene $\times$ Socioeconomic Status Interaction on Intelligence". *Psychological Science* 27 (2), pp. 138–149.

Turkheimer, E. (2000). "Three Laws of Behavior Genetics and What they Mean". *Current Directions in Psychological Science* 9 (5), pp. 160–164.

Van Doorslaer, E. and A. Jones (2003). "Inequalities in Self-reported Health: Validation of a New Approach to Measurement". *Journal of Health Economics* 22 (1), pp. 61–87.

Ware, E. B., L. L. Schmitz, J. Faul, A. Gard, C. Mitchell, J. A. Smith, W. Zhao, D. Weir, and S. L. Kardia (2017). "Heterogeneity in Polygenic Scores for Common Human Traits". *BioRxiv* 106062.

Wise, A. E. (1968). *Rich Schools, Poor Schools: The Promise of Equal Educational Opportunity*. Chicago: University of Chicago Press.

Wray, N. R., S. H. Lee, D. Mehta, A. A. Vinkhuyzen, F. Dudbridge, and C. M. Middeldorp (2014). "Research Review: Polygenic Methods and their Application to Psychiatric Traits". *Journal of Child Psychology and Psychiatry* 55 (10), pp. 1068–1087.

Yengo, L., J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood, M. N. Weedon, T. M. Frayling, J. Hirschhorn, J. Yang, P. M. Visscher, et al. (2018). "Meta-analysis of Genome-wide Association Studies for Height and Body Mass Index in ~700000 Individuals of European Ancestry". *Human Molecular Genetics* 27 (20), pp. 3641–3649.

Zumbuehl, M., T. Dohmen, and G. Pfann (2021). "Parental Investment and the Intergenerational Transmission of Economic Preferences and Attitudes". *Economic Journal* 131 (638), pp. 2642–2670.