Non-monetary and Monetary Returns to Education Throughout the Life-cycle

Five Empirical Essays in the Economics of Education

DISSERTATION

zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften
(Dr. rer. pol.)

durch die Fakultät für Wirtschaftswissenschaften der
Universität Duisburg-Essen
Campus Essen

vorgelegt von
Daniel Alexander Kamhöfer, M.Sc.
Econ aus Dortmund

Essen, 2018

Tag der mündlichen Prüfung:	22. März 2018
Erstgutachter:	Prof. Dr. Hendrik Schmitz (Universität Paderborn)
Zweitgutachter:	Prof. Dr. Martin Karlsson (Universiät Duisburg-Essen)

You can't always get what you want
But if you try sometimes well you might find
You get what you need

Sir Mike Jagger on life, the economy, and potentially
identification problems in non-experimental data

Acknowledgements

While numerous people affected in my life as well as my work in a very positive way over the course of the last years – and I am very grateful to all of them, I would like to single-out Hendrik Schmitz. As my supervisor, Hendrik, contributed not only to my development as an economist but taught me economic research. Without his constant support, encouragement, and guidance this dissertation would not have been possible. I would also like to thank Martin Karlsson. Besides kindly agreeing to be the co-supervisor of this dissertation, Martin constantly shared his research environment and network with me. Much of the research I have conducted in this thesis reflects the influences made possible through Martin. Furthermore, I am grateful to Daniel Avdic for providing support and engaging in fruitful discussions about my research that greatly improved this thesis. I would also like to use the opportunity to thank Reinhold Schnabel. From hiring me as a student assistant nearly ten years ago when I was in my undergraduate studies to supervising my Bachelor's and Master's theses, he gave me a glimpse into academia and encouraged me to take up doctoral studies.

Furthermore, I am indebted to my colleagues in Essen and Paderborn – not only for discussing our research ideas but also for sharing sorrows and laughs over lunch and during coffee breaks. While the list of colleagues I ought to thank is too long to name all of them, I would like to mention Martin Fischer, Matthias Westphal and Tobias Rühl. In an countless number of inspiring talks we had over the last years, we never missed a chance for discussing and improving each other's work. Besides improving my research through numerous working groups and seminars at my Alma Mater and at Paderborn, I moreover appreciate the experiences I gained through participating in many conferences and workshops in all parts of the world. I am especially grateful to Sarah Cattan for inviting me to the Institute for Fiscal Studies in London, to Therese Nilsson for having me at Lund University, and to Sonja Kassenboehmer for collaborating with me at Monash University, Melbourne. During these research stays I gained unique experiences and had the pleasure to meet and collaborate with incredible people.

Finally, I like to say thank you to my family and friends. In the time-demanding and often stressful research business they always put things into perspective. My parents, Angelika and Olaf, showed, like the rest of my family, unconditional support in all this time. The many days and evenings I spent with my friends, especially but not only in Bochum and Lette, is unmatched and was always a great help to relax. They always have an open door, an open ear, and an open heart and I am very grateful to them.

Thank you all!

Coauthors and publications

Chapter 2: Preschool education

- Coauthors: none
- Status: to be make available as working paper

Chapter 3: Elementary school education

- Coauthors:
  - Sarah Cattan (Institute for Fiscal Studies, London)
  - Martin Karlsson (University of Duisburg-Essen)
  - Therese Nilsson (Lund University)
- Status: IZA Discussion Paper #10995

Chapter 4: Upper-secondary school education

- Coauthor: Hendrik Schmitz (Paderborn University)
- Status: published in *Journal of Applied Econometrics* 2016(31): 865–872

Chapter 5: College education and skill and health returns

- Coauthors:
  - Hendrik Schmitz (Paderborn University)
  - Matthias Westphal (Paderborn University)
- Status: forthcoming in *Journal of the European Economic Association*

Chapter 6: College education and fertility preferences

- Coauthor: Matthias Westphal (Paderborn University)
- Status: Ruhr Economic Papers #717

Contents

Chapter 1

Introduction

1.1 The need for research in the economics of education

Among the many developments that shaped societies over the course of the last centuries, the surge in educational participation is certainly one of the most crucial ones. On the individual level, education has been found to increase the productivity and the returns to labor market activities. This makes education a powerful tool in the economists' "grand pursuit" to fight individual poverty and to enable a life in economic prosperity (see Nasar, 2012). The individual benefits of education do not only awake the interest of political decision-makers on a more aggregated level, but many aspects of education are directly in their hands. Legislators set the minimum school leaving age or the years of education an individual is required to have, the government decides on the supply and funding of schools and colleges and their autonomy, and it regulates the quality standards, such as caps for the student-teacher ration and the qualification of teachers. The presumably large returns and the government's leverage make educational policies an attractive instrument to fuel the development of societies further. Education is considered to address important challenges: satisfying the need for human capital set by new production technologies (Goldin and Katz, 2009), easing the burden of the demographic change by allowing individuals to cope old age better (Cervellati and Sunde, 2013; Rohwedder and Willis, 2010), and balancing inequalities by providing equal opportunities (Cunha and Heckman, 2009). Those high hopes for education link future policies to past experiences: What can we learn from the expansion of education in the past to improve the design of future policies?

While researchers have a rather good grasp of some returns to education (see, e.g., Card, 1999, 2001, for reviews), our understanding of other returns and their sources needs to be broadened in order to achieve of the full potential of educational policies. A growing body of research suggests that the scope of education is not limited to individuals' labor market productivity and monetary well-being but also affects non-pecuniary factors like the production of health, preferences and attractiveness for a partner, and even an individual's personality traits. To learn how education shapes those factors and whether this can contribute to our understanding of the monetary returns to education, it is important to disentangle the underlying causal pathway from a mere correlation. This doctoral dissertation aims at providing new quantitative evidence of how and which forms of education determine an individual's monetary and non-monetary well-being.

Both the potential and the need to learn from past experiences to shape further policies is perhaps best described by Figure 1.1. The figure depicts the percentage of the GDP that has been spent on education over time in the US and Germany, respectively. While the US has spent about 2.8 percent of the GDP on education in 1950, the corresponding number for Germany is 1.8 percent. More interesting than the cross-country comparison of the spending is its development over time: by 2014, the share of educational expenditures to the GDP rose to 5.6 percent in the US and to 5.0 percent in Germany. Put differently, within 65 years, the resources these societies are willing to devote to education have doubled. For other industrialized economies the corresponding numbers exhibit a similar trend. Putting these numbers into perspective, the importance of education becomes even more visible: the US, the world's supreme military power, spends less on its military (3.3 percent of the GDP, see World Bank, 2017) than on education. Likewise, Germany spends more on education that its flagship industry, car manufacturing, produces in a year (4.5 percent of the GDP in 2015, see German Federal Statistical Office, 2017) – although this may be taken with a grain of salt as the spending and the production approach are different ways of assessing the GDP. Moreover, it seems fair to assume that the spending on education has often higher returns than, for instance, the spending on the military or health care.

The high level of educational expenditures as well as the upward trend do not only affect legislators who pass educational policies but it also calls for a careful economic evaluation. The sheer size of the education systems opens the door for potentially substantial efficiency gains. Although it was not the zeitgeist in the era prior to the industrial revolution to consider education as an investment good, even Adam Smith was concerned with the efficiency of the education system when he wrote *The Wealth of Nations* in 1776. His main critique was the governmental involvement in education as a cause of inefficiency. The curriculum

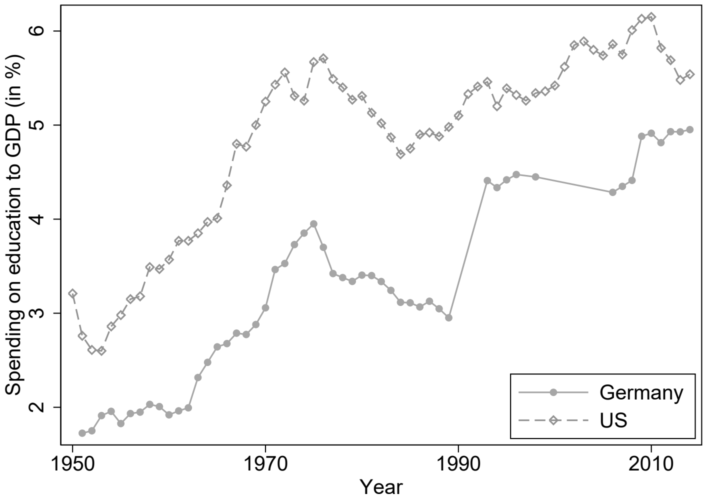


Figure 1.1: Educational expenditures as percentage of the GDP

Notes: Own illustration. Data for the US taken from US Government Spending (2017). For Germany, educational expenditures and GDP 1950–1990 is taken from Diebolt (2000) and German Federal Statistical Office (2016a), respectively, and refer to West Germany. After 1990, data refer to the unified Germany and are taken from World Bank (2017).

set by the government may, for instance, not reflect the individuals' or the market's demand for certain subjects and, thereby, causes an inefficient allocation of resources (see Diebolt, 2000). Although Smith's efficiency concern is nearly 250 years old, it still applies in today's knowledge-based economies. While the partly technology-driven increasing demand for human capital makes investments in education both necessary and potentially profitable; purely input-based investments, that do not account for incentives, are shown to be potentially inefficient (see, e.g., Hanushek, 2003, and Woessmann, 2016). Both from an individual and a policy point of view, the margin of education and the incentives in the education system are likely to be important determinants of the success of investments in further education. Should parents', for instance, invest a given amount of money in the preschool education of their youngest child or in the college education of their oldest child? Should governments likewise invest in improving preschool or college education? Research in the economics of education is potentially able to address such questions and the findings of some studies already made their way into legislation (see Gormley, 2011, for a more detailed account.)

To sum up, given the limited number of resources and the presumably low marginal returns to input-based educational policies, the trend depicted in Figure 1.1 cannot continue in the decades to come. Increasing the efficiency of the existing education systems and investing in educational interventions that work the best based on empirical evidence is necessary of a sustainable educational policy.

The growing importance of evidence-based policy advise in the economics of education and the closely related disciplines of labor, health, and family economics (often summarized as applied micro) can be linked to two developments in empirical research: the raise of the so-called "credibility revolution" aiming at disentangling the causal pathway of education and an increased focus on non-monetary returns to education (see Angrist and Pischke, 2010, and Oreopoulos and Salvanes, 2011, respectively, for reviews). This thesis can be understood as part of larger research efforts in applied micro that follows these stems of research. Before I summarize the single chapters, I therefore briefly describe the general development of these "new" methods and data in applied micro.

1.2 Recent developments in the field

1.2.1 The credibility revolution

Comparing the effect of a labor market training program on earnings in an experimental design – where individuals were randomly assigned into program participation – with the effect in an observational study, LaLonde (1986) finds a substantial difference in the estimated returns based on the research design. The reason for this is likely to be a selection of, for instance, more curious and intelligent individuals into the program in the observational study. However, only relying on an experimental research design that overcomes such selection would restrict the number of potential research question dramatically. It is often neither ethical nor feasible to randomly assign individuals to certain treatments (although there are some notable exceptions, for instance, the Perry Preschool Program discussed below). While this holds true in all fields of applied microeconomics, the returns to education are a primary example.

It is, for instance, not possible to randomly assign individuals to college education: this would require to force some individual to take several years of college education who do not want to, while withholding others from doing so. However, just comparing, say, the income of college graduates and non-graduates in observational data is likely to be biased as more intelligence individuals are not only more likely to study but would probably also earn more in absence of college education. Thus, even if the true effect of college education would be zero, not taking such selection into account may lead to the false conclusion that investments in college education are favorable as college education is associated with higher earnings. This example illustrates that it is not possible to draw reliable implications for individuals or policymakers based on evidence that is likely to suffer a severe selection. In order to address this selection problem in observational data nevertheless, the microeconometric toolkit has been expanded by the development of new methods and the adoption of methods from other fields like statistics and mathematics (see, e.g., Angrist and Pischke, 2009, for a detailed account). Most of those methods, like instrumental variables techniques, make use of a quasi-experimental change in an individual's environment that mimics the research design of a controlled experiment (see the chapters of this thesis for examples).

Besides methods that are, under the right circumstances, sufficient to deal with selection; the credibility revolution is characterized through a carefully investigation of those circumstances (Angrist and Pischke, 2010). That is, at the heart of the credibility revolution is the question whether the underlying identifying assumptions are justified based on the institutional background and, if possible, explorative empirical evidence. While, for instance, an instrumental variables approach using regional variation in the college availability (as, e.g., in the seminal work of Card, 1995) is per se able to overcome a selection of more motivated individuals with higher earnings potentials into college, it only does so if the original placement of the colleges was not in response to the individuals' desire to take college education. Thus, quasi-experimental methods can – and, under the right circumstances, do – provide a plausible case for the identification of the causal effects of education.

More recently, a growing number of studies also concerns with the external validity of the empirical strategies and their implementations. They criticize that exploiting exogenous variation may allow archiving a high internal validity; the range of the variation (or the group that experienced the variation) is, however, often too particular to generalize the resulting finding to a broader population (see, for instance, the discussion in Deaton, 2010, Imbens, 2010, and Heckman and Urzúa, 2010). Taking this aspect into consideration, the chapters in this thesis not only use state-of-the-art methods to archive a high internal validity, but the policy relevance of the analyzed setting and the external validity are also carefully discussed. For the exact methods, their implementation, and a discussion of their validity; see the chapters.

1.2.2 Non-monetary returns to education

While an unbiased effect of education, say, on earnings is a prerequisite of a credible policy implication, it is unlikely that an individual's earnings reflect the entire range of benefits. Gary Becker laid a theoretical foundation suggesting that individuals select themselves in all kinds of non-market outcomes[1] with the same rational as they maximize their monetary well-being (Becker, 1993). If more educated individuals earn more, this affects their opportunity costs of activities other than paid work. When engaging in criminal behavior and risking a two-years incarceration, more educated and productive individuals forgo more income in these two years than their less educated peers. This reduces their expectation value of committing a crime mechanically. Moreover, education may further alter preferences and increases the disutility from an incarceration. Similarly, education-induced time constraints and preference changes may affect an individual's decision for a partner, the number of children, and the utilization of health care and health behaviors (see Grossman, 1972, for the latter). As the distribution of such non-pecuniary outcomes is also in the interest of policymakers, considering the non-monetary as well as the monetary returns to education allows a more informed policy implication. A large education-induced increase in health behaviors may, for instance, compensate for a small earnings premium. Moreover, this information is not only valuable in its own right, but non-monetary returns may also constitute mechanisms through which education transmits into monetary returns.

Besides Becker's theoretical argument for the relevance of non-monetary returns to education, new and more detailed data sources additionally favor the analysis of non-monetary factors. This is especially true for cognitive and non-cognitive skills (that is, personality traits). While the assessment of intelligence and personality types has a long history in psychometric research, those factors are the textbook example of unobservable confounders of the education-earnings relationship in economics. Beginning with the early measures of skills in economic applications taken from military enlistment records, the quality of the information has rapidly increased. Most major socioeconomic surveys have incorporated test batteries to assess the respondents' competencies and personality in the last two decades. Germany's longest running individual-level survey, the Socioeconomic Panel Study, for instance, launched a Big Five personality test battery in its 22nd wave in 2005 and an "ultra-short intelligence test" was conducted two years later for the first time. Another rather recently available source of detailed information on education and the development of human capital are large-scale international assessment studies such as PISA and TIMSS. While early skill measures have been added to Mincer equations of earning on education to prevent an omitted variable bias, the availability of rich skill information allows – in line with Becker's theory – distinguishing education from human capital.

[1] I use the terms non-monetary, non-market, and non-pecuniary outcomes interchangeably, in spite of the view that some of the outcomes can be interpreted as the result of a market-kind bargaining, such as the marriage market.

Research on cognitive and non-cognitive skills was further advocated through the work of James Heckman and co-authors laying a theoretical foundation and providing compelling evidence for the efficiency of early childhood investments. At an early stage, the brain structure can be shape more effectively, while the longer time horizon and a self-fertilizing in skills – in that better skills in one period beget the acquisition of skills in subsequent periods – have the potential to make early investments particulary efficient (see, among others, Cunha et al., 2006; Cunha and Heckman, 2007; Heckman, 2007). That is, a child with more developed language skills has an advantage in learning reading and writing when entering elementary school compared to a child that first needs to learn the basics and the skill gap is likely to widen (so-called self-productivity of skills). Moreover, when the initial advantage is followed by a subsequent intervention, the resulting gain in skills may be higher for the first than for the second child (dynamic complementarity). Assessing such multipliers in the formation of skills empirically is rather data demanding as the children need to be followed over multiple periods. However, the availability of better data sources increasingly enables such analyses. A leading example for both the importance of educational interventions and the relevance of non-monetary returns is the analysis of the Perry Preschool Program in Michigan. Being one of the few examples for an experiment in the economics of education, treated children received high-quality care two years, while children in the control group only received informal care. Even after 30 years, treated children have a higher labor market attachment. Evidence suggests that the driving force behind the labor market returns is the development of non-cognitive skills – cognitive skills seem, in fact, less important (Heckman et al., 2013). Furthermore, the non-monetary returns (especially, a reduced probability of involvement in criminal activities) exceed the monetary returns to the intervention (Heckman et al., 2010a; Conti et al., 2016) and the overall annual internal rate of return is up to 10 percent (Heckman et al., 2010b).

These recent advancements in the research agenda in applied micro – the careful distinction of causation and selection, not only through state-of-the-arts methods but the plausibility of their assumptions, as well as the consideration of non-monetary outcomes – are key elements in all five essays of this thesis.

1.3 Overview and summary of this thesis

In the five essays of this thesis – each representing a stand-alone research paper – I analyze the non-monetary and possibly monetary returns to different margins of education. Figure 1.2 provides an overview over the margins, shows when individuals are affected, and presents the considered outcome variables. The dashed boxes in the center give the investigated margin of education. This is the variables of interest – reaching from the quality preschools to the decision to take higher education. As the margin of education is potentially subject to individual selection, I utilize information on factors that change the margin of education. For instance, how does the college availability change the decision to take higher education? Those changes in the education are given in the boxes on the left-hand side of the figure. One might think of them as interventions (although I refer to them more generally as "changes" because the student absence from school in Chapter 2 is strictly speaking not an intervention). The solid boxes below the margins indicate the considered short-term effects, while the solid boxes on the right-hand side give the long-term outcome variables, for instance, the effect of taking higher education on an individual's cognitive skills in prime age. The nature of the changes in education, how they affect the margin of education, and how this is used to address selection differs across the chapters and depends on the available data and the plausibility of the identifying assumptions. In the following I summarize the chapters (that is, the essays) in the order individuals are affected by the margin of education, starting with the youngest age group.

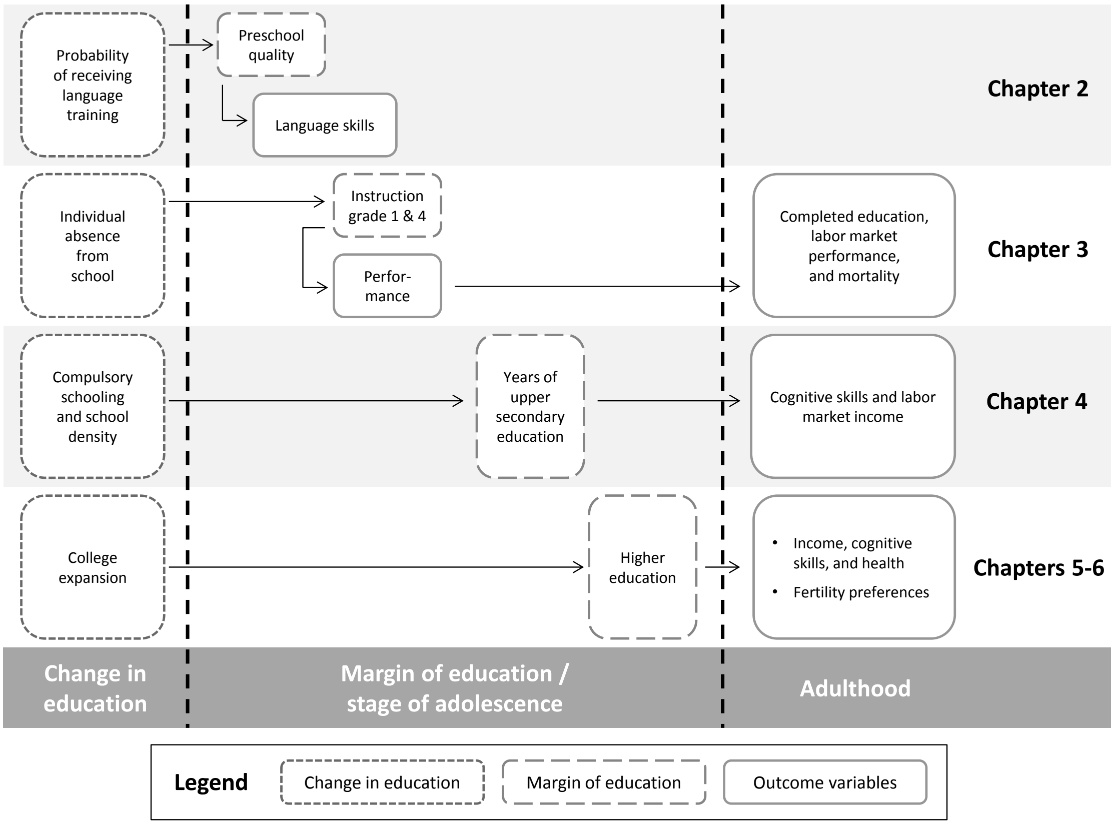

Probability of receiving language training
Preschool quality
Language skills
Chapter 2
Individual absence from school
Instruction grade 1 & 4
Perfor-mance
Completed education, labor market performance, and mortality
Chapter 3
Compulsory schooling and school density
Years of upper secondary education
Cognitive skills and labor market income
Chapter 4
College expansion
Higher education
• Income, cognitive skills, and health
• Fertility preferences
Chapters 5-6
Change in education
Margin of education / stage of adolescence
Adulthood
Legend
Change in education
Margin of education
Outcome variables

Figure 1.2: Overview over the chapters of this thesis

Notes: Own illustration.

In the **first essay** (Chapter 2) of the thesis, I consider the quality of preschool education in Germany at the children's age of three to four. About one-third of the about 200 preschools with complete information in the key variables in the data at hand, the National Educational Panel Study (NEPS), reports to rely on a curriculum-based language training in their education. The language education of the children in the remaining two-thirds of the preschools did not go beyond basic childcare. In this chapter, I aim at analyzing the effect of the quality of preschool (language) education, measured through the implementation of the curriculum-based language training, on the short- and medium-term formation for grammar skills – an important indicator for development and subsequent learning. To address that either preschools implement the language training is response to a greater need of the children or that parents enroll their child in a preschool with language training because of the child's individual need, I only compare children with and without language training, if they have the same probability of receiving the language training based on a large set of observable characteristics (that is, I employ a propensity score matching approach). The underlying identifying assumption is that those characteristics are sufficient to address all kinds of selection. Although this "selection on observables" is in many cases a rather strong assumption, the application at hand makes an arguably persuasive case. On the one hand, the NEPS includes a large number of information given by the parents, the preschool educator, and the preschool principal. One the other hand, using information on an individual's math skills available in the NEPS allows removing unobservable characteristics that affect grammar and math skills in the same way. The resulting effect of language training on the short-term formation of grammar skills is remarkable stable around 14 percent of a standard deviation. Compared to the literature this is a medium-to-large effect size. This finding holds for different sets of control variables (as chosen by an algorithm adopted from the machine learning literature in order facilitate the large array of potential confounders) and across the simple-difference and inter-skill differences-in-differences specifications. Subsampe evidence for children in grades 1 and 3 of elementary school indicates that the short-term effect is (at least in the first grade) persistent and even expanding in the medium-run perspective. Given that the analyzed language training is rather easy to implement, improving the quality of preschool education through such an intervention seems to provide an effective (and potentially efficient) way to boost the development of human capital from early childhood on.

In the **second essay** (Chapter 3), Sarah Cattan, Martin Karlsson, Therese Nilsson, and I analyze the effect of missed instructional time in grades 1 and 4 of elementary school on the short-term academic performance at the end of the school year and long-term completed education, employment, labor market income, and mortality. So far, only few studies consider the role of instructional time within a school year in the educational production function and the literature on individual absence from school is limited to short-term effects in the US. In order to consider long-term outcomes, we combine self-digitized historical school records for children born in the 1930s taken from Swedish archives with Census information from 1960 and 1970 as well as tax register data from 2002. To account for school and teacher characteristics as well as the students' genetic endowment that may cause a selection of children with certain backgrounds into absence, we rely on school, teacher, and siblings fixed effects. That is, we only use variation in days of absence and the outcome variables between siblings, while controlling for the school, the teacher, and other observable characteristics. The results indicate a moderate and robust short-term effect of absence on performance in school. In the long-run, this initial effect seems to fade away: while the results suggest that absence is still relevant for the employment status in 1960; afterwards, the effect size points toward the expected direction but is statistically indistinguishable from zero. This finding is in line with a broader literature that indicates an early-career effect of education that fades out over time.

The **third essay** (Chapter 4) takes a leap to students at the end of their secondary schooling and Hendrik Schmitz and I analyze the returns to an additional year of education at this margin in Germany. Analyzing the same margin, Pischke and von Wachter (2008) find that the introduction of an additional ninth grade in basic track secondary schools (the so-called compulsory schooling reform) between 1949 and 1969 (depending on the state), had no effect on the labor market earnings of the affected students. While Pischke and von Wachter (2008) can empirically rule out a number of reasons for the zero earnings returns, they can only conjecture that heterogeneity along different groups of students and the formation of skills play a role. As the reform rose the legal minimum years of schooling only for basic school students, it is questionable whether the zero returns can be generalized to students in intermediate and academic schools that had always more than eight years of schooling. To broaden the analysis to include intermediate and academic school students, we instrument their years of schooling with the relative supply of these schools. Controlling for the state and the birth year, regional and temporal variation in the school density allows measuring arguable exogenous variation in the decision to take additional years of schooling in others than basic track schools. The results indicate that the zero returns are persistent for intermediate and academic track students and cannot be related to the particular group of students affected by the compulsory schooling reform. To test Pischke and von Wachter's second conjecture for the zero returns, that is, the absence of a skill effect that mediates schooling into wages, we employ the same empirical strategy as before, but consider a measure of cognitive skills taken from the German Socio-economic Panel Study as outcome variable instead of labor market wages. Again, the results point toward a zero effect of years of secondary schooling on cognitive abilities for students in all school tracks. Given these results, it seems fair to argue that the formation of basic skills may take place before the ninth grade in Germany; that it does, however, not seem to take place after the eighth grade.

In the **fourth essay** (Chapter 5), joint work with Hendrik Schmitz and Matthias Westphal, we consider the effect of tertiary education in Germany on the labor market wage, cognitive skills, and health. To overcome a selection of, for instance, smarter or healthier individuals into college education, we instrument the probability of going to college using the expansion in the number of colleges and their sizes in the 1970s and '80s. Based on administrative data, the number of colleges doubled in the time under review and the number of students enrolled in higher education increased from below 200K to over 1M. Using German Micro Census data in an explorative analysis, college openings cannot be predicted based on local economic characteristics. This is in line with qualitative evidence that the college expansion was largely driven by political motives and did not systematically affect regions with a higher or lower average intelligence or health of the population. Given this powerful variation in the individuals' college decision at hand, we are able to go beyond conventional two-stage least square estimation and implement the so-called marginal treatment effect approach as developed by Heckman and Vytlacil (2005). This approach has the advantage that it allows estimating the returns to college education along the unobserved heterogeneity (that is, the part of the college decision that cannot be explained through observable characteristics). Combining the administrative data with information on adult individuals observed in the NEPS, the resulting range of marginal effects enables measuring the returns for distinct groups of individuals and is, thereby, more informative than a single point estimate. We find that individuals who study because of the college expansion although their observed probability of doing so is rather low, strongly benefit in terms of wage, skill, and health gains. In fact, individuals with a high observed probability of studying who comply with the increased availability of college spots benefit less from higher education (and some not at all) compared to their more eager peers. This heterogeneity in the returns to college education bears an important insight for policies that aim at increasing the share of college graduates in the population (such as the Higher Education Pact 2020 in Germany): it should not be taken for granted that college education is universally beneficial – depending on the individuals that are affected by the expansion of higher education, the returns may even be zero.

In the **fifth essay** (Chapter 6) of the thesis, Matthias Westphal and I use the same variation in the college availability as in Chapter 5 to investigate how college education affects a woman's probability of becoming a mother and, conditioning on being a mother, the number of children by the age of 40 (so-called completed fertility). The effect of education on fertility has been empirically analyzed ever since Becker suggested that families may trade off a large number of children with more educated children (the quality-quantity trade-off). Still, most studies overcome a preference-based selection by means of compulsory schooling reforms. For the margin of college education that falls well into the prime reproductive age, evidence on the education-fertility nexus is so far scarce. Combining administrative data on college availability with individual data taken from the NEPS, we find that women are about one-quarter less likely of becoming a mother because college education than non-college-educated women. However, once a college-educated woman has decided to have children, the number of children is slightly higher than the one of a woman without college education. Interestingly, college-educated mothers postpone their first birth about 6.5 years due to the college education. Given a usual duration of college education around 4.5 years, the effect size indicates that the overall postponement is not only caused by an "incarceration" in college but although through a further postponement in the years immediately after leaving college, that is, the early-career stage. To investigate the college effect on the timing of birth further, we unfold the probability of childbirth along the women's age. While the probability of becoming a mother increases slightly in the years after the college-educated women graduate, the probability of childbirth is still below the one of women without college education – reflecting the overall reduced probability of motherhood. Conditioning the sample to women who have at least one child by the age of 40, however, the results indicate a catch-up effect of college-educated mothers in the years after leaving college. This college-induced early-career effect for non-mothers indicates a limited compatibility between work and family life that may contribute to the overall negative effect of college education on fertility. Policies that aim at raising the compatibility, for instance, through more flexible working hours, are promising in unifying education-induced changes in career and fertility effects.

Chapter 2

Language Training in Preschool and the Formation of Grammar Skills

2.1 Introduction

The potential of early childhood interventions to shape the formation of human capital from early on and, thereby, laying the foundations for an individual's long-lasting economic and social well-being evokes the interest of legislators and researchers alike.[1] Yet, our understanding of why certain investments have higher returns than others is far from being conclusive. While the quality of care is frequently suspected to be an important determinant of the returns to the early investments, it is widely understudied. Large-scale universal childcare programs do often not meet the high quality of successful experimental interventions (Weiland and Yoshikawa, 2013) and exhibit lower and sometimes even negative returns (see, e.g., Ruhm and Waldfogel, 2012, and Cascio, 2015, for reviews). The study at hand aims at broadening the understanding of the role of the quality of care by investigating the effect of a curriculum-based language training program in preschool at the child's age of four on the short- and medium-term formation of skills.

Besides a high plasticity in brain development at this stage (Couperus and Nelson, 2008), the long time horizon and complementarities in the formation of skills provide strong theoretical arguments for interventions at an early age (Cunha and Heckman, 2007). The empirically most-compelling evidence on the merits of early investments comes from randomized controlled trials (RCTs), most-prominently the High/Scope Perry Preschool Program in Ypsilanti, Michigan,

[1] See, for instance, the Obama Administration's Zero to Five Plan and the 2011 *Science* Special Issue on "Investing Early in Education."

starting in 1962 (see, e.g., Currie, 2001, for a review). Treated children received 2.5 hours of formal care every weekday for seven months a year, for two years. The care consisted of a "child-centered, active learning curriculum" in small classes and teachers had advanced degree, were paid a bonus, and offered regular parent-teacher conferences (Mervis, 2011, p.953). The returns to this intervention are large and long-lasting: treated individuals had both higher cognitive (language and numeracy) and social-emotional skills and even up to 35 years after the treatment, they exhibit a higher completed education, large labor market returns, are healthier, and have a lower probability of being involved in crimes (see Heckman et al., 2010a, 2013, and Conti et al., 2016). The estimated internal rate of return is up to 10 percent p.a. (Heckman et al., 2010b). Compared to those benefits, the literature on the returns to universal childcare programs is much more ambiguous. In spite of a substantial correlation (see, e.g., OECD, 2010), the evidence from studies that seek to identify the causal relationship through exogenous variation in childcare availability is less clear (see Ruhm and Waldfogel, 2012) and may depend on the country-specific context (see, e.g., Cattan, 2016). While some studies suggest positive effects of universal preschool education on both the short and medium-term skill formation and long-term labor market performance[2], others indicate zero or even negative effects (see, e.g., Kühnle and Oberfichtner, 2017, for Germany; Cascio, 2009, for the US; Baker et al., 2008, and DeCicca and Smith, 2013, for Canada; and Datta Gupta and Simonsen, 2010, for Denmark).

A likely explanation for this divergence in the returns to RCT and universal childcare is the quality of care. Besides of having a higher absolute quality of care than large-scale universal childcare programs, the RCTs are also likely to differ from universal programs in the relative quality, that is, the gap in the quality between the formal care and the counterfactual informal care. The RCTs were targeted on children from disadvantaged families who may have also faced a lower-than-average quality of the learning environment at home or in other informal childcare arrangements (see Cascio and Schanzenbach, 2014, for the formal argument). This role of the relative quality is in line with recent evidence from Cornelissen et al. (2017) for Germany. Estimating the effect of preschool education on school entry examinations they find that "children from disadvantaged backgrounds are less likely to attend childcare than children from advantaged backgrounds but have larger treatment effects because of their worse outcome when not enrolled in childcare" (p.1). Considering complete education and labor market attachment in Norway, Havnes and Mogstad (2011) find heterogeneity in the preschool effect

[2] For school outcomes in adolescence, see, e.g., for the US Gormley and Gayer (2005) and Fitzpatrick (2008) for Oklahoma and Georgia, respectively; Black et al. (2014) and Drange and Havnes (2015) for Norway; and Berlinski et al. (2008, 2009) for Uruguay and Argentina, respectively. For adult labor market performance, see Dumas and Lefranc (2012) for France.

along the mother's education. Even without accounting for the gap in the relative care, adopting the quality standards of the RCTs in the existing universal childcare programs in order to boost the development of human capital from early on would nearly triple the average costs[3] and can therefore be seen as "prohibitively expensive" (Mervis, 2011, p.954). Evidence on how specific aspects of the quality of preschool education can potentially improve the success of the universal programs is rather scarce.

The study at hand aims at narrowing this gap. Using survey information on about 2,000 children (born between 2005 and 2006) and their families taught by more than 800 educators in more than 200 preschools in Germany I analyze the effect of the quality of the language education on the formation of grammar skills.[4] Although preschool education in Germany traditionally focuses on overseeing rather than teaching children, about one-third of the preschools at hand have implemented a curriculum-based language training. The curriculum is developed by linguists and educational psychologists and consists of extensive instructions for the educators and learning materials for the children, such as board and card games and CDs. To my knowledge, there is no study that investigates the returns to such a low-key but well-defined intervention. Most related to the scope of this paper is study by Weiland and Yoshikawa (2013), who apply a day-of-birth-induced regression discontinuity design to compare children in high-quality preschools with an entirely curriculum-based education and specially trained educators in Boston, Massachusetts, with children in other, lower quality care arrangements. They find positive effects on the short-term cognitive functioning and socio-emotional skills of treated children.[5] Moreover, linguistic studies provide – independent of the preschool context – evidence on the effectiveness of high-quality early childhood language training programs in improving language comprehension.[6]

To account for a potential selection of better preschools or children with greater need into the treatment, I make use of comprehensive and fine-grained survey

[3] The average per-child spending on early education in OECD countries is 6,800 dollars p.a. (OECD, 2013) compared to 18,000 dollar in the Perry Program (Heckman et al., 2010b).

[4] Grammar skills (that is, listening comprehension at sentence level) form together with receptive vocabulary (listening comprehension at word level) the broader concept of language skills and are seen as an important element for subsequent learning (NEPS, 2011).

[5] Moreover, Neidell and Waldfogel (2010) and Currie and Neidell (2007) analyze the composition of peers and the level of regional spending in Head Start preschool education in the US as determinants of quality, respectively. Other studies emphasize the importance of the quality of elementary schooling in the US. See, for instance, Hoxby (2000) for the student-teacher ratio, Chetty et al. (2014a,b) for teachers' qualification and incentives, and Chetty et al. (2011) for teacher quality and peer effects.

[6] See, for instance, the UK Nuffield Early Language Intervention Project, a targeted RCT run by linguists (Bowyer-Crane et al., 2008).

questionnaires – answered by the families, the preschool educators, and the principals – to compare only the grammar skills of children with and without language training who are equally likely to receive the treatment based on a large number of observable characteristics. To allow for unobserved heterogeneity, for instance, through innate abilities, I account for an individual's mathematical skills in a differences-in-differences setup. As long as potentially unobservable factors do not systematically differ between grammar and math skills, this strategy removes their confounding effect. The resulting estimator is a regression-adjusted differences-in-differences propensity score matching approach. To facilitate for hundreds variables and their interactions on that preschools and families potentially select into language training, I rely on machine learning techniques that enable to restrict the analysis to factors that matter for the selection empirically. The point estimates indicate that language training goes along with an increase in grammar skills by about 14 percent of a standard deviation. Remarkably, the estimate are consistent across the simple-difference and the differences-in-differences specifications. This consistency indicates that the observed characteristics are sufficient to account for skill-constant unobservable factors. This finding puts, moreover, hope to the notion that the same holds for skill-variant confounders that would otherwise bias the estimates. A placebo test (using other skill domains) and a formal bounding exercise further support the validity of the empirical strategy.

Moving beyond the short-term perspective, the positive association of language training is persistent and even expanding two to three years after the treatment at the children's age of six in the first grade of elementary school (even though skill test scores are only available for a subset of children). The increased magnitude may be attributed to both a prolonged exposure to the treatment and a self-fertilization of grammar skills – that is, higher skills in one period beget higher skills in subsequent periods (Cunha and Heckman, 2007). At the age of eight, in the third grade, the estimates still point toward better skills in a more broadly measured language competence test but are rather imprecise. Although the data does not allow to estimate the treatment effect beyond elementary school, Figure 2.1 exploits the association between language skills in adulthood and labor market performance using data on about 2,000 individuals born between 1944 and 1986.[7] The binned scatter plot gives the descriptive relationship between the percentile of language skills and the average income (indicated by the circles). In panel (a), the language test score and income are adjusted for basic sociode-

[7] Although the adults are unrelated to the children considered in the main analysis, the dataset is part of the same survey, the National Education Panel Study (NEPS), and similar but age-adjusted skill tests were conducted (see Figure O2.1 in the Online Appendix for empirical evidence).

mographic characteristics, such as gender, age, and parental education (see the figure note for details). The linear fit (depicted by the solid line) suggests that a one standard deviation increase in adjusted language skills is associated with an increase in the adjusted income by about 460 Euros (that is, 14 percent of the average income). Panel (b) repeats the exercises when additionally adjusting for an individual's math test score. The declining but positive slop of the linear fit suggests an income premium of better grammar skills in spite of an individual's math skills. While this back-of-the-envelope exploration does not address the selection problem or indicate that the treatment effect is persistent beyond the individual's age of eight, it does hint that, if there is a long-lasting effect of the language training, the increased grammar skills are also likely to be reflected in higher earnings.

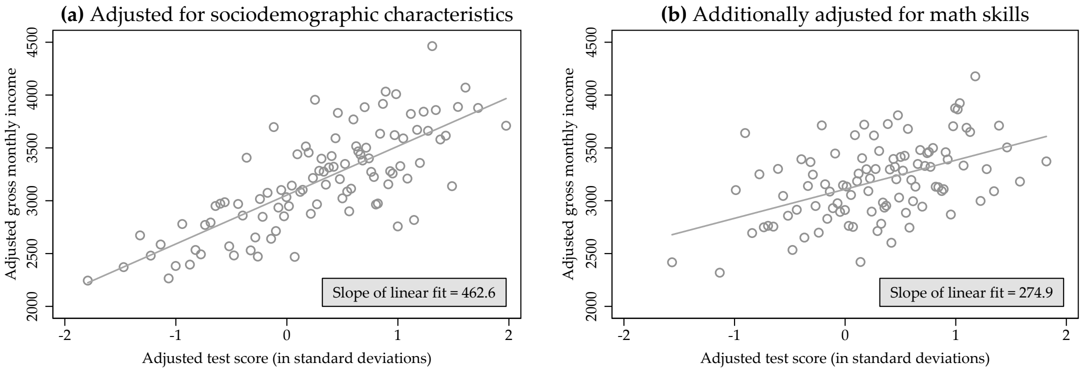


Figure 2.1: Association between language skills and income in adulthood

Notes: Own calculations based on NEPS–Starting Cohort 6 (wave 4 from 2011/12) using observations for 2,049 individuals with complete income and skill information. Following Chetty et al. (2011) the standardized language test score (mean 0, standard deviation 1) and monthly labor market income in Euros are, in an auxiliary step, regressed on a full set of age fixed effects and indicators for gender, East-German, native speaker, parental migration, and parental education. The axes give the residuals plus the unconditional mean. The circles depict the average income on the y-axis corresponding to the quantile of the language skills on the x-axis. The solid line gives the linear fit between the individual-level skill residual and the income residual using a simple linear regression. The figure is based on Michael Stepner's Stata ado-file `binscatter` (see Stepner, 2014). All errors are my own responsibility.

This study contributes to the literature is three ways: First, analyzing the formation of early language skills can be seen as a contribution in its own right given the importance of skills (see, e.g., Almond and Currie, 2011, for skills in general and Figure 2.1 for language skills in particular). Second, improving language education in preschool is a potentially efficient way to increase the overall quality of universal preschool programs that reduces the long-term fiscal strain of the government. Third, the careful analysis of the selection using machine learning techniques and partial identification reveals that a preschool's quality is already reflected in its socioeconomic environment and that adding individual-level information does not contribute to explaining the preschool's quality. Similar, family-level controls seem to be sufficient to account for skill-constant unobservable factors.

The paper proceeds as follows. Section 2.2 briefly summarizes early education in Germany and describes the treatment. Section 2.3 outlines the empirical strategy and the potential selection, while Section 2.4 introduces the survey data. Section 2.5 presents the results and Section 2.6 shows sensitivity checks. Section 2.7 exploits the effect persistency before Section 2.8 concludes.

2.2 Background

2.2.1 Day care in Germany

From the age of one until the age of three children in Germany are either in informal care through parents or other relatives (home care) or in family-day care (non-relative in-home care) or they are in formal care in *Kinderkrippe* day-care centers.[8] At the age of three, children usually enter *Kindergarten* day-care centers until the start of elementary school at the age of six. In Germany, *Kindergarten* centers are not considered as schools (and not part of the state-governed school system) as their focus is traditionally on overseeing children rather than curriculum-based education. In order to avoid confusion with kindergarten education in the US K–12 context, I nevertheless refer to the German *Kindergarten* centers as preschools. In the preschools children are cared for by educators. Educators may have but (unlike school teachers) do not need an advanced degree in education, and have at least completed a three-year part-time training-on-the-job, part-time vocational school apprenticeship. Figure 2.2 summarizes early education in Germany.

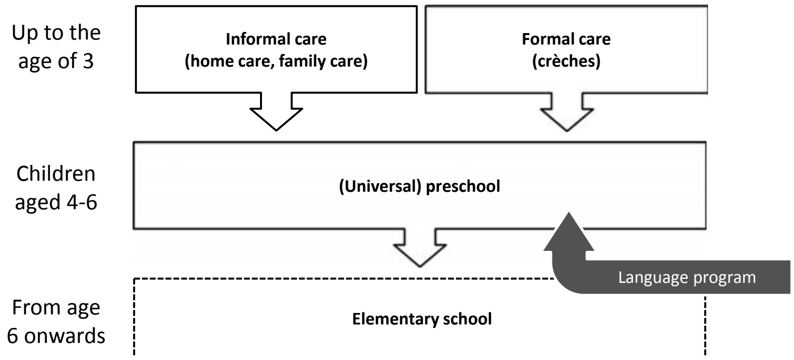


Figure 2.2: Day care in Germany

Notes: Own illustration. This is a broad and simplified depiction of early childhood day care in Germany. Crèches refers to the German *Kinderkrippe* and preschool to *Kindergarten*. Informal care includes home care as well as family day care (*Kindertagespflege*).

[8] This margin of early care is subject to ongoing reforms (aiming at increasing formal care) and research (see Felfe and Lalive, 2012, 2014).

Not being part of the state-organized school systems, preschools are run by a so-called provider (*Träger*). This is either a local non-profit organization, such as the church or the labor welfare, or one of the about 13,000 municipalities in Germany (see OECD, 2004, for details).[9] Although there is no central supervision or assessment, a joint framework by the 16 federal states and the states' youth welfare offices regulates the educators' qualification and defines a set of "educational objectives" for certain "areas of education." Preschool education should, for instance, contribute the "development of the child's physical, mental, emotional and social abilities" in "language, writing, communication" (KMK, 2015, p.103). However, the joint framework does neither define a curriculum nor criteria, when an object is met. Instead, parents can freely choose preschools in any municipality and may "vote with their feet" if they feel that a preschool does not live up to their expectations. Preschools are co-financed through the federal state, the municipality, the provider, and parental fees. The latter are either directly set or at least bounded by the state, are often means-tested, and constitute only a small part of the overall funding (as of 2014, on average, 14 percent, see OECD, 2004, 2006).

2.2.2 The treatment

The treatment under review, the curriculum-based language training, goes beyond the level of learning language that is inherent to overseeing children. The treatment effect I aim to identify is, therefore, the difference in the formation of grammar skills between the curriculum-based language training and the language skills that are acquired through regular childcare without any specific language education (e.g., by reading to the child).

Several learning programs, that provide a curriculum-based native language training, are available for preschools in Germany (Neugebauer and Becker-Mrotzek, 2013). Table 2.1 summarizes the most often named programs in the data at hand (see Section 2.4 for details). As the programs are rather similar to each other, most preschools have, if any, only one of them. Out of the 214 preschools with complete information in the data, 83 preschools have implemented at least one of the programs, while four preschools report to have two programs. The programs are offered by textbook companies and consist of instructions for educators how to implement the curriculum and often include learning materials, such as board games, word cards, CDs, and books, and can be purchased online separately or as a set. The first column in Table 2.1 gives the average costs of the programs for

[9] In-house day-care centers open for the children of employees of bigger companies and for-profit preschools do not play an important role for childcare in Germany.

Table 2.1: Overview of the curriculum-based language training programs

	Approx. costs	Avg. duration in months	Number of preschools	Number of children
Any program			83	718
Delfin 4[*]	€ 52	23	19	135
Hören, lauschen, lernen	€ 40–€ 110	17	12	88
Kon-Lab	€ 159–€ 428	24	10	112
Deutsch 240[**]	free	22	2	17
Other			40	366
No program			131	1,193
In total			214	1,911
Treated (in %)			37.6	38.8

Notes: The costs in column 1 refer to a basic set of the learning materials and are based on online research. For the costs of Delfin 4 two books with lessons and solution booklets are assumed. Columns 2 to 4 are based on own calculation using NEPS–Starting Cohort 2 data. Numbers refer to the final sample. In case a preschool reports to use more than one program (four preschools do so), I only consider the first-mentioned program.

[*] This program may be not confused with a assessment test for language competencies.

[**] Instructions are provided free of charge online by the Ministry of Education of the Stata of Bavaria and preschools may print some materials, such as word cards, themselves.

the instructions and a basic set of learning materials. Compared to the overall costs of running a preschool, the costs of the programs are rather humble and all preschools should be able to afford a program. In fact, comparing the programs it seems that the price is mainly driven by the amount learning materials they include – the curriculum and the instructions for the educators are quite similar. One of the most-often implemented programs, *Hören, lauschen, lernen* (this loosely translates to "listen and learn"), for instance, consists of a 40 Euros textbook that instructs educators when to schedule lessons and how to implement games and activities – that is, the curriculum. The program was developed by linguists and educational psychologists in order to promote the listening comprehension for the children. This comprehension, in turn, should foster the ability to learn how sentences are build and to apply this knowledge.[10] One of the most expansive programs, *Kon-Lab*, consists of several board games and CDs and sets are available ranging from 159 to 428 Euros, depending on the number of learning materials (see Figure O2.2 in the Online Appendix).

While the schedule of the lessons suggested by the developers may vary, the programs do not exhibit much variation in terms of the total time of instruction.

[10] The *Hören, lauschen, lernen* program has a fix schedule and takes about 10 minutes a day for 20 weeks. The educators are instructed to start every lesson at the same time and to implement a certain ritual, for instance, that the children first sing an opening song and then engage in games on a certain topic (e.g., the telephone game or finding words that rhyme with animal names). After each lesson the children get a stamp in a booklet and are awarded a "certificate" after completing all lessons.

However, the actual time children spent participating in the programs reveals some variation across the programs in the data at hand, see column 2 in Table 2.1. For instance, the program *Hören, lauschen, lernen* is, on average, used for 17 months while the developers suggest 12 months. This probably reflects that once the learning materials are available and educators and children know and maybe like some of the games, they will not stop engaging in them. It is more likely that children outgrow the content after a while.

For the analysis to come, it is helpful to distinguish some broad factors or channels why some preschools have implemented a language training program, while two-thirds did not. Although the list is not exhausting, four (not mutually exclusive) reasons may channel a preschool's decisions: (*i*) While a lack in financial means is a rather unlikely reason taken for itself, the preschool's *provider* might be more important. Church-run preschools may, for instance, prefer more religious-themed (language) learning materials even though if those materials have a less rigorous curriculum. (*ii*) Preschools may not see the *need of student body* for special language training because the children enrolled are from well-off families and already process good language skills when entering the preschool. (*iii*) Several preschools in close proximity may *compete* for children (who, in turn, affect financial resources) may set an incentive to increase the quality of the education by implementing a special language education. (*iv*) The *educators' and principal's engagement* will certainly affect the treatment status as they need to implement the language training. Educators might be reluctant to implement the language training because of the additional workload and principals may not be able to enforce the language training. On the other hand, a very experienced or engaged educator might be convinced to do a better job in supporting language skills than some program developed by academics.

Moreover, as families choose preschools, I can think of three broad groups of factors at this level: (*v*) The *availability* of language training is likely to depend on the child's month of birth and the enrollment into the preschool (although there a no systematically roll-out of the language training). Moreover, (*vi*) parents may decide for a preschool with language training because of the *child's individual need*, that is, they may want to compensate for a lag in or reinforce good skills; or (*vii*) because of their *preferences* for and their involvement in their offspring's education.

2.3 Empirical strategy

2.3.1 The selection problem

Depending on the treatment status, a child's grammar skills may take one out of two values: treated grammar skills (that is, the level of grammar skills if she receives the language training) or untreated grammar skills. The effect of language training participation is the difference between these two values. However, because a child either receives the language training or does not receive the training, only one of the two potential outcomes is realized – the unrealized and, thereby, unobserved outcome is the counterfactual. In order to overcome this identification problem, one would ideally wish to randomize language training participation, for instance, in form of a randomized controlled trial. If the assignment of and the participation in the language training is indeed random with respect to the level of grammar skills and only affects which level of skills is observed (the treated or the untreated skills), the level of grammar skills of the control group without the treatment (that can be observed) may serve as an appropriate proxy for the unobservable grammar skills of the treatment group in absence of the treatment.

However, in the observational data at hand, the language training is not (as-good-as) randomly assigned but probably subject to selection – as outlined above. This selection may emerge if children receive the language training because of unobserved potential grammar skills. In an extreme case, only children with particulary low grammar skills are enrolled in preschools that have implemented the language training. Comparing the mean values of the realized grammar skills of treated and untreated children would then underestimate the true effect of the language training. In the opposite extreme, only the children of parents who emphasize education receive the language training. If these children would also do better in absence of the language training, the simple mean comparison would overstate the true effect. In either case, the universe of untreated children does not serve as a sufficient control group to proxy the grammar skills of the treated children in absence of the language training.

2.3.2 Regression-adjusted propensity score matching

In order to guard against such a selection, as far as this is possible, I seek to compare only the grammar skills of children with and without language training who are identical in all other aspects. In other words, I aim at identifying the subset of untreated children who constitute a suitable control group. This can be done in two ways, either by controlling for all factors that are correlated with both the treatment status and grammar skills in a regression model or by estimating the probability of receiving language training (that is, the propensity score) and only comparing the grammar skills of treated and untreated children with a similar probability of receiving the language training in a propensity score matching (PSM) approach. Both methods only yield the causal effect of the language training if all confounding factors can be observed. This assumption is referred to as conditional independence assumption (CIA), unconfoundedness or selection-on-observables.

Although I will present linear regression results as a benchmark, in the preferred specification I rely on "more sophisticated [matching] methods for adjusting for differences in covariates" (Imbens and Wooldridge, 2009, p.24). To implement the matching approach I conduct a three step procedure: In the first step, I estimate the propensity score (PS) as the fitted value from an auxiliary probit model where language training participation is regressed on the covariates. Second, for each child in the treatment group I assign kernel weights (using an Epanechnikov kernel) to all untreated children indicating the distance in the estimated PS. The closer the estimated PS of the untreated child is compared to the estimated PS of the treated child, the higher the weight the untreated child receives. This procedure leaves me with a full set of weights for all untreated children for each child in the treatment group. If the PS estimation is based on all variables that confound the language training effect, reweighting the untreated children mimics the control group under random treatment assignment. To guard against outliers I only consider treated children and with a PS that is higher than the minimum and lower than the maximum PS of the untreated children. That is, I restrict the sample to children "on support." Following the suggestion of Bang and Robins (2005), I add a third step to the propensity score matching procedure: I take the final treatment effect (and the preschool-clustered standard errors) from a weighted regression of grammar skills on the language training indicator and the set of covariates that enters the PS estimation in the first step. This additional step ensures that the matching approach is "doubly robust" as it allows either the selection or the outcome equation to be misspecified as long as the other equation is correctly specified (Bang and Robins, 2005). Besides the identification and exclusion of outliers and the doubly robust specification, the PSM strategy has another advantage compared to a linear regression controlling for the same covariates: The matching approach is semi-parametric in that it only assumes a functional form between the covariates and the PS but not between the language training and grammar skills.

2.3.3 Differences-in-differences propensity score matching

To complement the (simple-difference) matching strategy, I additionally implement a regression-adjusted differences-in-differences propensity score matching (DD-PSM) approach following the idea of Heckman et al. (1997). While the data structure at hand prevents me from using pretreatment grammar skills to establish a before-after comparison, I instead follow Jürges et al. (2005) and others and compare the development of grammar skills with the formation of math skills.[11] Notwithstanding the semi-parametric nature of the PSM approach, the formation of grammar skills Y^{gra} might be represented through the following stylized skill production function[12] (suppressing individual and preschool subscripts):

$$Y^{\text{gra}} = \beta_0 + \beta_1 D + \boldsymbol{X}'\boldsymbol{\beta}_2^{\text{gra}} + \mu + \varepsilon^{\text{gra}}, \tag{2.1}$$

where β_1 is the effect of interest and ε denotes an i.i.d. error term. The $\boldsymbol{X}$ variables are observable and enter regression and the estimation of the PS, respectively. The term μ summarizes unobserved determinants of the skill formation such as innate abilities and preferences. If those factors are also correlated with the treatment status, neither the regression nor the matching approach would yield the causal estimate of β_1. Nevertheless, if the unobservables in μ affect the formation of grammar and math skills in the same way, establishing a differences-in-differences approach across both skill domains may still overcome the selection problem. To see this, the formation of mathematical skills Y^{math} (measured on the same scale as Y^{gra}) may be represented analogous to the formation of grammar skills:

$$Y^{\text{math}} = \beta_0 + \boldsymbol{X}'\boldsymbol{\beta}_2^{\text{math}} + \mu + \varepsilon^{\text{math}}. \tag{2.2}$$

If Eqs. (2.1) and (2.2) are correctly specified (that is, D does not enter Eq. (2.2) and μ is independent from the skill domain), subtracting math skills from grammar skills leads to the differences-in-differences notation:

$$Y^{\text{gra}} - Y^{\text{math}} = \beta_1 D + \boldsymbol{X}'(\boldsymbol{\beta}_2^{\text{gra}} - \boldsymbol{\beta}_2^{\text{math}}) + (\varepsilon^{\text{gra}} - \varepsilon^{\text{math}}). \tag{2.3}$$

[11] Compared to the DD-PSM approach suggested by Heckman et al. (1997), my strategy differs in two dimensions. Besides establishing the second difference across skills instead of time periods, continuously measured math skills do not allow to match on the exact value of math skills. Following Jürges et al. (2005) I instead use the skill difference as outcome variable in the baseline analysis. As a robustness check, I condense the math skill measure to take fewer values and condition on the exact value.

[12] Much research has been devoted to the arguments and the functional form of skill production functions (see, e.g., Todd and Wolpin, 2003, 2007). Here, I rely on a very basic version for the purpose of introducing the DD approach.

Here, the unobserved component μ is cancelled out. Given Eq. (2.3) is correct, this then yields the causal effect of language training. Taking the skill difference as dependent variable alters the CIA as follows: in order to identify the causal effect of language training all factors that simultaneously affect the treatment status and the difference between grammar and math skills need to be observed. In other words, while the simple-difference approach requires statistical independence of the *level* of the potential grammar skills and the treatment status (given controls); for the DD approach, it is sufficient that the *difference* between the potential outcomes (the grammar-math difference) is independent of the treatment status.[13]

Although the plausibility of the CIA eventually depends on the quality of the control variables (that is, how plausible it is that all confounders are account for), it might be useful to have a closer look what the CIA actually implies in the application at hand. First, the DD approach (and, thereby, the relaxed CIA) requires that language training does not enter the formation of math skills (D does not enter Eq. (2.2)). This rules, for instance, out that treated children understand the math test questions better because of better language skills. Although this assumption is not directly testable, the design of the skill tests I use (see the next section) and a placebo analysis using math skills as pseudo-outcome (see the sensitivity checks) provide some evidence that this assumption might be justified in the application at hand. The second requirement for the DD approach is that the unobservable factors μ affect grammar and math skills in the same way and are, thus, cancelled out. This boils down to the assumption that there are no skill-variant factors correlating with the treatment status after conditioning on the observable factors. In general, there are two categories of factors may violating this assumption: innate factors and nurtured factors. An innate difference would arise if some children are born with better language skills, while others have better innate math skills. If parents systematically choose a preschool with language training in response to such an innate difference, this would bias the estimated treatment effect in the DD approach. However, relying on research in psychology, such a gap in innate skills seems rather unlikely. Evidence indicates a substantial overlap across the innate skills in different domains.[14] This inter-correlation of ability tests allows

[13] This corresponds to the common trend assumption when establishing the DD approach by comparing the treatment and control group over time.

[14] Psychological research distinguishes two theories of the heritability of cognitive skills: modularity (that is, specific abilities are also genetically distinct) vs. molarity (that is, “people who do well on tests of one type of cognitive skill also tend to do well on tests of other cognitive abilities,” Wright, 1998, p.64). Genetic research indicates that the development of all kinds of skills is shaped through an underlying “general intelligence” (the so-called factor g in psychometrics, see American Psychological Association, 1995), which speaks for a molarity of cognitive skills (see, for instance, Plomin and DeFries, 1998 and Plomin and Spinath, 2002 for reviews).

summarizing an individual's performance using a single measure – the intelligence quotient. If this unity of the inherited skills holds, a genetic component confounding the relationship between language training and grammar skills is removed by taking the difference between grammar and math skills. The second concern, the existence of nurtured differences in the development of certain skills, may arise if parents prefer one kind of skills over another kind and choose the preschool accordingly. To accommodate for such a selection, I need to rely on the observable factors to capture parental and preschool preferences for either grammar or math skills.[15]

2.4 Data and variables

2.4.1 The National Education Panel Study

The data is taken from the German National Education Panel Study (NEPS), see Blossfeld et al. (2011).[16,17] The NEPS follows a multi-cohort sequence design, that is, it samples six different populations in a longitudinal manner. The six populations are newborns, children aged four, fifth graders, ninth graders, first-year university students, and adults. For each of the six samples, called "Starting Cohorts," the data exhibit a panel structure. Across all Starting Cohorts the data collection is organized along five dimensions: competence development, educational process, educational decisions, educational acquisition, and returns to education. The questions are designed by sociologists and psychologists to ensure comparability across the Starting Cohorts as far as this is possible.

For the purpose of the analysis, I use Starting Cohort 2 with data on children aged four at panel entry.[18] Staring with the first wave in 2010/11 about 3,000

[15] In the robustness checks I condition on variables assessing child's activities at home and in preschool. If parents, for instance, prefer language skills over math skills (or see a bigger need for improving language skills over math skills), this might be reflected in how often they read to their child. This piece of information this assessed in the questions on the child's activities.

[16] This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Kindergarten, doi:10.5157/NEPS:SC2:5.1.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

[17] Extensive documentation in German and English language may be found here: https://www.neps-data.de/en-us/datacenter/dataanddocumentation/startingcohortkindergarten/documentation.aspx. The examples of questions and test items I refer to are taken from the online resources.

[18] Figure 2.1 in the introduction is based on the adults sample of Starting Cohort 6 of the NEPS. The analysis benefits from the feature that the competence tests in all Stating Cohorts (but newborns) are designed to be comparable across the age groups (Weinert et al., 2011).

four-year-olds in 279 preschools were sampled and followed throughout their transition into elementary school (in the most recent release, wave 5 in 2014/15, children are in the third grade). While the children themselves only participate in competence tests, their parents and preschool educators answer extensive questionnaires about them and their own characteristics. Additionally, the preschools' principals provide information on themselves and the preschool. Principals report, for instance, in the first wave, whether a curriculum-based language training has been implemented and, if so, which program. This information is available for 214 of the preschools[19], resulting in about 2,000 observations on children of whom about one-third are in a preschool that has implemented the language training.

While the assessment of the treatment status in first wave only does not allow exploiting the panel structure to analyze a language training introduction, the longitudinal nature makes it possible to investigate the persistency of the treatment effect. However, due to the sampling of the preschools, elementary school information is only available for about 300 children.[20]

2.4.2 Skill measures

To assess several domains of cognitive competencies, extensive test procedures (about one hour in every wave) have been implemented in the NEPS, see Weinert et al. (2011). Due to the long test duration not each skill domain is tested in every wave, instead the different domains are usually repeated every other wave. The skill tests are designed along two principals: first, allowing a comparison across Starting Cohorts (that is, the difficulty and topics are adjusted for the age groups) and, second, the items are constructed to avoid overlapping between the different skill domains. In the Starting Cohort under review, the tests are conducted in one-to-one situations with specially trained interviewers who visit the preschool and elementary school of the child, respectively.

Grammar skills. The main outcome variable is the grammar test score in the first wave at the children's age of four. The grammar test takes 10 minutes and covers 48 items. The questions are based on the German version of the internationally established "Test for Reception of Grammar" (see Bishop, 1989, and Fox-Boyer, 2006, for the German adoption). An example of one of the items is depicted in

[19] Following the suggestion of Stuart (2010), missing values in the covariates are recoded to non-missing values and a variable indicating this is added to the analysis.

[20] The preschools are sampled through a register of all elementary schools and enter the sample if an elementary school reports to take children from the preschool. Only children enrolled in one of the originally sampled elementary schools are followed.

Figure O2.3 of the Online Appendix. In this example, the sentence "The cats are looking at the ball" is given by CD and the child is asked to point to one out of four pictures shown by the interviewer that depicts the situation described in the sentence the best. The resulting test score is the number of correct answers. A similar test, covering 40 items in 15 minutes, was conducted in wave 3 (first elementary school grade). In wave 5, the than-eight-year-olds answered a more broadly measured language test that covers reading speed and text understanding. Figure A2.1 in the Appendix depicts the distribution of the grammar test score in the first wave by treatment status. The figure suggests, on average, neither a reinforcing nor a compensating selection into the treatment.

Math skills. Math skills are assessed in the second wave (about six months after the first wave) alongside basic cognitive functioning (e.g., perceptional speed). The math test takes about 30 minutes and covers 26 items. In one of the questions, for example, the interviewer shows the child a bowl with four stones, covers the bowl under a blanket, and adds three stones. The child is asked to name the total number of stones in the bowl. Again, the total test score is the number of correct answers. An age-adjusted math test is repeated in second grade of elementary school (wave 4) at the children's age of seven.[21]

2.4.3 Potential confounders

Table A2.1 in the Appendix summarizes the potential confounders mapping the seven channels of a potential selection. Variables for the *provider* of the preschool (indicators for the municipality, the church or another non-profit as well as the number of children) and the *availability* (the child's gender, age-in-months fixed effects, an East Germany indicator[22], and the number of months the child visits the preschool) are rather easy to measure. Controlling for a selection on the other channels discussed above is less straightforward as many factors potentially capture those channels but only a few of the factors will actually confound the analysis. I address this by using a variable selection approach to reduce the dimensionality of the controls in the next section. Note, that the credibility of the

[21] Panel (a) of Figure A2.2 in the Appendix shows the raw correlation between the grammar test score in wave 1 and the math test scores in wave 2. As one might expect – and as it is a prerequisite for the DD strategy – an individual's grammar and math skills are highly correlated. This holds true even after adjusting the test score for the observables introduced in the next subsection in panel (b). This indicates that math skills capture indeed factors that are not reflected in the observables – however, those factors do not necessarily correlate with the treatment decision.

[22] While data protection makes it impossible to control for municipality fixed effects, the preschool-level variables on the composition of the socioeconomic background of the children – I am going to present in the following – should capture a lot of the variation that would also be removed through the inclusion of municipality fixed effects.

(DD-)CIA depends on the quality of the variables that can be controlled for and not necessarily their empirical relevance. In fact, the more variables are available and the fewer actually enter the model, the less likely it appears that unobservable factors confound the relationship of interest.

To capture the *need of the student body* for better language education, the NEPS data includes information on the share of children in the preschool from families with a low/middle/ high SES (as assessed by the principal), the share of children with migrational background, and the share of children from families where at least one parent has higher education. Information on the number of children and educators allows calculating the boys-girls as well as the child-educator ratio (similar to the student-teacher ratio in school). Moreover, information on the preschool's *competition* that may incentivize increasing the quality of education is available: the number of preschools within 5km and indicators for no, little, some, or strong competition. To capture the role of the *educators and the principal*, I make use of their sociodemographic characteristics (gender, age, migrational background) as well as variables that may reflect their engagement and quality (job experience, educational degree, number of times and hours spend in further training). An educator or principal having herself or himself a migrational background might, for instance, be more eager to implement the language training.

To assess the *child's individual need* for language training, I consider the child's, the parents', and the grandparents' mother tongue and their country of birth through several indicators (e.g., German, Turkish, Russian, etc., see Table A2.1 for the complete list). Using exact information on the mother tongue accounts for the "language distance" to German (see Isphording and Otten, 2013, 2014). To capture the child's general development, I include the child's birth weight and height as well as the weight and the height at the age of four. Especially birth weight and height are shows to be important proxies for the child's development and parental SES (see Currie, 2009). For this reason I also consider an indicator for premature birth. To capture *parental preferences* for education the data includes several information on the family's SES: family income, mother's and father's employment status at the child's age of four, indicators whether the parent's occupations require higher education, years of education, the number of younger and older siblings, the number of books in the household (see Brunello et al., 2017, for the potential relevance), and the living arrangements at child's age of four.

As evident from Table A2.1, preschools with the language training have, on average, more children with migrational background, less children from parents with higher education, more children coming from low-SES families (as assessed by the principal), and more educated educators. Given that the identification relies on the ability to control for all factors that confound the relationship between language training and grammar skills, having such a large number of variables is a desirable feature for the analysis.

2.5 Results

2.5.1 Covariate selection and propensity score estimation

While having many control variables at the disposal is desirable from an identification point of view, using all of the variables to estimate the PS provokes two pitfalls. First, not all potential confounders actually confound the relationship of interest. Fitting the PS using variables that enter the prediction with a non-zero coefficient by chance decreases the accuracy of the estimated probability of receiving language training. Such overfitting might additionally reduce of common support. Second, with only about 2,000 observations, including all interactions between the 100+ linear control variables is not feasible. To address these problems some kind of regularization is necessary. While many approaches of regularization have been suggested (see, e.g., the overview of Caliendo and Kopeinig, 2008), I apply the least absolute shrinkage and selection operator (Lasso) as a "modern" tool for variable selection (Belloni et al., 2014a, p.640). This is in line with an increasing number of studies that demonstrate how machine learning techniques originally developed for "big-data" problems can improve the analysis in small-data program evaluation applications (see, e.g., Belloni et al., 2014b, Athey, 2017, and Athey and Imbens, 2017). The Lasso variable selection is used in an auxiliary step before estimating the PS. The Lasso can be seen as penalized OLS estimation and has the feature to set the coefficients of variables, that do not contribute to the prediction of the PS, to exactly zero (Varian, 2014).[23] I refer to the variables with non-zero coefficients as "Lasso-chosen" or "Lasso-selected" variables. To protect against omitted variables, I follow Belloni et al. (2014a) and employ the Lasso strategy to select the predictors of the PS as well as of grammar skills. This approach is known as "double Lasso selection" and the union of the Lasso-chosen variables that affect the treatment or the outcome enter the analysis outlined in Section 2.3 as control variables.

[23] I rely on the Stata ado file `lassoShooting` provided online by Christian Hansen to conduced the Lasso analysis (see Hansen, 2014). All errors are my own responsibility. The penalty level, that affects how many coefficients are set to zero, is assessed through cross-validation, where the out-of-sample prediction error is minimized by the penalty level.

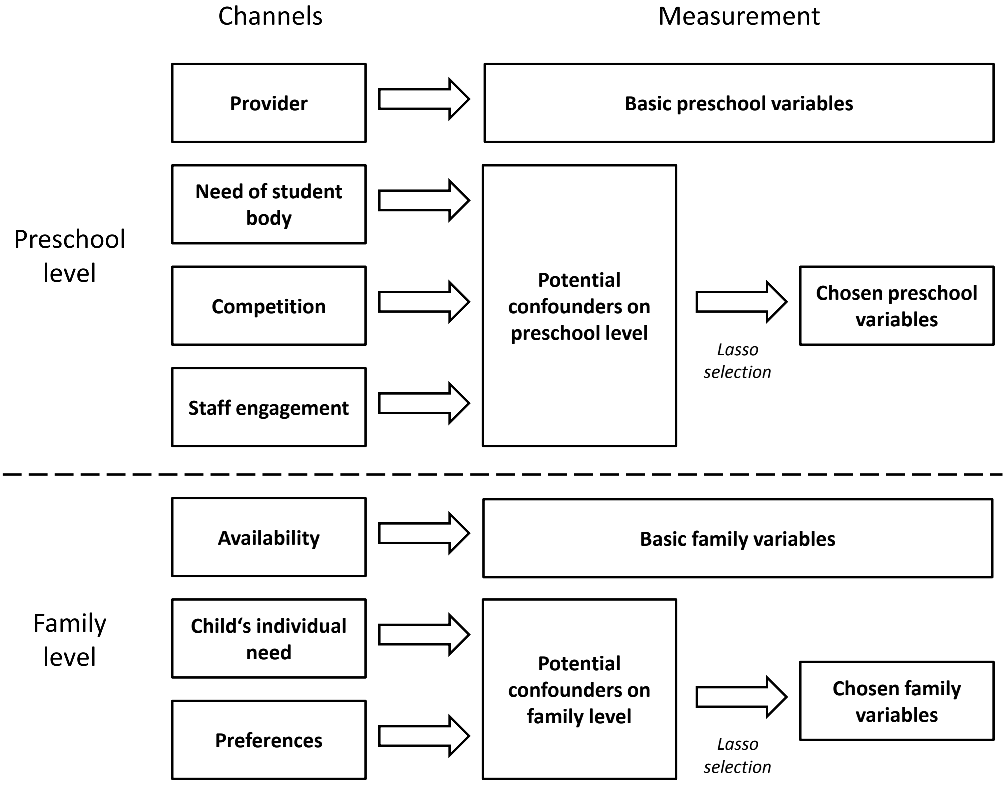


Figure 2.3: Channels of selection and their empirical implementation

Notes: Own illustration. The channels on the left-hand side correspond to channels (*i*)–(*vii*) discussed in Section 2.2.2. On preschool level, variables on the provider are treated as basic covariates that enter the model in spite of statistical considerations. On family level, covariates on the availability of language training, such as age-in-months indicators, are basic variables. Variables for the other five channels are chosen using Lasso regularization in order to avoid overfitting. Basic variables enter the model first, afterwards variables on preschool and family level are chosen by Lasso regression sequentially, starting with preschool variables and conditioning on the basic variables. The smaller boxes for the chosen variables indicate that only a subset of all variables is chosen to enter the final propensity score estimation.

Instead of using a "throwing-it-all-in approach" – that is, applying the double Lasso selection to all potential confounders at once, I follow the suggestion of Mullainathan and Spiess (2017) and combine the Lasso tool with knowledge of the decision-making process that determines the treatment status. Figure 2.3 shows the utilization of the covariates for the seven channels of selection. The *provider* and *availability* channels are not subject to the selection procedure. Variables that account for these channels are fairly easy to measure and I deem them as too important to be left out (one may think, for instance, about the importance of the child's age for grammar skills). Imbens and Rubin (2015) refer to such factors as basic variables. Given the basic variables, I then, in turn, first, choose the confounders on preschool level and, second, on family level using double Lasso selection. Finally, I select the pairwise interaction terms of the chosen linear variables in the same way (see Imbens, 2015).

Because the stepwise selection could, again, cause an overfitting problem, Figure 2.4 shows the accuracy of the PS estimation and the share of "off-support" observations along the Lasso-chosen covariate blocks. The prediction error rate is given on the left y-axis (see the figure note for the calculation) and the share of observations off-support is given on the right y-axis. The figure depicts the expected trade-off between a large number of included variables and a strong common support. Still, even in the most accurate specification that includes the highest number of variables (the model with interactions), less than 10 percent of the observations are off-support.[24] Table 2.2 shows both the selected linear covariates and their unmatched and matched sample means by treatment status. Following the structure of Figure 2.4, the table presents the variables in the order they enter the model, starting from the top. After including the basic covariates, I first choose the predictors of the treatment on preschool level. The Lasso approach indicates seven variables as important predictors (that is, the coefficients of these seven variables in the Lasso estimation are different from zero).[25] Only three variables on preschool level are chosen as predictors of the outcome. Given the nine variables on preschool level chosen in this step (the share of children with migrational background is chosen to predict the treatment and the outcome), I next select the covariates on family level. Interestingly, given the preschool-level variables, no family-level variable contributes to predicting the treatment status. In other words, there seems to be no parental selection of children into the language training after conditioning on preschool characteristics and some basic sociodemographics of the child. However, seven family-level characteristics still contribute to explaining the child's grammar skills.

If the matching is successful, the means of the covariates should not significantly differ across the treatment and the control group, given the estimated PS. To assess this, Table 2.2 compares the mean of treated and untreated individuals before and after the matching. Columns 3 and 6 report the p-values of a t-test of equal means for treated and untreated individuals before after the matching, respectively. The high p-values after the matching indicate that the null of hypothesis of equal means cannot be rejected at the conventional levels. Hence, conditioning on the estimated PS seems sufficient to balance the sample. The bottom of the table additionally gives the median standardized bias before and after the match-

[24] Figure A2.3 in the Appendix gives a similar plot when all variables introduced in Section 2.4.3 are taken for the PS estimation. The figure bears three lessons that underline the advantage of the Lasso approach compared to no regularization: the share of off-support observations in Figure A2.3 is higher, the fit is worse (in fact, the error rate indicates that family-level variables should not be used), and the large number of linear variables prevents including all pairwise higher-order terms.

[25] The Lasso algorithm is not able to reveal the underlying data generating process, different samples (or folds of the original sample) may yield a distinct sets of covariates (see Mullainathan and Spiess, 2017). For this reason I only discuss how many variables are chosen from the different blocks, but I do not interpret the choice in the light of some theory, why exactly those variables enter the model.

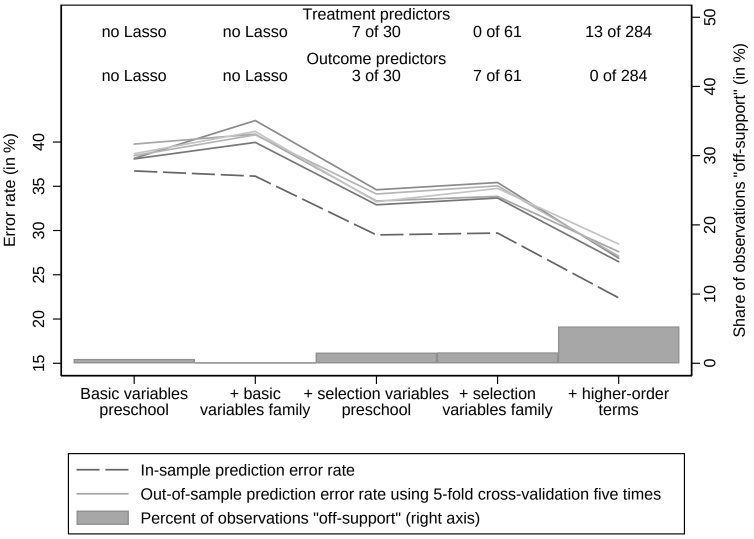


Figure 2.4: Error rate of the estimated propensity score across different sets of covariates

Notes: Own calculations based on NEPS–Starting Cohort 2 using 1,911 observations. For each of the five specifications the propensity score is estimated using a probit model. The error rate (left axis) is calculated as the share of falsely assigned treatment statuses, formally $Err = \frac{1}{n}\sum_{i=1}^{N} \mathbb{1}(D_i \neq \hat{D}_i)$, where $\mathbb{1}(\cdot)$ takes the value 1 if $D_i \neq \hat{D}_i$ and $\hat{D}_i$ is set to 1 if the estimated probability of being treated exceeds 0.5, otherwise $\hat{D}_i = 0$. Beside the in-sample error rate, the out-of-sample error rate is calculated using 5-fold cross-validation five times. For the use of cross-validated error rates in prediction, see, e.g., James et al. (2013, chapters 5-6). For the three specifications on the right-hand side double post-Lasso selection is applied before the probit model. The share of off-support observations (right axis) is calculated using Epanechnikov kernel matching with a bandwidth 0.06.

ing (see Rosenbaum and Rubin, 1985, for the calculation). Balancing the sample reduces the median standardized bias from 7.5 to 3.0 percent. Although there is no clear theoretical indication, a standardized bias below 5 is usually considered a successful matching approach. Figure A2.4 in the Appendix plots the resulting PS distributions by treatment status. As one might expect, untreated children are less likely to receive language training than treated children. The estimated PS has a common support up to 0.93 (that is, a 93 percent probability of receiving the treatment).

2.5.2 Baseline results

Table 2.3 presents the baseline results.[26] Columns 1 and 2 show the benchmark linear regression effect of language training on grammar skills and the difference between grammar and math skills, respectively. As the models rely on the Lasso-selected control variables, I refer to them as "post-Lasso" specifications. Both skill measures are standardized to mean 0 and standard deviation (SD) 1. The association between the treatment and grammar skills in column 1 is 14.5 percent of a

[26] Table O2.1 in the Online Appendix gives the full estimation output for the selection equation estimated with OLS and probit as well as for the outcome equation assessed by OLS and PSM.

Table 2.2: Descriptive statistics of selected linear covariates

	(1)	(2)	(3)	(4)	(5)	(6)
	Unmatched sample			Matched sample		
	Mean untreated	Mean treated	*p*-value diff.	Mean untreated	Mean treated	*p*-value diff.
Basic covariates preschool						
Num. of children	80.82	85.22	0.54	68.84	79.28	0.09
Provider: municiplaity	0.28	0.39	0.00	0.34	0.31	0.80
Provider: church	0.39	0.40	0.84	0.41	0.43	0.82
Provider: other	0.17	0.18	0.95	0.18	0.17	0.94
External supervision	0.25	0.27	0.19	0.19	0.24	0.47
Basic covariates family						
Age-in-months fixed effects *(omitted from output)*						
C: female	0.47	0.50	0.24	0.50	0.49	0.87
East-Germany	0.19	0.09	0.04	0.04	0.08	0.24
Current care	24.31	22.27	0.81	21.17	22.61	0.22
Preschool covariates selected for treatment						
Boys/girls ration	1.02	1.13	0.06	1.07	1.10	0.74
Share migration	18.77	26.97	0.05	20.81	21.44	0.89
Child/educator ratio	10.07	13.01	0.03	13.84	13.37	0.84
Competition: none	0.14	0.28	0.04	0.14	0.20	0.48
Competition: high	0.13	0.06	0.14	0.02	0.04	0.60
Princ.: further training	0.37	0.57	0.01	0.55	0.54	0.93
Princ.: college educ.	0.25	0.17	0.25	0.14	0.16	0.69
Preschool covariates selected for outcome						
Preschool: fee	95.89	87.57	0.30	92.01	93.68	0.88
Share higher education	17.66	14.65	0.32	12.56	15.38	0.39
Share migrants *(see above)*						
Family covariates selected for treatment (no variables selected)						
Family covariates selected for outcome						
Books at home: 11–25	0.09	0.10	0.49	0.09	0.08	0.86
M: years educ.	10.95	10.71	0.57	11.00	11.10	0.88
M: acad. job	0.17	0.14	0.22	0.16	0.16	0.78
F: acad. job	0.18	0.15	0.25	0.16	0.17	0.79
C: German	0.71	0.68	0.41	0.70	0.71	0.84
M: German	0.65	0.61	0.31	0.65	0.65	0.96
German mother–father	0.08	0.23	0.07	0.11	0.10	0.81
Median stand. bias			7.5			3.0

Notes: Own calculations based on NEPS–Starting Cohort 2 using 1,911 observations. Letters in front of the variables give the person to whom the variables refer: C=child, M=mother, F=father. Columns 1 and 2 report the mean values of the Lasso-chosen covariates for treated and untreated individuals, respectively, in the unmatched sample. Columns 4 and 5 state the corresponding numbers for the matched sample. The matched sample refers to the observations on-support weighted by the linear index of the propensity score. Columns 3 and 6 give the *p*-values of a *t*-test of equal means between the treated and the untreated individuals for the unmatched and the matched sample, respectively. A high *p*-value indicates that the null (equal means) cannot be rejected at the conventional levels. Age-in-months fixed effects, missing value indicators, and higher-order terms are omitted for brevity. The bottom of the table reports the median standardized bias for before and after the matching. The median standardized bias includes all variables.

SD and statistically significant different from zero at the 5 percent level. Going to the DD model in column 2, the coefficient is remarkable similar with 13.1 percent of a SD. Columns 3 and 4 of give the regression-adjusted simple-difference PSM and DD-PSM results, respectively. The point estimates are 14.1 and 13.8 percent of a SD, respectively.

Table 2.3: Baseline results

	(1)	(2)	(3)	(4)
	Post-Lasso regression		Regression-adj. matching	
	Simple diff.	DD	Simple diff.	DD
Effect of language training program				
Coefficient	0.148^{***}	0.130^{**}	0.141^{**}	0.138^{**}
S.E.	(0.055)	(0.063)	(0.059)	(0.054)

Notes: Own calculations based on NEPS–Starting Cohort 2 using 1,911 observations, 1,794 on-support. Every cell states the estimated effect of language training on the grammar skills (odd columns) and the grammar-math skill difference (even columns), respectively. Outcome variables are standardized to mean 0 and standard deviation 1. To perform the matching I use the Stata ado-file `psmatch2` by Edwin Leuven and Barbara Sianesi (see Leuven and Sianesi, 2003). All errors are my own responsibility. The matching algorithm is Epanechnikov kernel matching with a bandwidth of 0.06. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

While comparing the linear regression with the matching estimates reveals that there is no non-linear effect captured through the semi-parametric nature of the matching approach, the comparison of the simple-difference and DD results is more interesting as this sheds light on the role of the covariates. If skill-constant unobservable factors on individual level, such as motivation or general intelligence (if not entirely captured by parental characteristics), would bias the simple-difference estimate, the DD estimate would differ. As this is not the case, conditioning on the covariates seems sufficient for removing skill-constant unobservable factors. Although this comparison does not rule out the existence of skill-variant unobservable factors that would cause a bias, I interpret it as rather unlikely that the control variables are sufficient to capture skill-constant but not skill-variant unobservable factors. In other words, a large difference between the simple-difference and DD estimates would not only contradict the simple-difference CIA but would also cast doubt on the validity of the DD-CIA. This argument is referred to as subset-unconfoundedness (see Imbens and Rubin, 2015, chapter 21).

An effect size of about 14 percent of a SD seems moderate to large compared to the literature on the quality of education. Weiland and Yoshikawa (2013), find that one year of curriculum-based education in special preschools improves receptive vocabulary by 0.38–0.44 SD. Analysing students in grades 3–8, Chetty et al. (2014a) estimate that a one SD increase in the teacher value-added increases the reading test scores by 0.1 SD. Looking at grades 3–7, Rivkin et al. (2005) find that a one year increase in teacher experience increases the students' reading performance by 0.06 SD.

To verify the findings I conduct two kinds of robustness checks. First, Table O2.2 in the Online Appendix shows the estimates when no kind variable selection is used and when Lasso selection is applied to all potential confounders at once, respectively. Although the DD estimates exceed their OLS counterparts, both are rather close the baseline estimates. Second, Table O2.3 in the Online Appendix gives the results for 5-to-1 and 2-to-1 nearest neighbour caliper matching with replacement, respectively, and Epanechnikov kernel matching with a bandwidth of 0.02 and 0.10, respectively. The baseline results are stable across these specifications.

2.5.3 Effect heterogeneity

Program heterogeneity. As the treatment is not uniform over all preschools but the language programs vary in some aspects, program-specific effects might be more informative than the average effect presented above. With more than five programs and only about 700 treated children estimating separate effects using a treatment group that consists only of children that receive the same program does not seen reasonable. Instead, column 1 in Table A2.2 in the Appendix gives the estimates when regressing grammar skills on the treatment indicator and interaction terms between the treatment status and indicators of the three most common programs (and controls). Significant interaction terms (statistically or economically) would indicate an effect heterogeneity across the programs. However, this does not seem to be the case here. Column 2 of the table shows the heterogeneity when individuals are matched on the treatment status (without differentiating between programs) and interaction terms are included in the weighted regression. Again, there seems to be no heterogeneity.

Heterogeneity along math skills. In the baseline model, I use the grammar-math difference as outcome instead of conditioning on the exact math test score (as originally proposed by Heckman et al., 1997), because math skills are measured on a continuous scale. Before having a look at the heterogeneity along the math skills, column 1 in Table A2.3 in the Appendix, I show the DD-PSM results when conditioning on the tercile of the math test score (instead of using the grammar-math difference as outcome). That is, indicators for the math test score terciles enter the matching approach. As evident from Table A2.3, this does not change the interpretation of the baseline result. To access heterogeneity, columns 2 to 4 show the treatment effect when matching on the exact tercile. Unlike to column 1, here all individuals in the treatment and the control group are in the same math test score tercile. Interestingly, only individuals in the second tercile seem to benefit from the language training. For individuals in the lowest and highest terciles, the effect size is less than one-tenth and one-fifth, respectively. A potential reason for this pattern might be that children who perform well in the math test have also proper language skills and do not learn new things through the language training. On the other hand, children with poor math skills may struggle with learning in general and learning language even in a structured way is too hard or too fast for them as they would need an even more intensive care instead.

Heterogeneity along observables. Table A2.4 in the Appendix shows separate estimations by gender, mother tongue, and occupation of the father. Interestingly, boys seem to benefit stronger from language training than girls. In the regression and the matching models in panel (a), the treatment effect for boys is about 7 percentage points higher than for girls. The effect for girls is not statistically significant different from zero due to the smaller effect size (the standard errors are about the same for boys and girls). The opposite is the case when comparing German native speakers with non-native speakers in panel (b): the effect size is rather similar but the standard errors are higher for the latter group (this may be attributed to the smaller sample size). Otherwise, non-native speakers do not seem to benefit stronger than native speakers. Assessing a heterogeneity along the family's SES in panel (c) reveals that children from families, where the father has an occupation that does not require higher education, benefit stronger. Again, this indicates that the language training does not improve the skills of children who already process a high level of language skills.

2.6 Sensitivity analysis and discussion

So far, it seems fair to interpret the findings as suggestive evidence that the CIA is not violated: the data at hand allows accounting for all potential channels of selection, only few variables seem to confound the training-skill relationship empirically, and the point estimates are stable across the various specifications. To get a more complete picture, this section provides two kinds of supplementary evidence: first, I investigate the two additional assumptions necessary for establishing a plausible case for the DD approach (that is, math skills are independent of language training and the difference between grammar and math skills is constant in absence of the treatment) and, second, I go beyond point estimation toward partial identification.

2.6.1 Exogeneity of math skills

Math as pseudo-outcome. In column 1 in Table 2.4, I regresses math skills on the language training and controls (similar to column 1 in Table 2.3 for grammar skills) and receive a coefficient of -0.009. In column 2 in Table 2.4, I apply the PSM strategy (analogous to column 3 in Table 2.3) and get a coefficient of 0.003. I interpret these economically small estimates as evidence that language training has no direct effect on math skills.[27]

Table 2.4: Estimates for math skills as pseudo-outcome

	(1) Post-Lasso regression	(2) Regression-adj. matching
Effect of the language training program on math		
Coefficient	-0.009	0.003
S.E.	(0.064)	(0.059)

Notes: Own calculations based on NEPS–Starting Cohort 2 using 1,911 observations, 1,794 on-support. Math skills are standardized to mean 0 and standard deviation 1. The matching algorithm is Epanechnikov kernel matching with a bandwidth of 0.06. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Alternative DD specifications. Test scores on perceptual speed and reasoning skills in the NEPS enable establishing the DD approach using these so-called nonver-

[27] Similar to the baseline OLS estimates, the analysis may suffer an omitted variable bias through the unobservable component μ in Eq. (2.2). This leads to the possibility of a bigger effect that does not allow rejecting a non-zero effect. Therefore, the pseudo-outcome test should be taken as suggestive rather than firm evidence.

bal skills instead of math. Table A2.5 in the Appendix shows the regression and matching results of the DD specifications. Although the perceptual speed and reasoning tests are shorter and exhibit less variation (26 items in 90 seconds and 12 items in 6 minutes, respectively) and are based on a smaller sample, the coefficient of language training ranges between 9 and 11.5 percent of a SD. Even if one is not willing to assume that the difference between grammar and math skills is the same in absence of the treatment, if one is willing to assume that, for instance, parents do not systematically selected their child into the language training based on the difference between grammar and nonverbal skills, the interpretation of the treatment effect does not change. Put differently, if unaccounted nurturing (or genes) causes grammar and math skills to differ and is correlated with language training (that is, the DD-CIA is violated), the estimates of Table A2.5 indicate that such difference in the skills would not exist between perceptual speed/reasoning skills and math but between perceptual speed/reasoning skills and grammar skills.

Additional preference measures. To investigate skill-specific preferences, I add control variables for the child's activities at home and in the preschool (e.g., the question "How often does the child play games involving dices or cards?") as well as the parents' activities together with their child (e.g., "How often do you read to your child?"). Table O2.4 in the Online Appendix summarizes the additional preference measures. Table A2.6 in the Appendix shows the treatment effect after controlling for these factors. The effects are similar to the baseline results, this indicates that there is no selection on such preferences.[28]

All in all, the supplementary results presented here, do neither support that language training directly affects math skills nor that omitted preferences in the baseline specification cause a selection into the treatment. Thus, establishing the DD strategy via math skills in order to remove skill-constant unobservables seems to be empirically justified.

2.6.2 Selection and partial identification

Coefficient movement along covariate blocks. To gain better understanding of how the Lasso-chosen blocks of covariates change the estimated effect, Figure 2.5 depicts

[28] The activity and preference indicators might be better on the left-hand side than on the right-hand side, that is, as outcomes (see Pei et al., 2017) as they are potentially "bad controls." Tables O2.5 and O2.6 in the Online Appendix show the association between language training and the activities and preferences, respectively. As only the child's activities in the preschool are statistically related to the treatment, there seems to be no crowding-out or compensation on family level. In fact, the positive association between the treatment and the child's activities in the preschool indicate that preschools with language training put in general more emphasis on education and therefore are also better at teaching math.

the evolution of the treatment effect when the covariates are added stepwise. The raw correlation between language training and grammar skills in the topmost line in panel (a) is negative, indicating a compensating selection of poor-performing children into the treatment. Interestingly, the correlation becomes positive when either looking at the grammar-math difference in the topmost line in panel (b) or when adding covariates in panel (a). Adding, on the other hand, controls to the DD specification does not change the estimated effect much. This underlines that accounting for math skills does not remove a further selection into the treatment once the relationship is adjusted for the observable factors. Adding activity and preference indicators increases the precision of the estimate and slightly increases the R^2 (indicated by the size of the marker), but does change the economic interpretation.[29]

Bounding. To get a better grasp of how selection affects the treatment effect, I additionally implement a formal bounding exercise following the idea of Altonji et al. (2005) that the selection on observables serves as a guide for the selection on unobservables. I allow the treatment effect to be biased due to a selection on unobservable factors that is as strong as the selection on observed factors. To assess this, Oster (forthcoming) proposes the following bound:[30]

$$\beta^* \approx \tilde{\beta} - \delta(\dot{\beta} - \tilde{\beta}) \frac{R_{max} - \tilde{R}}{\tilde{R} - \dot{R}}, \tag{2.4}$$

where $\tilde{\beta}$ is the treatment effect in the baseline model that includes all Lasso-selected control variables. This effect is compared to the treatment effect of a restricted model, $\dot{\beta}$, where grammar skills are only regressed on the treatment. The movement of the language training coefficient between the models is evaluated against the corresponding change in the R^2. This change in the R^2 caused by omitting the control variables (that is, $\tilde{R} - \dot{R}$) is up-scaled by the highest plausible change in the R^2 ($R_{max} - \tilde{R}$). That is, I allow the unobserved variables to explain as much of the variation in grammar skills as an inclusion of math skills would

[29] Following the approach of Ichino et al. (2008) Figure O2.4 in the Online Appendix plots the contribution of each of the Lasso-chosen linear covariates to the overall selection and outcome effect. Because an overestimation of the treatment effect is potentially more harmful from a policy point of view than an underestimation, one may think of "dangerous" confounders as having a selection and outcome effect that goes in the same direction. For instance, if more needy children receive the language training less often, or well-off children being more often treated. Figure O2.4 shows that this is only the case for five of the covariates: the indicators for the church as provider, female children, no competition, and a principal with further training have a positive selection and outcome effect, while for the boys-girls ratio both effects are negative. This absence of a clear selection pattern in the confounders makes it unlikely, that there are unobserved confounders that matter more.

[30] This expression is only an approximation, see Oster (forthcoming) for the exact calculation.

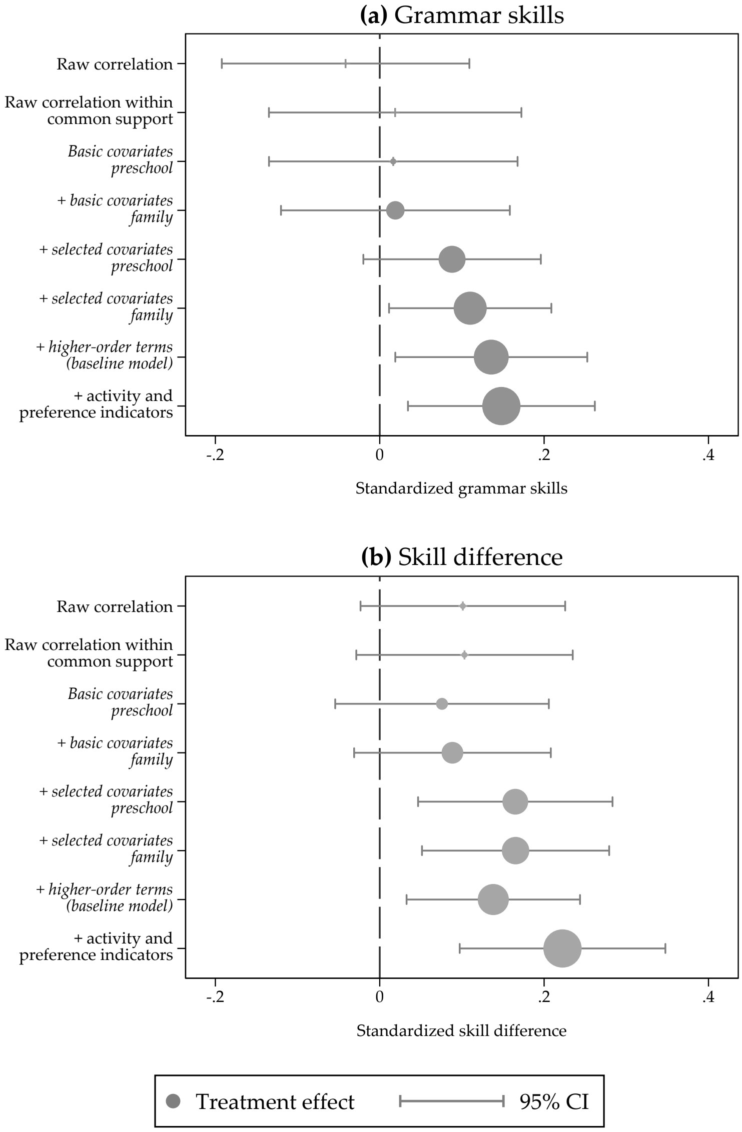


Figure 2.5: Coefficient movement along the covariate blocks

Notes: Own calculations based on NEPS–Starting Cohort 2 using 1,911 observations. Each marker gives the linear regression point estimate the of effect of language training on the grammar skills in panel (a) and the grammar-math skill difference in panel (b), respectively. The included covariate blocks are stated on the left-hand side. The Lasso-chosen blocks are in italics. The R^2 of the regression is indicated by the marker size and ranges from 0.0004 to 0.2817. Standard errors clustered on preschool-level. The spikes around the point estimates give their 95 percent confidence interval.

do.[31] The parameter δ in Eq. (2.4) denotes the degree of proportionality of selection on unobservable variables relative to the selection on the observed (and in $\dot{\beta}$ omitted) variables. As reasonable values for δ Oster suggests 1 and -1. That is, the selection on unobserved factors is exactly as strong as the (adverse) selection on observed factors. Given the extensive survey information at hand, it seems

[31] The resulting ceiling of the R^2, denoted by R_{max}, is the R^2 value of a regression of grammar skills on the treatment, the Lasso controls, and the individual's math skills. Oster (forthcoming) advises to ceil the R^2 as a value of 1, that is, the variation in grammar skills can be explained entirely, is rather unrealistic. Using an individual's math skills to ceil the movement in the R^2 is, for instance, in line with Durevall et al. (2015) who use the R^2 of an individual fixed effects model as R_{max}.

Table 2.5: Bounds

Covariates	(1) Restricted model $\dot{\beta}$	(2) Unrestricted model $\tilde{\beta}$	(3) Bound for $\delta = 1$	(4) Bound for $\delta = -1$
All covariates	0.0188 (0.0784) [0.0001]	0.1451 (0.0558) [0.2680]	0.2783	0.0647

Notes: Own calculations based on NEPS–Starting Cohort 2 using 1,911 observations. Columns 1 and 2 state the treatment effects for the restricted and the baseline model, respectively. Calculation of the bounds is based on Emily Oster's Stata ado-file `psacalc` (see Oster, 2017a). All errors are my own responsibility. Preschool-level clustered standard errors in parentheses. Model fit (R^2) in brackets.

plausible that the analysis does not miss more of the selection than it accounts for. Hence, a degree of proportionality of $\mid \delta \mid= 1$ should be appropriate.

Table 2.5 gives the estimated treatment effect for the restricted and the unrestricted model in column 1 and 2, respectively. As already indicated by Figure 2.5, the language training coefficient of the unrestricted (baseline) model exceeds its restricted counterpart. Put differently, the union of the control variables in the baseline model captures a compensating selection – that would cause an underestimation of the true effect if missed. Column 3 bounds the treatment effect when the same degree of compensating selection is missed. In this scenario, the treatment effect increases to 27.8 percent of a SD. Column 4 gives the bound when the selection on unobservables is adverse to the selection on observables (that is, the selection moves the treatment effect toward zero). The resulting estimate of the treatment effect still amounts to 6.5 percent of a SD. Thus, even if the baseline estimates miss as much selection as they account for, the treatment effect is still positive and rather large in size.

2.7 Effect persistency

Age-six grammar skills. Table 2.6 repeats the baseline analysis using grammar skills in the first grade of elementary school at age six. Although the sample size is much smaller (87 treated and 214 untreated children, see Section 2.4 for reasons), the positive association between the treatment and grammar skills increases in all specifications.[32] The estimated treatment effect in the simple-difference model increases from 14 percent of a SD in grammar skills in the baseline results to more than 30 percent in columns 1 and 3. The DD approach (the skill difference is

[32] Table O2.7 in the Online Appendix gives the baseline estimates for the this sample. The estimates are in line with Table 2.3.

established by subtracting wave-2 math skills from wave-3 grammar skills) exhibits a similar pattern. In fact, for the DD regression model in column 2, the coefficient even exceeds 40 percent of a SD. The increase in the effect size may be due to two (not mutually exclusive) reasons. First, a prolonged exposure to the treatment: once a preschool has the learning materials, children may use them until entering the elementary school at the age of six (or until they outgrow the content). Second, following the argumentation of, e.g., Heckman (2007), the increased treatment effect may reflect that a higher level of grammar skills begets the acquisition of subsequent grammar skills (for instance, because learning is easier with a sound understanding of the basics).

Table 2.6: Results for grammar skills at age six

	(1)	(2)	(3)	(4)
	Post-Lasso regression		Regression-adj. matching	
	Simple diff.	DD	Simple diff.	DD
Effect of language training program				
Coefficient	0.336**	0.444***	0.310***	0.357***
S.E.	(0.138)	(0.163)	(0.103)	(0.141)

Notes: Own calculations based on NEPS–Starting Cohort 2 using 301 observations, 221 on-support. Every cell states the estimated effect of language training on the grammar skills at age six (odd columns) and the grammar-math skill difference (even columns; grammar skills age six, math skills age four), respectively. Outcome variables are standardized to mean 0 and standard deviation 1. The matching algorithm is Epanechnikov kernel matching with a bandwidth of 0.06. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Placebo test. Similar to the previous section, it is possible to use math skills in the second grade as pseudo-outcome. Table A2.7 in the Appendix gives the results. Regressing math skills on the treatment yields a coefficient of -0.001; for the matching approach, the coefficient is 0.070. Although the matching coefficient is rather large, the linear regression indicates that there is no association between the treatment effect and math skills. Thus, if there are long-term complementarities between grammar skills and other skills (as suggested by Figure 2.1), those cross-skill complementarities seem to arise after the second grade of elementary school.

Age-eight language skills. Table A2.8 in the Appendix gives the treatment effect on text understanding and reading speed in the third grade – about four years after the treatment. The coefficients are still positive but smaller in size and not statistically different from zero at the 5 percent level. The drop in the precision may either indicate that the treatment effect is fading out over time (as suggested by Cascio and Schanzenbach, 2013) or simply reflect that the language test in wave 5 measures grammar skills more broadly. [33,34]

All in all, the medium-term perspective points toward a persistency of the pattern found for the short-run. Although only a subsample of observations is available, the impact of language training is quite precisely estimated and the effect size even exceeds the short-term one at the child's age of six. Even at the age of eight, the curriculum-based language training seems beneficial even though the analysis lacks the statistical power to rule out a zero effect.

2.8 Conclusions

Although the returns to preschool education are well-studied, it is not entirely understood why the benefits of the experimental interventions, such as the Perry Program, are much higher than the returns to large-scale roll-outs of universal programs where some studies find even zero and negative effects. Both the superior absolute quality and the higher relative quality of the formal childcare compared to the informal alternatives are conjectured to shape the differences in the long-lasting effects. Closing the gap between the benefits of the experimental programs and universal preschool programs by adopting the more comprehensive quality of the former is rather infeasible with average costs per child and year exceeding the sending on universal programs by the factor of three. Given the large potential that goes along with quality improvements, surprisingly little is known about the determinants of the overall preschool quality.

The study at hand aims at shedding light upon the quality of preschool education by analyzing the short- and medium-run returns to a curriculum-based language training. This not only contributes to the understanding of what shapes the returns to preschool education but it may also give an example of how policymakers can efficiently increase the quality of the existing universal preschool programs through a rather low-key but well-defined intervention. Using information on nearly 2,000 children in more than 200 preschools, I compare the standardized grammar test score of children who receive the language training with

[33] Table O2.8 in the Online Appendix gives the association between the treatment and performance in school. While the latter is originally assessed by the teacher on a 5-point scale, the outcome is a binary indicator that is 1 if an individual's performance is rated as above average, and 0 otherwise. Again, the findings point toward the expected direction but are statistically indifferent from zero.

[34] As Heckman et al. (2013) find that behavioral aspects mediate the long-lasting effects of the Perry Program, I also consider the children's Big Five personality traits as assessed by their parents and educators as outcomes (results are available upon request). However, there seems to be consistent pattern. This might be explained by more narrower cognitive focus of the language training intervention.

the one of children who do not receive it despite of having the same probability of doing so. Children who receive language training have, on average, about 14 percent of a standard deviation higher grammar test scores – a large-to-moderate effect size compared to the literature of school quality. This point estimate remains stable in a differences-in-differences approach established by considering an individual's difference between grammar and math skills. The constant effect size provides suggestive evidence that the observed probability of receiving language training captures all factors relevant for the treatment decision and is, thereby, able to successfully address selection. To deal with the large array of potential confounders I apply techniques adopted from the machine learning literature. A placebo regression of an individual's math skills on the language training indicator and controls suggests both that the treatment effect does not capture an unobserved selection of better-doing individuals into the treatment and that establishing the differences-in-differences approach using math skills is justified. Moving towards partial identification by means of a bounding exercises further indicates that even an omitted variable bias, that causes the estimates to miss as much of the selection as they account for, would not change the interpretation of the results. Various robustness checks confirm this finding.

Subsample evidence on the lasting effects of the curriculum-based language training in preschool on grammar skills at the time the children are in the first and third elementary school grade indicates a persistent relationship. In fact, the effect size even increases over time. This hints that early investments in preschool quality can indeed have long-lasting consequences and that the quality of the provided childcare may contribute to explaining the large returns found for experimental interventions like the Perry Program.

Appendix

Figures

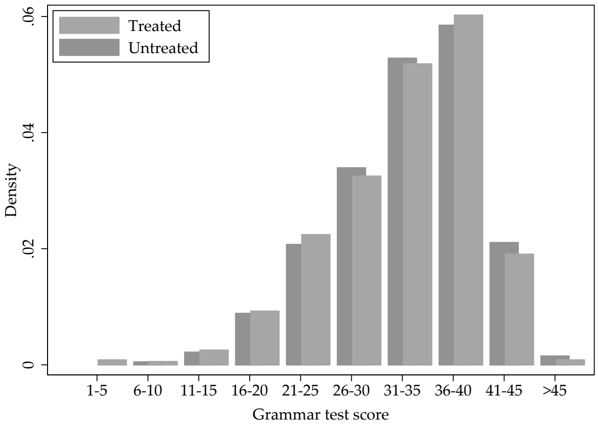


Figure A2.1: Distribution of grammar test score by treatment status

Notes: Own illustration based on NEPS–Starting Cohort 2 using 1,911 observations.

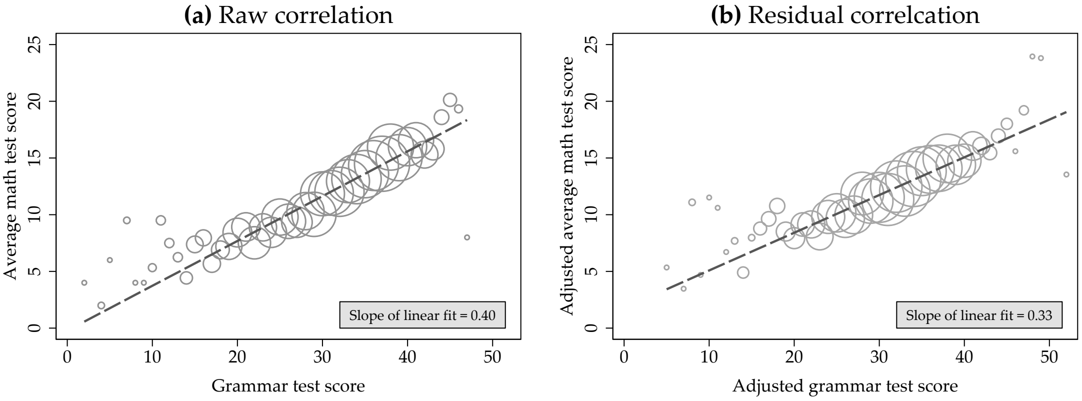


Figure A2.2: Correlation between grammar and math skills

Notes: Own calculations based on NEPS–Starting Cohort 2. Observations: 1,911. The left plot gives the raw correlation between the grammar and the math test score. For each grammar test score value on the x-axis the mean math test score is calculated and plotted on the y-axis. The fitted line results from a linear regression of the math test score on the grammar test score. For the right plot, both test scores are in an auxiliary step regressed on the observable characteristics that enter the baseline model and the unconditional mean is added to the residuals. The the residual grammar test score on the x-axis is rounded to the next integer and for each value the mean residual math test score is calculated and plotted on the y-axis. The size of the markers indicates the number of observations in the grammar test score bin. Following the Frisch-Waugh-Lovell Theorem (see, e.g., Lovell, 2008), the fitted regression line in right plot gives the gives the association between grammar and math skills when the math test score is regressed on the grammar test score and the control variables.

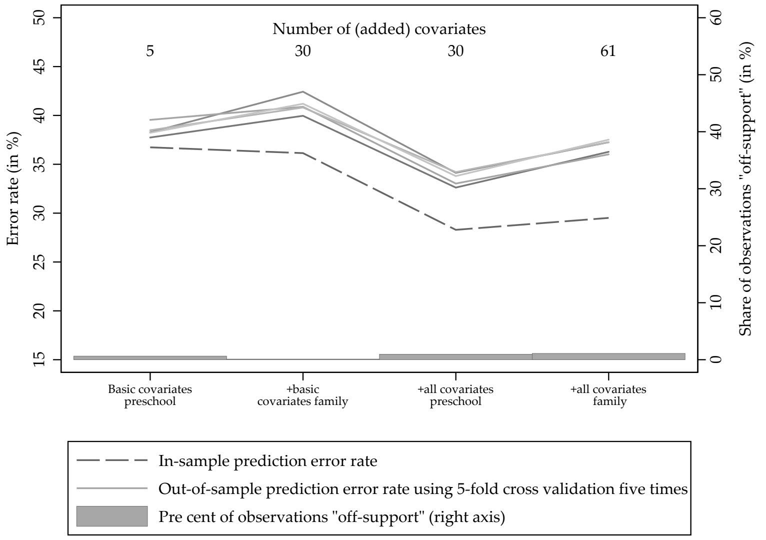


Figure A2.3: Error rate of the estimated propensity score across different sets of covariates without Lasso selection

Notes: Own calculations based on NEPS–Starting Cohort 2 using 1,911 observations. Epanechnikov kernel with bandwidth 0.06. used as matching algorithm.

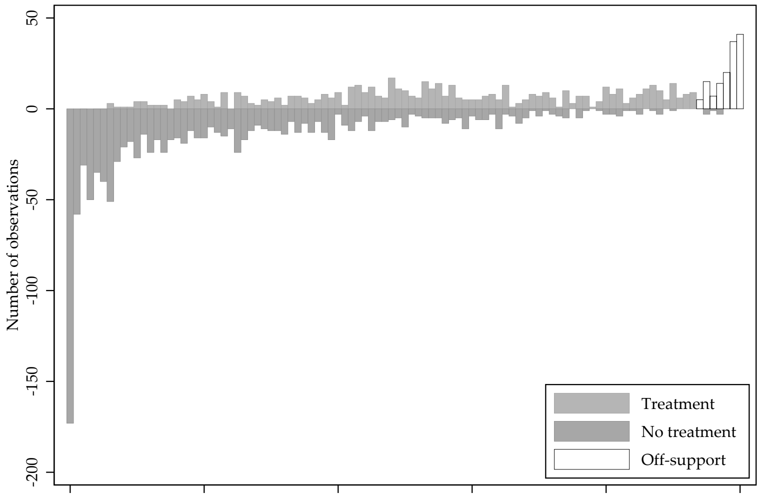


Figure A2.4: Distribution of the propensity score by treatment status

Notes: Own calculations based on NEPS–Starting Cohort 2 using 1,911 observations. Epanechnikov kernel with bandwidth 0.06. used as matching algorithm. Treated observations are off-support if the estimated PS exceeds the highest or is lower than the lowest PS of all untreated observations.

Tables

Table A2.1: Potential control variables and means by treatment status

Variable	Definition	Child	
		with treatment	w/o treatment
Preschool provider (assessed by the principal)			
Num. of children	Number of children in preschool	85.22	80.82
Provider: municipality	=1 if preschool is run by the municipality	0.34	0.28
Provider: church	=1 if preschool is run by the church	0.41	0.39
Provider: other	=1 if preschool is run by another provider	0.18	0.17
External supervision	=1 if preschool provider employs an external supervisor	0.27	0.25
Need of student body (assessed by the principal)			
Preschool: fee	Average parental fee preschool receives	87.57	95.89
Boys/girls ration	Ratio of male to female children in preschool	1.13	1.02
Share low SES	Share of children in preschool from families with a low socioeconomic background	18.00	14.28
Share middle SES	Share of children in preschool from families with a medium socioeconomic background	45.86	48.67
Share high SES	Share of children in preschool from families with a high socioeconomic background	12.12	13.72
Share higher education	Share of children in preschool from families where at least one parent has higher education	14.65	17.66
Share migration	Share of children in preschool from families with a migrational background	26.97	18.77
Competition (assessed by the principal)			
Child/educator ratio	Ratio of children to educators	13.01	10.07
Competition: none	=1 if preschool faces no competition	0.28	0.14
Competition: low	=1 if preschool faces low competition	0.45	0.50
Competition: middle	=1 if preschool faces medium competition	0.21	0.23
Competition: high	=1 if preschool faces high competition	0.06	0.13
Num. preschools within 5km	Number of preschools within 5km	6.08	6.45
Educator and principal			
Educ.: female	=1 if educator is female	0.91	0.86
Educ.: age	Age of the educator	37.41	37.17
Educ.: middle school	=1 if educator has middle school education (German *Realschule*)	0.49	0.51

Continued on next page

Table A2.1 – *continued*

Variable	Definition	Child	
		with treat-ment	w/o treat-ment
Educ.: high school	=1 if educator has high school education (German *Gymnasium*)	0.34	0.25
Educ.: other educ.	=1 if educator has other school degree	0.03	0.02
Educ.: principal	=1 if educator is also the principal of the preschool	0.09	0.08
Educ.: migration	=1 if educator has migrational background	0.04	0.04
Educ.: further training 1	=1 if educator participated in further training once in the 12 months prior to the interview	0.16	0.18
Educ.: further training 2+	=1 if educator participated in further training twice or more in the 12 months prior to the interview	0.04	0.02
Educ.: further training ¿20 hrs.	=1 if educator participated in further training with a total duration of more than 20 hours in the 12 months prior to the interview	0.42	0.37
Princ.: female	=1 if principal is female	0.96	0.97
Princ.: age	Age of the principal	48.87	50.13
Princ.: migration 1st gen.	=1 if principal has a migrational background	0.03	0.06
Princ.: migration 2nd gen.	=1 if principal's parents have a migrational background	0.05	0.04
Princ.: yrs. experience	Years of experience as principal	15.91	14.67
Princ.: further training	=1 if educator participated in further training in the 12 months prior to the interview	0.57	0.37
Princ.: college education	=1 if principal has a college degree	0.17	0.25
Availability (C=child)			
C: age in years*	Child's age in years	4.11	4.17
C: female	=1 if child is female	0.50	0.47
East Germany	=1 if family lives in East Germany	0.09	0.19
Current care	Number of months in current preschool	22.27	24.31
Child's individual need (C=child, M=mother, F=father)			
C: German	=1 if child's mother tongue is German	0.68	0.71
C: Russian	=1 if child's mother tongue is Russian	0.03	0.02
C: Turkish	=1 if child's mother tongue is Turkish	0.04	0.03
C: other lang.	=1 if child has another mother tongue	0.25	0.24
M: German	=1 if mother's mother tongue is German	0.61	0.65
M: Russian	=1 if mother's mother tongue is Russian	0.06	0.03
M: Turkish	=1 if mother's mother tongue is Turkish	0.04	0.03
M: other lang.	=1 if mother has another mother tongue	0.28	0.28
F: German	=1 if father's mother tongue is German	0.52	0.57

Continued on next page

Table A2.1 – *continued*

Variable	Definition	Child	
		with treatment	w/o treatment
F: Russian	=1 if father's mother tongue is Russian	0.04	0.02
F: Turkish	=1 if father's mother tongue is Turkish	0.05	0.03
F: other lang.	=1 if father has another mother tongue	0.39	0.38
Other language with mother	=1 if spoken language between the child and the mother is manly not German	0.08	0.05
Other language with father	=1 if spoken language between the child and the father is manly not German	0.09	0.08
Other language mother–father	=1 if spoken language between the mother and the father is manly not German	0.11	0.08
Other language with siblings	=1 if spoken language between the child and siblings is manly not German	0.02	0.02
C: German-born	=1 if child was born in Germany	0.80	0.79
M: German-born	=1 if mother was born in Germany	0.64	0.67
M: Arab-born	=1 if mother was born in Arabic country	0.01	0.02
M: Polish-born	=1 if mother was born in Poland	0.01	0.02
M: Russian-born	=1 if mother was born in Russia	0.02	0.02
M: Turkish-born	=1 if mother was born in Turkey	0.04	0.02
M: other	=1 if mother was born in another country	0.28	0.26
F: German-born	=1 if father was born in Germany	0.54	0.58
F: Arab-born	=1 if father was born in Arabic country	0.02	0.02
F: Polish-born	=1 if father was born in Poland	0.01	0.01
F: Russian-born	=1 if father was born in Russia	0.02	0.01
F: Turkish-born	=1 if father was born in Turkey	0.05	0.02
F: other	=1 if father was born in another country	0.37	0.35
M: mother non-German	=1 if mother's mother was not born in Germany	0.06	0.05
M: father non-German	=1 if mother's father was not born in Germany	0.05	0.06
F: mother non-German	=1 if father's mother was not born in Germany	0.04	0.04
F: father non-German	=1 if father's father was not born in Germany	0.03	0.04
Preschool fee: family	Preschool fee paid by the family	72.38	74.08
C: current weight (in kg)	Child's weight at the age of four	15.08	14.88
C: current height (in cm)	Child's height at the age of four	87.55	87.80
C: birth weight (in g)	Child's birth weight	2635	2638
C: birth height (in cm)	Child's birth height	39.80	39.59
C: premature birth	=1 if child was born prematurely	0.11	0.08
Parental preferences			
Family income	Family income in 100 Euros	20.90	22.95

Continued on next page

Table A2.1 – *continued*

Variable	Definition	Child	
		with treatment	w/o treatment
Num. younger sibl.	Number of younger siblings	0.50	0.51
Num. older sibl.	Number of older siblings	0.36	0.33
Books at home: 0–10	=1 if number of books at home: 0–10	0.04	0.03
Books at home: 11–25	=1 if number of books at home: 11–25	0.10	0.09
Books at home: 26–100	=1 if number of books at home: 26–100	0.27	0.25
Books at home: 101–200	=1 if number of books at home: 101–200	0.17	0.17
Books at home: 201–500	=1 if number of books at home: 201–500	0.14	0.18
Books at home: ¿500	=1 if number of books at home: ¿500	0.08	0.08
M: years educ.	Mother's years of education	10.71	10.95
F: years educ.	Father's years of education	9.40	9.77
Both parents live at home	=1 if both parents live in household	0.84	0.84
Mother's partner is not father	=1 if the partner of the mother is not the child's father	0.05	0.06
M: age	Mother's age in years	28.92	29.10
F: age	Father's age in years	27.49	27.78
M: employed	=1 if mother is employed at the child's age of four	0.46	0.49
F: employed	=1 if father is employed at the child's age of four	0.65	0.65
M: acad. job	=1 if mother has an occupation that requires higher education (ISCO code's firth digit is 2)	0.14	0.17
F: acad. job	=1 if father has an occupation that requires higher education (ISCO code's firth digit is 2)	0.15	0.18
Care: mths. in other preschool	Number of months the child spent in other formal care arrangement (also *Kinderkrippe* centers)	23.28	25.08
Care: mths. family day care	Number of months the child spent in family day care	3.95	3.77
Care: mths. home care	Number of months the child spent in home care	7.58	7.60
Number of observations		718	1,193

Notes: Own calculations based on NEPS–Starting Cohort 2. This table shows all variables that enter the Lasso specification. If a variable is chosen the enter the post-Lasso analysis, I also include a variable indicating whether missing information where replaced with non-missing values in order to allow considering the observation. Please note that I use the terms "mother" and "father" for simplicity. I refer to the interviewee as "mother" because it is almost always the mother. However, it could also be the father or another legal guardian. I refer to the partner of the interviewee as "father." Because I control for the exact status, this simplification only affects the labels of the mean values but has no consequences for the analysis.

*In the analysis I control for the child's age using indicators for each month.

Table A2.2: Heterogeneity along treatment programs

	(1) Post-Lasso regression	(2) Regression-adj. matching
Treatment status	0.141**	0.153**
	(0.065)	(0.066)
Treatment×Deflin 4	0.001	−0.019
	(0.090)	(0.105)
Treatment×Hören	−0.024	0.032
	(0.106)	(0.103)
Treatment×Kon-Lab	0.049	−0.073
	(0.116)	(0.152)

Notes: Own calculations based on NEPS–Starting Cohort 2. 1,911 observations, 1,794 on-support. Outcome variables are standardized to mean 0 and standard deviation 1. The matching algorithm is Epanechnikov kernel matching with a bandwidth of 0.06. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table A2.3: Heterogeneity along math skills

	(1)	(2)	(3)	(4)
		Regression-adj. DiD model		
		Tercile of math test score		
	All	1^{st}	2^{nd}	3^{rd}
Effect of language training program				
Coefficient	0.107^{**}	0.021	0.261^{***}	0.036
S.E.	(0.054)	(0.098)	(0.075)	(0.073)

Notes: Own calculations based on NEPS–Starting Cohort 2. 1,911 observations in total, observations on-support 1,529 (all), 483 (1^{st} tercile), 542 (2^{nd} tercile), 504 (3^{rd} tercile). The outcome is the grammar test score standardized to mean 0 and standard deviation 1. The matching algorithm is Epanechnikov kernel matching with a bandwidth of 0.06. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table A2.4: Heterogeneity along observable characteristics

	(1) Post-Lasso regression	(2) Regression-adj. matching	(3) Post-Lasso regression	(4) Regression-adj. matching
	(a) Gender		**(b) Mother tongue**	
	Female		German	
Treatment	0.107	0.064	0.161^{**}	0.166^{***}
	(0.080)	(0.071)	(0.080)	(0.071)
Observations	855	853	1277	1234
	Male		Other	
Treatment	0.170^{**}	0.135^{*}	0.146	0.180
	(0.074)	(0.080)	(0.126)	(0.119)
Observations	947	814	569	514
	(c) Academic occupation			
	Yes			
Treatment	0.053	0.085		
	(0.128)	(0.103)		
Observations	287	256		
	No			
Treatment	0.172^{**}	0.217^{***}		
	(0.080)	(0.069)		
Observations	842	738		

Notes: Own calculations based on NEPS–Starting Cohort 2. The number of observations in the column 1 refers to observations with complete information, column 2 gives observations on-support. The father's occupation refers to a first-digit ISCO classification of 2 of the partner of the interviewed parent, which is however usually the father. When the ISCO classification is missing, individuals are excluded. The outcome variable is the grammar test score standardized to mean 0 and standard deviation 1. The matching algorithm is Epanechnikov kernel matching with a bandwidth of 0.06. Preschool-level clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table A2.5: DiD approach using perceptual speed and reasoning

Difference between grammar and...	(1) Post-Lasso regression	(2) Regression-adj. matching
perceptual speed	0.106^{*}	0.087^{**}
	(0.061)	(0.060)
reasoning	0.102	0.116^{*}
	(0.065)	(0.060)
Observations	1905	1651

Notes: Own calculations based on NEPS–Starting Cohort 2. Specifications as in the baseline results but indicator for language- and math-related activities in preschool and at home are taken into account as additional covariates. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table A2.6: Treatment effect when controlling for language- and math-related activities

	(1)	(2)	(3)	(4)
	Post-Lasso regression		Regression-adj. matching	
	Simple diff.	DiD	Simple diff.	DiD
Effect of language training program				
Coefficient	0.130^{**}	0.188^{***}	0.142^{**}	0.218^{***}
S.E.	(0.060)	(0.072)	(0.058)	(0.064)
Observations	1341	1341	1168	1168

Notes: Own calculations based on NEPS–Starting Cohort 2. Specifications as in the baseline results but indicator for language- and math-related activities in preschool and at home are taken into account as additional covariates. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table A2.7: Estimates for grade-2 math skills as pseudo-outcome

	(1) Post-Lasso regression	(2) Regression-adj. matching
Effect of the language training program on math		
Coefficient	-0.001	0.070
S.E.	(0.142)	(0.119)

Notes: Own calculations based on NEPS–Starting Cohort 2 using 303 observations, 236 on-support. Math skills are standardized to mean 0 and standard deviation 1. The matching algorithm is Epanechnikov kernel matching with a bandwidth of 0.06. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

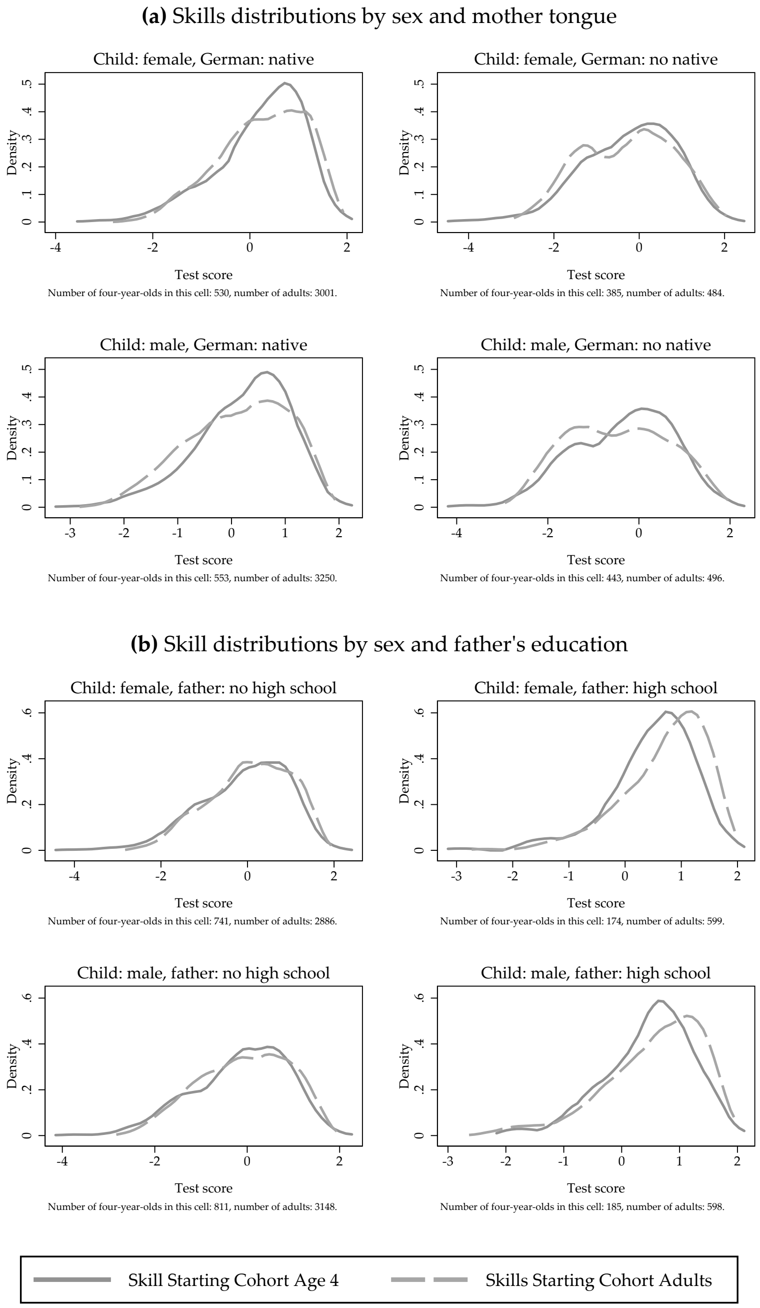
Table A2.8: Results for language skills at the age of eight (grade 3)

	(1) Post-Lasso regression	(2) Regression-adj. matching
Text understanding		
Coefficient	0.049	0.100
S.E.	(0.138)	(0.102)
Reading speed		
Coefficient	0.112	0.173^{*}
S.E.	(0.144)	(0.103)

Notes: Own calculations based on NEPS–Starting Cohort 2 using 276 observations, 250 on-support. Every cell states the estimated effect of language training on the grammar skills at age 6 (odd columns) and the grammar-math skill difference (even columns; grammar skills age 6, math skills age 4), respectively. Outcome variables are standardized to mean 0 and standard deviation 1. The matching algorithm is Epanechnikov kernel matching with a bandwidth of 0.06. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Online Appendix

Figures



(a) Skills distributions by sex and mother tongue

Child: female, German: native

Number of four-year-olds in this cell: 530, number of adults: 3001.

Child: female, German: no native

Number of four-year-olds in this cell: 385, number of adults: 484.

Child: male, German: native

Number of four-year-olds in this cell: 553, number of adults: 3250.

Child: male, German: no native

Number of four-year-olds in this cell: 443, number of adults: 496.

(b) Skill distributions by sex and father's education

Child: female, father: no high school

Number of four-year-olds in this cell: 741, number of adults: 2886.

Child: female, father: high school

Number of four-year-olds in this cell: 174, number of adults: 599.

Child: male, father: no high school

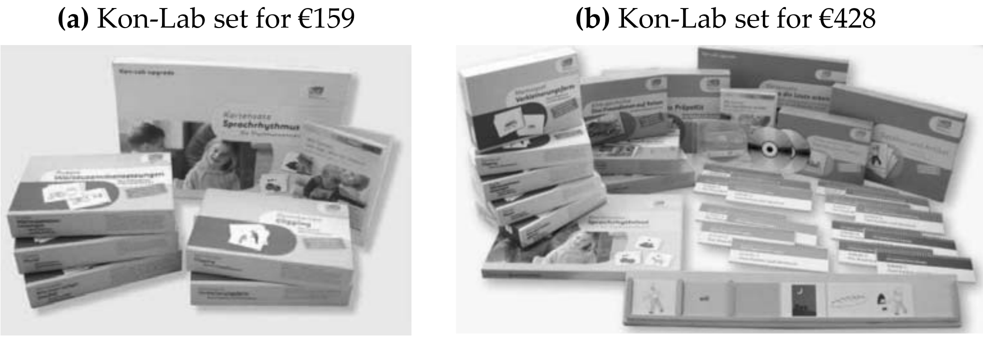
Number of four-year-olds in this cell: 811, number of adults: 3148.

Child: male, father: high school

Number of four-year-olds in this cell: 185, number of adults: 598.

Figure O2.1: Conditional distributions of grammar skills in four-year-olds and adult samples

Notes: Own calculations based on NEPS–Starting Cohort 2 and NEPS–Starting Cohort 6.

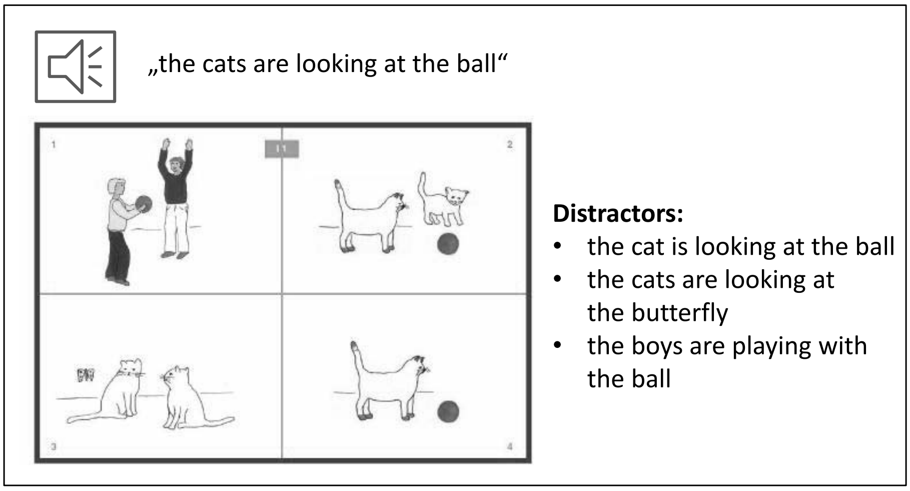


(a) Kon-Lab set for €159

(b) Kon-Lab set for €428

Figure O2.2: Example of the Kon-Lab language training program

Notes: Illustrations taken from LOGO (2014).



„the cats are looking at the ball“

Distractors:

- the cat is looking at the ball
- the cats are looking at the butterfly
- the boys are playing with the ball

Figure O2.3: Example of a grammar test question implemented in NEPS

Notes: Illustration taken from Skopek et al. (2013). Example based on the German version of the "Test for Reception of Grammar" (see Bishop, 1989, and Fox-Boyer, 2006, for the German version).

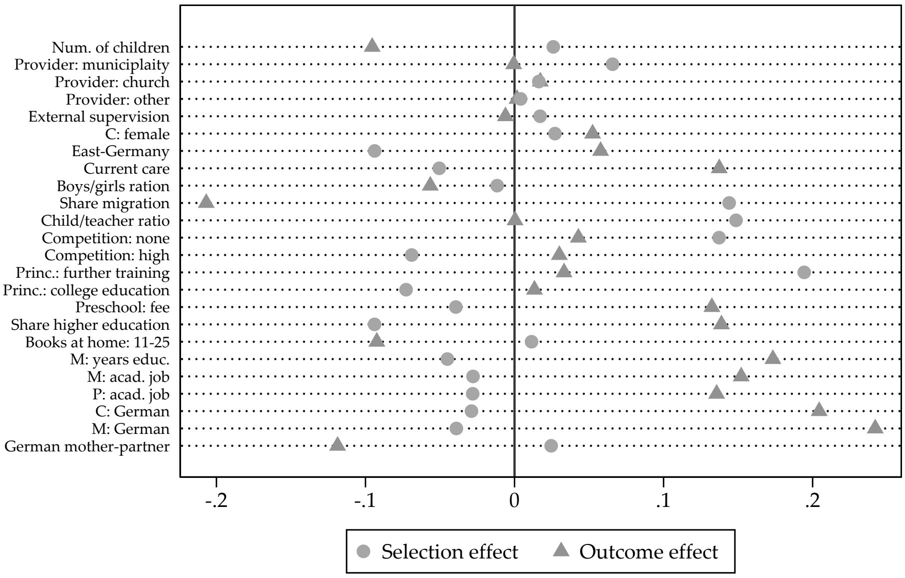


Figure O2.4: Selection and outcome effect of the covariates

Notes: Own calculations based on NEPS–Starting Cohort 2 using 1,911 observations. Following Ichino et al. (2008), the selection effect is calculated as $s = \big(Pr(U = 1 \mid D = 1, Y = 0) \times Pr(Y = 0 \mid D = 1) + Pr(U = 1 \mid D = 1, Y = 1) \times Pr(Y = 1 \mid D = 1)\big) - \big(Pr(U = 1 \mid D = 0, Y = 0) \times Pr(Y = 0 \mid D = 0) + Pr(U = 1 \mid D = 0, Y = 1) \times Pr(Y = 1 \mid D = 0)\big)$, where U is the binary confounder of interest, D the treatment, and Y the binary outcome variable. The outcome effect is $d = Pr(U \mid D = 0, Y = 1) - Pr(U \mid D = 0, Y = 0)$. Non-binary variables are recoded to 1 if their value exceeds the mean values, and 0 otherwise.

Tables

Table O2.1: Full regression output for outcome and selection regressions

	Outcome equation		Selection equation	
	Grammar skills	Skill diff.	Treatment	Treatment
	OLS	OLS	OLS	Probit
Num. of children	0.001	0.000	−0.003	−0.014
	(0.001)	(0.001)	(0.001)	(0.005)
Provider: municipality	−0.217	−0.061	0.328	−0.072
	(0.145)	(0.167)	(0.192)	(0.503)
Provider: church				−1.994
				(0.866)
Provider: other	−0.214	−0.255	0.089	−0.868
	(0.167)	(0.201)	(0.215)	(0.768)
External supervision	−0.113	−0.152	0.040	0.215
	(0.065)	(0.072)	(0.088)	(0.298)
C: female	0.066	0.258	0.018	0.084
	(0.038)	(0.046)	(0.016)	(0.061)
East Germany	0.072	0.047	−0.170	−1.938
	(0.162)	(0.143)	(0.120)	(0.720)
Current care	0.004	0.001	0.000	−0.001
	(0.003)	(0.003)	(0.001)	(0.005)
Boys/girls ration	0.149	0.118	−0.346	−1.546
	(0.108)	(0.127)	(0.147)	(0.583)
Share migration	−0.006	−0.002	−0.001	−0.003
	(0.001)	(0.002)	(0.002)	(0.006)
Child/teacher ratio	−0.001	−0.013	0.013	0.044
	(0.004)	(0.004)	(0.006)	(0.019)
Competition: none	0.323	−0.168	−0.552	−3.344
	(0.181)	(0.156)	(0.191)	(1.055)
Competition: high	0.151	0.233	−1.032	−4.992
	(0.326)	(0.437)	(0.394)	(1.916)
Princ.: further training	0.056	−0.061	−0.046	−0.578
	(0.113)	(0.135)	(0.117)	(0.531)
Princ.: college education	0.304	0.277	−0.725	−3.233
	(0.205)	(0.207)	(0.223)	(0.944)
Preschool: fee	0.001	0.000	−0.001	−0.001
	(0.000)	(0.001)	(0.001)	(0.002)
Share higher education	0.006	0.002	−0.002	−0.008
	(0.002)	(0.002)	(0.002)	(0.008)
Books at home: 11–25	−0.301	0.022	−0.005	−0.024
	(0.074)	(0.084)	(0.039)	(0.148)
M: years educ.	0.032	0.002	−0.003	−0.010

Continued on next page

Table O2.1 – *continued*

	Outcome equation		Selection equation	
	Grammar skills	Skill diff.	Treat- ment	Treat- ment
	OLS	OLS	OLS	Probit
	(0.012)	(0.014)	(0.005)	(0.018)
M: acad. job	0.127	0.011	−0.004	−0.045
	(0.059)	(0.075)	(0.029)	(0.103)
F: acad. job	0.206	−0.167	−0.027	−0.116
	(0.056)	(0.068)	(0.030)	(0.107)
C: German	0.192	0.304	−0.014	0.003
	(0.130)	(0.141)	(0.054)	(0.215)
M: German	0.294	0.189	−0.048	−0.234
	(0.086)	(0.112)	(0.044)	(0.151)
German mother–partner	−0.179	0.155	0.004	−0.016
	(0.129)	(0.136)	(0.060)	(0.236)

Notes: Own calculations based on NEPS–Starting Cohort 2 using 1,911 observations. Columns 1 and 3 give the regression results of grammar skills and the grammar-math difference on the language training indicator, respectively. In column 3, the language training indicator is regressed on the covariates. Column 4 reports the corresponding coefficients of the probit model used for estimating the propensity score. All specification in include the basic covariates as well as the Lasso-chosen covariates (see text). Age-in-months fixed effects, indicators for missing values, and higher-order terms are not reported for brevity.

Table O2.2: Robustness checks for Lasso variable selection

	No variable selection			
	(1)	(2)	(3)	(4)
	Post-Lasso regression		Regression-adj. matching	
	Simple diff.	DiD	Simple diff.	DiD
Effect of language training program				
Coefficient	0.081^{*}	0.171^{***}	0.093^{**}	0.189^{***}
S.E.	(0.047)	(0.061)	(0.045)	(0.056)
Observations	1,894	1,894	1,894	1,894

	Throwing-it-all-in double Lasso selection			
	(5)	(6)	(7)	(8)
	Post-Lasso regression		Regression-adj. matching	
	Simple diff.	DiD	Simple diff.	DiD
Effect of language training program				
Coefficient	0.090^{*}	0.170^{***}	0.080	0.208^{***}
S.E.	(0.052)	(0.064)	(0.052)	(0.061)
Observations	1,862	1,862	1,862	1,862

Notes: Own calculations based on NEPS–Starting Cohort 2. The given number of observations refers to observations on-support, the total number of observations is 1,911. Every cell states the estimated effect of language training on the grammar skills (odd columns) and the grammar-math skill difference (even columns), respectively. Outcome variables are standardized to mean 0 and standard deviation 1. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table O2.3: Robustness checks for matching specifications

	Nearest neighbor matching			
	(1)	(2)	(3)	(4)
	5-to-1 NN matching		2-to-1 NN matching	
	Simple diff.	DiD	Simple diff.	DiD
Effect of language training program				
Coefficient	0.138**	0.149**	0.210***	0.141**
S.E.	(0.064)	(0.061)	(0.069)	(0.063)
Observations	1,180	1,180	1,000	1,000

	Kernel matching			
	(5)	(6)	(7)	(8)
	bandwidth 0.02		bandwidth 0.10	
	Simple diff.	DiD	Simple diff.	DiD
Effect of language training program				
Coefficient	0.141**	0.130**	0.139***	0.143***
S.E.	(0.061)	(0.056)	(0.059)	(0.054)
Observations	1,473	1,473	1,794	1,794

Notes: Own calculations based on NEPS–Starting Cohort 2. The given number of observations refers to observations on-support, the total number of observations is 1,911. Every cell states the estimated effect of language training on the grammar skills (odd columns) and the grammar-math skill difference (even columns), respectively. Outcome variables are standardized to mean 0 and standard deviation 1. Nearest neighbor (NN) matching was conducted with replacement with a caliper of 0.25 standard deviations of the estimated propensity score. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table O2.4: Questions about child activities and parental preferences

	Share at least daily	
Question	treated	untreated
Child's activities		
In preschool (assessed by teacher)		
How often does the child occupy itself with picture books, word games and the like?	0.53	0.48
How often does the child occupy itself with number games, dice and the like?	0.39	0.31
At home (assessed by parents)		
How often does the child use picture books, word puzzles and similar things?	0.73	0.75
How often does the child use number games, dice and similar things?	0.38	0.38
Parental preferences (activities together with child)		
How often do you or someone else in the household read aloud to the child at home?	0.79	0.79
How often do you or someone else in the household show the child individual letters or the ABC, for example when looking at picture books?	0.39	0.38
How often do you or someone else in the household practice individual numbers or counting with the child, for example when playing with a dice or cards?	0.42	0.41

Notes: Own calculations based on NEPS–Starting Cohort 2.

Table O2.5: Crowding-out through child activities

	(1)	(2)
	Activity	
Effect of the treatment on...	in preschool	at home
language-related activities	0.078* (0.042)	-0.022 (0.025)
math-related activity	0.103*** (0.038)	-0.021 (0.029)
Observations	1829	1528

Notes: Own calculations based on NEPS–Starting Cohort 2. Every cell states the estimated effect of language training on the activity given in the row. The outcome variables take the value 1 if the child engages in the activity at least once a day, and 0 otherwise. The mean values are: language-related activities in preschool 0.50 and at home 0.74, math-related activities in preschool 0.34 and at home 0.38. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table O2.6: Crowding-out through parental preferences

	(1)	(2)	(3)
		Activities with child	
	Reading to child	Teaching alphabet	Playing dice or cards
Treatment	0.014 (0.026)	0.009 (0.031)	0.016 (0.033)
Observations	1529	1529	1526

Notes: Own calculations based on NEPS–Starting Cohort 2. Every cell states the estimated effect of language training on the activity given in the row. The outcome variables take the value 1 if the child engages in the activity at least once a day, and 0 otherwise. The mean values are: reading to child 0.79, teaching alphabet 0.39, and playing dice or cards 0.41. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table O2.7: Baseline results for the wave-3 sample

	(1)	(2)	(3)	(4)
	Post-Lasso regression		Regression-adj. matching	
	Simple diff.	DD	Simple diff.	DD
Effect of language training program				
Coefficient	0.063	0.165	0.122	0.148
S.E.	(0.140)	(0.117)	(0.103)	(0.097)

Notes: Own calculations based on NEPS–Starting Cohort 2 using 281 observations, 224 on-support, for that wave-3 information are available. Every cell states the estimated effect of language training on the grammar skills (odd columns) and the grammar-math skill difference (even columns), respectively. Outcome variables are standardized to mean 0 and standard deviation 1. The matching algorithm is Epanechnikov kernel matching with a bandwidth of 0.06. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: ${}^{*}p \leq 0.1$, ${}^{**}p \leq 0.05$, ${}^{***}p \leq 0.01$.

Table O2.8: Results for performance in elementary school

	(1)	(2)
	Post-Lasso regression	Regression-adj. matching
Overachieving language skills in grade 2		
Coefficient	0.059	0.073
S.E.	(0.063)	(0.047)
Overachieving language skills in grade 3		
Coefficient	−0.051	−0.027
S.E.	(0.069)	(0.061)
Overachieving writing skills in grade 2		
Coefficient	0.089	0.132***
S.E.	(0.080)	(0.046)
Overachieving writing skills in grade 3		
Coefficient	0.013	0.048***
S.E.	(0.067)	(0.051)

Notes: Own calculations based on NEPS–Starting Cohort 2 using 234 observations for language skills in grade 2 and 203 in grade 3 as well as 221 observations for writing skills in grade 2 and 205 in grade 3. The binary outcome variables take the value 1 if the teacher rates the child's perfromance as overachieving. The matching algorithm is Epanechnikov kernel matching with a bandwidth of 0.06. Preschool-level clustered standard errors (S.E.) in parentheses. Significance: ${}^{*}p \leq 0.1$, ${}^{**}p \leq 0.05$, ${}^{***}p \leq 0.01$.

Chapter 3

The Short- and Long-term Effects of Student Absence: Evidence from Sweden

Joint work with Sarah Cattan, Martin Karlsson, and Therese Nilsson

3.1 Introduction

Student absence from school is pervasive around the world. In 2015, 19 percent of fourth-graders in the US were absent from school for three or more days in the last month. Students from low-income backgrounds are more likely to be absent than their more affluent peers, and this is the case for both excused and unexcused absences (Child Trends, 2015). As school attendance correlates with academic achievement and is generally viewed as an important input in the education production function, reducing school absences has become a challenging matter for schools and a high priority for local and national governments. Indeed, absence has reached such an alarming level in some schools that commentators talk about an "empty-desk epidemic."[1]

Despite the relatively uncontested importance of reducing school absence in the policy arena, there is little causal evidence of the effect of absence on achievement and beyond. Identifying such impact is difficult for several reasons. First, it requires individual-level panel data on school absences, school performance, and any other outcome of interest. Such data rarely exist as many countries only started collecting absence records recently. Second, it requires a credible

[1] See Chicago Tribune (2012).

strategy to identify the causal impact of absence from the vast array of unobserved confounding factors. Students who miss school may be less motivated, in poorer health, or attending schools that fail to promote student engagement, which could lead to spurious correlations between absence and achievement.

The few papers that credibly estimate the causal effect of student absence focus on standardized test scores in the US context (Goodman, 2014, Aucejo and Romano, 2016). To our knowledge, there does not exist comparable evidence for a context outside the US, although high rates of student absences are prevalent in many countries. Moreover, these studies focus on the short-term impact of student absence on academic outcomes. Yet, in the presence of dynamic complementarities in the production function of human capital, the adverse effects of absence on the formation of skills could persist and even widen over the long-run (Cunha and Heckman, 2007). Assessing the long-term effects of individual absences is thus crucial to assess the potential benefits of policies aimed at reducing student absences.

An analysis of such long-run effects will meet several challenges. First, data that link student absence and performance to later labor market outcomes are rare to find. Second, the follow-up period needs to be long enough to allow for a reasonable approximation of life-cycle earnings. This second point is important given that there are several examples in the literature of early-career advantages either fading relatively fast (such as the effect of the business cycle on earnings, cf. Genda et al., 2010; Oreopoulos et al., 2012; Altonji et al., 2016) or becoming more pronounced at higher ages (such as the effects of schooling, cf. Bhuller et al., 2011). An analysis based only on early-career labor market outcomes may thus be biased in an unknown direction.

This paper fills this gap by providing evidence of the short- and long-term impact of student absence using a unique panel following a representative sample of cohorts born between 1930 and 1935 in Sweden. This novel dataset links digitized school records of absence and performance to adult socio-economic outcomes measured up to 60 years later from Census and tax register data. This combination of historical and administrative data allows us to investigate a wide array of outcomes. Specifically, we analyze the effect of student absence in grade 1 and grade 4 (at ages 7–8 and 10–11, respectively) on student performance in these grades, as well as its effect on final education, employment (at ages 25–30 and 35–40), labor market income (at ages 35–40), pensions from past labor market activity (measured at ages 67–72) and mortality.

To deal with the potential endogeneity of absence, we exploit two features of the data. First, the sample includes pairs of siblings, which we use to implement a sibling fixed effect (FE) strategy and control for all time-invariant, family-level characteristics that could simultaneously drive absence and our outcomes of interest. Second, we exploit the fact that absence and achievement were collected for two grades (grade 1 and grade 4) to control for individual FE when looking at short-term impacts. Finally, we also implement two approaches as sensitivity checks: a bounding approach following Altonji et al. (2005) and an instrumental variable (IV) strategy exploiting local and temporal changes in weather conditions as source of exogenous variation for absence.

In line with the existing literature, we find a negative and significant impact of student absence on academic performance in elementary school equivalent to 3.3 percent of a standard deviation for ten days of absence (the average number of absences in our sample). To address the arbitrariness of test score scales, we take advantage of our long panel to translate the effect on performance in school into its association with adult earnings. Anchoring the test scores to long-term income this way confirms a moderate effect size.

Our findings for long-term effects suggest that the consequences of absence in elementary school fade out over time. While absence negatively correlates with final educational achievement, employment, income and longevity, only the relationship between employment at ages 25–30 and school absence remains strongly significant when we include sibling fixed effects. In this case, the impact of absence is rather large, as ten days of absence lead to a 4 percent reduction in employment. Ten years later however, the impact of absence on employment is less precisely estimated and we cannot reject that its effect is no longer distinct from zero.

Our paper makes several contributions to a broad literature examining the impact of instructional time on educational achievement and later socio-economic outcomes. Although school absence is an important determinant of the total individual amount of time spent in school, most existing studies exploit exogenous variation in the length of the school year as source of exogenous variation in instructional time. Among others, such studies use laws and law changes that cause variation in the school year length (e.g., Leuven et al., 2010, Pischke, 2007, Sims, 2008, Agüero and Beleche, 2013, and Fischer et al., 2016)[2]; variation in test dates, where the total amount of education the students receive is eventually the same but some students are tested earlier than others (see, e.g., Carlsson et al., 2015, and Fitzpatrick et al., 2011); and unscheduled school closures resulting from ex-

[2] Other examples include Battistin and Meroni (2016, evidence for Italy), Huebener and Marcus (2015, Germany) and Bellei (2009, Chile) who use structural reforms that expand instructional time.

treme weather events (e.g., Marcotte, 2007, Marcotte and Hemelt, 2008, Marcotte and Hansen, 2010, and Hansen, 2011).

When it comes to school absence, two recent studies, Goodman (2014) and Aucejo and Romano (2016), analyze short-run effects of individual school absence in the US. Using Massachusetts data (school years 2003–2010) for students attending grade 3 onwards and North Carolina data (school years 2006–2010) for grade 3 to 5 students, respectively, they show that school results are negatively affected by absence. Both studies control for institutional heterogeneity using school, teacher and individual fixed effects. To corroborate their results, both studies also implement an IV approach using local variation in snowfall (Goodman, 2014) and infectious diseases (Aucejo and Romano, 2016) to instrument school absence.

Our paper contributes to the above literature by providing new evidence on the effect of student absence as one determinant of instructional time. Our paper is the first to present estimates of the impact of days of absence on long-term outcomes, including final education, labor market outcomes, and mortality. Moreover, we study individual-level changes in instructional time in a context outside the US. The literature examining the effect of region- or school-level changes in instructional time suggests that the educational system is an important factor for the observed effects, but individual changes in instructional time have not yet been analyzed outside the US.

Our results show that these innovations matter for our understanding of the impact of school absences. In fact, considering effects throughout the life-cycle sheds new light on previous findings regarding the role of school absence. Our short-term point estimates are remarkably close to those of Goodman (2014) and Aucejo and Romano (2016) – even though we analyze the relationship in another country and in another decade. The long-term estimates indicate that the short-term effects – although substantial at the beginning – slowly fade away over time. This highlights the importance of having outcomes measured at different points of the career, as impacts measured early in the career do not reflect impacts found later in the working life. A declining effect of missed instruction in school throughout the life-cycle is in line with Pischke (2007) who finds a negative effect of school years with reduced instructional time on subsequent schooling but no long-lasting labor market consequences.[3]

The remainder of the paper proceeds as follows. Section 3.2 provides some background on the schooling system in Sweden in the 1930s. Section 3.3 describes the

[3] More broadly, the absence of long-lasting consequences of the amount of schooling an individual receives is also in line with studies that find zero returns to compulsory schooling, e.g., Stephens and Yang (2014) for the US and Pischke and von Wachter (2008) for Germany.

data and some descriptive statistics. Section 3.4 discusses our empirical strategy, while Section 3.5 presents our results. Section 3.6 includes our sensitivity analysis and Section 3.7 concludes.

3.2 Background

3.2.1 Elementary education in Sweden in the 1930s and 1940s

The Swedish education system as it appeared in the 1930s has a long history. Compulsory schooling was introduced in 1842 when all parishes of the country had to offer basic education. In the 1930s and 1940s all children went to a common, public, and free school, *Folkskola*[4], and the country was divided into 2,400 school districts responsible for primary education. It was compulsory to enter the first grade at the age of seven and complete at least six years of schooling.

A clear majority of school districts offered six years of compulsory schooling, but a clause introduced in 1921 allowed school districts to introduce seven years of compulsory schooling. The clause was followed by a government decision on July 1, 1936 to increase compulsory schooling by one year over a twelve year period. Accordingly, a mandatory seventh grade was introduced stepwise across districts in the following years. Similarly, the length of the school year corresponded to 34.5 weeks in most districts, but in the period under review, the school year length was increased stepwise to 39 weeks.[5]

Although the responsibility for providing primary education was decentralized, the Ministry of Ecclesiastical Affairs provided clear nationwide standards that applied to all school districts. The most central decree was the 1919 Educational Plan (*Utbildningsplanen*), which included the full curriculum of the *Folkskola*. Students attended elementary school full time, six days a week.[6] Instruction was generally done in classes separated by grade. When the number of students was

[4] We use the terms *Folkskola* and elementary school interchangeably.

[5] See Fischer et al. (2013) for an analysis of the seven-year reform, and Fischer et al. (2016) for an examination of the changes in term length. In principle it would be possible to compare the effect of more instructional time due to the school year length increase with the effect of less instructional time because of absence in school. However, we would not expect the former to affect performance in the same grade. Teachers could have adjusted their expectations because all students were affected by the school year length expansion. Moreover, we do not expect long-term effects because the curriculum remained unchanged.

[6] Instruction ended at noon on Saturdays. Following an exception rule, schools in rural areas had the possibility to offer half-time reading (students went to school every second day or only during certain periods of the year) but this option was very limited in the 1930s and only 0.5 percent of our sample took half-time reading.

low, schools were also allowed to pool students in different grades into one classroom, so that a teacher instructed, for instance, students of grade 1 and grade 2 in the same room during the same lesson. The content of the education was grade-specific, however, as stated in the Educational Plan.

The educational system of the 1930s exhibited several features of a modern educational system – like absence of tuition fees and joint instruction of boys and girls at all educational levels (Erikson and Jonsson, 1993) – but education was very selective (Fischer et al., 2016). Students who decided to take more than compulsory education followed a tracking system and generally left *Folkskola* after grade 4 to enter lower secondary school (*Realskola*). All other students remained in *Folkskola* until they reached the compulsory years of schooling. From 1939 and onwards the admission to *Realskola* was based on grades received in elementary school. The system also offered a second alternative where a student could proceed to lower secondary school after finishing *Folkskola*. After four or five years of lower secondary schooling, students either entered upper secondary school (*Gymnasium*) or finished their educational career. In the birth cohorts that we consider 87.5 percent of students only have compulsory education[7], and until the 1940s only about 5 percent of a cohort continued with upper secondary schooling (Fredriksson, 1971).

3.2.2 Historical records of student absence and achievement

As their main organizational tool teachers kept daily records in an exam catalog called *Dagbok med examenskatalog* (see Appendix Figure A3.1 for a picture). In these catalogs, the teachers recorded students' performance and absences, and noted whether absences were due to sickness, natural obstacles (e.g., heavy snowfall), inappropriate clothes and shoes, other valid reasons for absence, or no valid excuses (that is, truancy). They also included general information about the school and the school year length.

Regarding student performance teachers were encouraged to take notes on the student's performance throughout the entire school year. At the end of the school year, the teachers summarized the days of absence by reason and the final grades by subject in a separate column for end-of-school-year information. Unlike tests that take place on a certain date, the frequent recording of student performance ensures that teaching-to-the-test behavior of teachers and factors on the day of a test did not affect the grades. Moreover, regular record-keeping makes recall bias of the teacher unlikely.

[7] Based on own calculation for the birth cohorts 1930–1935 using the Census 1970, see Table 3.3 in the next section.

3.2.3 Grading standards

Three theoretical subjects were taught in *Folkskola*: math; reading and speaking; and writing. Although grades recorded in exam catalogs were not based on standardized tests and hence may partly reflect teachers' subjective impressions of students, a 1940 Royal Commission established precise guidelines for teachers to evaluate and grade their students' performance relative to that of the classmates. For example, to assess a student's math performance, teachers were to take both the ability to solve "standard problems" and more sophisticated ones into account. For reading and speaking, grades were supposed to reflect loud and silent reading and the ability to express a familiar topic in own words. For writing, grades were supposed to assess both the form and content of essays.

The highest possible grade was A ("passed with great distinction") and the poorest grade was C ("not passed"). Teachers were also allowed to add a plus or minus sign in order to express the strength or weakness of the grade. While the grading scheme remained unchanged in the time under review, the grading guidelines changed slightly. From the school year 1940/41 onwards, teachers were advised to award the grade BA ("passed with credit") for an average performance. One-third of the students in the class should receive a better grade and one-third a poorer grade. Before the school year 1940/41, teachers were more likely to award a student with the grade B for an average performance. The highest grade A was reserved for exceptional students and less than 1 percent of all students should receive this grade.[8] As we show in the next section, the distribution of grades observed in our sample is remarkably in line with the Royal Commission's guidelines. This gives us confidence that, even though our main measure of academic performance is not a standardized test score, it is a valid measure to compare students' achievement with each other.

3.3 Data and descriptive statistics

The data we use for analysis combines several historical and administrative data sources. This section provides information about these sources and presents some descriptive statistics on student absence and the main outcomes of interest.

[8] In the empirical analysis, the point estimates between specifications with only parish and school fixed effects and with additional teacher fixed effects (that account for subjective grading) do not differ noteworthily. We also change the baseline outcome (performance measured on the 15-point grading scale) to the 7-point grading scale and into a binary indicator that takes the change in the Royal Commission's guidelines into account. The findings do not change our interpretation of the results.

3.3.1 Data sources

Base data The base foundation for our dataset is individual-level data from administrative church records covering all 30,150 children born between 1930 and 1935 in a representative sample of 133 out of about 2,500 Swedish parishes.[9] Figure 3.1 presents the spatial distribution of the sample parishes across Sweden. The church records contain individual information on name, gender, date of birth and parish of birth. The records also provide information on the child's parents' birth date, whether the child was born in a hospital (8 percent of the individuals in our sample), whether the birth was a twin birth (4.2 percent), and whether the child was born out of wedlock (4.4 percent). We also know the occupation of the parents at the time of birth. For the empirical analysis we generate an indicator for mothers being employed (2.4 percent) and a set of indicators for the family's socio-economic status based on the main category of the father's occupation according to the first digit of the Historical International Standard Classification of Occupations (HISCO) code[10] (see Table A3.1 in the Appendix).

Schooling data Individual schooling information was collected in local archives. Specifically, we collected the exam catalogs in which teachers made systematic notes about types of absence and reported grades for each student, for each elementary school of the 133 parishes in our base dataset. As shown in Figure A3.1 each student is listed with their first name, surname, date of birth and parents' name. Using this information, we merged the schooling information onto the base dataset. We were able to match schooling information for 17,999 out of the 30,150 children in at least grade 1 or grade 4.[11]

[9] The base dataset was originally collected and digitized to evaluate an infant and maternal health program that the Swedish Government introduced between 1931 and 1933. See Bhalotra et al. (forthcoming) for details on the construction and representativeness of the data.

[10] The HICSO code is historical version of today's International Standard Classification of Occupations (ISCO) code, see van Leeuwen et al. (2002). The HISCO occupations coding does not allow ranking jobs according to their prestige or any other criterium. The only group of occupations that can be related to a higher socio-economic status is service workers. If the father had an agricultural occupation, we additionally consider whether he was a farmer, fisherman or hunter (one HISCO category) because this is potentially related to both the family's subsistence as well as the need that children in the household help with reaping the harvest (although we find no evidence that was systematically the case).

[11] The reasons why we are not able to get a perfect match are that (1) exam catalogs were destroyed or cannot be found in the archives, (2) there is insufficient information for identifying an individual, (3) an individual left the sample parish and moved before the age of seven, and/or (4) an individual passed away before reaching school age. The first two reasons are due to the data collection and operationalization and not subject to individual selection. The decision to move and an early death are, however, likely non-random with respect to (sickness) absence and skills. If an early death is health-related, attrition due to mortality may bias the estimates. However, the long-term effect of absence on mortality does not exhibit a noteworthy association between the two factors: see the results section. To address selection due to moving we tried to

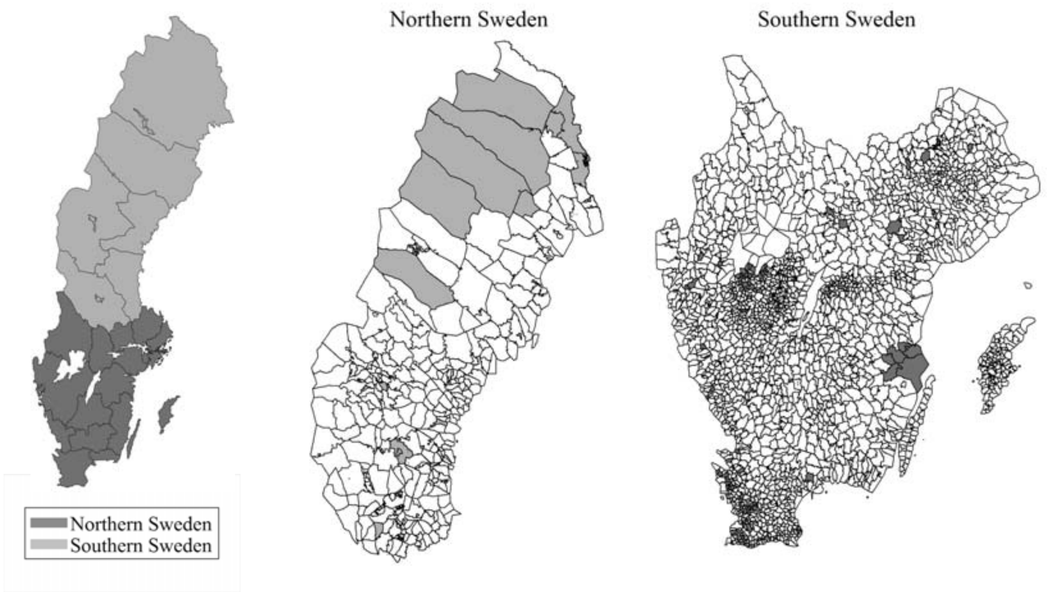


Figure 3.1: Spatial distribution of 133 sample parishes within Sweden

Notes: Own illustration. The plot on the left shows the map of Sweden in its regions (*Län*) and the plots in the center and on the right show Northern and Southern Sweden, respectively, in parishes in the time under review. The left plot indicates which regions belong to the Northern and Southern Sweden in the plots in the center and on the right. Parishes belonging to our sample are depicted darker in the plots in the center and on the right.

We focus on grade 1 and grade 4 (the last grade in which all students attend *Folkskola*) and digitize the end-of-school-year summary information of the exam catalogs.[12] With cohorts born in 1930–1935 the schooling data covers the school years 1936/37 to 1946/47.[13] Table 3.1 gives an overview over the data structure and corresponding sample sizes by birth cohort. Out of the 30,150 individuals born in the sampled parishes in 1930–1935, we have complete exam catalog records for about 14,000 individuals in either grade 1 or 4. For about 10,000 individuals we have both grades. Reassuringly there is no difference in the matching quality with respect to the birth cohort or the school grade. Using information on the par-

trace down exam catalogs for individuals who have moved to a different parish before enrolling into *Folkskola* using official registers on movers. For the very few children leaving Sweden before enrolling into *Folkskola* we have no information after they left the country. The assumption we have to make is that the decision to migrate out of Sweden is unrelated to absence in school and educational performance given the socio-economic background. The Online Appendix Table O9 compares the mean value of characteristics in the church records data (that are available of all individuals) between the full sample of all 30,150 individual and the subsample of the individual with schooling information. The results do not indicate a systematic difference in socio-economic factors. Table O10 gives the baseline results separately for individuals that moved between the birth and schooling and individuals that did not move. The coefficients are similar.

[12] Therefore, the data at hand do not allow us to identify the length of absence spells but only the total number of days missed per grade.

[13] The WWII falls in the time under review. Sweden was neutral in the war and we have not found any historical sources suggesting that the war caused major disruptions in education, nor do the war years reduce the probability that we found exam catalogs in the local archives. In fact, children from Finland were sent to and educated in Sweden because Sweden was less affected by the war, see Santavirta (2012).

Table 3.1: Number of individuals by birth cohort and sample

# of individuals...	Birth cohort						Total
	1930	1931	1932	1933	1934	1935	
...born in sample	5,355	5,095	5,116	4,743	4,775	5,066	30,150
...in grade 1	2,513	2,438	2,491	2,409	2,360	2,136	14,347
...in grade 4	2,734	2,653	2,647	2,315	2,378	1,864	14,591
...in grades 1 and 4	1,929	1,875	1,887	1,805	1,735	1,448	10,679
...with sibling info.	821	748	818	777	738	567	4,469

Notes: Own calculations based on church records and exam catalog information. For 17,999 out of the 30,150 children born in our sample parishes children we could at least find exam catalog information on one grade. For 10,679 individuals exam catalog information on both grades are available (that is, the individual panel consists 21,358 observations). 4,469 of these individuals have siblings we also observe in both grades (the siblings panel includes 8,938 observations).

Table 3.2: Grading scale

	Grade														
	Passed...														
	with great distinction		with distinction		with great credit			with credit			without credit			Not passed	
Observed symbols	A	A-	a	a-	AB+	AB	AB-	BA+	BA	BA-	B+	B	B-	BC	C
15-points scale	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
7-points scale	7		6		5			4			3			2	1

Notes: Own illustration based on historical records. The first line states the original grade as denoted in the exam catalog. Lines 2 and 3 give our translation into numerical values on a 15-point and 7-point scale, respectively. The baseline models use the 15-point scale, the Online Appendix includes results for the 7-point scale.

ents, we can also identify sibling pairs born between 1930–1935. Our final sample includes 4,469 siblings for whom we have information on both grades (resulting in 8,938 observations).

As discussed in Section 3.2, educational performance is measured with the grades teachers assign to students at the end of the school year. Each grade is assigned a numerical value which we refer to as grade point. In our baseline specification we use a scale that takes into account that teachers could assign a plus and a minus sign to a student's grade, ranging from 1 (poorest grade) to 15 (excellent grade). Table 3.2 gives the mapping of the potentially ordinal grades into cardinal grade points. To facilitate interpretation we standardize the grade points to have mean 0 and standard deviation 1. In our baseline specification, we measure achievement as the average grade across all subjects. While all students had to take math and reading and speaking, writing was not always graded in the first school year. For the 31.3 percent of students in our sample with missing writing grade points in the first grade, we calculate the average grade points using the grade points in the other two subjects.

To gain an economically meaningful interpretation of the effect size we "anchor" the raw grades in later-life earnings potential (see Bond and Lang, 2013). Performance anchored in earnings potential should not be confused with the effect of absence on income. The anchored effect of absence still gives the short-term effect on educational performance, but scaled in units of Swedish krona (SEK, in values of 2002) instead of the somewhat hard-to-interpret numerical grade points. In the analysis we exchange the grade points as dependent variable with the fitted value of the following auxiliary anchoring regression[14]:

$$y_{ig}^{\text{anchor}} = \omega_{0g} + \sum_{j=1}^{13} \omega_{1g,j} \text{math}_{ig} + \sum_{j=1}^{13} \omega_{2g,j} \text{reading}_{ig} + \sum_{j=1}^{13} \omega_{3g,j} \text{writing}_{ig} + \xi_{ig},$$

where y_{ig}^{anchor} is individual i's pension income in 2002, math_{ig}, reading_{ig} and writing_{ig} are her grade points in the particular subject in school grade g, and ξ denotes the estimation error.[15] The anchoring is performed separately for grade-1 grade points and grade-4 grade points. The estimates for the anchoring regressions are reported in the Online Appendix.

Subsequent education data Information on subsequent education beyond elementary school is taken from the highest educational degree as stated in the 1970 Census. Given that individuals are aged 35–40 in 1970, this reflects final education. In our baseline specifications we measure educational attainment with an indicator that takes the value 1 if an individual attains a more advanced track than *Folkskola*, and 0 otherwise.[16]

Labor market data We follow our sample over the life-cycle by using information from the 1960 and 1970 population censuses and tax registers available from 2002 onwards. These sources give long-term information on employment, earnings and pensions. Specifically we identify 11,570 of the individuals in the schooling dataset in the 1960 Census which includes information on individual occupation, and 10,246 individuals in the 1970 Census where we observe employment status and labor market income.[17] While the tax register data include several types of income, we use the income coming from labor market activity.

[14] The grade points in each subject enter the regression through full sets of dummy variables. Grade points of 14 and 15 are omitted as these grades are very rare.

[15] Using the labor market income from the Census 1970 for anchoring the performance in school does not change our interpretation of the findings throughout the analysis.

[16] That is, the indicator is 1 if an individual leaves *Folkskola* after grade 4 and attends *Realskola* or if the individual leaves *Folkskola* after the compulsory years of schooling and enrolls into secondary education afterwards. *Realskola* dropouts, *Folkhögskola* or *Folkskola* with 8 or 9 years of compulsory schooling are treated as *Folkskola*, see Fischer et al. (2016).

[17] The effective number of observations used in the final analysis of the long-term effects is lower as we restrict ourselves to empirical strategies that require information on both grades and only consider individuals with siblings in the data.

As individuals in our sample are well into their retirement ages, this income reflects pensions and, thereby, constitutes a proxy for lifetime earnings.[18] Missing information on labor market outcomes might be due to individuals passing away or migrating from Sweden before 1960. We try to trace down individuals that have migrated (see Appendix) and directly investigate mortality.

Mortality data The exact date of death is taken from the Swedish Death Index of Federation of Swedish Genealogical Societies (see Federation of Swedish Genealogical Societies, 2014). The data includes information on all individuals that passed away between 1901 and 2013.

3.3.2 Descriptive statistics

Our main explanatory variable of interest is the number of missed school days in grade 1 and in grade 4. The data allows us to distinguish between absences due to sickness and absences due to other reasons.[19] Figure O3.1 shows the distribution of individual days of absence and sickness absence in grade 1 and 4, respectively. In grade 1, 64 percent of all students miss less than 10 days and 6 percent of all students have no absence. The average number of missed days in grade 1 is 11 days (median 7 days). In grade 4, students tend to miss slightly more days (mean 11.6 days, median 8.5 days). 59 percent of all students miss 10 or less days and 5 percent never miss school. Despite a very different context and time period, the distribution of total days of absence is comparable with that reported in recent US studies (Goodman, 2014; Aucejo and Romano, 2016). We observe a slightly higher density of very high number of absent days than these studies report, but unlike Goodman (2014) who excludes observations with more than 60 days of absence, we do not cap absence days.

Figure O3.1 illustrates that most absences are sickness absences. Compared to sickness absence, other types of absences only play a minor role – the average number of missed days is 1.6 in grade 1 and 3.3 in grade 4. In grade 1 and 4, 60 percent and 38 percent of all students never miss a day for other reasons than sickness, respectively.[20]

[18] For the cohorts considered here, full pensions require thirty years of contributions and the level of the pension is based on the best fifteen years (Sundén, 2006). Widows were in some cases entitled to a certain share of their spouse's earnings after their death and these widow pensions represent the most important deviation from the general rules.

[19] Although the exam catalogs include columns for several reasons for non-health related absence, teachers often only noted other absence without naming the reason.

[20] The Online Appendix additionally plots the within-family and within-individual distributions of total days of absence and days of sickness absence.

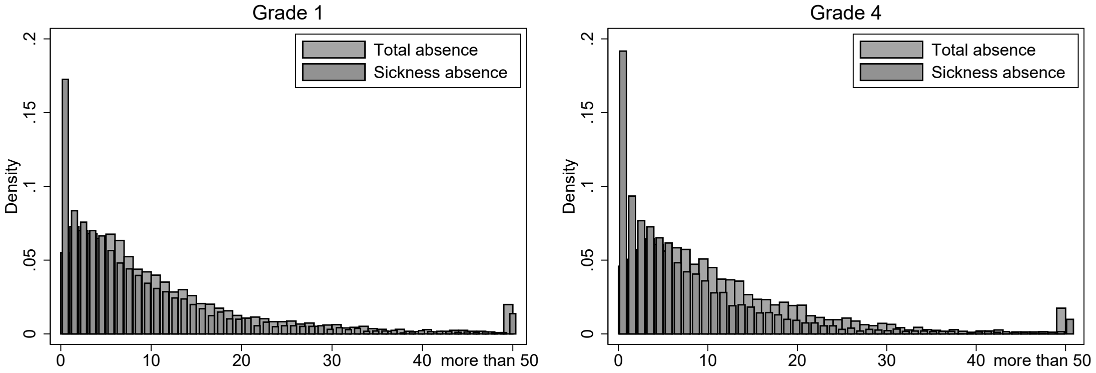


Figure 3.2: Distribution of (sickness) absence by grade

Notes: Own calculations based on exam catalog information. 8,938 observations.

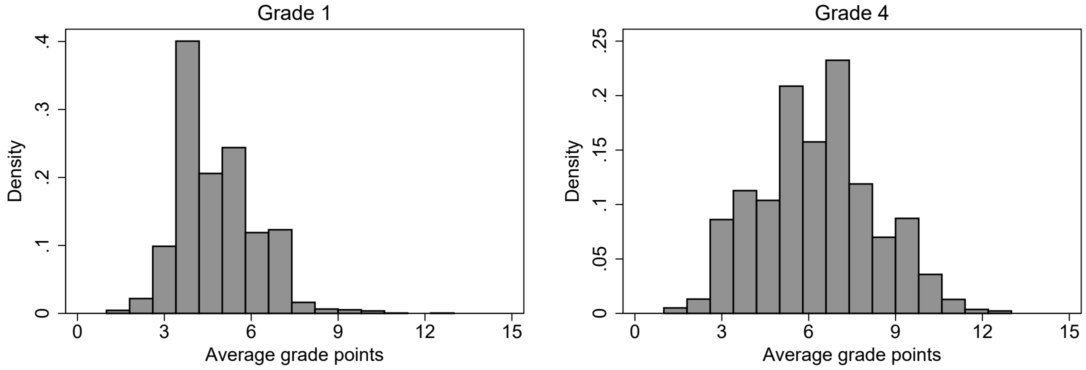


Figure 3.3: Distribution of average grade points across math, reading and speaking, and writing

Notes: Own calculations based on exam catalog information. 8,938 observations.

Turning to school achievement, Figure 3.3 shows the distribution of the raw average grade points over math, reading and speaking, and writing by school grade. In line with the suggestion of the Royal Commission, only a few students receive a very low or a very high grade point and the variance of the grade points is higher in grade 4 than in grade 1.[21]

Table 3.3 presents the long-term outcomes. The highly selective nature of the education system in the time under review, is reflected in only 13 percent of the individuals in our sample having more than *Folkskola*. Interestingly, this number does not differ by gender. Employment is measured in 1960 and 1970, when our sample is aged 25–30 and 35–40 respectively, and corresponds to a binary indicator equal to 1 if an individual is employed.

[21] The Online Appendix shows the distributions of grade points by subject and school grade.

Table 3.3: Descriptive statistics on long-term outcomes

		Mean				
	Age range	All	Female	Male	# obs	% female
Education						
More than *Folkskola* (in %)		12.51	12.99	12.02	3,565	50.29
Employment status						
in 1960 (in %)	25–30	65.08	36.51	93.96	4,129	50.28
in 1970 (in %)	35–40	57.93	43.00	72.89	4,469	50.06
Earnings						
in 1970	35–40	23,924	14,989	30,555	2,932	42.60
in 2002	67–72	150,816	128,175	175,138	3,072	51.79
Mortality at age 70						
passed away (in %)		20.70	16.45	24.96	3,072	50.06

Notes: Own calculations based on the final sample of siblings. Age range gives the individual's age at which the variable is measured. Education is taken from the Census 1970 but is likely to refer to completed schooling for most individuals. Employment in 1960 and 1970 is taken from the Census information in these years. Labor market income 1970 and pensions 2002 are based on Census 1970 and tax registers, respectively, and measured in Swedish krona in the year the information refers to. The mortality information is taken from the Swedish Death Index.

Income measures are available for 1970 and 2002. The 1970 income measure refers to the labor market income recorded in the 1970 Census when individuals in our sample were in prime working age (35–40 years old). Table 3.3 states the original values in SEK in the year in which income is measured. With individuals born 1930–1935, the 2002 income measure refers to pension income mirroring previous labor market participation (in the baseline specification we do not consider non-labor market income). Looking at longevity, 16 percent of women and 25 percent of men in our final sample passed away before reaching the age of 70.

3.3.3 Correlations between absence, academic and socio-economic outcomes

To set the stage for the empirical analysis we document the associations between the number of days of absences and the outcomes of interest. As expected, the correlation between absence and academic performance is negative, see Figure 3.4. The linear fits indicate that it is more strongly negative for sickness absence than total absence.

Figure 3.5 shows the distribution of the income measures by grouped days of absence. While the visual difference between the income distributions for individuals who have missed below 5 days and between 5 and 20 days is rather small,

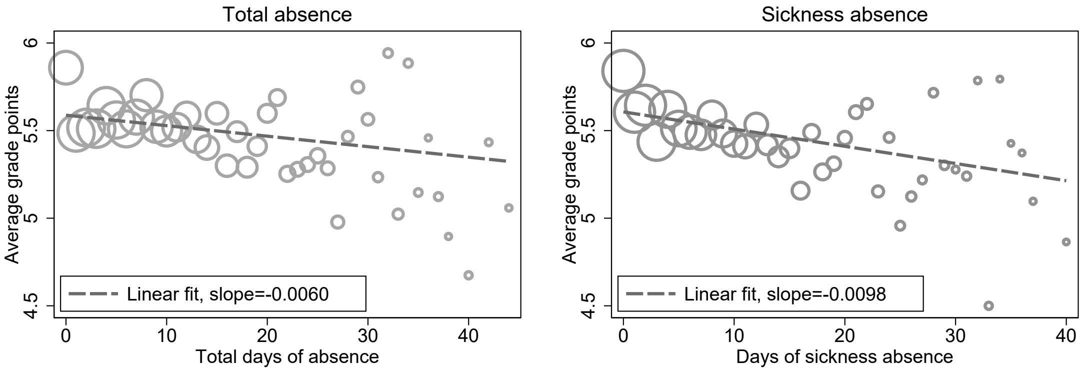


Figure 3.4: Descriptive relationship between (sickness) absence and performance

Notes: Own calculations based on exam catalog information. 8,938 observations. Grade points are collapsed on the integer of the days of absence. The size of the marker indicates the relative number of observations in the days-of-absence cell. Only cells with 15 or more observations are plotted. The fitted line is taken from a simple linear regression of performance on total absence and sickness absence, respectively, without restricting to the number of observations per cell.

individuals who missed more than 20 days because of sickness seem to earn less later in life. A Kolmogorov-Smirnov test for the equality of the distributions indicates that all conditional distributions but the 5-to-20-days and more-than-20-days distributions for income 1970 differ significantly at the 10 percent level (see note to the figure). Figure 3.6 shows the survival rate of individuals who have missed less than 5 days, 5 to 20 days or more than 20 days. The differences between the lines are small (and statistically insignificant), although individuals who missed more than 20 days seem more likely to die younger.

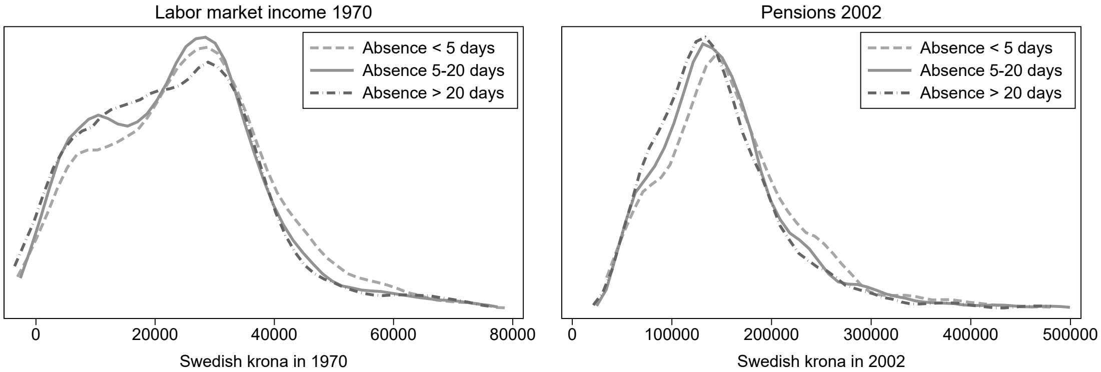


Figure 3.5: Income distributions by total days of absence

Notes: Own calculation based on exam catalog, Census 1970 and tax register 2002 information. The Census labor market income is limited to values> 0. Using a Kolmogorov-Smirnov test for the equality of the distributions yields that the <5-days distribution of the 1970 income is statistically different from the 5–20-days distribution (corrected *p*-value 0.025) and the >20-days distribution (*p*-value 0.006). The 5–20-days and the >20-days distributions do not differ at the conventional levels (*p*-value 0.223). For 2002 pensions, all three conditional distributions differ statistically significant at the 10 per cent level (*p*-values for <5 days and 5–20 days: 0.007, <5 days and >20 days: <0.001, 5–20 days and >20 days: 0.086).

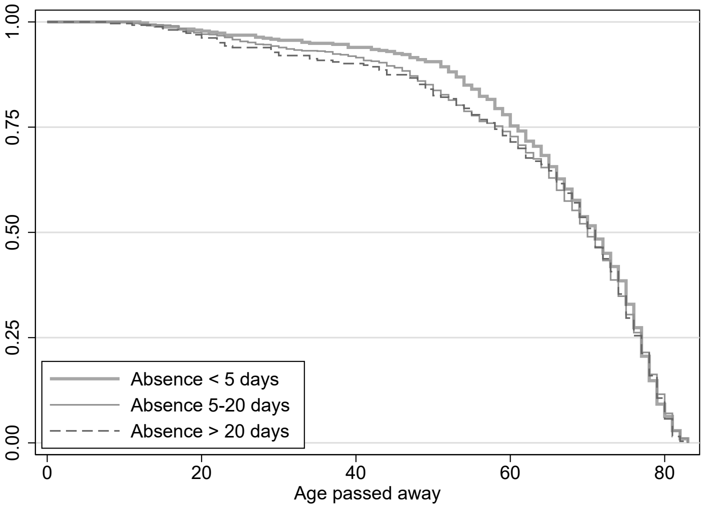


Figure 3.6: Kaplan-Meier survival function by total days of absence

Notes: Own calculations based on exam catalog and Swedish Death Index information. A Kolmogorov-Smirnov test for the equality of the distributions indicates that the conditional distributions to not differ significantly (corrected p-values for <5 days and 5–20 days: 0.209, <5 days and >20 days: 0.985, 5–20 days and >20 days: >0.999).

These figures show raw correlations and we should refrain from interpreting them as evidence of a causal link. Indeed, students who are more likely to miss school may also be those of lower ability or those of frailer nature. To start exploring the extent to which such selection may exist, Table 3.4 reports the average number of days of absence across groups of students defined by observable characteristics. Students whose father is a service worker are more likely to be absent than children whose fathers are agricultural and production workers. Children who have fewer siblings are also more likely to be absent. Based on the available observables it is not obvious whether we should expect students to select positively or negatively into absence. We now turn to the empirical strategy we propose to deal with the potential selection on unobservables.

3.4 Empirical strategy

3.4.1 The effect of absence on short- and long-term outcomes

The aim of our analysis is to estimate the causal effect of absence during elementary school on later outcomes, but absence is inherently endogenous as it likely relates to individual unobservable characteristics, including personal health. In addition, our ‘treatment variable’ – number of days of absence – is a count variable, implying varying treatment intensity. As noted above, most absence days are due to illness, which means that our ‘treatment’ is in fact typically defined as the combination (*sick*, *absent from school*). Taken together this calls for a careful

Table 3.4: Summary statistics of absence by type and individual characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Grade 1				Grade 4			
	All absences		Sickness absence		All absences		Sickness absence	
Overall	10.9	(12.0)	9.3	(10.8)	11.6	(11.9)	8.3	(10.3)
Gender								
Female	11.1	(11.9)	9.4	(10.7)	11.9	(12.8)	8.8	(11.3)
Male	10.8	(12.1)	9.2	(11.0)	11.2	(10.9)	7.8	(9.1)
Year of birth								
1930	10.2	(12.5)	8.7	(11.5)	12.0	(12.4)	8.5	(11.2)
1931	9.5	(11.4)	7.8	(9.4)	11.6	(11.6)	7.8	(10.3)
1932	10.9	(10.9)	8.9	(9.6)	12.3	(12.6)	7.9	(10.9)
1933	14.3	(13.2)	12.0	(12.0)	12.8	(12.5)	8.9	(10.1)
1934	11.2	(12.0)	9.2	(10.7)	11.6	(11.5)	8.3	(9.7)
1935	9.2	(11.2)	9.2	(11.2)	8.0	(9.1)	8.0	(9.1)
Occupation of father								
Agricultural worker	10.8	(11.5)	8.7	(10.0)	12.5	(12.0)	8.0	(10.1)
Production worker	10.8	(11.5)	8.7	(10.0)	12.5	(12.0)	8.0	(10.1)
Service worker	12.7	(14.1)	11.2	(13.3)	11.5	(10.7)	9.2	(10.0)
Number of siblings in sample								
0[a]	12.3	(13.8)	11.0	(13.3)	11.1	(11.6)	9.1	(10.7)
1	11.4	(12.2)	9.8	(11.0)	11.6	(11.8)	8.6	(10.1)
2 or more	10.3	(11.8)	8.7	(10.7)	11.5	(12.0)	7.9	(10.5)
Born out of wedlock								
yes	13.7	(13.8)	11.2	(11.6)	14.2	(15.0)	9.7	(11.4)
no	10.8	(11.9)	9.2	(10.8)	11.5	(11.7)	8.2	(10.3)
Born in hospital								
yes	11.3	(12.5)	9.9	(11.7)	11.4	(11.1)	9.0	(9.2)
no	10.9	(12.0)	9.3	(10.8)	11.6	(12.0)	8.2	(10.4)

Notes: Own calculations based on church records and exam catalog information. Observations: 8,938. Columns 1 and 5 give the mean value of the days of absence in total (that is, for all reasons) in grade 1 and 4, respectively. Columns 3 and 7 give the mean value of days of sickness absence in grade 1 and 4, respectively. Standard deviations are given in parentheses in even columns refer to the mean in the odd column on the left.

[a] Information based on the individual panel not restricted to siblings.

definition of the treatment effect (and counterfactual treatment) we are seeking to estimate.

In a standard model of a situation with a multi-valued treatment, we would denote potential outcomes under different treatment intensities w by $Y_i(w)$ (cf. Athey and Imbens, 2017), from which we may derive various treatment effects τ_{w_1,w_2} for different levels of treatment w_1 and w_2. Such a specification would require the assumption that the potential outcomes $Y_i(w)$ are insensitive to the source of variation in w. This is a reasonable approximation in many cases, but when most absence is due to illness this assumption may not be warranted. We therefore introduce a second argument, s, in the potential outcome function $Y_i(w, s)$ where s is the number of days of illness during the school year.[22]

Having defined potential outcomes, we may define the causal effect we seek to estimate. In all specifications, we seek to estimate the incremental effect of one additional day of absence from school within a school year. This causal effect corresponds to

$$\tau = \sum_{w=1}^{W} \pi_{w-1} \mathbb{E}\left[Y_i(w, s_w) - Y_i(w-1, s_{w-1})\right], \tag{3.1}$$

where π_{w-1} represents the empirical frequency of total absence days being equal to $w-1$. τ captures the effect of one additional day of absence averaged over the entire distribution of absence. However, we have not yet made any assumptions regarding s, the number of days the student is ill. Should we keep s constant when comparing different levels of absence, or should we allow it to adjust? When short-term perspectives are concerned, we probably do not want to keep s constant between different levels of w. Doing so would lead to the policy question "Should children go to school when ill?" rather than the seemingly more relevant policy question "Should we try to keep children healthy so that they do not miss school?". Put differently: in the short-term perspective, we may think of s as generating variation in absence days, based on which we can estimate the effects of absence.

For long-term outcomes, it is less clear that we want to allow s to vary in the definition of the treatment effect. Indeed, in the long-run perspective, we are more concerned that a health shock during elementary school may have persistent effects on health, which in turn would affect adult outcomes.[23] In terms of potential outcomes, we would have $Y_i(w, s) \neq Y_i(w, s')$ for $s \neq s'$, and any attempt to use variation in s to identify the effect of absence from school would also pick up an indirect effect operating via the dependence of adult health on childhood health.

Thus, for long-term outcomes, we would prefer to define the incremental effect of a day of absence as $Y_i(w, s) - Y_i(w-1, s)$ for some suitably chosen s. However, if the child's health is the main source of variation in w, it will be difficult to

[22] The days of illness s should not be confused with the days of sickness *absence*, which also depend on the choice of going to school or not when ill.

[23] The literature on the dynamics of child health suggests that shocks to a child's health have persistent effects. See for example: Currie and Stabile (2003), Contoyannis and Li (2011), Fletcher and Wolfe (2014) and Conti (2013).

estimate such an effect in the data – because irrespective of the level at which we fix s, some combinations of (w, s) will be very rare in the data. In order to address this potential issue, we try to rule out the possibility that short-term variation in health has an independent effect on outcomes by comparing the estimated effects of absence due to different reasons. If sickness absence has a similar impact on outcomes as other types of absence, it seems safe to conclude that the main component of the treatment is not poor health, but rather the absence. Such an interpretation is plausible despite persistence in health as long as health persistence is related to unobservables (such as genetic traits or family background) that our empirical strategy adequately controls for.

3.4.2 Estimation

In order to estimate the incremental effect of absence days, denoted τ in equation (3.1), a natural starting point is to estimate a model in which the achievement of an individual i in grade g, denoted by y_{ig}, is assumed to depend linearly on the number of days he or she was absent from school in grade g, denoted by W_{ig}, a set of individual-specific controls $X_{1,i}$, a set of of school-specific controls $X_{2,ig}$, and a vector of parish fixed effects P_{ig}:

$$y_{ig} = \beta_0 + \tau W_{ig} + \beta_1 X_{1,i} + \beta_2 X_{2,ig} + P_{ig} + \varepsilon_{ig}, \tag{3.2}$$

where ε_{ig} captures the unobservables affecting student performance. Given our data, the vector X_1 includes students' characteristics taken from the church records: gender, full sets of year and month of birth dummies as well as interaction terms between the year and the month of birth, age-in-month fixed effects, mother and father's year of birth dummies, father's occupation at the time of birth, and indicators for whether the child was born out of wedlock, whether the child was born in hospital and whether he or she has a twin. The vector X_2 of school-specific factors includes an indicator for grade 4 (in the pooled specification), class size as well as lowest and highest grade taught to students in the same classroom.[24] Finally, the vector of parish fixed effects P_{ig} controls for time-invariant factors that are common to all students going to school in the same parish and that affect their performance in school, for instance compulsory years of education and term length.

The key problem with interpreting the OLS estimates of τ in equation (3.2) as the causal effect of days of absence is that days of absence likely correlate with the

[24] Class size is taken into account through spline variables. That is, we include variables that group the number of classmates in bins of five, where the bins for more than five classmates only give the marginal number relative to the previous bin.

unobservables ε_{ig} and the exogeneity assumption $\mathbb{E}(\varepsilon|W, X_1, X_2, P) = 0$, which is necessary to interpret $\hat{\tau}$ as a causal effect, is likely to be violated. For example, days of absence may be correlated with unmeasured school factors, such as school resources and teacher quality, which we are not able to control. These factors are presumably positively correlated with performance and negatively correlated with absence as they determine students' engagement in school. Neglecting them would therefore overestimate the impact of absence in the OLS model.

A common approach to address this concern is to augment the above equation with school and teacher fixed effects, thus effectively relating the absences and performance of students attending the same school and taught by the same teacher. This is one of the strategies implemented by Goodman (2014) and Aucejo and Romano (2016). While this approach controls for all school-specific and teacher-specific time-invariant factors that may be confounding the effect of absence on performance, there may well be other individual-specific unobservable characteristics that distinguish students who are more frequently absent than others. If, conditional on the observables included in the model, these characteristics are correlated with performance in school, the effect of days of absence on performance will still be biased. Students who are less able, less motivated or whose parents place less emphasis on education may be absent more frequently. If these students also perform worse in school, then this unobserved difference result in a downward bias in the effect of days of absence on performance.

To address this further concern, we take advantage of two key features of our dataset: that we observe sibling pairs and that we observe students' absence and performance twice. Exploiting the fact that we observe sibling pairs we augment equation (3.2) not only with school and teacher fixed effects, but also with family fixed effects. That is, our estimating equation becomes the following siblings fixed effect model:

$$\begin{aligned} y_{i(f),g} &= \beta_0 + \tau W_{i(f),g} + \beta_1 X_{1,i(f),} + \beta_2 X_{2,i(f),g} + P_{i(f),g} + S_{i(f),g} + T_{i(f),g} \\ &\quad + \lambda_f + \varepsilon_{i(f),g}, \end{aligned} \tag{3.3}$$

where $S_{i(f),g}$ is the school fixed effect, $T_{i(f),g}$ the teacher fixed effect, and λ_f the family fixed effect for individual i in family f. This design controls for any unobserved individual characteristics that has the same additive effect on outcomes of both siblings. While siblings fixed effects remove innate genes and other family-constant factors with certainty, parental involvement could per se differ between siblings. Given that siblings in our sample are born in a relatively tight time span of five years (1930 to 1935) the underlying parenting style is less likely to differ across offspring compared to siblings born farther apart. Moreover, in the time period we study, parental involvement in their offspring's education was quite low in Sweden (Fredriksson, 1971).

When analyzing the impact of absences on short-term attainment, we strengthen the strategy even further by exploiting the fact that we observe student's absence and performance twice. We pool observations on grade 1 and grade 4 for each individual and include individual fixed effects in the estimating equation:

$$\begin{aligned} y_{i(f),g} &= \beta_0 + \tau W_{i(f),g} + \beta_1 X_{1,i} + \beta_2 X_{2,i(f),g} + P_{i(f),g} + S_{i(f),g} + T_{i(f),g} \\ &\quad +\alpha_{i(f)} + \varepsilon_{i(f),g}, \end{aligned} \tag{3.4}$$

where α_i is an individual fixed effect. This design controls for any unobserved individual characteristic that has the same linear effect on achievement in grade 1 and grade 4. Even if unobserved ability, motivation or parental taste for education differs between siblings, α_i will absorb this as long as the difference is constant between grades 1 and 4.

To implement this equation we effectively estimate the following within-student model:

$$\Delta y_{i(f)} = \tau^{FE} \Delta W_{i(f)} + \beta_2^{FE} \Delta X_{2i(f)} + \Delta P_{i(f)}^{FE} + \Delta S_{i(f)}^{FE} + \Delta T_{i(f)}^{FE} + \Delta \varepsilon_{i(f)},$$

with $\Delta y_{i(f)} \equiv y_{i(f),4} - y_{i(f),1}$, $\Delta W_{i(f)} \equiv W_{i(f),4} - W_{i(f),1}$, etc., and $\Delta \varepsilon_{i(f)} \equiv \varepsilon_{i(f),4} - \varepsilon_{i(f),1}$. The intercept β_0, the vector of time-constant observables X_1 as well as the time-constant unobservables α_i will be removed from the estimation. The parish, school and teacher fixed effects will only be identified from students that move to another parish, switch schools and/or are assigned to a new teacher between grades 1 and 4. As an individual always belongs to the same family, the individual fixed effects model nests siblings fixed effects at the same time. As long as $\mathbb{E}(\Delta\varepsilon_i | \Delta(W, X_2, P, S, T)) = 0$, $\hat{\tau}$ will be unbiased.

While the individual fixed effect strategy is arguably more valid than a strategy only relying on within-school or within-teacher variation, it is not without limitations. First, it requires us to assume that the effect of absence on performance is the same in grade 1 and grade 4. *A priori*, it is unclear if these effects are the same, but in Section 3.5 we present suggestive evidence supporting this assumption.

Second, the individual (siblings) FE estimates do not recover τ if there are individual-specific (family-specific) factors of student achievement that vary over time (across siblings) and are correlated with the student's absence. An example of such a threat to the identification would be changes in class size – which may lead to increased absence and to changes in student performance. In order to address this potential issue, we include class size as a control variable. Another issue is dynamic parental investments: parents may adjust their own inputs in response to a child's absence which can lead to biased estimates. In our case, however, spillover effects of this kind can only occur in the rare cases that two siblings are observed in the same school year. Nevertheless, we cannot rule out all threats to identification. To address this, we provide in Section 3.6 a comprehensive sensitivity analysis where we bound our estimates against unobservable confounders and complement our main analysis with an instrumental variables (IV) strategy.

3.5 Estimation results

3.5.1 Short-term effects of absence in school

Table 3.5 reports the pooled estimates of the effect of days of absence in grade 1 and grade 4 on performance in the same grade.[25] The rows of the table indicate the different ways of measuring average performance. In column 1, we regress average performance on days of absence and control variables (including parish FE, but not family or individual FE). These estimates show that one additional day of absence is significantly associated with a 0.35 percent of a standard deviation (SD) decrease in average performance. Once school, teacher and siblings fixed effects are added in the model (column 2), the negative effect of days of absence becomes slightly larger in magnitude, 0.40 percent of a SD. This is in line with the hypothesis that students whose parents invest less in their children's education may also be more likely to miss school, but the fact that these estimates are so close to the OLS estimates suggests that selection on family unobservables may not be very important in this context.[26] When including individual fixed effects to the model (column 3), the point estimate returns to the magnitude of 0.33 percent of a SD.[27]

The effect of absence is of moderate size and statistically significant at the 1 percent level in all specifications. Assuming linearity, the effect of 10 days of absence – about the average in our sample – corresponds to around 3 percent of a SD in

[25] Appendix Table A3.2 reports the coefficient estimates associated with all the control variables included in these specifications.

[26] The same seems to hold true if we include school and teacher FE stepwise in Table A3.2 in the Appendix. Given the conditioning variables in the OLS model, particularly the full sets of year-of-birth, age and parish indicators, the coefficient of absence does not change noteworthily across the school, teacher and siblings FE specifications.

[27] We restrict the individual FE sample to contain individuals with siblings in the dataset. The Online Appendix shows the effect when all observations are used. The point estimates are quite similar, suggesting that siblings and singletons react in the same way to an additional day of absence.

Table 3.5: Baseline results

	(1) OLS	(2) Sibl. FE	(3) Indi. FE
Average grade points in units of SD (mean: 0, SD 1)			
Days of absence	-0.0035^{***}	-0.0040^{***}	-0.0033^{***}
	(0.0009)	(0.0011)	(0.0012)
Average grade points in units of pension 2002 (mean 2002 pension in sample: 150,816 SEK)			
Days of absence	-56.1501^{***}	-39.7389^{*}	-45.1403^{**}
	(14.7496)	(20.2815)	(20.5733)
Conventional controls			
Time-variant	✓	✓	✓
Time-invariant	✓	✓	
Fixed effects			
Socio-economics	✓	✓	
Parish	✓	✓	✓
School		✓	✓
Teacher		✓	✓
Siblings		✓	
Individual			✓
# observations	8,938	8,938	8,938
# families/individuals		1,988	4,469

Notes: Each cell states the coefficient of days of absence for a separate regression. The rows give different measures of the dependent variable average grade points. In the first row average performance over math, reading and speaking, and writing is standardized with mean 0 and standard deviation 1. The second row measures average grade points in units of pensions 2002, see the data description in the text for details. Time-variant conditional variables: grade, range of grades instructed in the same classroom, length of the school year in weeks. Time-invariant conditional variables: female, born out of wedlock, twin birth, mother employed at the time of birth, born in hospital. Socio-economics fixed effects include full sets of fixed effects for the year and month of birth, year and month interactions, age, parent's year of birth, and the family's socio-economic status based on the first-digit HISCO code of the father. Parish-clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

student performance. Interestingly, despite analyzing absence in a very different context and literally in another century, our results measured in SD units are comparable to those in Goodman (2014) and Aucejo and Romano (2016). Using a similar identification strategy with recent US data Goodman (2014) finds an effect 0.8 percent of a SD in math and English and Aucejo and Romano (2016) find effects of 0.55 percent of a SD in math and 0.29 percent in reading in their preferred specifications.

This effect size is also comparable to what has been found in the literature examining the effect of teacher quality on performance. For instance, Chetty et al. (2014a) find that a one SD increase in teacher Value Added improves students' math test scores by 0.14 SD and English test scores by 0.1 SD. Rivkin et al. (2005) find that a one-year increase in teacher experience increases student performance by up to 0.13 SD in math and 0.06 SD in reading.

An advantage of our study is that we can anchor student performance, which is measured on a somewhat arbitrary scale, to adult outcomes. In other words, we can translate the short-term effect of absence on school performance into its effect on earnings potential. This still measures the short-term effect of absence, but in a unit (SEK) that is economically more meaningful than standardized grade points. In the individual FE specification the impact of ten additional days of absence on school performance translates into a decrease in earnings potential of 451 SEK (in values of 2002). Given that the average pension is about 150,000 SEK, this effect seems rather humble.

3.5.2 Long-term effects of absence in school

Table 3.6 reports the effects on our long-term outcomes of the average number of absences across grades 1 and 4 in columns 1 (OLS) and 2 (siblings FE) as well as the effects of the number of absences in each grade by grade in columns 3 (OLS) and 4 (siblings FE), respectively.

Our estimates suggest there is a robust negative effect of absence on secondary school enrollment, but we are unable to attribute it to a certain school grade. The point estimates are negative for both grades and of at most 0.1 percentage points – which can be compared to a baseline probability of 12.5 percent. For comparison, our estimates of the effect of school performance, as measured by average grade points, on secondary schooling enrollment range between 0.05 and 0.13 (results available upon request). Multiplying this estimate with the estimated effect of absence on school performance of around -0.003 (cf. Table 3.5 above), we would expect an effect of absence on enrollment of between -0.0003 and -0.0005. Our estimates of the impact of absence on enrollment are in general larger and typically twice as large as this indirect estimate, even though the differences between the two are insignificant.

Turning to employment, the results are mixed. When it comes to early-career employment (as measured in the 1960 Census when individuals are between 25 and 30 years old), our results suggest that there is a negative relationship driven by absences in grade 1. Indeed, when siblings FE are included in the model, ten days of absence in grade 1 leads to a decrease in the probability of being employed at ages 25–30 by 2.6 percentage points, and this estimate is strongly significant. Relative to the average employment at the time (65 percent), this corresponds to a 4 percent reduction, which is a rather large effect. On the other hand, for employment in 1970 (at ages 35–40) the estimates again have the expected negative sign, but the point estimates are smaller and noisier. The 95 percent confidence interval for the effect of ten days of absence (across both grades) on employment in 1970 ranges between -0.017 and -0.002, suggesting that there may still be a negative effect of absences on employment at 35–40, but this effect is unlikely to be large and is rather getting smaller over time.

With respect to labor market earnings, the association between income and absence is negative for most specifications. For 1970 labor market income the estimates are not statistically significant and rather small in size – in most specifications ten days of absence correspond to less than 1 percent of the average income. Besides the long time horizon (27 to 32 years after grade 1 and 24 to 29 years after grade 4), the lack of a relationship between absence and income at ages 35–40 may also be due to the Swedish wage structure being extremely compressed at this time, so that individual productivity had a very limited impact on earnings (Bhalotra et al., 2016). For 2002 pensions the pooled and fourth-grade estimations are negative (while the coefficients for grade 1 absences are close to zero). The OLS estimate of the effect of grade 4 absence on pensions 2002 is significant at the 10 percent level. When adding siblings fixed effects, the point estimate increases in magnitude, but becomes less precise.[28] Finally, for mortality the coefficients alternate around zero and do not exceed 0.01 percentage points. This indicates that there is no effect on mortality.[29]

Overall, the results presented so far suggest that absence in elementary school has a robust negative impact on short-term performance of small but non-negligible magnitude. Absences in elementary school also have a detrimental impact on early-career employment, but this effect fades out with increased labor market experience. Broadly speaking, the findings are very much in line with those of Pischke (2007) and Dustmann et al. (forthcoming) who find that initial differences in the quantity and quality of schooling have no long-lasting labor market effects. The fact that the effect of absence on labor market outcomes are significant early in the career but not later on underscores the value of having access to data on outcomes at different points of the life-cycle in order to get an unbiased perspective of the full impacts of absence.

That being said, it is possible that the above results mask some heterogeneity between certain groups or for specific causes of absences (e.g., sickness absences)

[28] In the baseline specification we only consider income and pension values >0. The Online Appendix gives the results for alternative income measures. The alternative measures do not change our interpretation of the baseline results.

[29] The missing association between absence and longevity indicates that sample attrition due to selective mortality is not a major concern when interpreting the other long-term effects.

Table 3.6: Long-term effect of absence in school

	(1)	(2)	(3)	(4)
	Pooled effect		Separate effects	
	OLS	Sibl. FE	OLS	Sibl. FE
More than Folkskola *(1=yes)*				
Total abs. (avg. both grades)	−0.0012	−0.0013		
	(0.0010)	(0.0009)		
Total absence in grade 1			−0.0003	−0.0011*
			(0.0006)	(0.0007)
Total absence in grade 4			−0.0010*	−0.0002
			(0.0006)	(0.0005)
Employment 1960 (1=yes)				
Total abs. (avg. both grades)	−0.0016**	−0.0016*		
	(0.0007)	(0.0009)		
Total absence in grade 1			−0.0014**	−0.0026***
			(0.0005)	(0.0006)
Total absence in grade 4			−0.0003	0.0008
			(0.0007)	(0.0008)
Employment 1970 (1=yes)				
Total abs. (avg. both grades)	−0.0010*	−0.0020		
	(0.0006)	(0.0015)		
Total absence in grade 1			−0.0006	−0.0008
			(0.0005)	(0.0009)
Total absence in grade 4			−0.0004	−0.0012
			(0.0006)	(0.0009)
Labor market income 1970				
Total abs. (avg. both grades)	−54.4802	−13.9569		
	(36.9846)	(57.2241)		
Total absence in grade 1			−18.5468	−1.3469
			(21.2133)	(40.3369)
Total absence in grade 4			−36.0149	−12.2763
			(23.2125)	(28.5027)
Pensions 2002				
Total abs. (avg. both grades)	−124.6255	−186.0209		
	(126.1594)	(264.0324)		
Total absence in grade 1			0.1536	46.2308
			(125.6788)	(169.5014)
Total absence in grade 4			−120.9218*	−215.0359
			(72.6683)	(147.7769)
Passed away before the age of 70 (1=yes)				
Total abs. (avg. both grades)	−0.0001	0.0011		
	(0.0008)	(0.0008)		
Total absence in grade 1			0.0002	0.0008
			(0.0005)	(0.0006)
Total absence in grade 4			−0.0003	0.0003
			(0.0006)	(0.0006)

Notes: Number of observations: More than *Folkskola* 3,019 (in 1,373 families), employment 1960 3,902 (1,750), employment 1970 4,469 (1,988), income 1970 2,137 (985), pensions 2002 2,363 (1,080), passed away before age 70 4,469 (1,988). Parish-clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

or that absence only has adverse effects if it occurs with large frequency (non-linearities). We now explore these different margins.

3.5.3 Heterogeneity

Subgroup analysis Table 3.7 reports the estimates for academic performance in elementary school where we allow the impact of the total number of days of absence to vary between males and females, as well as between children of agricultural and non-agricultural workers. For each panel, the first row reports the main effect of the number of days of absence, while the second row reports the coefficient on the interaction. Overall, we find little evidence of heterogeneous impacts. With respect to gender heterogeneity, the individual FE model estimates suggest that the effect of absences may be worse for men than for women's achievement, but the difference between the two groups is statistically not different from zero.[30] While absence is more strongly negatively correlated with the performance of children of agricultural workers than with the that of other children, differences in the impact of absences wash away once we account for unobserved heterogeneity at the family level. This lack of a clear effect heterogeneity along the socio-economic status is in line with Goodman (2014), while Aucejo and Romano (2016) find evidence of some heterogeneous impacts between students of different abilities. This result also underlines the educational and societal context of the analysis. In the setting we investigate, textbooks were, for instance, provided by the parish if families could not afford them otherwise.

Table 3.8 reports the results of a similar subgroup analysis for our long-term outcomes. As the difference in labor market participation and earnings between men and women is quite substantial in the time under review (see Table 3.3), interacting pooled days of absence with the female indicator yields rather remarkable findings. Although not statistically significant, ten days of absence associate with a decrease in probability of being employed in 1960 for women by 6.6 percent, while the corresponding number males is less than one-tenth.[31] The long-term effect of absence on pensions in 2002 seems to be stronger for females as well. Looking at the social gradient along the father's occupation indicates some heterogeneity on this dimension, where the effect of absences on 1970 income is significantly larger for children of agricultural workers than other children.

[30] One should keep in mind the gender difference is only identified through a variation in days of absence in the individual FE model. Similarly, the interaction term between father's occupation and absence in the individual FE model is only identified through variation in absence.

[31] For women the effect is $\left((10 \text{ days} \times -0.0004) + (10 \text{ days} \times -0.0020)\right)/0.3651$ baseline probability $= -0.0657 \approx -6.6$ percent and for men $(10 \times -0.0004)/0.9396 = -0.0043 \approx -0.4$ percent.

Table 3.7: Heterogeneity in the short-term effects by subgroup

	(1) OLS	(2) Sibl. FE	(3) Indi. FE
Gender			
Absence	-0.0028^{*}	-0.0037^{**}	-0.0043^{***}
	(0.0014)	(0.0016)	(0.0013)
Absence×female	-0.0013	-0.0004	0.0017
	(0.0016)	(0.0018)	(0.0016)
Father's occupation			
Absence	-0.0024^{**}	-0.0041^{***}	-0.0031^{**}
	(0.0011)	(0.0012)	(0.0016)
Absence×agri. worker	-0.0027^{*}	0.0002	-0.0005
	(0.0015)	(0.0017)	(0.0026)
Grade			
Absence	-0.0040^{***}	-0.0047^{***}	
	(0.0014)	(0.0014)	
Absence×grade 1	0.0009	0.0014	
	(0.0016)	(0.0017)	

Notes: Each panel states the coefficient of total days of absence as well as of an interaction between total days of absence and the subgroup indicator. 8,938 observations. Control variables as in the baseline specification. Parish-clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Grade-specific effects An assumption underlying our individual FE strategy is that the impact of absence in grade 1 and in grade 4 is the same. While we cannot test for this assumption in the context of the individual FE model, we present suggestive evidence that this assumption holds by estimating a model where we allow for different effects of absence in grade 1 and grade 4 and control for teacher, school and siblings FE. The results presented in the bottom panel of Table 3.7 indicate that the effect of absence in grade 1 cannot be statistically distinguished from the effect of absence in grade 4 on any academic performance measure.

3.5.4 Sickness vs. non-sickness absences

Our data allows us to distinguish between absences due to sickness and absences due to other reasons. Table 3.9 reports the estimates of the short-term effects of days of absences on attainment, when we allow the effect to be different for absences due to sickness and absences due to other reasons. At the bottom of the table, we report in brackets the p-value of a F-test that the two coefficients of interest are equal to each other. Comparing the effect of sickness absence with non-

Table 3.8: Heterogeneity in the long-term effects by subgroup

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent variable					
	> *Folk-skola*	Empl. 1960	Empl. 1970	Income 1970	Pensions 2002	Passed away≤70
Gender						
Absence	−0.0023	−0.0004	−0.0021	−8.9994	−28.2702	0.0011
	(0.0016)	(0.0011)	(0.0020)	(56.4558)	(327.9810)	(0.0016)
Abs.×female	0.0016	−0.0020	0.0002	−10.0480	−277.5613	0.0001
	(0.0017)	(0.0017)	(0.0018)	(64.3459)	(287.1122)	(0.0017)
Father's occupation						
Absence	−0.0008	-0.0021^{**}	−0.0022	70.2782	−315.6315	0.0015
	(0.0013)	(0.0010)	(0.0016)	(60.2351)	(280.5521)	(0.0012)
Abs.×agri.	−0.0011	0.0011	0.0005	-192.8888^{***}	328.3218	−0.0010
	(0.0014)	(0.0017)	(0.0033)	(54.5382)	(357.2552)	(0.0019)

Notes: Each panel states the coefficient of total days of absence (average over grades 1 and 4) as well as of an interaction between total days of absence and the subgroup indicator. Number of observations: More than *Folkskola* 3,087 (in 1,396 families), employment 1960 3,904 (1,751), employment 1970 4,471 (1,989), income 1970 2,139 (986), pensions 2002 2,365 (1,081), passed away before age 70 4,471 (1,989). Dependent variables defined as in the baseline long-term results. Parish-clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

sickness absence across the different specifications (including school and teacher FE) reveals an interesting pattern. In the OLS model and the individual FE model (columns 1 and 5 in Table 3.9), the estimated effects of both types of absence are similar in magnitude and statistically undistinguishable from each other. If we only compare students who have the same teacher (and, thereby, are in the same school and generally in the same class) to each other in column 3, the association between non-sickness absence and performance is more than twice as strong as the association between sickness absence and performance (the coefficients differ at the 10 percent level). This pattern suggests that the association between non-sickness absence and performance is driven by family-level or individual factors.

A candidate for such a factor may be behavioral issues that cause truancy (which is reported as non-sickness absence). Accounting for unobserved behavioral problems either through within-family or within-individual comparison yields rather similar results – at least when compared to the effect in the teacher FE specification. If there are no time-varying behavioral problems (or other time-varying unobservable confounders of non-sickness absence), the coefficient of days of non-sickness absence in column 5 may as well be a reasonable approximation for the effect absence days w while holding the number of days in illness s constant. Taking up the discussion of what an ideal experiment may look like in Section 3.4.1,

Table 3.9: Short-term effects – total absence vs. sickness absence

	(1)	(2)	(3)	(4)	(5)
	OLS	School FE	Teacher FE	Sibl. FE	Indi. FE
Average grade points in units of SD					
Days of sickness absence	-0.0037^{***}	-0.0042^{***}	-0.0036^{***}	-0.0044^{***}	-0.0034^{**}
	(0.0009)	(0.0011)	(0.0010)	(0.0011)	(0.0014)
Days of non-sickness absence	-0.0027	-0.0039^{**}	-0.0075^{***}	-0.0017	-0.0031
	(0.0018)	(0.0017)	(0.0022)	(0.0024)	(0.0025)
	[0.5789]	[0.8444]	[0.0793]	[0.2651]	[0.9319]
# observations	8,938	8,938	8,938	8,938	8,938
# individuals/families		749	1,259	1,988	4,469

Notes: See note to the baseline results table. The brackets at the bottom of the table give the p-value of a F-test of equality of the coefficients of sickness and non-sickness absence. Parish-clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

the similar coefficients of sickness and non-sickness absence in column 5 (the p-value is close to 1) thus lends support to the notion that the reduced performance associated with absence is driven by the absence in itself, and not by the student's health. The fact that only the impact of sickness absence is significantly different from zero might be driven by the fact that there is a lot more variation in this variable than there is in days of non-sickness absence.[32]

A similar exercise for our long-term outcomes reveals very comparable patterns. Table 3.10 reports the OLS (first panel) and siblings FE (second panel) coefficients of sickness absence and non-sickness absence for grades 1 and 4. Interestingly, the negative association between total absence and having more than *Folkskola* education seems driven by non-sickness absence – even though sickness is the main cause for overall absence. This supports the view that non-sickness absence is driven by behavioral problems, which teachers account for when recommending students for *Realskola* enrollment. Once behavioral problems on family level are partialled out (through the inclusion of siblings FE), the negative association between absences with education vanishes.

Both the OLS and siblings FE specifications point to the fact that both types of absences significantly decrease early employment. Although the effect of non-sickness absences is more negative than that of sickness absences, we cannot distinguish the two from each other at conventional levels of significance. Regarding

[32] Aucejo and Romano (2016) distinguish between excused and unexcused absence. They do not compare the effect size across different FE models, but in line with our findings, their estimated effect of unexcused absence exceeds the effect of excused absence.

Table 3.10: Long-term effects – total absence vs. sickness absence

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent variable					
	$>$ *Folk-skola*	Empl. 1960	Empl. 1970	Income 1970	Pensions 2002	Passed away$\leq$70
OLS estimation						
Sick. abs. gr. 1	0.0002	-0.0011^{*}	−0.0005	−16.0216	−24.3926	−0.0000
	(0.0007)	(0.0007)	(0.0006)	(17.1981)	(139.5187)	(0.0005)
Non-sick. abs. gr. 1	-0.0025^{**}	-0.0028^{*}	−0.0013	−33.0191	158.5347	0.0014
	(0.0010)	(0.0014)	(0.0015)	(78.3599)	(380.7130)	(0.0013)
	[0.0289]	[0.3363]	[0.6589]	[0.8195]	[0.6619]	[0.3175]
Sick. abs. gr. 4	−0.0003	0.0002	−0.0005	−36.6376	−88.3797	0.0004
	(0.0007)	(0.0007)	(0.0005)	(27.8351)	(91.9464)	(0.0006)
Non-sick. abs. gr. 4	-0.0035^{***}	−0.0020	−0.0002	−32.5383	−255.0109	-0.0031^{***}
	(0.0011)	(0.0013)	(0.0015)	(50.2889)	(238.1689)	(0.0009)
	[0.0253]	[0.1076]	[0.8157]	[0.9462]	[0.5525]	[0.0004]
	{0.0280}	{0.1316}	{0.9465}	{0.8493}	{0.8239}	{0.0012}
Siblings FE						
Sick. abs. gr. 1	−0.0013	-0.0022^{***}	−0.0008	−25.7318	−15.8283	0.0004
	(0.0008)	(0.0008)	(0.0010)	(42.5099)	(198.1435)	(0.0006)
Non-sick. abs. gr. 1	−0.0004	-0.0055^{**}	−0.0009	138.0634	443.4685	0.0032^{**}
	(0.0010)	(0.0022)	(0.0024)	(101.3751)	(397.9748)	(0.0016)
	[0.5476]	[0.2390]	[0.9675]	[0.1317]	[0.3465]	[0.0775]
Sick. abs. gr. 4	−0.0006	0.0011	−0.0011	−39.8570	−272.8657	0.0008
	(0.0007)	(0.0008)	(0.0009)	(25.4238)	(174.1033)	(0.0007)
Non-sick. abs. gr. 4	0.0016^{*}	−0.0009	−0.0017	81.6146	56.3521	−0.0018
	(0.0010)	(0.0027)	(0.0027)	(65.2114)	(249.8072)	(0.0015)
	[0.1149]	[0.4702]	[0.8370]	[0.0692]	[0.3066]	[0.1160]
	{0.1381}	{0.0003}	{0.9703}	{0.1736}	{0.2434}	{0.2317}

Notes: Number of observations: More than *Folkskola* 3,087 (in 1,396 families), employment 1960 3,904 (1,751), employment 1970 4,471 (1,989), income 1970 2,139 (986), pensions 2002 2,365 (1,081), passed away before age 70 4,471 (1,989). The brackets give the p-value of a F-test of equality of the coefficients of sickness and non-sickness absence in the respective grade. The braces state the p-value of a F-test of equality of all four coefficients. Parish-clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

employment in 1970, the absence coefficients are negative but neither statistically different from each other nor from zero. The income measures are only negatively associated with non-sickness absence in the OLS estimates. To the extent that non-sickness absence reflects a bolder behavior, these results that suggest there might be a wage premium for being "pushy."

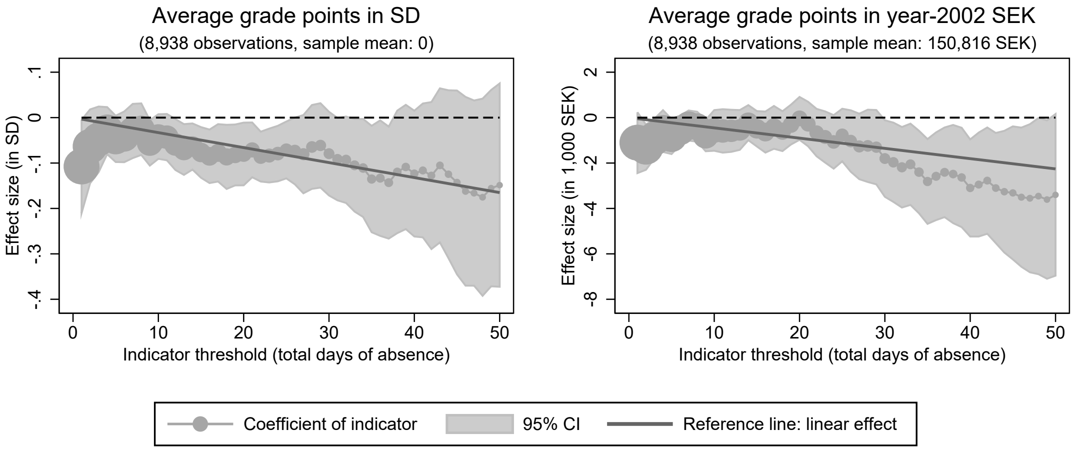


Figure 3.7: Non-linearities in the short-term effect of absence for different threshold values

Notes: This graph plots the coefficient of a regression of performance in school on a binary indicator for total days of absence. In the left plot performance is measured in units of standard deviations and in the right plot in year-2002 SEK. The indicator threshold is given on the x-axis. The size of the coefficient markers depicted through orange dots is proportional to the number of observations for that the indicator is 1. Out of the 8,942 student-grade observations 95 per cent have at least one day of absence (leftmost orange dot in the plots) and 2 per cent missed 50 or more days in one school year (rightmost orange dot in the plots). The gray area indicates the 95 per cent significance band of the coefficient estimates. The red line depicts the linear effect of an additional day of absence taken from the baseline model.

3.5.5 Non-linearities

While a student may be able to compensate a few days of absence, this may not be possible for a longer period of absence. This would result in a non-linear relationship between absence and educational performance. To investigate the presence of non-linearities we run the individual FE specification (similar to column 3 in Table 3.5) where we define the treatment as an indicator for whether the number of days of absence exceeds a certain threshold, where we vary this threshold between 1 and 50. The dots in Figure 3.7 give the coefficients associated with such indicator along the different threshold values on the x-axis.[33] The size of the dots indicates the relative number of observations for which the indicator is 1. While 95 percent out of the c.a. 9,000 student-grade observations exhibit at least one day of absence, less than 2 percent of the student-grade observations have 50 or more days. Naturally, the 95 percent confidence interval of the point estimators – depicted in gray – increases with the threshold value. The red reference line depicts

[33] Because the number of students that miss a large number of days in one school year is often rather low (e.g., only 12 students miss exactly 45 days), it is not meaningful to regress performance on a full set of binary indicators for each number of days in a single regression. The Online Appendix gives the results of a regression using indicator variables that bin days of absence. The coefficients of the indicator variables lie around the linear effect reference line.

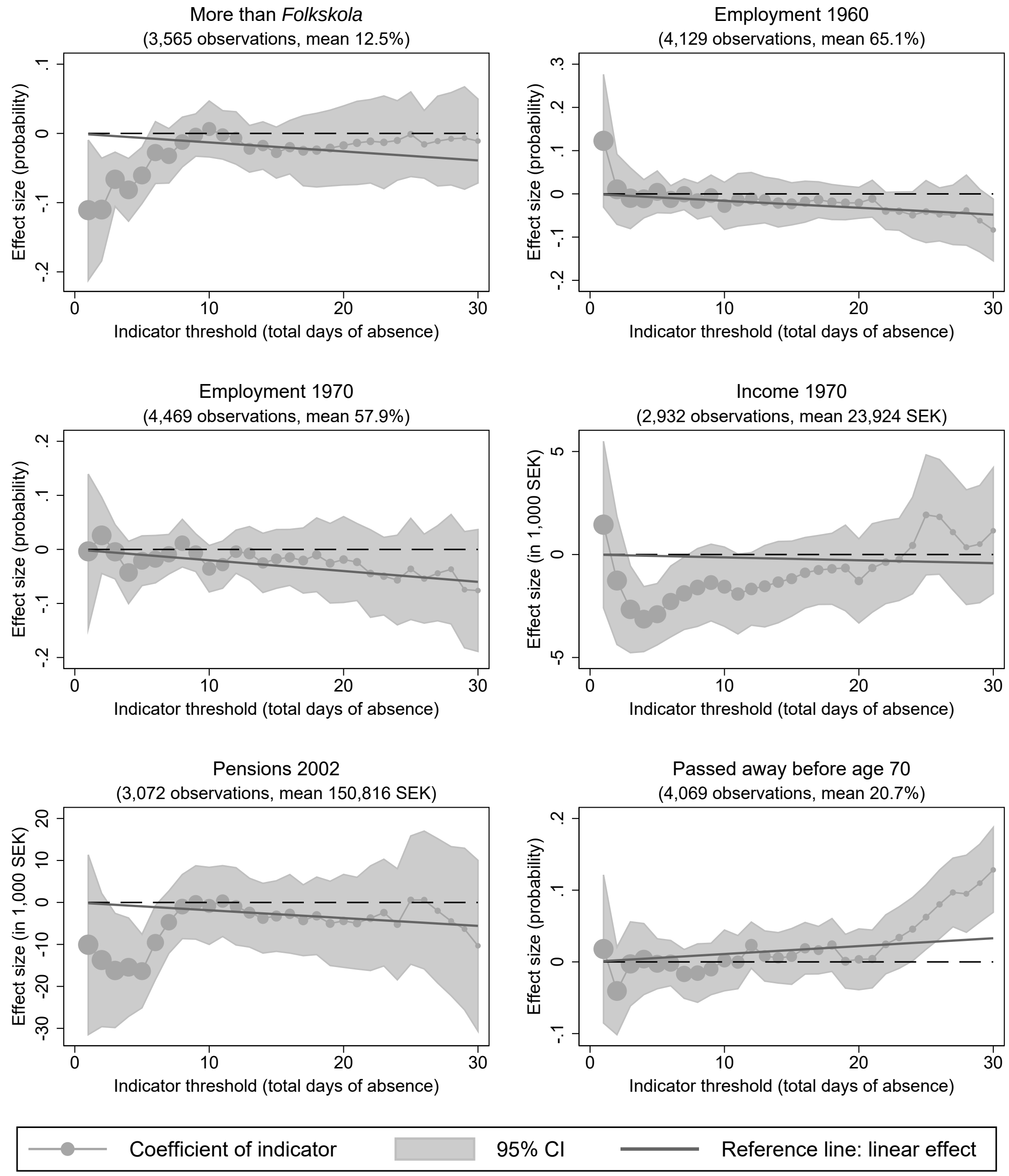


Figure 3.8: Non-linearities in the effect of average days of absence in both grades on long-term outcomes using individual fixed effects

Notes: This graph plots the coefficient of a regression of the long-term on a binary indicator for average days of total absence over grades 1 and 4. The indicator threshold is given on the *x*-axis. The size of the orange coefficient plot is proportional to the number of observations for that the indicator is 1. The gray area indicates the significance band of the coefficient estimates. The red line depicts the linear effect of an additional day of absence in the baseline specification.

the linear effect of the baseline model as of Table 3.5 multiplied by the number of absent days.

A comparison of the orange and red lines indicates that the per-day effect in SD (left plot) of the non-linear estimations using the binary indicators does not substantially differ from the linear effect. If short-term performance in school is measured in SEK in the year 2002, the non-linearly estimated effect exceeds the linear effect for more than 30 days of sickness absence. That is, only if an individual is absent more than 30 days in one school year, the effect of absence increases disproportionately. Regardless of how we measure the outcome variable, the deviation from the linear trend is never significantly different. Given that less than 6 percent of all student-grade observations have 30 or more days of absence, non-linearities do not seem to play an important role. This finding is in line with Aucejo and Romano (2016) who neither find evidence of non-linearities.

Figure 3.8 reports the results of a similar exercise for our long-term outcomes.[34] Overall, there is no strong evidence that absence has non-linear impacts on final educational achievement, adult employment, income or mortality.

3.6 Sensitivity analysis

The estimates in Tables 3.5 and 3.9 are remarkably stable across specifications, suggesting that selection on unobservables into absence may not be a salient feature. However, the FE estimates could still be biased if there are unobservable factors that correlate with individual absence and student outcomes *and* that vary over time or across siblings. To address this caveat and show the robustness of our results, we implement a bounding approach where we bound the influence of omitted variables. We also present results from an instrumental variables approach using local weather shocks as instrument for absences.

3.6.1 Bounds

We employ the bounding approach suggested by Oster (forthcoming), building on the idea of Altonji et al. (2005). Our goal is to bound the effect of absence assuming that the selection on unobservables is as strong as the selection on observables. We consider the case where the selection on unobservables is in the same or the opposite direction as the selection on observables, thus allowing the true

[34] Because of the fewer observations for the long-term outcomes, we only run the absence indicator up to the threshold of 30 or more days of absence.

effect to be overestimated or underestimated. The exercise is only helpful if the observables are informative with respect to the selection, so we control for a large array of control variables and the full set of siblings fixed effects. This removes factors such as constant family resources, parental preferences, and genetic endowment that are likely negatively correlated with absence and positively correlated with performance. As omitting these factors would likely cause an upward bias that challenges our implications, the bounding approach seems particularly useful in the application at hand.

The starting point is to compare the coefficient of absence in the baseline model ($\tilde{\beta}$), with the coefficient of absence in a simple linear regression of the dependent variable on absence and an intercept ($\dot{\beta}$). Formally, the bound around the coefficient of absence β^* is:[35]

$$\beta^* \approx \tilde{\beta} - \delta(\dot{\beta} - \tilde{\beta}) \frac{R_{max} - \tilde{R}}{\tilde{R} - \dot{R}},$$

where the degree of proportionality of selection on observables to selection on unobservables, δ, is either set to 1 (unobservable selection goes into the same direction) or -1 (unobservable selection is into the adverse direction). In a second step, the movement in the coefficient of absence, $\dot{\beta} - \tilde{\beta}$, is re-scaled by the movement in the R^2 relative to the potential change in the R^2 (where $\tilde{R}$ and $\dot{R}$ denote the R^2 of the baseline model and the simple regression, respectively, and R_{max} denotes the highest possible value of the R^2).[36]

Table 3.11 shows the bound estimates for the short- and long-term effects of days of absence. For the baseline model estimate $\tilde{\beta}$ we take the siblings FE specification in order to be able to calculate the bounds separately for each grade.[37]

For the short-term effect on performance in grade 1 the estimated β^* for $\delta = 1$ is -0.0078. Thus, if unobservable selection is proportional to and goes in the same direction as the observable selection an additional day of absence associates with a decrease in performance by 0.78 percent of a SD. If instead assuming the same amount of selection but in the opposite direction ($\delta = -1$), the upper bound is -0.0017. Even if our baseline model fails to account for selection on unobservables that affect the outcome as strongly as family-level time-invariant characteristics,

[35] This expression is only an approximation, see Oster (forthcoming) for the exact calculation. To calculate the bounds we use the Stata ado-file `psacalc` provided online by Emily Oster. All errors are our own responsibility.

[36] Following Hener et al. (2016) we consider as $R_{max} = \min(2.2 \times \tilde{R}, 1)$.

[37] To estimate $\tilde{\beta}$ we stratify the sample by grade and run separate regressions. For the short-term estimates for grade heterogeneity in Table 3.7 we use instead an interaction term, for the long-term estimates by grade in Table 3.6 we regress the outcome on absence in both grades in the same regression. Thereby, the $\tilde{\beta}$ estimates used here are not exactly the same as in the analysis before, but they are quite similar, see column 2 of Table 3.11.

the effect of ten days of absence would be a decrease performance by 1.7 percent of a SD. For the short-term effects for absence in the fourth grade the bounds have the same sign for $\delta = 1$ and $\delta = -1$ and both bounds are barely distinguishable from the baseline estimate. This strongly supports our interpretation that the baseline results do not indicate an omitted variable bias.

For all long-term outcome variables all bounds for the adverse selection case point towards the same direction as in the baseline model (absence in school is harmful). For $\delta = 1$, some of the bound estimates point to a positive effect of absence. However, these are the cases when conditioning on the observables reduces the absence coefficient in absolute terms and omitting the variables causes an underestimation. The bounds, again, reinforce our belief that serious omitted variable bias is unlikely.

Table 3.11: Coefficient bounds for $\delta = 1$ and $\delta = -1$ selection

		(1)	(2)	(3)	(4)
Dependent variable	Absence	Restricted model $\dot{\beta}$	Baseline model $\tilde{\beta}$	Bound β^* for $\delta = 1$	Bound β^* for $\delta = -1$
Short-term outcome					
Average performance	Grade 1	−0.0032 (0.0007) [0.0029]	−0.0028 (0.0015) [0.3721]	−0.00777	−0.00191
	Grade 4	−0.0032 (0.0019) [0.0031]	−0.0046 (0.0018) [0.3940]	−0.00496	−0.00456
Long-term outcomes					
More than *Folkskola*	Grade 1	−0.0004 (0.0005) [0.0002]	−0.0012 (0.0007) [0.1562]	−0.0012	−0.0011
	Grade 4	−0.0010 (0.0006) [0.0013]	−0.0003 (0.0006) [0.1548]	−0.0000	−0.0004
Employment 1960	Grade 1	−0.0014 (0.0006) [0.0012]	−0.0026 (0.0006) [0.4712]	−0.0017	−0.0032
	Grade 4	−0.0011 (0.0008) [0.0007]	0.0007 (0.0008) [0.4688]	0.0022	−0.0003
Employment 1970	Grade 1	−0.0008 (0.0006) [0.0004]	−0.0008 (0.0010) [0.2104]	−0.0004	−0.0011
	Grade 4	−0.0007 (0.0007)	−0.0012 (0.0009)	−0.0010	−0.0014

Continued on next page

Table 3.11 – *continued*

		(1)	(2)	(3)	(4)
Dependent variable	Absence	Restricted model $\dot{\beta}$	Baseline model $\tilde{\beta}$	Bound β^* for $\delta = 1$	Bound β^* for $\delta = -1$
		[0.0003]	[0.2108]		
Income 1970	Grade 1	-22.6056	-1.6248	55.9162	-28.4313
		(16.4742)	(39.6379)		
		[0.0004]	[0.4697]		
	Grade 4	-48.0319	-12.3049	95.0982	-60.4634
		(23.0625)	(28.2772)		
		[0.0015]	[0.4698]		
Pensions 2002	Grade 1	-191.7764	35.9554	171.1624	-20.6282
		(103.5173)	(170.6840)		
		[0.0012]	[0.3100]		
	Grade 4	-258.6889	-213.1113	171.4723	-401.7276
		(79.8756)	(148.5325)		
		[0.0022]	[0.3111]		
Passed away $\leq$ age 70	Grade 1	0.0004	0.0008	0.0002	0.0013
		(0.0004)	(0.0006)		
		[0.0001]	[0.1456]		
	Grade 4	0.0001	0.0004	-0.0011	0.0014
		(0.0006)	(0.0006)		
		[0.0000]	[0.1453]		

Notes: Column 1 gives the coefficient of absence in the restricted model where the outcome variable is regressed on absence and an intercept. The unrestricted model in column 2 is similar to the baseline results for the short- and long-term effects. Column 3 states the bound for proportional unobservable selection that goes in the same direction as observable selection. In column 4 the unobserved selection is proportional but in an adverse direction. Parish-clustered standard errors for the regression coefficients in parentheses. The resulting R^2s are given in brackets.

3.6.2 Instrumental variables estimates

Short-term IV estimates Next, we follow Goodman (2014) and Aucejo and Romano (2016) and employ an instrumental variables strategy. In line with Goodman (2014), Marcotte (2007) and Marcotte and Hemelt (2008), we exploit changes in weather conditions using local meteorological time series data on the temperature collected from Matsuura and Willmott (2012).[38] In the simplest feasible specification we instrument yearly absence using a count variable that gives the number of "benign" months in the school year. Following Bruckner et al. (2014),

[38] The data includes monthly temperature information interpolated to a 0.5 degree by 0.5 degree latitude/longitude grid and we assign each school parish to the closest grid node using latitude and longitude information on the parish center as well as the grid node centroid.

we define a month in a certain parish as benign if the average temperature in this parish is within the upper quartile of the temperatures measured in all parishes and all years for this month. The Online Appendix Figure O4 shows the average temperature for the school years 1936/37–1947/48 across regions and the spatial variation in the average number of benign months.[39]

The exclusion restriction of the instrument is that the number of benign months only affects an individual's performance in school through days of absence. As pointed out by Goodman (2014) weather may affect both individual absence and school closures. If the number of benign months correlates with weather-related school closures, this would violate the exclusion restriction. While we have no school-level information on closures, the context of our analysis makes weather-related school closures unlikely. The northern part of Sweden is often covered with snow from October to April, and it seems fair to assume that schools, parents and students were well-adapted to the situation. Moreover, schools were usually at walking distance from a student's home and historical sources do not mention snow-related school closures. Another concern is that weather-related teacher and classmate absence affects performance. There is no direct information on teacher absence, but teachers generally lived in one part of the school building, and students were provided with a substitute teacher if the teacher was sick.

We consider the two-stage least square (2SLS) estimator. To identify the impact of weather on days of absence, we run a first-stage regression of days of absence W_{ig} on the number of benign months experienced by student i when he or she is in grade g. In the second stage, we regress the short-term outcome measure on fitted days of absence:

$$\begin{aligned} W_{ig} &= b_0 + b_1 \,\text{\#benign months}_{ig} + b_2 X_{1i} + b_3 X_{2ig} + P_{ig} + v_{ig} \quad \text{and} \\ y_{ig} &= \alpha + \tau \widehat{W}_{ig} + \beta_1 X_{1i} + \beta_2 X_{2ig} + P_{ig} + \varepsilon_{ig} \quad \text{for } g = 1, 4, \end{aligned}$$

where b_1 is the effect of the number of benign months on days of absence.[40] The instrument needs to satisfy the independence assumption that $Y_i(w, s) \perp\!\!\!\perp$ $\text{\#benign months}_{ig} \mid X_{1i}, X_{2ig}, P_{ig}$. Given the inclusion of parish fixed effects that

[39] Naturally the parishes in the north with very low average temperatures are also the ones that barely experience months with benign temperatures. This has limited consequences for our analysis. Even in the north there is variation in the number of benign months and most parishes in our sample are located in the more populated southern part, where the variation in benign months is higher.

[40] In order to keep as much variation as possible in the first stage we only rely on parish fixed effects. Moreover, we do not restrict our sample to siblings observed in both grades as neither individual nor siblings FE seem necessary for the exclusion restriction to hold. This allows us to use all individuals for that we have grade-1 and/or grade-4 information.

account for more educated families living into certain regions (e.g., warmer, more industrial regions), we believe this assumption is likely to hold.

Additionally we consider a two-sample two-stage least square (TS2SLS) approach.[41] The universe of the weather and absence information across grades constitute the first-stage sample that allows a pooled estimation of the first-stage effect. We use this first-stage sample to estimate the effect of the number of benign months on days of absence in either grade. In the reduced-form sample, we then predict the days of absence based on the effect of benign months in the first-stage sample. Instead of using the effect of weather in grade 1 or grade 4 on absence in the same grade to predict absence in this grade, we thus use the coefficient of the pooled estimation to fit absence in either grade.[42] To get standard errors of the second-stage TS2SLS estimates we apply the Delta method.

Regardless of the IV estimator, two remarks should be kept in mind. First, our weather instrument is more plausibly providing exogenous variation in absences due to sickness than to other reasons such as truancy. To the extent that the effect of absence is heterogeneous across individuals missing school due to sickness or non-sickness reasons, the IV estimates of τ will be the LATE for individuals whose absence is affected by weather. Second, as remarked by Goodman: "estimates of the impact of weather on student achievement can be quite sensitive to the chosen definition of the instrument" (see Goodman, 2014, p.15). The effect of temperature on days of absence seems to be non-linear, and mapping monthly information on weather conditions in days of absence per school year without losing the variation that triggers absence is not trivial. For this reason, we think our IV estimates provide complementary (rather than superior) evidence to our individual fixed effects estimates.

Table 3.12 shows the IV results for performance. Columns 1 and 2 give the 2SLS effects for grade 1 and 4, and columns 3 and 4 state the corresponding TS2SLS effects when the first stage is estimated jointly. Regardless how the first stage is estimated, an additional month with benign temperatures reduces, on average, the number of days of absence by about 0.35. The assumption that the first-stage

[41] When implementing 2SLS, we can pool grade-1 and grade-4 information to receive average first- and second-stage effects or conduct the analysis separately by grade. Pooling has the advantage that we use more information at once and receive more precise estimates. Yet, a separate estimation allows the first- and second-stage effects to differ across grades. To circumvent this trade-off we consider the TS2SLS approach.

[42] The underlying assumption is that the first-stage effect does not differ across grades. Comparing the separate 2SLS first-stage estimates with the jointly estimated TS2SLS first-stage estimate allows us to investigate the plausibility of this assumption.

Table 3.12: Short-term IV results for total days of absence

	(1)	(2)	(3)	(4)
	Specification			
	2SLS		TS2SLS	
	grade 1	grade 4	grade 1	grade 4
First-stage effect of weather on absence				
# benign months	-0.3610^{***}	-0.3553^{***}	-0.3494^{***}	-0.3494^{***}
	(0.0080)	(0.1632)	(0.0701)	(0.0701)
Second-stage effect of fitted absence on performance				
Absence	-0.0038	-0.0220	-0.0039	-0.0252
	(0.0222)	(0.0252)	(0.0653)	(0.0654)
# observations first stage	13,884	14,152	28,036	28,036
# observations second stage	13,884	14,152	13,884	14,152
F-statistic instrument	7.91	4.74	24.86	24.86

Notes: Each cell (but the first stage in column 3 and 4) states a separate regression. In the two-sample two-stage least square (TS2SLS) specification the first stage is estimated jointly over both grades. All standard errors in parentheses. The standard errors of the second stage in columns 1 and 2 are Stata's default standard errors clustered on parish level. In columns 3 and 4 the standard errors of the second stage are calculated using the Delta method and clustered on parish level. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

relationship is the same across grade 1 and grade 4 seems justified. Noteworthily, only the F-statistic of the pooled specification exceeds the Staiger and Stock (1997) rule-of-thumb value of 10. Turning to the structural estimates, all coefficients indicate a negative relationship between absence and performance in school. While the second-stage coefficients exhibit a large difference across grades, the decision of the IV estimator (2SLS vs. TS2SLS) makes no noteworthy difference. All second-stage estimates have rather large standard errors and we cannot reject that the coefficient is zero. Still, the point estimates are negative and lower than the baseline result. In fourth grade, an additional day of absence decreases performance by 2 percent of a SD.[43] Using a similar strategy, the IV results of Goodman (2014) are also up to 2 percent of a SD.

All in all, we interpret the short-term IV results as confirmation of the baseline estimates. Given the limited external validity of the LATE estimates and the assignment mechanism, the IV approach is not a silver bullet to address identification. Still, an additional day of absence seems negatively related to academic performance.

[43] A reason for the big effect size may be that the IV strategy yields LATE and not the population average treatment effect.

Long-term IV estimates For the long-term outcomes we extend the TS2SLS approach to all seven grades of compulsory education.[44] As before, we estimate the first-stage effect of benign months on absence in grades 1 and 4:

$$W_{ig} \quad = \quad b_0 + b_1 \,\text{\#benign months}_{ig} + b_2 X_{1i} + P_i + v_{ig} \quad \text{for } g = 1, 4,$$

but we calculate the fitted days of absence for all seven grades. This is possible due to two features of the data. First, weather information is available for all years. Second, knowing when students were supposed to visit a certain grade based on the year of birth allows us to assign the instrument value. The first-stage sample of the TS2SLS approach is the same as before (weather and absence information for grades 1 and 4). The reduced-form sample is now the universe of the weather information for all grades and long-term outcomes. In the second stage we regress the long-term outcome variable $y_i^{\text{long-term}}$ on the predicted number of days of absence for all seven grades:

$$y_i^{\text{long-term}} \quad = \quad \alpha + \sum_{g=1}^{7} \tau_g \widehat{W}_{ig} + \beta_1 X_{1i} + P_i + \varepsilon_i.$$

Socio-economic controls X_1 come from the church records while the information on the parish of residence during childhood P_i comes from the exam catalog. We restrict the analysis to individuals that did not move between grades 1 and 4 and assume that the school parish was the same in all grades.

This strategy relies on the assumption that weather conditions affect absence in the same way in all seven grades. As indicated in Table 3.12, at least for the observed grades 1 and 4, this assumption seems justified. We estimate the first stage using all absence information wherefore the first-stage estimates for the long-term TS2SLS approach are (nearly) those stated in Table 3.12.[45] Table 3.13 presents the second-stage estimates for long-term outcomes. The point estimates have the expected sign, but despite the strong first stage, the estimated structural effects of absence in school fail to reach statistical significance. For employment

[44] It is not possible to employ the 2SLS implementation of the IV approach to the long-term outcomes. The following example illustrates the problem: individual A is hit by a weather shock in grade 4 in 1938, individual B is in grade 4 in 1939 and his or her absence in grade 4 is not affected by the weather shock in 1938. However, individual B's absence in grade 3 in 1938 is affected by the weather shock. This is problematic as absences in grades 3 and 4 probably affect the long-term outcomes similarly, but we do not observe absence in grade 3 (or any other grade but grades 1 and 4).

[45] Because we limit the first stage to individuals that did not move between grade 1 and 4, we lose about 100 observations for the first stage. The point estimates barely change compared to Table 3.12. The Online Appendix reports the first-stage estimates for the final long-term sample. Moreover, we present first-stage estimates by grade when the first-stage sample is restricted to individuals with available second-stage information. This does not affect our results and underlines similar first-stage effect across both grades.

Table 3.13: Long-term IV using the TS2SLS approach – Second-stage results for total days of absence

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent variable					
	> *Folk-skola*	Empl. 1960	Empl. 1970	Income 1970	Pensions 2002	Passed away≤70
Fitted days of sickness absence in						
grade 1	−0.0033	−0.0096	−0.0217	−39.36	−248.18	0.0091
	(0.0238)	(0.0254)	(0.0256)	(623.10)	(4127.05)	(0.0201)
grade 4	−0.0054	−0.0055	0.0006	154.13	−379.79	0.0039
	(0.0269)	(0.0149)	(0.0213)	(941.11)	(4325.53)	(0.0158)
# observations	22231	25657	27913	18459	19009	27913
# individuals	13814	15980	15980	11483	11848	17422
F-statistic instr.	19.72	19.22	25.59	18.38	13.54	25.59

Notes: Each cell states a separate regression. Parish-clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

in 1960 the point estimates clearly exceed the baseline estimates and the grade-1 effect is, again, bigger than the one in grade 4. A similar pattern holds for employment in 1970 but the grade-4 effect is close to zero with a positive sign. All in all, as for the short-term effects, the long-term IV estimates exceed the baseline FE estimates but do not indicate that omitted variables cause an overestimation in the baseline model.

To overcome the assumption that weather conditions have the same effect on absence in all grades, we additionally consider a model where days of absence in grades 1 and 4 are treated as two endogenous variables and weather conditions in grades 1 and 4 as two instruments. Controlling for the number of benign months in the other five grades, solves the problem of weather shocks in unobserved grades while, at the same time, does not rely on assuming equal first-stage effects. The Online Appendix shows the first- and second-stage estimations using sickness absences as endogenous variables. The first-stage F-statistics are well-above zero, but below 10. The reason for this is that we use two variables containing of number of benign months in grade 1 and in grade 4, respectively, in each first-stage regression. Given the high serial correlation of weather conditions in both grades and that we would only expect contemporary weather to affect absence, we may not overstate the importance of the F-statistics here. In fact, the results indicate, that weather conditions in the same grade have an effect on days of sickness absence (as also indicated by the short-term IV and long-term

TS2SLS results). The second-stage results do not exhibit a different pattern than for the TS2SLS estimates.

3.7 Conclusions

Student absence from school is an important but often overlooked determinant of instructional time. To date, little is known about the long-run impact of students missing school, and the only studies providing causal evidence of the impact of student absence on academic performance focus on the US. The major contribution of this paper is to estimate the impact of student absence in elementary school on short- and long-term outcomes for a non-US context by using a unique combination of historical records and administrative datasets from Sweden.

Our analysis shows that absence in elementary school has a significant impact on student performance: ten days of absence over a school year leads to a reduction in grade point average of 3.3 percent of a standard deviation. The estimated effect is very robust across empirical strategies and comparable in magnitude to results found for the US. This immediate impact on school performance spills over into secondary school admissions, which are based on elementary school performance. Our estimated effect of absence on secondary schooling admissions is at least as large as one would expect based on the effect of absence on performance – even though we are unable to attribute it to a certain school grade. For the other long-term outcomes, the effect of student absence in elementary school is only pronounced for early-career employment. For employment and income at age 35–40, pension income, and mortality, our sibling fixed effect estimates indicate that the negative effect of absence is undistinguishable from zero.

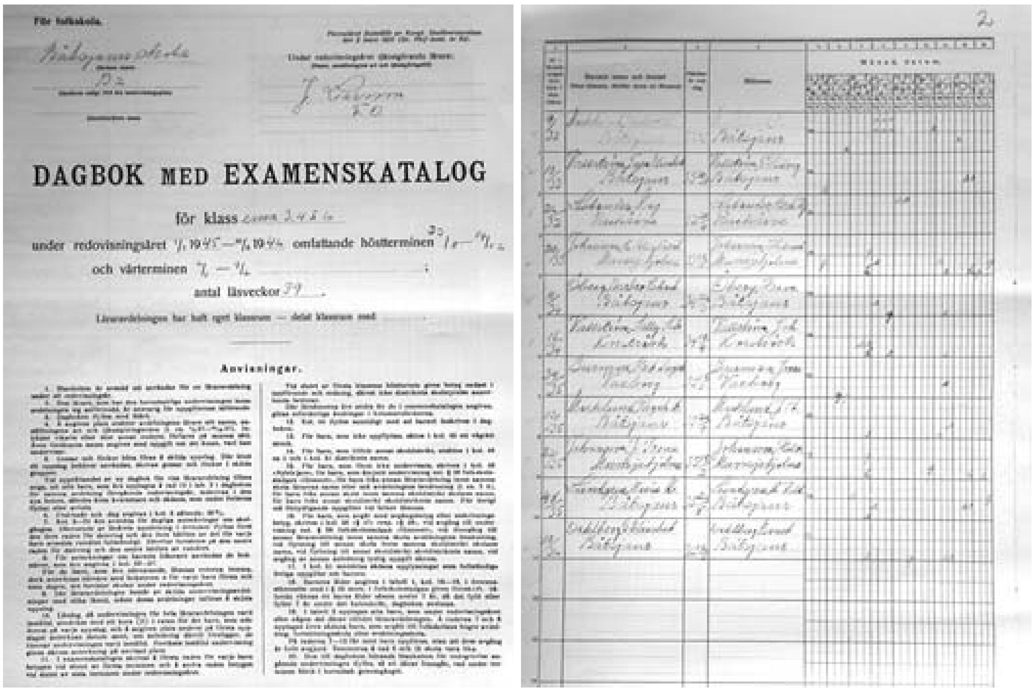
Our findings for the short-term effects of absence on school performance deliver very robust results and consistently suggest that the existence of an omitted variable bias is rather unlikely. Nevertheless, we are careful in interpreting the causality of our long-term effects. When considering long-term outcomes, it becomes more difficult to define what the alternative to the 'treatment' is. A large majority of absence days are due to illness, and we cannot rule out *a priori* that a persistent health shock from elementary school has an independent effect on long-term outcomes. As a result, we exploit the fact that our data has information on reasons for absence to compare the long-term effects of absence due to sickness with those of absence due to other reasons. We find that there are no important differences between the two, which we interpret as evidence that our long-term effects most likely capture the impact of reduced instructional time.

Our findings are obviously specific to a particular context, but as the first study providing evidence of the long-term effects of student absence, we believe that it is highly relevant to both academics and policy-makers concerned with high rates of absence around the world. Our results suggest that although student absence leads to worse performance at the end of the school year, the associated penalty on the labor market eventually fades out. Thus, the reduction in instructional time resulting from individual absence does not prevent those missing school from acquiring the skills that determine their long-run productivity. The fact that absence only affects employment at age 25–30, but no outcome later on is consistent with the labor market using educational performance as a signal of ability early on and progressively learning about workers' true productivity (Altonji and Pierret, 2001).

Our findings are strikingly in the same vein as those of a small but growing number of studies interested in the long-run consequences of variation in instructional time through changes in the school year length (Pischke, 2007) or in number of years of compulsory schooling (Stephens and Yang, 2014 and Pischke and von Wachter, 2008). One possible explanation for the patterns we find is that students are able to compensate for the educational content they miss over the next few years in school and/or that teachers are effective at helping students catch up the skills that have the most return in the labor market (though not being able to help them catch up on the whole curriculum, as reflected by the negative effect of absence on grade point average and secondary schooling enrollment). At a time when policy-makers around the world are paying increasing attention to school absence, our findings indicate that policies aimed at reducing student absence may not be particularly effective at increasing productivity in the economy. At least in the context that we study, students – perhaps with the support of their teachers and/or parents – seem able to compensate for any shortfall in learning associated with their absence in a way that does not affect their long-run productivity.

Appendix

Appendix figures



För folkskola.

DAGBOK MED EXAMENSKATALOG

för klass

under redovisningsåret 1945 — 1946 omfattande höstterminen

och vårterminen

antal läsveckor

Läraravdelningen har haft eget klassrum — delat klassrum med

Anvisningar.

Figure A3.1: Example of an exam catalog

Notes: Pictures of an exam catalog taken in an archive in Sweden.

Appendix tables

Table A3.1: Summary statistics for control variables

Time-invariant variables	Mean
Female (in %)	50.06
Number of siblings	1.56
Year of birth (in %)	
1930	18.37
1931	16.74
1932	18.30
1933	17.39
1934	16.51
1935	12.69
(we additionally control for the month of birth and interaction terms between the year and the month of birth)	
Occupation of the parents at the time of birth (in %)	
Father: farmer, fisherman, hunter	42.13
Father: agricultural worker	34.59
Father: service and sales worker	7.21
Father: production workers	49.18
Mother: employed (binary)	2.39
Living at the time of time and birth conditions (in %)	
Born out of wedlock (in %)	4.36
Born in hospital (in %)	7.97
Twin birth (in %)	4.16
Mother's year of birth	1902
Father's year of birth	1897
(mother's and father's year is controlled for by using 10-year dummies)	

	Mean grade	
Time-variant variables	1	4
Age (in years)	8.13	11.27
(included through age-in-months fixed effects)		
Class characteristics		
All classmates in same grade (in %)	34.93	29.38
Some classmates in lower grade (in %)	0.00	64.85
Some classmates in higher grade (in %)	65.07	30.81
Class size (number of students)	13.84	15.64
(measured through 5-day splines from 0 to 25)		

Notes: Own calculation based on church records and exam catalog information. Sample restricted to individuals with available sibling information when all siblings are observed in grade 1 and grade 4. Observations: 8,938. Mutually exclusive indicators may not add up to 100 per cent because of missing information. For the estimations, missing information are coded as separate category taken into account that the reason for the missing information might be meaningful in its own right (e.g., the father's occupation is missing because the father is unknown).

Table A3.2: Full estimation output for all fixed effects specifications

	(1)	(2)	(3)	(4)	(5)
	OLS	School FE	Teacher FE	Sibl. FE	Indi. FE
Total days of absence	-0.0035^{***}	-0.0043^{***}	-0.0043^{***}	-0.0041^{***}	-0.0035^{***}
	(0.0009)	(0.0010)	(0.0010)	(0.0011)	(0.0013)
Female	0.2761^{***}	0.2794^{***}	0.2863^{***}	0.3256^{***}	
	(0.0318)	(0.0313)	(0.0328)	(0.0397)	
Birth year: 1931	-0.6899^{***}	-0.9647^{***}	-0.7462^{***}	-0.2439	
	(0.2360)	(0.1333)	(0.2010)	(0.1929)	
Birth year: 1932	-0.8483^{***}	-0.5598^{***}	-0.6575^{***}	-0.7286	
	(0.1765)	(0.1308)	(0.1894)	(0.4565)	
Birth year: 1933	2.4848^{***}	-0.0635	-0.0588	0.0116	
	(0.4723)	(0.1455)	(0.1706)	(0.2730)	
Birth year: 1934	-0.6706^{**}	-0.9025^{***}	-0.6238^{***}	-0.4792	
	(0.2738)	(0.1762)	(0.2176)	(0.6530)	
Birth year: 1935	2.4016^{***}	2.6556^{***}	3.4850^{***}	-0.0263	
	(0.3734)	(0.3098)	(0.6624)	(0.3861)	
Occup. father: agriculture	0.0616	0.0051	0.0053	0.0491	
	(0.0622)	(0.0609)	(0.0638)	(0.0994)	
Occup. father: services	0.1276^{*}	0.1637^{**}	0.1675^{**}	0.0057	
	(0.0723)	(0.0802)	(0.0818)	(0.0911)	
Occup. father: farmer	0.0071	0.0533	0.0554	-0.0920	
	(0.0595)	(0.0607)	(0.0586)	(0.1011)	
Occup. father: unknown	0.0411	0.1140^{**}	0.1393^{**}	0.1571^{**}	
	(0.0576)	(0.0573)	(0.0578)	(0.0664)	
Mother employed	-0.1685	-0.1324	-0.1503	-0.0723	
	(0.1094)	(0.1021)	(0.1026)	(0.0902)	
Wedlock	0.1362	0.1202	0.0684	0.0647	
	(0.0946)	(0.0848)	(0.0757)	(0.0980)	
Born in hospital	0.1455^{**}	0.1322^{**}	0.1134^{*}	0.1111^{*}	
	(0.0648)	(0.0662)	(0.0641)	(0.0646)	
Twin	-0.2138^{***}	-0.2091^{***}	-0.2080^{**}	-0.1821	
	(0.0721)	(0.0742)	(0.0831)	(0.1362)	
Grade 4	1.2066^{***}	0.9332^{***}	1.0493^{***}	0.7943^{***}	0.5899^{***}

Notes: See note to the baseline results table. Fixed effects are suppressed. Parish-clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table A3.2 – *continued*

	(1) OLS	(2) School FE	(3) Teacher FE	(4) Sibl.	(5) Indi.
	(0.2107)	(0.1712)	(0.2313)	(0.2286)	(0.0344)
Classmates in lower grade	0.0451	0.0757	0.1033^{*}	0.1355^{***}	0.1445^{**}
	(0.0505)	(0.0461)	(0.0552)	(0.0440)	(0.0621)
Classmates in higher grade	0.0556	-0.0411	-0.0792	-0.0588	-0.0574
	(0.0378)	(0.0480)	(0.0702)	(0.0664)	(0.0553)
Class size 1–5	-0.0546^{**}	-0.0176	-0.0015	-0.0151	0.0258
	(0.0259)	(0.0297)	(0.0351)	(0.0371)	(0.0414)
Class size 6–10	0.0169	0.0148	0.0142	0.0030	0.0024
	(0.0129)	(0.0157)	(0.0164)	(0.0118)	(0.0233)
Class size 11–15	-0.0062	-0.0043	0.0032	0.0123	0.0204
	(0.0092)	(0.0098)	(0.0098)	(0.0119)	(0.0177)
Class size 16–20	-0.0009	0.0132	-0.0052	-0.0018	0.0051
	(0.0086)	(0.0097)	(0.0192)	(0.0154)	(0.0172)
Class size 21–25	0.0083	-0.0024	0.0137	0.0100	0.0149
	(0.0122)	(0.0137)	(0.0166)	(0.0166)	(0.0233)
Class size >25	0.0008	0.0137^{***}	0.0142^{***}	0.0121^{**}	0.0118^{*}
	(0.0041)	(0.0040)	(0.0047)	(0.0054)	(0.0068)
Class size (missing)	0.0029	0.4751^{***}	0.5963^{***}	0.5285^{***}	0.7646^{***}
	(0.1297)	(0.1591)	(0.1686)	(0.1791)	(0.1854)
# observations	8,942	8,942	8,942	8,942	8,942
# units		748	1,259	1,989	4,471

Notes: See note to the baseline results table. Fixed effects are suppressed. Parish-clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Online Appendix

Figures

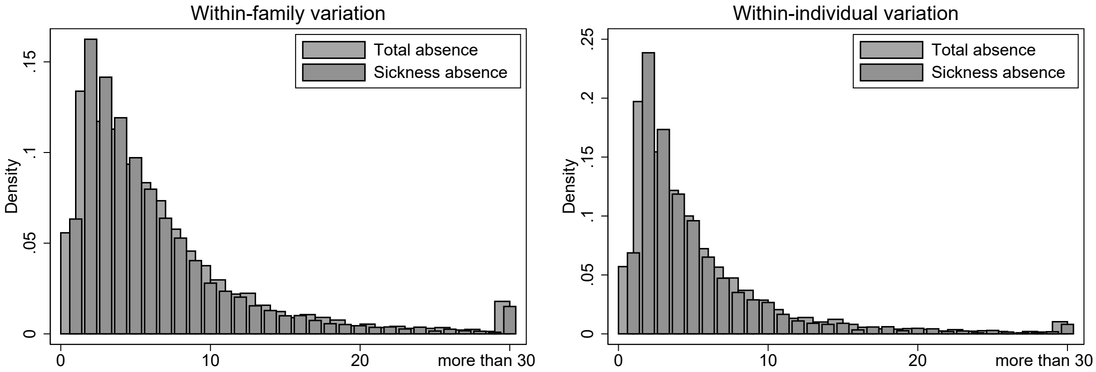


Figure O3.1: Distribution of the within-family and within-individual variation in (sickness) absence

Notes: Own calculations based on exam catalog information. 8,938 observations.

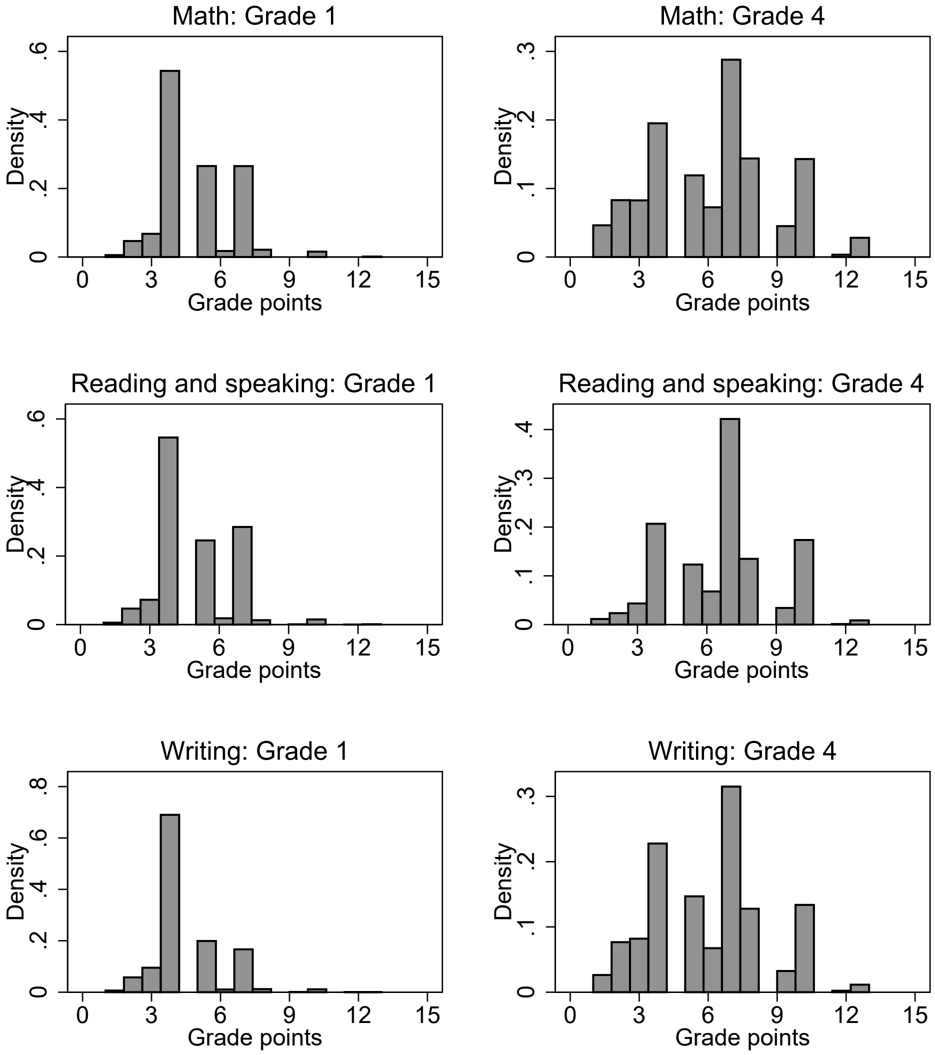


Figure O3.2: Distribution of grades by subject

Notes: Own calculations based on exam catalog information.

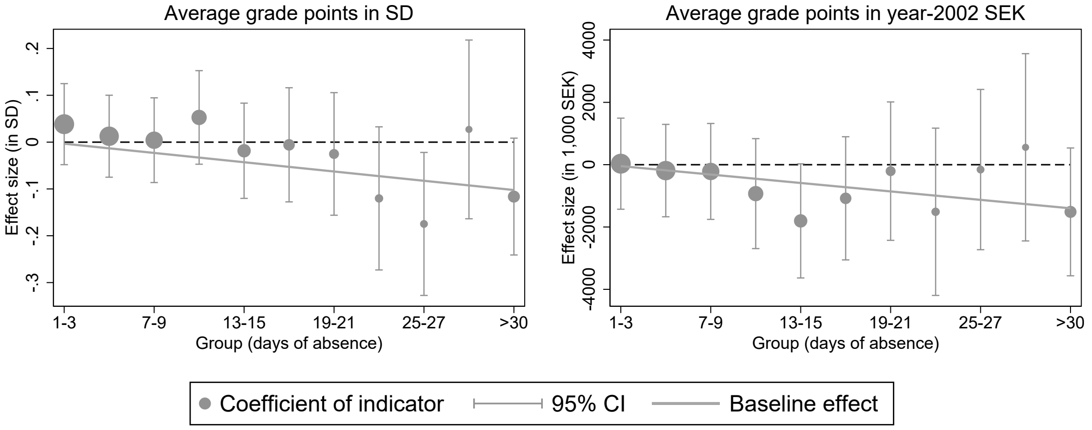


Figure O3.3: Nonlinearities in the short-term effect of grouped sickness absence

Notes: To detect nonlinearities in the effect of sickness absence we regress performance on indicator variables giving the number of days of sickness absence in groups of 3. This graph plots the coefficients of the indicator variables. The size of the marker given the relative number of observations for which the group indicator is 1. In total grade 1 and 4 information on 10,682 individuals is used. The spikes around the markers state the 95 per cent confidence interval. The orange line depicts the linear effect of an additional day of absence in the baseline results.

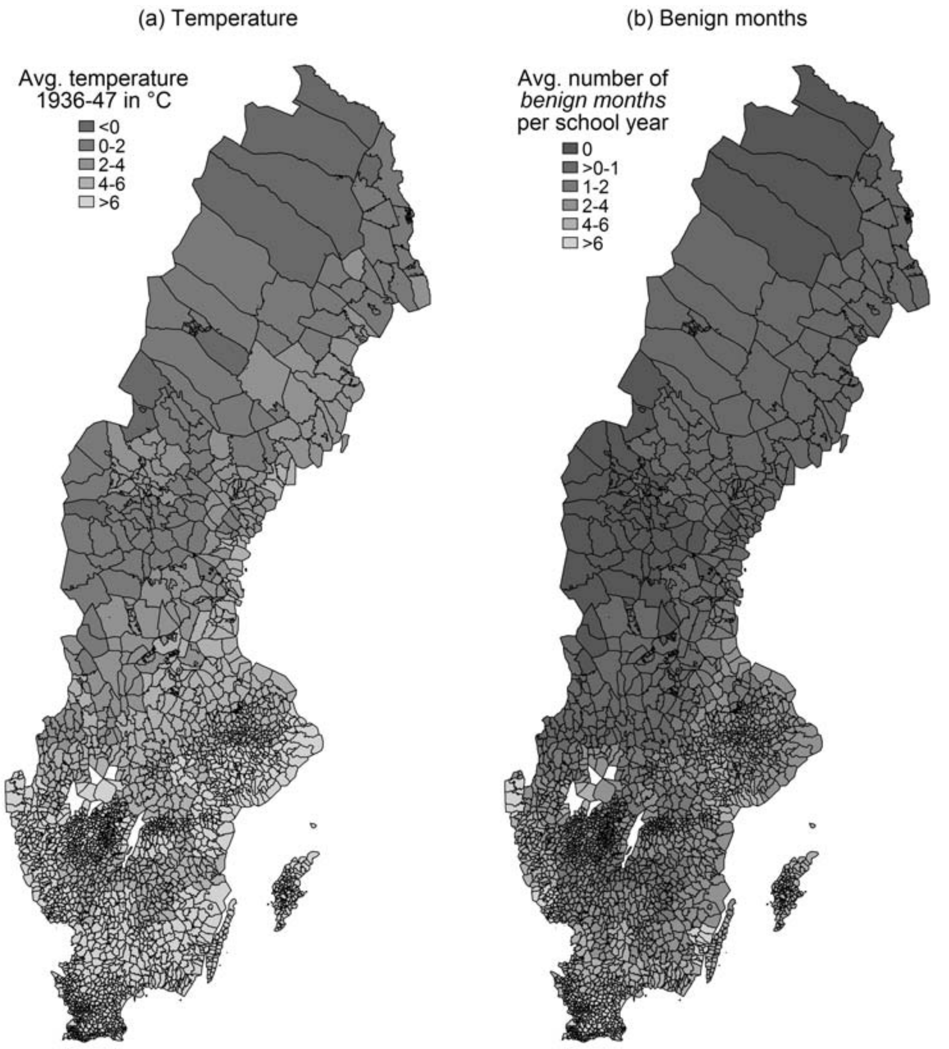


Figure O3.4: Spatial temperature distribution

Notes: Own illustration. Data on monthly temperatures are taken from Matsuura and Willmott (2012).

Tables

Table O3.1: Estimation of earnings potential

	(1)	(2)
	Earnings potential in pensions 2002	
	Grade 1	Grade 4
Math points 1	−96266.3***	19367.1***
	(32873.0)	(6668.3)
Math points 2	−21115.4**	−27084.9***
	(9183.3)	(5701.6)
Math points 3	−25267.2***	−24420.8***
	(7611.9)	(5470.9)
Math points 4	−21651.7***	−16571.9***
	(4400.2)	(3985.3)
Math points 5	−13210.3***	−5952.7
	(4608.6)	(4489.5)
Math points 6	−28367.5**	−10389.4**
	(13940.7)	(5242.2)
Math points 8	−15765.1	3756.6
	(10779.7)	(4195.3)
Math points 9	*omitted*	11150.2*
		(6752.2)
Math points 10	−5489.0	21965.2***
	(13303.2)	(4368.4)
Math points 13	26109.9	38576.6***
	(65184.5)	(8914.4)
Reading points 1	21166.6	−6229.5
	(26801.0)	(12175.4)
Reading points 2	−14140.2	12179.3
	(8862.1)	(9248.4)
Reading points 3	−10991.7	6074.4
	(7257.6)	(7185.5)
Reading points 4	−7235.8*	4654.9
	(4331.4)	(3980.4)
Reading points 5	−304.0	4637.7
	(4683.1)	(4363.7)
Reading points 6	−11745.8	9996.8*
	(12893.1)	(5364.9)
Reading points 8	−6410.5	−334.9
	(13150.2)	(4167.6)
Reading points 9	19499.8	−4478.6
	(37223.7)	(7066.7)
Reading points 10	6939.5	5217.8
	(14187.8)	(4251.6)
Reading points 13	−64483.0	10075.0
	(47935.8)	(15234.3)
Writing points 1	91143.2***	−7982.3

Continued on next page

Table O3.1 – *continued*

	(1)	(2)
	Earnings potential in pensions 2002	
	Grade 1	Grade 4
	(29263.6)	(9019.7)
Writing points 2	15566.0*	620.5
	(9243.1)	(6484.4)
Writing points 3	17213.0**	8285.3
	(7209.3)	(5955.1)
Writing points 4	13766.5***	7793.7*
	(4945.8)	(4049.7)
Writing points 5	8260.5	8874.8**
	(5427.1)	(4209.8)
Writing points 6	6036.5	−1645.7
	(16110.9)	(5563.9)
Writing points 8	13814.6	−2000.9
	(14438.2)	(4343.8)
Writing points 9	74533.5	−12542.6
	(48343.4)	(7671.2)
Writing points 10	−35879.9*	−7030.0
	(21749.1)	(4852.7)
Writing points 13	−14917.4	8006.3
	(102333.7)	(13230.3)
Constant	158638.4***	148308.6***
	(3915.1)	(3075.5)

Notes: Dependent variable: pensions taken from tax registers 2002. Explanatory variables: binary indicators of the points in math, reading and speaking and writing (15-point scale, reference category is 7 points). Grade points of 14 and 15 are not considered as such grades are barely awarded. Standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table O3.2: Baseline results using the full sample

	(1)	(2)
	OLS	Indi. FE
Average grade points in units of SD		
Days of absence	-0.0035^{***}	-0.0047^{***}
	(0.0005)	(0.0007)
Average grade points in units of pension 2002		
Days of absence	-46.9136^{***}	-74.0147^{***}
	(6.9904)	(9.3396)
# observations	28946	28946
# individuals		10682

Notes: See note to the baseline results table. Parish-clustered standard errors in parentheses. Significance: ${}^{*}p \leq 0.1$, ${}^{**}p \leq 0.05$, ${}^{***}p \leq 0.01$.

Table O3.3: Short-term effects measuring performance on a 7-point grading scale

	(1)	(2)	(3)
	OLS	Sibl. FE	Indi. FE
Average grade points (7-point scale) in units of SD			
Days of absence	-0.0035^{***}	-0.0049^{***}	-0.0045^{***}
	(0.0009)	(0.0011)	(0.0014)

Notes: See note to the baseline results table. Parish-clustered standard errors in parentheses. Significance: ${}^{*}p \leq 0.1$, ${}^{**}p \leq 0.05$, ${}^{***}p \leq 0.01$.

Table O3.4: Long-term effect of absence in school – alternative outcome measures

	(1)	(2)	(3)	(4)
	Pooled effect		Separate effects	
	OLS	Sibl. FE	OLS	Sibl. FE
Years of schooling				
Total absence (avg. both grades)	−0.0040	−0.0010		
	(0.0038)	(0.0035)		
Total absence in grade 1			0.0004	−0.0010
			(0.0025)	(0.0027)
Total absence in grade 4			-0.0043^{**}	−0.0000
			(0.0021)	(0.0020)
Gymnasium-*track education (1=yes)*				
Total abs. (avg. both grades)	−0.0001	−0.0001		
	(0.0006)	(0.0005)		
Total absence in grade 1			0.0003	0.0001
			(0.0004)	(0.0004)
Total absence in grade 4			−0.0004	−0.0001
			(0.0003)	(0.0004)
Labor market income 1970 (incl. zero incomes)				
Total abs. (avg. both grades)	-54.4394^{**}	−59.2415		
	(25.7378)	(49.9878)		
Total absence in grade 1			-29.8358^{**}	−25.4044
			(14.1129)	(24.7863)
Total absence in grade 4			−24.6486	−33.2657
			(16.7865)	(31.0227)
Pensions 2002 (incl. zero pensions)				
Total abs. (avg. both grades)	-235.8604^{*}	−229.8668		
	(119.0779)	(234.4330)		
Total absence in grade 1			−63.5785	5.0779
			(123.0447)	(170.3856)
Total absence in grade 4			-167.9506^{**}	-216.4590^{*}
			(75.2518)	(124.7150)
Pensions 2002 (incl. non-labor market income)				
Total abs. (avg. both grades)	−305.5247	−265.5650		
	(184.8329)	(254.3537)		
Total absence in grade 1			−143.9064	−19.9836
			(171.0246)	(184.6732)
Total absence in grade 4			−160.9356	−228.0035
			(110.7283)	(161.8834)

Notes: Number of observations: education measures 3,019 (in 1,373 families), income 1970 3,092 (1,398), income 1970 2,137 (985), pensions 2002 measures 2,468 (1,129). Parish-clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table O3.5: Short-term IV results for days of sickness absence

	(1)	(2)	(3)	(4)
	Specification			
	2SLS		TS2SLS	
	grade 1	grade 4	grade 1	grade 4
First-stage effect of weather on absence				
# benign months	-0.4033^{***}	-0.3231^{**}	-0.3744^{***}	-0.3744^{***}
	(0.1238)	(0.1285)	(0.0617)	(0.0617)
Second-stage effect of fitted absence on performance				
Absence	-0.0034	-0.0242	-0.0036	-0.0209
	(0.0199)	(0.0284)	(0.0568)	(0.0566)
# observations first stage	13,884	14,152	28,036	28,036
# observations second stage	13,884	14,152	13,884	14,152
F-statistic instrument	10.61	6.32	36.83	36.83

Notes: Each cell (but the first stage in column 3 and 4) states a separate regression. In the two-sample two-stage least square (TS2SLS) specification the first stage is estimated jointly over both grades. All standard errors in parentheses. The standard errors of the second stage in columns 1 and 2 are Stata's default standard errors clustered on parish level. In columns 3 and 4 the standard errors of the second stage are calculated using the Delta method and clustered on parish level. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table O3.6: Long-term IV using the TS2SLS approach – First-stage results for total days of absence by grade

	(1)	(2)	(3)	(4)	(5)	(6)
	Sample for					
	$>$ *Folk-skola*	Empl. 1960	Empl. 1970	Income 1970	Pensions 2002	Passed away$\leq$70
Grade 1 only (and only complete information)						
# benign months	-0.3206^{**}	-0.3215^{**}	-0.3431^{***}	-0.2797^{**}	-0.2780^{*}	-0.3431^{***}
	(0.1303)	(0.1342)	(0.1257)	(0.1127)	(0.1424)	(0.1257)
# observations	10,978	12,664	13,821	9,121	9,380	13,821
F-statistic instr.	6.05	5.74	7.46	6.16	3.81	7.46
Grade 4 only (and only complete information)						
# benign months	-0.2922^{*}	-0.2951^{*}	-0.3435^{**}	-0.2904^{*}	-0.3324^{**}	-0.3435^{**}
	(0.1491)	(0.1562)	(0.1581)	(0.1521)	(0.1554)	(0.1581)
# observations	11,253	12,993	14,092	9,338	9,629	14,092
F-statistic instr.	3.84	3.57	4.72	3.65	4.58	4.72
T2SSLS first-stage results (all first-stage information used)						
# benign months	-0.3527^{***}	-0.3527^{***}	-0.3527^{***}	-0.3527^{***}	-0.3527^{***}	-0.3527^{***}
	(0.0749)	(0.0749)	(0.0749)	(0.0749)	(0.0749)	(0.0749)
# observations	27,913	27,913	27,913	27,913	27,913	27,913
F-statistic instr.	22.19	22.19	22.19	22.19	22.19	22.19

Notes: Each cell states a separate regression. In rows 1 and 2 only observations with complete second-stage information were used to estimate the first stage. Parish-clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table O3.7: Long-term IV strategy treating sickness absences in grades 1 and 4 as two endogenous variables – First-stage results

	(1)	(2)	(3)	(4)	(5)	(6)
	Sample for					
	$>$ *Folk-skola*	Empl. 1960	Empl. 1970	Income 1970	Pensions 2002	Passed away$\leq$70
Endogenous variable: first-grade sickness absence						
# benign mon. gr. 1	-0.2834^{**}	-0.3592^{**}	-0.3706^{***}	-0.2983^{***}	-0.3872^{**}	-0.3706^{***}
	(0.1317)	(0.1451)	(0.1403)	(0.1059)	(0.1769)	(0.1403)
# benign mon. gr. 4	-0.1958	-0.1959	-0.2282	-0.2019	-0.2306	-0.2282
	(0.1833)	(0.1848)	(0.1811)	(0.1745)	(0.2229)	(0.1811)
# observations	10,978	12,664	13,821	9,121	9,380	13,821
F-statistic instr.	1.49	3.08	4.99	2.39	2.87	4.99
Endogenous variable: fourth-grade sickness absence						
# benign mon. gr. 1	0.1471	0.1533	0.1258	0.1503	0.1856	0.1258
	(0.1391)	(0.1338)	(0.1250)	(0.1355)	(0.1484)	(0.1250)
# benign mon. gr. 4	-0.1794	-0.1877^{*}	-0.2074^{*}	-0.1515	-0.1919^{*}	-0.2074^{*}
	(0.1118)	(0.1111)	(0.1090)	(0.1038)	(0.1150)	(0.1090)
# observations	11,253	12,993	14,092	9,338	9,629	14,092
F-statistic instr.	3.63	5.19	5.15	2.86	4.13	5.15

Notes: Each cell states a separate regression. The first-stage F-statistic of the instrument is calculated using the method suggested by Sanderson and Windmeijer (2016). Parish-clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table O3.8: Long-term IV strategy treating sickness absences in grades 1 and 4 as two endogenous variables – Second-stage results

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent variable					
	$>$ *Folk-skola*	Empl. 1960	Empl. 1970	Income 1970	Pensions 2002	Passed away$\leq$70
Sickness abs. gr. 1	0.0051	-0.0100	-0.0137	185.65	-695.65	0.0097
	(0.0243)	(0.0162)	(0.0143)	(570.65)	(2759.25)	(0.0105)
Sickness abs. gr. 4	-0.0152	-0.0003	0.0090	73.05	-2104.72	0.0042
	(0.0151)	(0.0084)	(0.0129)	(554.44)	(2826.38)	(0.0126)
# observations	8,417	9,677	10,491	6,976	7,161	10,491

Notes: Each cell states a separate regression. Parish-clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Sample selection

The socio-economic characteristics assessed in the church books used to gather the base dataset are available for all 30,150 individuals born in the sampled parishes. If exam catalog information are missing at random, the mean value of those characteristics between individuals we are able to trace down in school should equal the mean of the full sample of all 30,150 individuals. Table O3.9 shows the results. The first two columns show the means and standard deviations over all individuals, while the second two columns give the means and SD of the subsample of individuals for that we have exam catalog information. The right-most column indicates whether the difference of the means is statistically significant at any of the conventional levels. Only the share of individuals born in certain years differs occasionally at the 5 percent level. Individuals we are able to trace down are more likely to be born in 1935. However, the absolute difference is quite small and we do not think that this is somehow correlated with the relationship between absence and performance. A likely reason of the difference is that exam catalogs are often missing for entire schools and school years so that the data are missing for a larger number of individuals. All in all, Table O3.9 does not indicate systematic sample selection.

Still, to investigate this further Table O3.10 gives the baseline short-term effects separately for individuals who did not move parishes between birth and grade 4 and individuals who have moved. If moving is selective with respect to absence and performance in school, the effects would differ between the samples. This does not seem to be the case. In fact, for the siblings FE model, the point estimates are the same, even thus the association is only statistically different from zero for same-parish matches due to the fewer movers observations.

Table O3.9: Balancing check for church and school data samples

	(1)	(2)	(3)	(4)	(5)
	Background variable in				
	full sample		exam catalog sample		Difference
Variable	mean	SD	mean	SD	significant
Female	0.49	(0.50)	0.47	(0.50)	
Year of birth: 1930	0.18	(0.38)	0.17	(0.38)	*
Year of birth: 1931	0.17	(0.38)	0.16	(0.37)	**
Year of birth: 1932	0.17	(0.38)	0.16	(0.37)	**
Year of birth: 1933	0.16	(0.36)	0.16	(0.37)	
Year of birth: 1934	0.16	(0.37)	0.15	(0.36)	**
Year of birth: 1935	0.15	(0.36)	0.19	(0.39)	
Father: farmer, fisherman, hunter	0.32	(0.47)	0.26	(0.44)	
Father: agricultural worker	0.27	(0.44)	0.22	(0.42)	
Father: service and sales worker	0.09	(0.29)	0.12	(0.32)	
Father: production worker	0.57	(0.50)	0.60	(0.49)	
Father: occupation unknown	0.23	(0.42)	0.31	(0.46)	
Mother employed	0.04	(0.19)	0.08	(0.27)	
Born out of wedlock	0.08	(0.28)	0.15	(0.36)	
Born in hospital	0.11	(0.32)	0.13	(0.34)	
Twin birth	0.02	(0.15)	0.04	(0.19)	
Observations	30,150		17,771		

Notes: Own calculations on church records. Columns 1 and 2 gives the mean value and the standard deviation (SD), respectively, for the full sample. Columns 3 and 4 state the corresponding values for the sample restricted to individuals for that we are able to find exam catalog information. Column 5 indicates whether the difference in the means is statistically significant based on the p-value of a t-test of equal means. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table O3.10: Short-term effects for same-parish matches and movers

	(1)	(2)	(3)
	OLS	Sibl. FE	Indi. FE
Same-parish matches			
Average grade points in units of SD			
Days of absence	-0.0038^{***}	-0.0039^{***}	-0.0031^{**}
	(0.0009)	(0.0012)	(0.0013)
Average grade points in units of income 1970			
Days of absence	-8.5387^{***}	-6.8373^{**}	-6.6752
	(2.9067)	(3.2566)	(4.1866)
Average grade points in units of pension 2002			
Days of absence	-55.7055^{***}	-37.7550^{*}	-44.8117^{**}
	(15.1663)	(21.7071)	(20.6301)
# observations	8173	8173	8173
# individuals/families	4110	1851	4110
Movers			
Average grade points in units of SD			
Days of absence	-0.0025	-0.0039	-0.0119^{**}
	(0.0031)	(0.0035)	(0.0059)
Average grade points in units of income 1970			
Days of absence	-9.4861	-2.1847	5.1823
	(10.4066)	(6.9127)	(6.4421)
Average grade points in units of pension 2002			
Days of absence	-43.2471	-34.0697	-85.7676
	(43.2815)	(62.9909)	(105.1412)
# observations	769	769	769
# individuals/families	408	202	408

Notes: See note to the baseline results table. Parish-clustered standard errors in parentheses. Significance: $^{*}p \leq 0.1$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Chapter 4

Reanalyzing Zero Returns to Education in Germany

Joint work with Hendrik Schmitz

4.1 Introduction

A recent study by Pischke and von Wachter (2008) (PW henceforth) finds zero earnings returns to (additional compulsory) schooling in Germany. This result is in contrast to standard findings from the UK, the US, and Canada, where causal returns to schooling are often estimated to be in the range of 10–15 percent per year (e.g. Oreopoulos, 2006), and also to much of the work focussing on Continental Europe.[1] PW can rule out wage rigidities and the prominent role of apprenticeships in Germany as explanations for the zero returns. They hypothesize (but get only indirect evidence) that the extra year of schooling did not enhance labor market relevant skills which are formed earlier in the school life in Germany than in the US.

This study contributes to the literature in three ways: (1) We replicate PW's finding using a different data set (the German Socio-Economic Panel, SOEP) and a slightly different sample selection. (2) We, then, extend their analysis by using different instruments in order to get a broader picture of returns to schooling. This allows us to widen the picture of returns to schooling by evaluating whether different groups react heterogeneously. Effects for basic track students

[1] For the Netherland and France, Oosterbeek and Webbink (2007) and Grenet (2013) find effects close to zero. Meghir and Palme (2005) find heterogeneous effects in Sweden, where the highest returns are 6.7 percent. For Norway, the estimate is 9.4 percent (Aakvik et al., 2010).

(the compliers in PW) do not necessarily translate to other groups. Intermediate and academic schools differ by the curriculum, the degrees awarded, and the career paths suggesting different returns. On the other hand, there is some evidence for approximately linear returns to years of schooling in Germany (e.g. Pischke and Krueger, 1995) which might suggest transferability of the PW results to other groups. Ultimately, it is an empirical question what the effects are for other groups of students. (3) Finally, we directly test PW's conjecture that a lack of skill formation could be a reason for zero returns in Germany by estimating the causal effect of education on cognitive abilities.

Using the SOEP and exogenous variation in education, our results reinforce the PW findings. Applying compulsory years of schooling and, additionally, the supply of schools in different tracks to instrument education, we do not find any significant causal effect on wages in Germany – neither for basic nor for higher tracks. In a second step, we find that years of education, regardless of the track, do not significantly affect a word fluency score used as ability measure. This supports the hypothesis that wage relevant skills are learned at an earlier stage in Germany.

4.2 The effect of education on wages

4.2.1 Data and variables

Starting in 1984, the SOEP is the most important representative German longitudinal survey containing yearly information on about 22,000 individuals in over 11,000 households (Wagner et al., 2007). We use the 2006 wave because, apart from information on educational background and wages, it is the only wave that also includes cognitive skills measures employed later in Section 4.3.

As a measure of earnings we use the log of hourly gross wage in 2006, calculated as in PW. Years of education, are also computed as in PW by using the regular length of the track, taking the compulsory reform in the case of basic schools into account.

For the wage regression sample, we start with SOEP information on over 12,000 individuals who participate in the labor market in 2006. We restrict our sample to West-German non-city states due to the different school system in the former GDR (also see Footnote 6). Moreover, we consider the birth cohorts 1940–1970. After also dropping individuals with missing values in covariates, the estimation sample includes 5,500 people.[2] Control variables in the baseline regressions are gender and birth cohort as well as state fixed effects.

4.2.2 Identification

The first instrument (and the one used by PW) is the increase in compulsory years of education in the basic track schools (*Hauptschulen*).[3] Basic track schools covered grades 5 to 8 before the compulsory schooling reform in Germany and included a ninth grade afterwards. While some states introduced a compulsory ninth grade at an early stage, the majority of the states only introduced an additional year of schooling due to the Hamburg Accord (*Hamburger Abkommen*) in 1964; see Table A4.1 of the Online Appendix A for regional variation in years of implementation.

Since the results of instrumental variables estimations depend on the instrument and the external validity of one local average treatment effect (LATE, see Imbens and Angrist, 1994) may be considered limited (see e.g. Heckman, 2010), we use – in an extension to PW – two more instruments which capture the effect for more kinds of students in the German educational system. While the compliers of the compulsory schooling reform are only basic track students, we use the supply of schools in the two other tracks (academic schools and intermediate schools[4]) to get LATEs for students of these schools.

We measure the supply as the number of both intermediate and academic schools, respectively, per 1,000 square km in the state of residence at the student's age of 10. An increased supply of intermediate (or academic) schools enables more students to visit such a school due to increased capacities and reduced average travel time.[5] The compliers with the school supply instruments are most likely individuals at the margin of visiting a higher track. The construction of new schools

[2] A table of descriptive statistics is available in the Online Appendix. Even though we use another data source than PW, the characteristics of the samples are similar. PW use information on individuals born 1930–1960 and consider all West-German states.

[3] Enrollment into elementary school is at the child's age of six in all states. After grade four, students visit a secondary school of one out of three possible tracks. Which track a student is assigned to, basically depends on the performance in elementary school.

[4] In academic schools (*Gymnasien*) students receive a degree qualifying for university entrance (*Abitur*) after grade 13. Afterwards, many students decide to have university studies (in our sample 78 percent, see Table A4.2). In intermediate schools (*Realschulen*), students reach the leaving degree after grade 10. Even if the degree is different from the basic track degree, students usually enter vocational training afterwards (nearly 89 percent do so). Additional to *Realschulen* some states offer a comprehensive school track (*Gesamtschule*). Since comprehensive schools play only a minor role and most students leave after grade 10, we count comprehensive schools as intermediate schools, too. Even when leaving comprehensive schools out, the results remain unchanged.

[5] Academic school supply was used by Jürges et al. (2011) as an instrument in the context of health and health behaviors. The idea of instrumenting education by the availability of educational institutions goes back to Card (1995).

reduces the distance and, thus, the costs of attending a higher track school. This should be most relevant in rural areas. Certainly, the share of compliers to all students within a school track is substantially lower than for the compulsory reform.[6]

Information on the supply of schools is taken from several issues of the German Statistical Yearbook (German Federal Statistical Office, 1992). Figure A4.1 shows that there is a lot of variation among and across states in schools per 1,000 square km. This generates exogenous variation in school supply that can be used to identify effects of schooling on wages and skills. Since we use a full set of year of birth and federal state dummies we basically exploit state level deviations from the national trend in increased school supply. With the educational expansion in the 60s and 70s, all states increased the number of schools but starting points and intensity varied across states.[7] E.g., the number of academic schools per square km increased in the state of Rhineland-Palatinate by 5 percent and in Baden-Württemberg by nearly 50 percent.

The identifying assumption regarding school supply is that the variation in the timing of the educational expansion is independent from wage (and skill) expectations. This assumption would be violated if individuals with lower income expectations (or worse skills) demand more schools to improve their (or their children's) chances on the labor market relative to individuals from other states. This is unlikely to be the case and, if so, should largely be taken into account by the state fixed effects. More likely reasons for the different timing are electoral cycles and political preferences (see Hadjar and Becker, 2006 and Jürges et al., 2011).

Another concern might be the weighting of the number of schools with the state's area instead of, e.g., the state's cohort size.[8] Here we argue that three reasons challenge the use of schools per students as instrument (see Jürges et al., 2011). First, the cohort size is more volatile and would thus mainly drive the instrument's variation instead of the number of schools. Second, the cohort size probably affects earnings directly (see e.g. Freeman, 1979 and Welch, 1979), which threatens

[6] This argument does not hold for the three German city states which have a substantially higher school density than the other states. Unsurprisingly, including these reduced the first-stage *F* statistics of the instruments and, moreover, led to very unstable results. We, thus, dropped these states from the analysis.

[7] Moreover, the educational expansion did not affect intermediate and academic schools in the same way in each state.

[8] Since even a partial correlation with state-specific cohort sizes may bias the IV results towards zero, Figure A4.2 depicts the variation in school supply adjusted by the cohort sizes. The figure still shows a fair amount of variation in the construction of intermediate and academic school over time and across states. We would interpret this as a hint that the cohort size does not drive the variation of the school supply instruments.

the validity. Third, the schools per students instrument would no longer take the average distance to school into account.

We apply all instruments one by one. Doing so, this paper adds to knowledge on heterogeneity of the effects. Note, however, that, since the years of schooling measure is based on degrees the main effect of the school supply instrument is to shift individuals between tracks.

4.2.3 Results: the effect of education on wages

Table 4.1 reports the coefficients of the regressions of log hourly gross wages on years of education and control variables. Each of the seven cells is the result of one different regression. For the sake of clarity, we only report the coefficients of the instruments in the first stage regressions and of years of education in the second stage regressions. The first column shows results from simple OLS regressions, thereby neglecting any endogeneity problems. Column 2 shows results of IV regressions with compulsory schooling as an instrument, as used in PW. This instrument refers to the basic track. Column 3 uses the number of intermediate track schools per 1,000 square km in the state and column 4 the number of academic track schools per state as instrument. Formal regression equations of each specification are shown in Online Appendix B.

The first-stage results are presented in the first line of Table 4.1. Students affected by the compulsory schooling reform attend school on average 0.91 years longer due to the reform. When the availability of intermediate schools increases by one school per 1,000 square km, the average length of education increases by 0.09 years. In case of academic schools, the effect is 0.17 years. The average number of intermediate schools per 1,000 square km is 3.1 in 1950 and 11.2 in 1980 over all states under review. For academic schools the mean is 6.3 in 1950 and 9.6 in 1980, see Table A4.2. Thus, a one unit increase is large in relative terms and together with the small coefficients this hints at the amount of compliers with the school extensions being much smaller than with the compulsory schooling reform.

Regarding the structural equation and wages as dependent variable, the OLS coefficient is statistically different from zero. An increase by one year of education goes along with about 6.9 percent higher wages on average. The magnitude is in line with the one in PW, Table 2, although they use different data sources. Contrary to the OLS case, the IV coefficients of education are not only statistically but also economically insignificant (again in line with PW for the compulsory schooling instrument). Using the instruments for intermediate and academic schooling one by one, the effect of additional education is practically zero. This result goes

Table 4.1: Estimation results: Wages

Dependent variable	OLS	IV		
		Instrument referring to		
		Basic	Inter.	Acad.
First stage results				
Years of education	$--$	0.9079^{***} (0.1894)	0.0919^{***} (0.0246)	0.1672^{***} (0.0426)
Second stage results				
Log hourly gross wage	0.0686^{***} (0.0020)	-0.0004 (0.0276)	-0.0002 (0.0373)	0.0010 (0.0360)
First-stage F-statistic instruments	$--$	22.99	13.94	15.40

Notes: Own calculations based on SOEP data. Numbers of observations: 5,499. Control variables: female, as well as state and birth cohort fixed effects. State of schooling × year aged 10-clustered standard errors in parentheses. Coefficients in the first stage refer to the respective instruments. Coefficients in the second stage refer to years of schooling. Coefficients of other control variables not reported here but available upon request. Significance: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

beyond PW's finding of zero returns to compulsory schooling. While compliers to their reform are basic track students, these results suggest that schooling also does not pay off for higher track students in Germany. The finding that the effect of an additional year of education is the same for all levels of schooling in Germany is in line with Pischke and Krueger (1995). They find a linear wage increase by years of education using OLS.

The Online Appendix provides results of several robustness checks where we, (1) use net instead of gross wages as outcome variable, (2), limit years of education to primary and secondary schooling (no post-secondary education), (3), add more control variables that were left out in the preferred specification due to potential "bad control" problems,[9] (4), also control for the average number of students per school by track, (5), add interaction terms for gender and the cohort and state fixed effects, and, (6) estimate the reduced-form coefficients. In no specification do we find a significant effect of education on wages for either instrument. All in all, the robustness checks underline the baseline finding of zero returns.

[9] The added controls are dummy variables for mother's/father's education (at least intermediate school degree), number of siblings, dummies for at least good self-assessed health status, obesity (Body Mass Index ¿ 30), migrational background, university degree, completed apprenticeship training, and an ISCO scale-based measure of the skill level demanded by the respondent's job.

Table 4.2: Estimation results: Cognitive abilities

Dependent variable	OLS	IV		
		Instrument referring to		
		Basic	Inter.	Acad.
First stage results				
Years of education	$--$	1.0211*** (0.2701)	0.0618* (0.0341)	0.1650*** (0.0597)
Second stage results				
Log crystallized intelligence test score	0.0456*** (0.0039)	-0.0290 (0.0541)	0.0078 (0.0966)	-0.0197 (0.0661)

Notes: Own calculations based on SOEP data. Numbers of observations: 2,464. Control variables: female, as well as state and birth cohort fixed effects. State of schooling × year aged 10-clustered standard errors in parentheses. Coefficients in the first stage refer to the respective instruments. Coefficients in the second stage refer to years of schooling. Coefficients of other control variables not reported here but available upon request. Significance: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

4.3 Explanations: effects of education on cognitive skills

PW conjecture – but cannot directly test – that the potential reason for zero returns to (additional compulsory) schooling is that German students have learned the important skills already before the ninth grade. To measure skills we use the word fluency test score included in the SOEP. While overall cognitive abilities – also referred to as intelligence – incorporates many components, word fluency, or "crystallized intelligence", is the component attributed to environmental factors like education (Anderson, 2007). In the SOEP, word fluency is assessed by an ultra-short intelligence test where respondents have to name as many animals as possible in 90 seconds.

The test score is the number of correct unique answers. Lang et al. (2007a) show that the ultra-short intelligence test applied in the SOEP is comparable to more extensive ones used in psychology. In order to simplify the interpretation of the crystallized intelligence test score, we use its log value. For a documentation see Schupp et al. (2008). For the ultra-short intelligence test, a computer assisted personal interview (CAPI) was needed and only one third of all SOEP respondents were randomly asked to participate in the tests. Therefore, our cognitive test sample includes only about 2,500 observations.

The same identification issues as with wages appear to be relevant in this case (see Heckman and Vytlacil, 2001), hence, we again prefer IV results with the same instruments as before over benchmark OLS results in Table 4.2. The same picture as with wages emerges. Individuals with more years of schooling have higher intelligence test scores – about 4.8 percent with one more year of schooling in the OLS regressions. However, once accounting for the endogeneity of school length, the coefficients clearly approach zero and become insignificant (with larger confidence bands, however).

If one compares the first-stage results of the wage and skills regressions, the relevance of the supply of intermediate schools decreases for skills. The instrument is only significant at the 10 percent level, probably due the lower sample size in the skills regression. The other instruments remain highly significant. Even if one would lose faith in the IV results for intermediate tracks, there is no reason to believe that the effect should drastically differ from those in the other two tracks.

Like for the wage regression, we carry out a series of robustness checks in the Online Appendix. The specifications are the same as those for wages as outcome variable (excluding gross/net comparison). The results are basically in line with the baseline results. In a handful of estimations in the robustness checks, we find somewhat higher coefficients than before (both in positive and negative direction). They are never significant and, if any, do not systematically point into one direction. We conclude that it is fair to say that "zero is not a bad number" to describe the effect of schooling on cognitive skills in Germany.[10]

To conclude, we find no systematic and significant effect of education on cognitive abilities. Hence, a lack of skills learned in school is an explanation for the zero wage returns to additional education that is consistent with the evidence shown in this paper.

4.4 Conclusions

This paper replicates Pischke and von Wachter's (2008) study on returns to compulsory schooling in Germany using a different data set. After reinforcing their results we extend their analysis in two dimensions. First, we also use other instruments and, thus, estimate the effect of schooling on earnings for different groups of compliers. The group of compliers for the new instruments is much smaller, however.

Second, we test a hypothesis by PW for their finding of zero returns to compulsory schooling, namely that basic skill formation – relevant for the labor market at

[10] This is a classic statement of James Heckman on the effect of training programs for unemployed, published on p.23 of the 6 April 1996 edition of *The Economist*, also cited by, e.g., Lechner et al. (2011).

least – takes places before the ninth grade in Germany. This is done by estimating the causal effect of education on cognitive skills.

We do find no causal effect of schooling on earnings for any group of compliers. Moreover, we find no significant effect of education on cognitive abilities which is consistent with the mentioned explanation for no effects of schooling on wages. Of course, this does not prove that basic skill formation *does* indeed take place *before* the ninth grade in Germany. It is, however, some evidence that it *might not* take place *after* the eighth grade. Both positive correlations of education with earnings and skills seem to be mainly driven by selection of higher skilled individuals into more years of schooling.

Online Appendix

Figure and tables

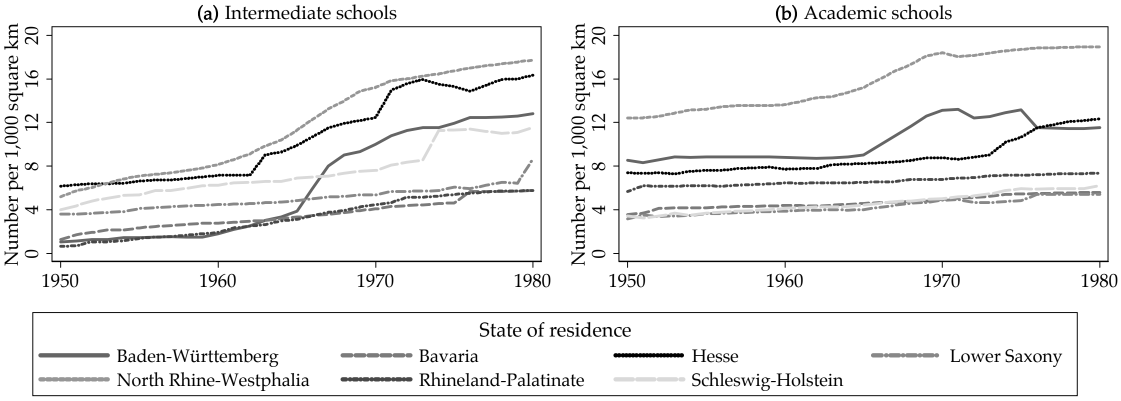


Figure A4.1: Number of intermediate and academic schools per 1,000 square km

Notes: Own calculations, data taken from the German Statistical Yearbook (German Federal Statistical Office, 1992). Since birth cohorts 1940–1970 are used and school supply is measured at the respondent's age of 10, we use information on school supply from 1950–1980.

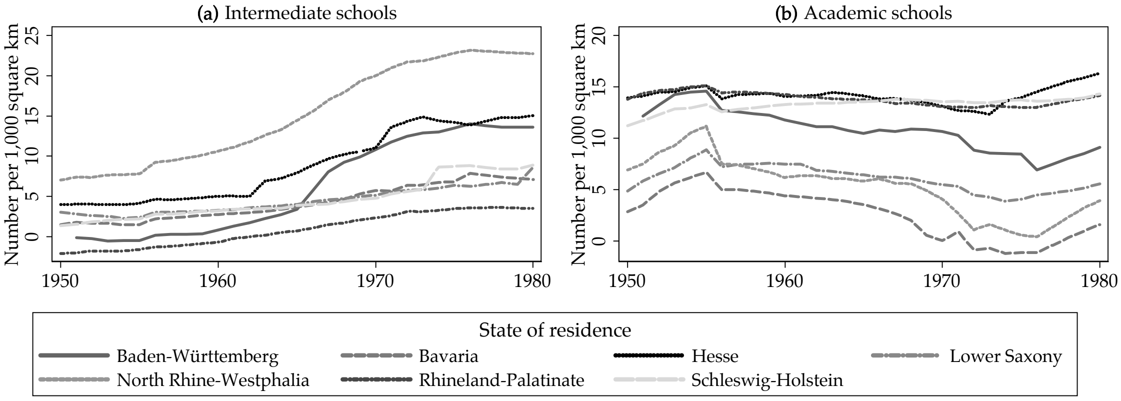


Figure A4.2: Number of intermediate and academic schools per 1,000 square km adjusted by cohort size

Notes: Own calculations, data taken from the German Statistical Yearbook (German Federal Statistical Office, 1992). Since birth cohorts 1940–1970 are used and school supply is measured at the respondent's age of 10, we use information on school supply from 1950–1980. We create this figure using a two-step procedure. In the first step, we regress the number of intermediate and academic schools, respectively, on state and year indicators and the state and year-specific cohort size. In the second step, we use the coefficients of the first step to predict the cohort size-adjusted number of schools.

Table A4.1: Summary statistics for the instruments

Panel A: Size of the states and introduction of compulsory schooling

	Area (in 1,000 square km)	Mandatory ninth grade for basic track schools		
		Year of introduction	First cohort affected	Share of students affected
Schleswig-Holstein	15.799	1956	1941	76.0%
Lower Saxony	47.635	1962	1947	80.5%
North Rhine-Westphalia	34.088	1967	1953	74.6%
Hesse	21.115	1967	1953	71.4%
Rhineland-Palatinate	18.854	1967	1953	78.3%
Baden-Württemberg	35.751	1967	1953	72.6%
Bavaria	70.550	1969	1955	78.5%
Average				76.0%

Panel B: Number of intermediate schools per 1,000 square km by selected years

	1950	1960	1970	1980
Schleswig-Holstein	3.988	6.266	7.595	11.583
Lower Saxony	3.590	4.492	5.353	8.670
North Rhine-Westphalia	5.192	8.155	15.225	17.748
Hesse	6.157	7.151	12.456	16.339
Rhineland-Palatinate	0.636	1.909	4.455	5.781
Baden-Württemberg	1.063	1.790	9.986	12.811
Bavaria	1.247	2.764	4.068	5.755
Average	3.125	4.647	8.448	11.241

Panel C: Number of academic schools per 1,000 square km by selected years

	1950	1960	1970	1980
Schleswig-Holstein	3.418	4.114	5.000	6.203
Lower Saxony	3.149	3.863	4.891	5.416
North Rhine-Westphalia	12.409	13.641	18.394	18.922
Hesse	7.388	7.720	8.762	12.361
Rhineland-Palatinate	5.675	6.471	6.789	7.372
Baden-Württemberg	8.559	8.811	13.119	11.552
Bavaria	3.558	4.394	4.947	5.599
Average	6.308	7.002	8.843	9.632

Notes: Years and birth cohorts affected by the compulsory schooling reform are taken from Pischke and von Wachter (2005), all other information are form the German Statistical Yearbook (German Federal Statistical Office, 1992). The nature of the variation in compulsory schooling and school supply is distinct. In general, the German Constitution guarantees the autonomy of the Federal States in educational policy. After the WWII, basic track schools offered 8 years of education in total. For reasons described in the text, some states decided to introduce a mandatory ninth grade at an early stage (see panel A of this table). In 1964 the prime ministers of the states agreed on the Hamburg Accord (*Hamburger Abkommen*) in order to unify some key characteristics of the educational systems in the German states. Besides the introduction of a mandatory ninth grade in all states by 1967, the Hamburg Accord mainly regulated the start of the school year, see Pischke (2007). The number of schools per track remained unaffected by the Hamburg Accord or any other agreement. In other words, while changes in compulsory schooling are the consequence of a small degree of unification, the variation in the school construction reflects the states' autonomy in educational issues.

Table A4.2: Means of selected variables by track

	Basic	Inter.	Acad.	Total
Income				
Gross hourly wage (in €)	14.64	17.65	23.10	17.84
Gross monthly wage (in €)	2,474	3,053	4,269	3,132
Education				
Years of education	10.26	12.17	16.73	12.57
University degree (in %)	5.00	15.84	78.15	27.42
Apprenticeship (in %)	79.00	88.97	34.23	70.99
Cognitive skills				
Raw crystallized intelligence test score	23.21	28.69	31.64	26.56
Socio-demographic characteristics				
Female (in%)	39.50	47.31	40.36	42.41
Age (in years)	48.32	46.54	47.54	47.51
Mother has intermediate school degree (in %)	6.07	15.91	42.08	19.07
Father has intermediate school degree (in %)	7.64	22.14	51.48	24.36
Number of siblings	2.35	1.67	1.47	1.89
Self-assessed health stats at least good (in %)	46.18	55.60	60.61	53.11
Obesity: Body Mass Index ¿ 30 (in %)	22.00	14.52	11.60	16.77
Migratinal background (in %)	29.09	7.39	7.40	16.08
Measure of skills needed for job[a]	2.19	2.73	3.50	2.74
Observations[b]	2,200	1,894	1,405	5,499
Share (in %)	40.01	34.44	25.55	100

Notes: Own calculations based on SOEP data. For cognitive skills the number of observations varies from the number given at the bottom of the table.

[a] ISCO scale-based measured of the skill level demanded by the respondents job, scale ranges from 1 (low skills needed) to 4 (high skills needed), see International Labour Organization (2012).

[b] Based on wage information.

Table A4.3: Robustness for wage as outcome variable

Specification	OLS	IV		
		Instruments referring to		
		Basic	Inter.	Acad.
First stage results				
Log net hourly wage	– –	0.908***	0.092***	0.167***
		(0.189)	(0.025)	(0.043)
Only school years	– –	0.373***	0.047***	0.098***
		(0.115)	(0.015)	(0.026)
Socio-economic controls	– –	0.902***	0.083***	0.119***
		(0.101)	(0.014)	(0.024)
Institutional controls	– –	0.864***	0.097***	0.208***
		(0.207)	(0.033)	(0.055)
Female specification	– –	0.924***	0.097***	0.172***
		(0.191)	(0.024)	(0.041)
Second stage results				
Log net hourly wage	0.067***	0.008	0.019	0.000
	(0.002)	(0.028)	(0.035)	(0.038)
Only school years	0.102***	−0.031	−0.020	−0.012
	(0.004)	(0.068)	(0.075)	(0.063)
Socio-economic controls	0.029***	0.017	−0.018	−0.043
	(0.004)	(0.026)	(0.039)	(0.053)
Institutional controls	0.069***	0.015	0.039	0.033
	(0.002)	(0.030)	(0.042)	(0.035)
Female specification	0.068***	0.000	0.008	0.005
	(0.002)	(0.027)	(0.035)	(0.035)
Reduced form	– –	0.000	−0.000	0.001
		(0.025)	(0.017)	(0.030)

Notes: Own calculations based on SOEP data. Control variables: female, as well as state and birth cohort fixed effects. State of schooling × year aged 10-clustered standard errors in parentheses. Significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Explanations: Log net hourly wage: dependent variable is the net instead of the gross hourly wage in logs. Observations: 5,499. Only school years: endogenous explanatory variable is limited to primary and secondary education. Observations: 5,268. Socio-economic controls: additional control variables: dummy variables for mother's/father's education (at least intermediate school degree), number of siblings, dummy variables for at least good self-assessed health status, obesity (Body Mass Index ¿ 30), migrational background, university degree, completed apprenticeship training, and an ISCO scale-based measure of the skill level demanded by the respondent's job. Leaving the potentially endogenous variables university degree and completed apprenticeship training out, does not change the pattern. Observations: 4,666. Institutional controls: additional control variables (starting with the baseline model) for the average size of the schools per track by year and federal state. Observations: 5,499. Female interaction terms: additional interaction terms between female and the state and birth cohort fixed effects are included. Observations: 5,499. Reduced form: instrument directly plugged into the the wage equation instead of instrumented years of education. Observations: 5,499.

Table A4.4: Robustness for crystallized intelligence as outcome variable

Specification	OLS	IV		
		Instruments referring to		
		Basic	Inter.	Acad.
First stage results				
Only school years	––	0.600^{***}	0.062^{***}	0.137^{***}
		(0.182)	(0.022)	(0.039)
Socio-economic controls	––	1.129^{***}	0.125^{***}	0.169^{***}
		(0.174)	(0.025)	(0.043)
Institutional controls	––	1.118^{***}	0.103^{**}	0.234^{***}
		(0.276)	(0.046)	(0.072)
Female specification	––	0.991^{***}	0.062^{*}	0.157^{***}
		(0.255)	(0.034)	(0.060)
Second stage results				
Only school years	0.067^{***}	-0.072	-0.023	-0.033
	(0.007)	(0.096)	(0.099)	(0.082)
Socio-economic controls	0.023^{*}	-0.018	0.038	-0.044
	(0.012)	(0.051)	(0.073)	(0.103)
Institutional controls	0.046^{***}	-0.024	-0.024	0.009
	(0.004)	(0.054)	(0.079)	(0.054)
Female specification	0.045^{***}	-0.035	0.007	-0.014
	(0.004)	(0.054)	(0.095)	(0.069)
Reduced form	––	-0.030	0.002	-0.016
		(0.051)	(0.031)	(0.053)

Notes: Own calculations based on SOEP data. Control variables: female, as well as state and birth cohort fixed effects. State of schooling × year aged 10-clustered standard errors in parentheses. Significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Explanations: Only school years: endogenous explanatory variable is limited to primary and secondary education. Observations: 2,355. Socio-economic controls: additional control variables: dummy variables for mother's/father's education (at least intermediate school degree), number of siblings, dummy variables for at least good self-assessed health status, obesity (Body Mass Index ¿ 30), migrational background, university degree, completed apprenticeship training, and an ISCO scale-based measure of the skill level demanded by the respondent's job. Leaving the potentially endogenous variables university degree and completed apprenticeship training out, does not change the pattern. Observations: 1,259. Institutional controls: additional control variables (starting with the baseline model) for the average size of the schools per track by year and federal state. Observations: 2,464. Female interaction terms: additional interaction terms between female and the state and birth cohort fixed effects are included. Observations: 2,464. Reduced form: instrument directly plugged into the the wage equation instead of instrumented years of education. Observations: 2,464.

Formal Description of the Model

The conventional OLS estimation of the relationship between education and the log of hourly gross wages is the following:

$$y_i = \beta_0 + \beta_1 educ_i + \boldsymbol{\beta_2 X}_i + u_i. \tag{A4.2}$$

y_i is the log hourly gross wage of person i (with $i = 1, 2, ..., N$ observations). *educ* denotes the years of education measured as years in school determined by chosen track plus years of further education due to an apprenticeship or university studies. $\boldsymbol{X}_i$ is a matrix of control variables. In the baseline specification these variables are gender as well as state of schooling and year of birth fixed effects.[11] Further variables – which might depend on education and are therefore left out in the preferred specification – are added later on to check the robustness. u_i is the error term. If there is a selection of individuals with higher skills into more education and better paid jobs, β_1 would be biasedly estimated by OLS.

To overcome this problem we instrument years of education using different instruments. The first-stage equation is

$$educ_i = \delta \text{ Instrument}_i + \gamma \boldsymbol{X}_i + \varepsilon_i \tag{A4.2}$$

where ε_i denotes the error term. The instrument is one of the following three variables: a dummy variable which is 1 if person i was affect by the compulsory schooling reform, and 0 otherwise; and the number of either intermediate or academic schools per 1,000 square km in the state of residence at the respondent's age of 10. Thus, we run three regressions with one instrument at a time.

The formal model depicted here is the same when the outcome variable is the log value of the cognitive skills test score as in Section 4.3 of the text. In this case, however, y in Eq. (A4.2) states the log crystallized intelligence test score of the SOEP.

Cognitive skill measures and previous findings

In the psychological literature the commonly used test procedure to measure intelligence is the Wechsler Adult Intelligence Scale (WAIS). It covers seven distinct

[11] As Jürges et al. (2011) we do not include state-specific trends. Regarding our IV estimations, this would discard variation and reduce the explanatory power of the school supply instruments considerably. The second-stage results would, however, not change qualitatively. I.e., the second-stage coefficients are not large and significant after including state-specific trends, thus not leading to different conclusions.

skill components. For two of those components the SOEP includes a short test especially designed for the conduction in the survey. According to psychological insights (see e.g. **?**), the measure we use, crystallized intelligence, is determined by environmental factors, e.g., education. The other intelligence measure in the SOEP refers to fluid intelligence – a component of the overall intelligence that is attributed to inherited genes.

We are only aware of four studies on cognitive skill returns to secondary education. These studies based on samples for which a positive earning returns to education were established. The findings of the studies are ambiguous but indicate positive skill returns if any. Moreover, the choice of the intelligence component seems to matter.

Glymour et al. (2008) and Banks and Mazzonna (2012) instrument education using changes in compulsory schooling in the US and the UK, respectively. Using similar law changes in Continental European countries, Schneeweis et al. (2014) provide pooled evidence on the schooling-skills relationship in the Survey of Health, Ageing and Retirement in Europe (SHARE). Mazzonna (2012) uses compulsory schooling and the birth order to instrument years of education in the SHARE data. While the sets of countries analysed by Schneeweis et al. (2014) and Mazzonna (2012) include Germany, they pool the observations affected by the German compulsory schooling reform with observations from other countries (for which there is evidence on positive wage returns to education). Therefore, our study provides the first separate analysis of the effect of years of secondary education on cognitive skills.

Chapter 5

Heterogeneity in Marginal Non-monetary Returns to Higher Education

Joint work with Hendrik Schmitz and Matthias Westphal

5.1 Introduction

"The whole world is going to university – Is it worth it?" *The Economist*'s headline read in March 2015.[1] While convincing causal evidence on positive labor market returns to higher education is still rare and nearly exclusively available for the US, even less is known about the non-monetary returns to college education (see Barrow and Malamud, 2015 and Oreopoulos and Petronijevic, 2013). Although non-monetary factors are acknowledged to be important outcomes of education (Oreopoulos and Salvanes, 2011), evidence on the effect of college education is so far limited to health behaviors (see below). We estimate the long-lasting marginal returns to college education in Germany decades after leaving college. As a benchmark, we start by looking at wage returns to higher education but the paper's focus is on the non-monetary returns which might also be seen as mediators of the more often studied effect of education on wages. These non-monetary returns are cognitive abilities and health.

Cognitive abilities and health belong to the most important non-monetary determinants of individual well-being. Moreover, the stock of both factors also influences the economy as a whole (see, among many others, Heckman et al., 1999,

[1] *The Economist*, edition March 28th to April 3rd 2015.

and Cawley et al., 2001, for cognitive abilities and Acemoglu and Johnson, 2007, Cervellati and Sunde, 2005, and Costa, 2015, for health). Yet, non-monetary returns to college education are not fully understood (Oreopoulos and Salvanes, 2011). Psychological research broadly distinguishes between effects of education on the long-term cognitive ability differential that are either due to a change in the cognitive reserve (i.e., the cognitive capacity) or due to an altered age-related decline (see, e.g., Stern, 2012). Still, even the compound manifestation of the overall effect has rarely been studied for college education over a short-term horizon[2] and – as far as we are aware – it has never been assessed for the long run. Few studies analyze the returns to college education on health behaviors (Currie and Moretti, 2003, Grimard and Parent, 2007, de Walque, 2007).

We use a slightly modified version of the marginal treatment effect approach introduced and forwarded by Björklund and Moffitt (1987) and Heckman and Vytlacil (2005). The main feature of this approach is to explicitly model the choice for education, thus turning back from a mere statistical view of exploiting exogenous variation in education to identify casual effects towards a description of the behavior of economic agents. Translated into our research question, the MTE is the effect of education on different outcomes for individuals at the margin of taking higher education. The MTE can be used to generate all conventional treatment parameters, such as the average treatment effect (ATE). On top of this, comparing the marginal effects along the probability of taking higher education is also informative in its own right: different marginal effects do not just reveal effect heterogeneity but also some of its underlying structure (for instance, selection into gains). This is be an important property that the local average treatment effect – LATE, as identified by conventional two stage least squares methods – would miss.

The individuals in our sample made their college decision between 1958 and 1990 and graduated in the case of college education between 1963 and 1995. Our outcome variables (wages, standardized measures of cognitive abilities[3] and mental and physical health) are assessed between 2010 and 2012, thus, 20 to 54 years after the college decision. Our instrument is a measure of the relative availability of college spots (operationalized by the number of enrolled students divided by the number of inhabitants) in the area of residence at the time of the secondary school graduation. Using detailed information on the arguably exogenous expansions

[2] Hansen et al. (2004) use a control function approach to adjust for education in the short-term development of cognitive abilities. Carneiro et al. (2001, 2003) analyze the short-term effects of college education. Glymour et al. (2008), Banks and Mazzonna (2012), Schneeweis et al. (2014), and Kamhöfer and Schmitz (2016) analyze the effects of secondary schooling on long-term cognitive skills.

[3] See Section 5.4 for a detailed definition of cognitive abilities. We use the terms “cognitive abilities”, “cognitive skills”, and “skills” interchangeably.

of college capacities in all 326 West German districts (cities or rural areas) during the so-called "educational expansion" between the 1960s and 1980s generates variation in the availability of higher education.

By deriving treatment effects over the entire support of the probability of college attendance, this paper contributes to the literature mainly in two important ways. First, this is the first study that analyzes the long-term effect of college education on cognitive abilities and general health measures (instead of specific health behaviors). Long-run effects on skills are crucial in showing the sustainability of human capital investments after the age of 19. Along this line, this outcome can complement existing evidence in identifying the fundamental value of college education since – unlike studies on monetary returns – effects on cognitive skills do neither directly exhibit signaling (see the debate on discrepancy between private and social returns as in Clark and Martorell, 2014) nor adverse general equilibrium effects (as skills are not determined by both, forces of demand and supply). Second, by going beyond the point estimate of the LATE, we provide a more comprehensive picture in an environment of essential heterogeneity.

The results suggest positive average returns to college education for wages, cognitive abilities, and physical health. Yet, the returns are heterogeneous – thus, we find evidence for selection into gains – and even close to zero for the around 30 percent of individuals with the lowest desire to study. Mental health effects are zero throughout the population. Thus, or findings can be interpreted as evidence for remarkable positive average returns for those who took college education in the past. Yet, a further expansion in college education, as sometimes called for, is likely not to pay off as this would mostly affect individuals in the part of the distribution that are not found to be positively affected by education. We also try to substantiate our results by looking at potential mechanisms of the average effects. Although we cannot causally differentiate all channels and the data allow us to provide suggestive evidence only, our findings may be interpreted as follows. Mentally more demanding jobs, jobs with a less health deteriorating effects and better health behaviors probably add to the explanation of skill and health returns to education.

The paper is organized as follows. Section 5.2 briefly introduces the German educational system and describes the exogenous variation we exploit. Section 5.3 outlines the empirical approach. Section 5.4 presents the data. The main results are reported in Section 5.5 while Section 5.6 addresses some of its potential underlying pathways. Section 5.7 concludes.

5.2 Institutional background and exogenous variation

5.2.1 The German higher educational system

After graduating from secondary school, adolescents in Germany either enroll into higher education or start an apprenticeship. The latter is part-time training-on-the-job and part-time schooling. This vocational training usually takes three years and individuals often enter the firm (or another firm in the sector) as a full-time employee afterwards. To be eligible for higher education in Germany, individuals need a university entrance degree. In the years under review, only academic secondary schools (*Gymnasien*) with 13 years of schooling in total award this degree (*Abitur*). Although the tracking from elementary schools to secondary schools takes place rather early at the age of 10, students can switch secondary school tracks in every grade. It is also possible to enroll into academic schools after graduating from basic or intermediate schools in order to receive a university entrance degree.

In Germany, mainly two institutions offer higher education: universities/colleges[4] and universities of applied science (*Fachhochschulen*). The regular time to receive the formerly common *Diplom* degree (master's equivalent) was 4.5 years at both institutions. Colleges are usually large institutions that offer degrees in various subjects. The other type of higher educational institutions, universities of applied science, are usually smaller than colleges and often specialized in one field of study (e.g., business schools). Moreover, universities of applied science have a less theoretical curriculum and a teaching structure that is similar to schools. Nearly all institutions of higher education in Germany do not charge any tuition fees. However, students have to cover their own costs of living. On the other hand, their peers in apprenticeship training earn a small salary. Possible budget constraints (e.g., transaction costs arising through the need to move to another city in order to go to college) are likely determinants of the decision to enroll into higher education.

5.2.2 Exogenous variation in college education over time

While the higher educational system as described in Section 5.2.1 did not change in the years under review, the accessibility (in terms of mere quantity but also dis-

[4] We use the words university and college as synonyms to refer to German *Universitäten* and closely-related institutions like technical universities (*Technische Universitäten/Technische Hochschulen*), an institutional type that combines features of colleges and universities applied science (*Gesamthochschulen*) and universities of the armed forces (*Bundeswehruniversitäten/Bundeswehrhochschulen*).

tribution within Germany) of tertiary education changed significantly, providing us with a source of exogenous variation. This so called "educational expansion" falls well into the period of study (1958–1990). Within this period, the shrinking transaction costs of studying may have changed incentives and the mere presence of new or growing colleges could also have nudged individuals towards higher education that otherwise would not have studied. In this paper, we consider two processes in order to quantify the educational expansion. The first is the openings of new colleges, the second is the extension in capacity of all colleges (we refer to both as college availability).[5] College availability as an instrument for higher education was introduced to the literature by Card (1995) and has frequently been employed since then (e.g., Currie and Moretti, 2003), also to estimate the MTE (e.g., Carneiro et al., 2011, and Nybom, 2017). We exploit the rapid increase in the number of new colleges and in the number of available spots to study as exogenous variation in the college decision.

Between 1958 (the earliest secondary school graduation year in our sample) and 1990 the number of colleges in Germany doubled from 33 to 66.[6] In particular, the opening of new colleges introduced discrete discontinuities in choice sets. As an example, students had to travel 50 kilometers, on average, to the closest college before a college was opened in their district (measured from district centroid to centroid), see Figure 5.1. Figure A4.1 in the Appendix gives an impression of the spatial variation in college availability over time.

There was an increase in the size of existing colleges and, therefore, in the number of available spots to study as well. The average number of students per college was 5,013 in 1958 and 15,438 in 1990. Of the 33 colleges in 1958, 30 still existed in 1990 and had an average size of 23,099 students. The total number of students increased from 155,000 in 1958 to 1 million in 1990. Figure 5.2 shows the trends in college openings and enrolled students (normalized by the number of inhabitants) for the five most-populated German states. While the actual numbers used in the regressions vary on the much smaller district level, the state level figures simplify the visualization of the pattern.

Factors that have driven the increase in the number of colleges and their size can briefly be summarized into four groups: (i) The large majority of the population

[5] The working paper version Kamhöfer et al. (2015) also uses the introduction of a student loan program as further source exogenous variation. Using this instrument does not affect the findings at all but is not considered here for the sake of legibility of the paper.

[6] All data are taken from the German Statistical Yearbooks, 1959-1991, see German Federal Statistical Office (1991). We only use colleges and no other higher educational institutes described in Section 5.2 (e.g., universities of applied science). Administrative data on openings and the number of students are not available for other institutions than colleges. However, since other higher educational institutions are small in size and highly specialized, they should be less relevant for the higher education decision and, thus, neglecting them should not affect the results.

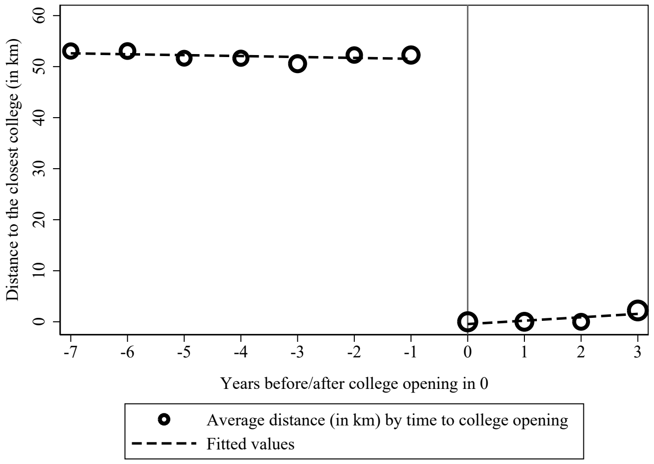


Figure 5.1: Average distance to the closest college over time for districts with a college opening

Notes: Own illustration. Information on colleges are taken from the German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, 1991). The distances (in km) between the districts are calculated using district centroids. These distances are weighted by the number of individuals observed in the particular district-year cells in our estimation sample of the NEPS-Starting Cohort 6 data. The resulting average distances are depicted by green circles. Note that prior to time period 0, the average distance changes over time either due to sample composition or a college opening in a neighboring district. Only districts with a college opening are taken into account.

had a low level of education. This did not only result from WWII but also from the "anti-intellectualism" (Picht, 1964, p.66) in the Third Reich, and the notion of education in imperial Germany before, befitting the social status of certain individuals only (ii) An increase in the number of academic secondary schools at the same time (as analyzed in Kamhöfer and Schmitz, 2016, and Jürges et al., 2011, for instance) qualified a larger share of school graduates to enroll into higher education (Bartz, 2007). (iii) A change in production technologies led to an increase in firm's demand for high-skilled workers – especially, given the low level of educational participation (Weisser, 2005). (iv) Political decision makers were afraid that "without an increase in the number of skilled graduates the West German economy would not be able to compete with communist rivals" (Jürges et al., 2011, p.846, in reference to Picht, 1964).

Although these reasons (maybe except for the firm's demand for more educated workers) affected the 10 West German federal states – that are in charge of educational policy – in the same way, the measures taken and the timing of actions differed widely between states. Because of local politics (e.g., the balancing of regional interests and avoiding clusters of colleges) there was also a large amount of variation in college openings within the federal state. See the Supplementary Materials A to the paper for a much more detailed description of the political process involved.

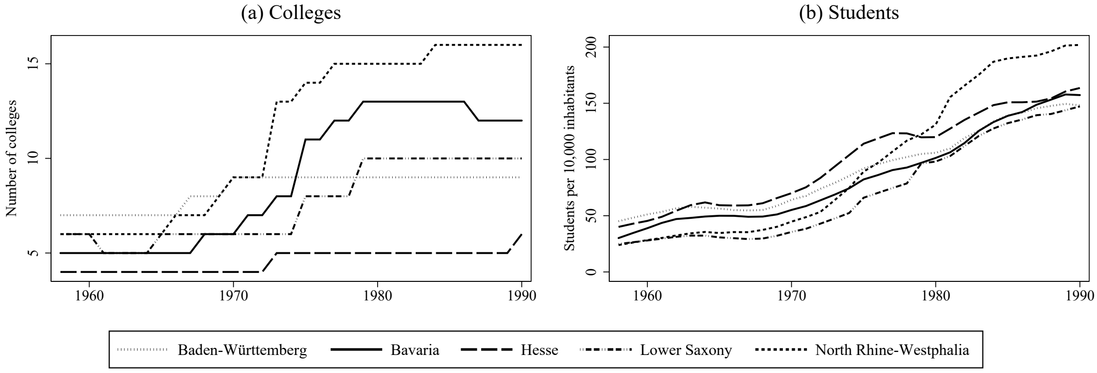


Figure 5.2: Number of colleges and students over the time in selected states

Notes: Own illustration. College opening and size information are taken from the German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, 1991). Yearly information on the district-specific population size is based on personal correspondence with the statistical offices of the federal states. For sake of lucidity the trends are only plotted for the five most-populated states.

A major concern for instrument validity is that, even though the political process did not follow a unified structure and included some randomness in the final choice of locations and timing of openings, regions where colleges were opened differed from those that already had colleges before (or that never established any). Table 5.1 reports some numbers on the regional level as of the year 1962 (the earliest possible year available to us with representative data).[7] Regions that already had colleges before did not differ in terms of socio-demographics (except for population densities, as mostly large cities had colleges before) but were somewhat stronger in terms of socio-economic indices. The differences were not large however. Given that we include district fixed-effects and a large set of socioeconomic controls (including the socioeconomic environment before the college decision, see Section 5.4), this should not be a problematic issue.

Yet, changes in district characteristics that are potentially related to the outcome variables might be a more important problem. There could, for instance, be changes in the population structure that both induce a higher demand for college education and go along with improved cognitive abilities and health. This could be the case if the regions with college openings were more "dynamic" with a younger and potentially increasing population. Table 5.1 shows a decline in the population density by 6 percent between 1962 and 1990 in the areas that opened colleges while there were no average changes in the areas with preexisting colleges and a 10 percent increase in the areas that never opened any. This reflects different regional trends in population ageing. As one example, the Ruhr Area in the west, where three colleges were opened, experienced a population decline

[7] Table 5.1 uses a different data source than the main analysis and the local level is slightly broader than districts, see the notes to the table.

Table 5.1: Comparison of regions with and without college openings before college opens using administrative data

	(1)	(2)	(3)	(4)	(5)	(6)
	College opening...					
	before 1958		between 1958-1990		later than 1990 or never	
	mean	s.d.	mean	s.d.	mean	s.d.
Observations						
Number of regions	27		30		190	
Sociodemographic characteristics						
Female (in %)	53.0	*(2.0)*	53.0	*(1.4)*	52.9	*(4.3)*
Average age (in years)	37.2	*(1.1)*	37.0	*(1.1)*	36.6	*(1.9)*
Singles (in %)	38.8	*(2.5)*	37.7	*(2.3)*	38.9	*(4.6)*
Population density per km^2 in 1962	1381.9	*(1076.7)*	1170.1	*(1047.3)*	327.1	*(479.7)*
Change in population density 1962 to 1990	1.6	*(186.3)*	-71.0	*(202.8)*	31.5	*(98.5)*
Migrational background (in %)	2.7	*(3.0)*	1.6	*(1.5)*	2.1	*(2.3)*
Socioeconomic characteristics						
Share of employees to all individuals (in %)	47.0	*(3.6)*	45.3	*(4.2)*	46.2	*(5.2)*
Employees with an income>600 DM (in %)	27.3	*(3.8)*	24.8	*(5.3)*	25.9	*(6.4)*
Employees by industry (in %)						
– primary	2.1	*(5.2)*	5.2	*(5.2)*	2.8	*(5.5)*
– secondary	52.9	*(8.4)*	54.7	*(6.2)*	54.3	*(8.9)*
– tertiary	45.0	*(9.3)*	40.1	*(8.3)*	42.9	*(9.6)*
Employees in blue collar occup. (in %)	53.6	*(9.4)*	59.0	*(7.9)*	56.5	*(9.3)*
Employees in academic occup. (in %)	22.0	*(4.4)*	17.5	*(4.3)*	20.3	*(5.9)*

Notes: Own calculations based on Micro Census 1962, see Lengerer et al. (2008). Regions are defined through administrative Regierungsbezirk entries and the degree urbanization (Gemeindegrößenklasse) and may cover more than one district. College information is aggregated at regional level and a region is considered to have a college if at least one of its districts has a college. Calculations for population density and change in population density based on district-level data acquired through personal correspondence with the statistical offices of the federal states. Data are available on request. The variables "employees in blue collar occup." and "employees in academic occup." state the shares of employees in the region in an occupation that is usually conducted by a blue collar worker/a college graduate, respectively. Standard deviations (s.d.) are given in italics in parentheses.

and comparably stronger population ageing over time. Again, these differences are not dramatically large, but we might be worried of different trends in health and cognitive abilities that are correlated with college expansion. If this was the case – more expansion in areas that have a more ageing population with deteriorating health and cognitive abilities – we might underestimate the effect of college eduction on these outcomes. We include a district-specific time trend to account for this in the analysis.

The expansion in secondary schooling noted above was unrelated to the college expansion. While college expansion naturally took place in a small number of districts, expansion in secondary schooling was across all regions. In addition, Kamhöfer and Schmitz (2016) do not find any local average treatment effects of school expansion on cognitive abilities and wages. Thus, it seems unlikely that selective increases in cognitive abilities due to secondary school expansion invalidate the instrument. Nevertheless, again, district-specific time trends should capture large parts if this was a problem.

So essentially, what we do is the following: we look within each district and attribute changes in the college (graduation/enrollment) rate from the general trend (by controlling for cohort FE) and the district specific trend (which might be due to continually increased access to higher secondary education) to either changes in the college spots or a new opening of a college nearby. We use discontinuities in college access over time that cannot be exploited using data on individuals that make the college decision at the same point in time (for instance cohort studies) as some of the previous literature that used college availability as an instrument did. Details on how we exploit the variation in college availability in the empirical specification are discussed in Section 5.4.4 after presenting the data.

5.3 Empirical Strategy

Our estimation framework widely builds on Heckman and Vytlacil (2005) and Carneiro et al. (2011). Derivations and in-depth discussion of most issues can be found there. We start with the potential outcome model, where Y^1 and Y^0 are the potential outcomes with and without treatment. The observed outcome Y either equals Y^1 in case an individual received a treatment – which is college education here – or Y^0 in the absence of treatment (the individual identifer i is implied). Obviously, treatment participation is voluntary, rendering a treatment dummy D in a simple linear regression endogenous. In the marginal treatment effect framework, this is explicitly modeled by using a choice equation, that is, we specify the following latent index model:

$$\begin{aligned}
Y^1 &= X'\beta_1 + U_1 && (5.1)\\
Y^0 &= X'\beta_0 + U_0 && (5.2)\\
D^* &= Z'\delta - V \qquad \text{where } D = \mathbf{1}[D^* \geq 0] = \mathbf{1}[Z'\delta \geq V] && (5.3)
\end{aligned}$$

The vector X contains observable, and U_1, U_0 unobservable factors that affect the potential outcomes.[8] D^* is the latent desire to take up college education which depends on observed variables Z and unobservables V. Z includes all variables in X plus the instruments. Whenever D^* exceeds a threshold (set to zero without loss of generality), the individual opts for college education, otherwise she does not. U_1, U_0, V are potentially correlated, inducing the endogeneity problem (as well as heterogenous returns) as we observe $Y(= DY^1 + (1-D)Y^0), D, X, Z$, but not U_1, U_0, V.

Following this model, individuals are indifferent between between higher education and directly entering the labor market (e.g., through an apprenticeship) whenever the index of observables $Z'\delta$ is equal to the unobservables V. Thus, if we knew the switching point (point of indifference) and its corresponding value of the observables, we could make sharp restriction on the value of the unobservables. This property is exploited in the estimation. Since for every value of the index $Z'\delta$ one needs individuals with and without higher education, it is important to meaningfully aggregate the index by a monotonous transformation that for example returns the quantiles of $Z'\delta$ and V. One such rank-preserving transformation is done by the cumulative distribution function that returns the propensity score $P(Z)$ (quantiles of Z) and U_D (quantiles of V).[9]

If we vary the excluded instruments in $Z'\delta$ from the lowest to the highest value while holding the covariates X constant, more and more individuals will select into higher education. Those who react to this shift also reveal their rank in the unobservable distribution. Thus, the unobservables are fixed given the propensity score and it is feasible to evaluate any outcome for those who select into treatment at any quantile U_D that is identified by the instrument-induced change of the higher education choice. In general, estimating marginal effects by U_D does not require stronger assumptions than those required by the LATE since Vytlacil

[8] Note that the general derivation does not require linear indices. However, it is standard to assume linearity when it comes to estimation.

[9] By applying, for instance, the standard normal distribution to the left and the right of the equation: $Z'\delta \geq V \Leftrightarrow \Phi(Z'\delta) \geq \Phi(V) \Leftrightarrow P(Z) \geq U_D$ where $P(Z) \equiv P(D=1|Z) = \Phi(Z'\delta)$.

(2002) showed its equivalence.[10] Yet, strong instruments are beneficial for robustly identifying effects over the support of $P(Z)$. This, however, is testable.

The marginal treatment effect (MTE), then, is the marginal (gross) benefit of taking the treatment for those who are just indifferent between taking and not-taking it and can be expressed as

$$MTE(x, u_D) = \frac{\partial E(Y|x,p)}{\partial p}.$$

This is the effect of an incremental increase in the propensity score on the observed outcome. The MTE varies along the line of U_D in case of heterogeneous treatment effects which arise if individuals self-select into the treatment based on their expected idiosyncratic gains. This is a situation Heckman et al. (2006) call "essential heterogeneity". This is an important structural property that the MTE can recover: If individuals already react at low values of the instrument, where the observed part of the latent desire of selecting into higher education ($P(Z)$) is still very low, a prerequisite for yet going to college is that V is marginally lower. These individuals could choose college against all (observed) odds because they are more intrinsically talented or motivated as indicated by a low V. If this is translated into higher future gains $(U_1 - U_0)$, the MTE would exhibit a significant negative slope: As $P(Z)$ rises, marginal individuals need less and less compensation in terms of unobserved and expected returns to yet choose college – this is called selection into gains. As Basu (2011, 2014) notes, essential heterogeneity is not restricted to active sorting into gains but is always an issue if selection is based on factors that are not completely independent of the gains. Thus, in health economic applications, where gains are arguably harder to predict for the individual than, say, monetary returns, essential heterogeneity is also an important phenomenon.

In this case the common treatment parameters ATE, ATT, and LATE do not coincide. The MTE can be interpreted as a more fundamental parameter than the usual ones as it unfolds all local switching effects by intrinsic 'willingness' to study and not only some weighted average of those.[11]

The main component for estimating the MTE is the conditional expectation $E(Y|X, p)$. Heckman and Vytlacil (2007) show that if we plug in the counterfactuals

[10] In this model the exclusion restriction is implicit since Z has an effect on D^* but not on Y^1, Y^0. Monotonicity is implied by the choice equation since D^* monotonously either increases are decreases the higher the values of Z.

[11] To make this explicit, all treatment parameters ($TE_j(x)$) can be decomposed into a weight ($h_j(x, u_D)$) and the MTE: $TE_j(x) = \int_0^1 MTE(x, u_D) h_j(x, u_D) du_D$. See, e.g. Heckman and Vytlacil (2007) for the exact expressions of the weights for common parameters.

in (5.1) and (5.2) in the potential outcome equation, rearrange and apply the expectation $E(.|X, p)$ to all expressions and impose an exclusion restriction of p on Y (exposed below), we get an expression that can be estimated:

$$\begin{aligned} E(Y|X, p) &= X'\beta_0 + X'(\beta_1 - \beta_0) \cdot p + E(U_1 - U_0|D = 1, X) \cdot p \\ &= X'\beta_0 + X'(\beta_1 - \beta_0) \cdot p + K(p) \end{aligned} \tag{5.4}$$

where $K(p)$ is some not further specified function of the propensity score if one wants to avoid distributional assumptions of the error terms. Thus, the estimation of the MTE involves estimating the propensity score in order to estimate Equation (5.4) and, finally, taking its derivative with respect to p. Note that this derivative – and hence the effect of college education – depends on heterogeneity due to observed components X and unobserved components $K(p)$, since this structure was imposed by Equations (5.1) and (5.2):

$$\frac{\partial E(Y|X, p)}{\partial p} = X'(\beta_1 - \beta_0) + \frac{\partial K(p)}{\partial p} \tag{5.5}$$

To achieve non-parametric identification of the terms in Equation (5.5), the Conditional Independence Assumption has to be imposed on the instrument.

$$(U_1, U_0, V) \perp\!\!\!\perp Z|X$$

meaning that the error terms are independent of Z given X. That is, after conditioning on X a shift in the instruments Z (or the single index $P(Z)$) has no effect on the potential outcome distributions.

Non-parametrically estimating separate MTEs for every data cell determined by X is hardly ever feasible due to a lack of observations and powerful instruments within each such cell. Yet, in case of parametric or semiparametric specifications a conditional independence assumption is not sufficient to decompose the effect into observed and unobserved sources of heterogeneity. To separately identify the right hand side of Equation (5.5) unconditional independence is required: $(U_1, U_0, V) \perp\!\!\!\perp Z, X$ (Carneiro et al., 2011, for more details consult the Supplementary Materials).[12]

In a pragmatic approach, one can now either follow Brinch et al. (2017) or Cornelissen et al. (2017) who do not aim at causally separating the causes of the effect heterogeneity. In this case a conventional exclusion restriction on the instruments suffices for estimating the overall level and the curvature of the MTE. Our solu-

[12] Essentially, this is equivalent to a simple 2SLS case. If one wants to identify observable effect heterogeneity (that is, interact the treatment indicator with control variables in the regression model) the instrument needs to be independent unconditional of these controls.

tion in bringing the empirical framework to the data without too strong assumptions, is to estimate marginal effects that only vary over the unobservables while fixing the X-effects at mean value. This means to deviate from (5.4) by restricting $\beta_1 = \beta_0 = \beta$ except for the intercepts α_1, α_0 in (5.1) and (5.2) such that $E(Y|X,p)$ becomes:

$$E(Y|X,p) \quad = \quad X'\beta + (\alpha_1 - \alpha_0) \cdot p + K(p) \tag{5.6}$$

Thus, we allow for different levels of potential outcomes, while we keep conditioning on X. This might look like a strong restriction at first sight but is no more different than the predominant approach in empirical economics of trying to identify average treatment effects where the treatment indicator is typically not interacted with other observables. Certainly, this does not rule out that the MTE varies by observable characteristics.

Even with the true population effects that are varying over X, note that the derivative of Equation (5.4) w.r.t. the propensity score is constant in X. Hence, only the level of the MTE changes for certain subpopulations determined by X, the curvature remains unaffected. Thus, estimation of Equation (5.6) delivers an MTE that has a level which is averaged over all subpopulations without changing the curvature. In this way all crucial elements of the MTE are preserved, since we are interested in the average effect and its heterogeneity with respect to the unobservables for the whole population. How this heterogeneity is varying for certain subpopulations is of less importance and also the literature has focused on MTEs where the X-part is averaged out. On the other hand we gain with this approach by considerably relaxing our identifying assumption from an unconditional to a conditional independence of the instrument. One advantage in not estimating heterogeneity in the observables can arise if X contains many variables that each take many different values. In this case, problems of weak instruments can inflate the results.[13]

In estimating (5.6), we follow Carneiro et al. (2010, 2011) again and use semi-parametric techniques as suggested by Robinson (1988).[14] Standard errors are

[13] On the other hand, estimating with heterogeneity in the observables can lead to an efficiency gain.

[14] Semi-parametrically, the MTE can only be identified over the support of P. The greater the variation in Z (conditional on X) and, thus $P(Z)$, the larger the range over which the MTE can be identified. This may be considered a drawback of the MTE approach, in particular, because treatment parameters that have weight unequal to zero outside the support of the propensity score are not identified using semi-parameteric techniques. This is sometimes called the "identification at infinity" requirement (see Heckman, 1990) of the MTE. However, we argue that the MTE over the support of P is already very informative. We use semi-parametric estimates of the MTE and restrict the results to the empirical ATE or ATT that are identified for those individuals who are in the sample (see Basu et al., 2007). Alternatively one might use a flexible approximation of $K(p)$

clustered at the district level and were generated by bootstrapping the entire procedure using 200 replications.

5.4 Data

5.4.1 Sample selection and college education

Our main data source are individual level data from the German National Educational Panel Study (NEPS), see Blossfeld et al. (2011). The NEPS data map the educational trajectories of more than 60,000 individuals in total. The data set consists of a multi-cohort sequence design and covers six age groups, called "starting cohorts": newborns and their parents, pre-school children, children in school grades 5 and 9, college freshmen students, and adults. Within each starting cohort the data are organized in a longitudinal manner, i.e., individuals are interviewed repeatedly. For each starting cohort, the interviews cover extensive information on competence development, learning environments, educational decisions, migrational background, and socioeconomic outcomes.

We aim at analyzing longer term effects of college education and, therefore, restrict the analysis to the "adults starting cohort". For this age group six waves are available with interviews conducted between 2007/2008 (wave 1) and 2013 (wave 6), see LIfBi (2015). Moreover, the NEPS includes detailed retrospective information on the educational and occupational history as well as the living conditions at the age of 15 – about three years before individuals decide for higher education. From the originally 17,000 respondents in the adults starting cohort, born between 1944 and 1989, we exclude observations for four reasons: First, we focus on individuals from West Germany due to the different educational system in the former German Democratic Republic (GDR), thereby dropping 3,500 individuals living in the GDR at the age of the college decision. Second, to allow for long term effects we make a cut-off at college attendance before 1990 and drop 2,800 individuals who graduated from secondary school in 1990 or later. Third, we drop 1,000 individuals with missing geographic information. An attractive (and for our analysis necessary) feature of the NEPS data is that they include information on the district (German *Kreis*) of residence during secondary schooling which is used in assigning the instrument in the selection equation. The fourth reason for losing observations is that the dependent variables are not available for

based on a polynomial of the propensity score as done by Basu et al. (2007). This amounts to estimating $E(Y|X,p) = X'\beta + (\alpha_1 - \alpha_0) \cdot p + \sum_{j=1}^{k} \phi_j p^j$ by OLS and using the estimated coefficients to calculate $\widehat{MTE}(x,p) = (\widehat{\alpha}_1 - \widehat{\alpha}_0) + \sum_{j=1}^{k} \widehat{\phi}_j j p^{j-1}$.

each respondent, see below. Our final sample includes between 2,904 and 4,813 individuals, depending on the outcome variable.

The explanatory variable "college degree" takes on the value 1 if an individual has any higher educational degree, and 0 otherwise. Dropouts are treated as all other individuals without college education. More than one fourth of the sample has a college degree, while three fourths do not.

5.4.2 Dependent variables

Wages

The data set covers a wide range of individual employment information such as monthly income and weekly hours worked. We calculate the hourly gross wage for 2013 (wave 6) by dividing the monthly gross labor market income by the actual weekly working hours (including extra hours) times the average number of weeks per month, 4.3. A similar strategy is, e.g., applied by Pischke and von Wachter (2008) to calculate hourly wages using German data.

For this outcome variable, we restrict our sample to individuals in working age up to 65 years and drop observations with hourly wages below 5 Euros and above the 99th quantile (77.52 Euros) as this might result from misreporting. Table 5.2 reports descriptive statistics and reveals considerably higher hourly wages for individuals with college degree. The full distribution of wages (and the other outcomes) for both groups is shown in Figure **??** in the Appendix. In the regression analysis we use log gross hourly wages.

Health

Two variables from the health domain are used as outcome measures: the Physical Health Component Summary Score (PCS) and the Mental Health Component Summary Score (MCS), both from 2011/2012 (wave 4).[15] These summary scores are based on the SF12 questionnaire, which is an internationally standardized set of 12 items regarding eight dimensions of the individual health status. The eight dimensions comprise physical functioning, physical role functioning, bodily pain, general health perceptions, vitality, social role functioning, emotional role functioning and mental health. A scale ranging from 0 to 100 is calculated for each of these eight dimensions. The eight dimensions or subscales are then aggregated to the two main dimensions mental and physical health, using explorative factor analysis (Andersen et al., 2007). For our regression analysis, we standardize the aggregated scales (MCS and PCS) to have mean 0 and standard

[15] The working paper version also considers health satisfaction with results very similar to PCS (Kamhöfer et al., 2015).

Table 5.2: Descriptive statistics dependent variables

	(1)	(2)	(3)	(4)	(5)	(6)
	Gross hourly wage	Health measure		Cognitive ability component		
		PCS	MCS	Read. speed	Read. comp.	Math liter.
Observations	3,378	4,813	4,813	3,995	4,576	2,904
with college degree (in %)	31.0	28.1	28.1	27.8	28.1	28.0
Raw values						
Mean with degree	27.95	53.31	51.15	39.69	29.76	13.37
Mean without degree	19.35	50.39	50.53	35.99	22.75	9.36
Maximum possible value	– –[a]	100	100	51	39	22
Transformed values						
Mean with degree	3.25	0.23	0.04	0.32	0.63	0.61
Mean without degree	2.88	−0.09	−0.02	−0.12	−0.25	−0.24

Notes: Own calculations based on NEPS-Starting Cohort 6 data. Gross hourly wage given in Euros. Gross hourly wage is transformed to its log value, the other variables are transformed in units of standard deviation with mean 0 and standard deviation 1.

[a] The gross hourly wage is truncated below at 5 Euros and above at the highest quantile (77.52 Euros).

deviation 1, where higher values indicate better health. Columns (2) to (3) of Table 5.2 report sample means of the health measures across individuals by college graduation. Those with college degree have, on average, a better physical health score. With respect to mental health, both groups differ only marginally.

Cognitive abilities

Cognitive abilities summarize the "ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought" (American Psychological Association, 1995), where the sum of these abilities is referred to as intelligence. Psychologists distinguish several concepts of intelligence with different cognitive abilities. However, they all include measures of verbal comprehension, memory and recall as well as processing speed.

Although comprehensive cognitive intelligence tests take hours, a growing number of socioeconomic surveys includes much shorter proxies that measure specific skill components. The short ability tests are usually designed by psychologists and the results are highly correlated with the results of more comprehensive intelligence tests (cf. Lang et al., 2007b, for a comparison of cognitive skill tests in the German Socio-economic Panel with larger psychological test batteries). The NEPS includes three kinds of competence tests which cover various domains of cognitive functioning: reading speed, reading competence, and mathematical competence.[16] All competence tests were conducted once in 2010/2011 (wave 3) or 2012/2013 (wave 5), respectively, as paper and pencil tests under the supervision of a trained interviewer and the test language was German.

The first test measures reading speed.[17] The participants receive a booklet consisting of 51 short true-or-false questions and the test duration is 2 minutes. Each question has between 5 and 18 words. The participants have to answer as many questions as possible in the given window. The test score is the number of correct answers. Since the test aims at the answering speed, the questions only deal with general knowledge and use easy language. One question/statement, for example, reads "There is a bath tub in every garage." The mean number of correct answers in our estimation sample is 39.69 (out of 51) for college graduates and 35.99 for others, see Table 5.2. For more information, see Zimmermann et al. (2014).

The reading competence test measures understanding of texts. It lasts 28 minutes and covers 32 items. The test consists of three different tasks. First, participants have to answer multiple choice questions about the content of a text, where only one out of four possible answers is right. In a decision-making task, the participants are asked whether statements are right or wrong according to the text. In a third task, participants need to assign possible titles out of a list to sections of the text. The test includes several types of texts, e.g., comments, instructions, and advertising texts (LIfBi, 2011). Again, the test score reflects the number of correct answers. Participants with college degree score on average 29.76 and without 22.75 (out of 39).[18]

The mathematical literacy test evaluates "recognizing and [...] applying [of] mathematics in realistic, mainly extra-mathematical situations" (LIfBi, 2011, p.8). The test has 22 items and takes 28 minutes. It follows the principle of the OECD-PISA tests and consists of the areas quantity, space and shape, change and relations, as well as data and change, and measures the cognitive competencies in the areas of application of skills, modelling, arguing, communicating, representing, as well as problem solving; see LIfBi (2011). Individuals without college degree score on average 9.36 (out of 22) and persons who graduated from college receive 4 points more.

[16] For a general overview over test designs and applications in the NEPS, see Weinert et al. (2011).

[17] The test measures the "assessment of automatized reading processes", where a "low degree of automation in decoding [...] will hinder the comprehension process", i.e., understanding of texts (Zimmermann et al., 2014, p.1). The test was newly designed for NEPS but based on the well-established Salzburg reading screening test design principles (LIfBi, 2011).

[18] The total number of possible points exceeds 32 because some items were worth more than one point.

Due to the rather long test duration given the total interview time, not every respondent had to do all three tests. Similarly to the OECD-PISA tests for high school students, individuals were randomly assigned a booklet with either all three or two out of the three tests. 3,995 individuals did the reading speed test, 4,576 the reading competence test, and 2,904 math. Since the tests measure different competencies that refer to distinct cognitive abilities, we may not combine the different test scores into an overall score but give the results separately (see Anderson, 2007).

5.4.3 Control variables

Individuals in our sample made their college decision between 1958 and 1990. The NEPS allows us to consider important socioeconomic characteristics that probably affect both the college education decision as well as the outcomes today (variables denoted with X in Section 5.3). This is *general demographic information* such gender, migrational background, and family structure, *parental characteristics* like parent's educational background. Moreover, we include two blocks of controls that were determined before the educational decision was made. *Pre-college living conditions* include family structure, parental job situation and household income at the age of 15, while *pre-college education* includes educational achievements (number of repeated grades and secondary school graduation mark).

Table A4.1 in the Appendix provides more detailed descriptions of all variables and reports the sample means by treatment status. Apart from higher wages, abilities and a better physical health status (as seen in Table 5.2), individuals with a college degree are more likely to be males from an urban district without a migrational background. Moreover, they are more likely to have healthy parents (in terms of mortality). Other variables seem to differ less between both groups. We also account for cohort effects of mother and father, district fixed effects as well as district-specific time trends (see Mazumder, 2008, and Stephens and Yang, 2014, for the importance of the latter).

5.4.4 Instrument

The processes of college expansion discussed in Section 5.2.2 probably shifted individuals also with a lower desire to study into college education. Such powerful exogenous variation is beneficial for our approach as we try to identify the MTE along the distribution of the desire to study. We assign each individual the college availability as instrument (that is, a variable in Z but not in X). In doing so, we use the information on the district of the secondary school graduation and the year of the college decision, which is the year of secondary school graduation. The district – there are 326 districts in West Germany – is either a city or a certain rural area.

The question is how to exploit the regional variation in openings and spots most efficiently as it is almost infeasible to control for all distances to all colleges simultaneously. Our approach to this question is to create an index that best reflects the educational environment in Germany and combines the distance with the number of college spots:

$$Z_{it} = \sum_{j}^{326} K(dist_{ij}) \times \left(\frac{\#students_{jt}}{\#inhabitants_{jt}} \right). \tag{5.7}$$

The college availability instrument Z_{it} basically includes the total number of college spots (measured by the number of students) per inhabitant in district j (out of the 326 districts in total) individual i faces in year t weighted by the distance between i's home district and district j. Weighting the number of students by the population of the district takes into account that districts with the same number of inhabitants might have colleges of a different size. This local availability is then weighted by the Gaussian kernel distance $K(dist_j)$ between the centroid of the home district and the centroid of district j. The kernel puts a lot of weight to close colleges and a very small weight to distant ones. Since individuals can choose between many districts with colleges, we calculate the sum of all district-specific college availabilities within the kernel bandwidth. Using a bandwidth of 250km, this basically amounts to $K(dist_j) = \phi(dist_j/250)$ where ϕ is the standard normal pdf. While 250km sounds like a large bandwidth, this implies that colleges in the same district receive a weight of 0.4, while the weight for colleges that are 100km away is 0.37, but it is reduced to 0.24 for 250km. Colleges that are 500km away only get a very low weight of 0.05. A smaller bandwidth of, say, 100km would mean that already colleges that are 250km away receive a weight of 0.02 which implies the assumption that individuals basically do not take them into account at all. Most likely this does not reflect actual behavior. As a robustness check, however, we carry out all estimations with bandwidths between 100 and 250km and the results are remarkably stable, see Figure S.C1 in the Supplementary Materials. Table 5.3 presents the descriptive statistics. We also provide background information on certain descriptive measures on distance and student density.

The instrument jointly uses college openings and increases in size. Size is measured in enrollment as there is no available information on actual college spots.

Table 5.3: Descriptive statistics of instruments and background information

	(1)	(2)	(3)	(4)
	Statistics			
	Mean	SD	Min	Max
Instrument: College availability	0.459	0.262	0.046	1.131
Background information on college availability (implicitly included in the instrument)				
Distance to nearest college	27.580	26.184	0	172.269
At least one college in district	0.130	0.337	0	1
Colleges within 100km	5.860	3.401	0	16
College spots per inhabitant within 100km	0.034	0.019	0	0.166

Notes: Own calculations based on NEPS-Starting Cohort 6 data and German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, 1991). Distances are calculated as the Euclidean distance between two respective district centroids.

This might be considered worrisome as enrollment might reflect demand factors that are potentially endogenous. While we believe that this is not a major problem as most study programs in the colleges where used to capacity, we also, as a robustness check, neglect information on enrollment and merely exploit information on college openings by using

$$Z_{it} = \sum_{j}^{326} K(dist_{ij}) \times \mathbb{1}[\text{college avaiable}_{jt}] \tag{5.8}$$

where $\mathbb{1}[\cdot]$ is the indicator function. The results when using this instrument are comparable, with minor differences, to those from the baseline specification as shown in Figure A4.3 in the Appendix. Certainly, the overall findings and conclusions are not affected by this choice. We prefer the combined instrument as this uses information from both aspects of the educational expansion.

5.5 Results

5.5.1 OLS

Although we are primarily interested in analyzing the returns to college education for the marginal individuals, we start with ordinary least squares (OLS) estimations as a benchmark. Column (1) in Table 5.4, Panel A, reports results for hourly wages, columns (2) and (3) for the two health measures, while columns (4) to (6) do the same for the three measures of cognitive abilities. Each cell reports the coefficient of college education from a separate regression. After conditioning on observables, individuals with a college degree earn approximately 28 percent higher wages, on average. While PCS is higher by around 0.3 of a standard deviation – recall that all outcomes but wages are standardized –, there is no significant relation with MCS. Individuals with a college degree read, on average, 0.4 SD faster than those without college education. Moreover, they approximately have a by 0.7 SD better text understanding and mathematical literacy. All in all, the results are pretty much in line with the differences in standardized means as shown in Table 5.2, slightly attenuated, however, due to the inclusion of control variables.

Table 5.4: Regression results for OLS and first stage estimations

	(1)	(2)	(3)	(4)	(5)	(6)
	Gross hourly wage	Health measure		Cognitive ability component		
		PCS	MCS	Read. speed	Read. comp.	Math liter.
Panel A: OLS results						
College degree	0.277***	0.277***	0.003	0.398***	0.729***	0.653***
	(0.019)	(0.033)	(0.036)	(0.037)	(0.032)	(0.044)
Panel B: 2SLS first-stage results						
College availability	2.368***	2.576***	2.576***	2.521***	2.327***	2.454***
	(0.132)	(0.122)	(0.122)	(0.132)	(0.119)	(0.159)
Observations	3,378	4,813	4,813	3,995	4,576	2,904

Notes: Own calculations based on NEPS-Starting Cohort 6 data. Regressions also include a full set of control variables as well as year-of-birth and district fixed effects, and district-specific linear trends. District clustered standard errors in parentheses; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Panel B of Table 5.4 reports the first stage results of the 2SLS estimations. The coefficients of the instrument point into the expected direction and are individually significant. As to be expected, they barely change across the outcome variables (as the first-stage specifications only differ in the number of observations across the columns).

In order to get a feeling for the effect size of college availability in the first-stage, we consider, as an example, the college opening in the city of Essen in 1972. In 1978, about 11,000 students studied there. To illustrate the effect of the opening, we assume a constant population size of 700,000 inhabitants. The kernel weight of new spots in the same district is 0.4 ($= K(0)$). According to Equation (5.7), the instrument value increases by 0.006 (rounded). Given the coefficient of college availability of 2.4, an individual who made the college decision in Essen in 1978 had a 1.44 percentage points higher probability to go to college due to the opening of the college in Essen (compared to an individual who made the college decision in 1971). This seems to be a plausible effect. The effect of the college opening in Essen on individuals who live in districts other than Essen is smaller, depending on the distance to Essen.

5.5.2 Marginal treatment effects

Figure 5.3a shows the distribution of the propensity scores used in estimating the MTE by treatment and control group. They are obtained by logit regressions of the college degree on all Z and X variables. Full regression results of the first and the second stage of the 2SLS estimations are reported in the Supplementary Materials. For both groups, the propensity score varies from 0 to about 1. Moreover, there is a common support of the propensity score almost on the unit interval. Variation in the propensity score where the effects of the X variables are integrated out is used to identify local effects.

This variation is presented in Figure 5.3b. It shows the conditional support of P when the influence of the linear X-index of observables on the propensity score is integrated out ($\int f_{P(Z,X)} dX$). Here, the support ranges nearly from 0 to 0.8 only caused by variation in the instrument – the identifying variation. This is important in the semiparametric estimation since it shows the regions in which we can credibly identify (conditional on our assumptions) marginal effects without having to rely on inter- or extrapolations to regions where we do not have identifying variation.

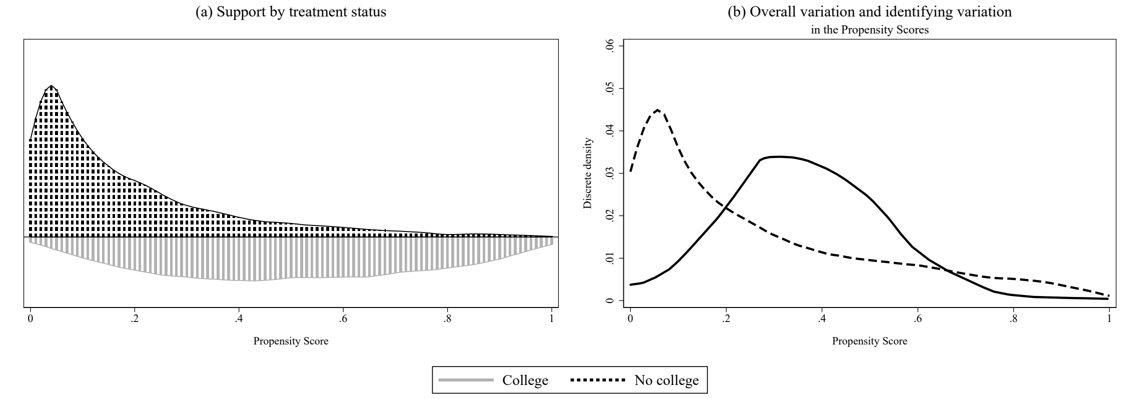


Figure 5.3: Distribution of propensity scores

Notes: Own illustration based on NEPS-Starting Cohort 6 data. The left panel shows the propensity score (PS) density by treatment status. The right panel illustrates the joint PS density (dashed line). The solid line shows the PS variation solely caused by variation in Z, since the X-effects have been integrated out. Further note that in the right panel the densities were both normalized such that they sum up to one over the 250 points where we evaluate the density.

We calculate the MTE using a local linear regression with a bandwidth that ranges from 0.10 to 0.16 depending on the outcome variable.[19] We calculate the marginal effects along the quantiles U_D by evaluating the derivative of the treatment effect with respect to the propensity score (see Equation (5.6) in Section 5.3).

Figure 5.4 shows the MTE for all outcome variables. The upper left panel presents the MTE for wages. We find that individuals with low values of U_D have the highest monetary returns to college education. Low values of U_D mean that these are the individuals who are very likely to study as already small values of $P(z)$ exceed U_D, see the transformed choice equation in Section 5.3. The returns are close to 80 percent for the smallest values of U_D and then approach 0 at $U_D \approx 0.7$. Thus, we tend to interpret these findings as clear and strong positive returns for the 70 percent of individuals with the highest desire to study, while there is no clear evidence for any returns for the remaining 30 percent. Hence, there is obviously selection into gains with respect to wages, where individuals with higher (realized) returns self-select into more education. This reflects the notion that individuals make choices based on their expected gains.

The curve of marginal treatment effects resembles the one found by Carneiro et al. (2011) for the US with the main difference that we do not find negative effects (but just zero) for a part of the distribution. The effect sizes are also comparable although ours are somewhat smaller. For instance, Carneiro et al. (2011) find highest returns of 28 percent per year of college, while we find 80 percent for the college degree which, on average, takes 4.5 years to be earned.

What could explain these wage returns? Two potential channels of higher earnings could be better cognitive skills and/or better health due to increased education. The findings on skills and health that we discuss in the following could, thus, be read as investigations into mechanisms for the positive wage returns. However, at least for health, this would only be one potential interpretation as health might also be directly affected by income.

The right column of Figure 5.4 plots the results for cognitive skills. The distribution of marginal treatment effects is remarkably similar to the one for wages. We see that, also in terms of cognitive skills, not everybody benefits from more education. Some individuals, again those with high desire to study, strongly benefit, while the effects approach zero for individuals with $U_D > 0.6$. This holds for reading speed, reading competence, as well as mathematical literacy. The largest returns are as high as 2 to 3 standard deviations, again, for the small group with

[19] We assess the optimal bandwidth in the local linear regression using Stata's `lpoly` rule of thumb. Our results are also robust to the inclusion of higher order polynomials in the local (polynomial) regression. The optimal, exact bandwidths are: wage 0.10, PCS 0.13, MCS 0.16, reading competence 0.10, for reading speed 0.11, math score 0.12.

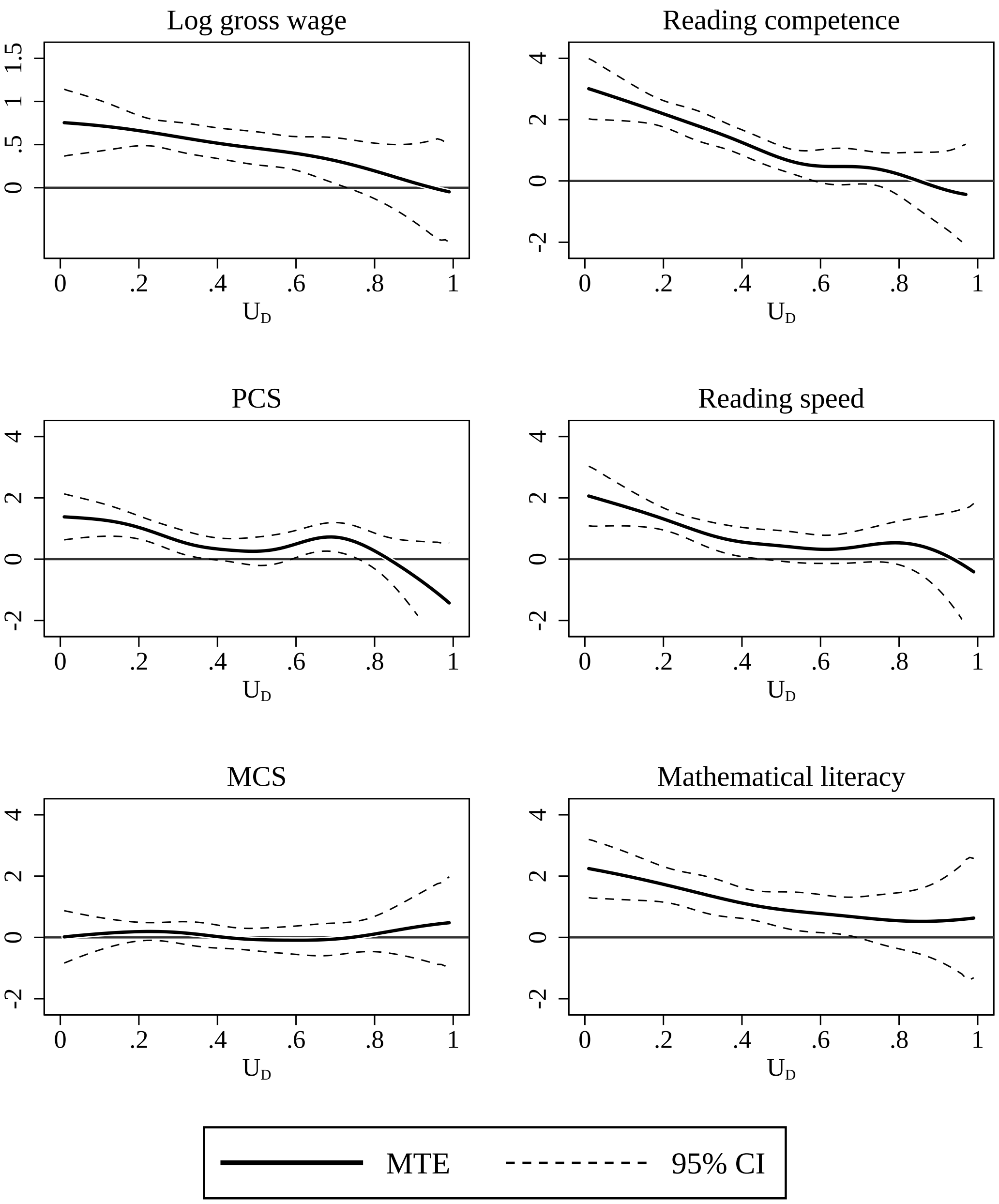


Figure 5.4: Marginal Treatment Effects for cognitive abilities and health

Notes: Own illustration based on NEPS-Starting Cohort 6 data. For gross hourly wage, the log value is taken. Health and cognitive skill outcomes are standardized to mean 0 and standard deviation 1. The MTE (vertical axis) is measured in logs for wage and in units of standard deviations of the health and cognitive skill outcomes. The dashed lines give the 95 percent confidence intervals based on clustered bootstrapped standard errors with 200 replications. Calculations based on a local linear regression where the influence of the control variables was isolated using a semiparametric Robinson estimator (Robinson, 1988) for each outcome variable. The optimal, exact bandwidths for the local linear regressions are: for wage 0.10, PCS 0.13, MCS 0.16, reading competence 0.10, for reading speed 0.11, math score 0.12.

highest college readiness only. Thus, we observe the same selection into gains as with wages and the findings could be interpreted as returns to cognitive abilities from education being a potential pathway for positive earnings returns.

The findings are somewhat different for health, as seen in the lower left part of Figure 5.4. First of all, the returns are much more homogeneous then those for wages and skills. While there is still some heterogeneity in returns to physical health (though to a smaller degree than before) returns are completely homogeneous for mental health. Moreover, the returns are zero throughout for mental health. Physical health effects are positive (although not always statistically significant) for around 75 percent of the individuals while they are close zero for the 25 percent with the lowest desire to study.

The main findings of this paper can be summarized as follows:

- Education leads to higher wages and cognitive abilities for the same approx. 60 percent of individuals. This can also be read as suggestive evidence for cognitive abilities being a channel for the effect of education on wages.
- Education does not pay off for everybody. However, in no case are the effects negative. Thus, education does never harm in terms of gross wages, skills and health. (Obviously, this view only considers potential benefits and disregards costs - thus, net benefits might well be negative for some individuals.)
- There are clear signs of selection into gains. Those individuals who realize the highest returns to education are those who are most ready to take it.

With policy initiatives such as the "Higher Education Pact 2020" Germany continuously increases participation in higher education in order to meet OECD standards (see OECD, 2015b,a). Our results imply that this might not pay off, at least in terms of productivity (measured by wages), cognitive abilities, and health. Without fully simulating the results of further increased numbers of students in Germany, it is save to assume that additional students would be those with higher values of U_D as those with the high desire to study are in large parts already enrolled. But these additional students are the ones that do not seem to benefit from college education. However, this projection needs to be taken with a grain of salt as our findings are based on education in the 1960s to 1980s and current education might yield different effects.

We carry out two kinds of robustness checks with respect to the definition of the instrument (see Section 5.4.4). Figure A4.3 in the Appendix reports the findings when the instrument definition does not consider the increases in college size. The MTE curves do not exactly stay the same as before but the main conclusions are unchanged. Wage returns are slightly more homogeneous. The results for reading competence and mathematical literacy are virtually the same while for reading speed homogeneously positive effects are found. However, the confidence bands of the curves for both definitions of the instrument widely overlap. This also holds for the health measures. The MTE curve for MCS is slightly shifted upwards and the one for PCS is more homogeneous but the difference in the curves across both kinds of instruments are not significant. While the likelihood that two valid instruments exactly deliver the same results is fairly low in any application (and basically zero when so many points are evaluated as is the case here), the broad picture that leads to the conclusions above is invariant to the change in the instrument definition.

In the Supplementary Materials C, we report the results of robustness check where we use different kernel bandwidths to weight the college distance (bandwidths between 100km and 250km). Here the differences are indeed widely absent. Although the condensation of college availability in Equation (5.7) seems somewhat arbitrary, these robustness checks show that the specification of the instrument does not affect our conclusions.

5.5.3 Treatment parameters

Table 5.5 reports the conventional treatment parameters estimated using the MTE and the respective weights as described above and more formally derived and explained in, for example, Heckman et al. (2006). In particular, we calculate the average treatment effect (ATE), the average treatment effect on the treated (ATT), the average treatment effect on the untreated (ATU) and the local average treatment effect (LATE). The estimated weights applied to the returns for each U_D on the MTE curve are shown in Figure 5.5. Whereas the local average treatment effect is an average effect weighted by the conditional density of the instrument, the ATT (vice versa for the ATU) for example gives more weight to those individuals that select already into higher education at low U_D values (indicating low intrinsic reluctance for higher education). The reason is that their likelihood of being in any ‘treatment group’ is higher compared to individuals with higher values of U_D. The ATE places equal weight over the whole support.

In all cases but mental health and reading speed, the LATE parameters in column (4) approximately double compared to the OLS estimates. Increasing local average treatment effects (compared to OLS) seem to be counterintuitive as one often expects OLS to overestimate the true effects. Yet, this is not an uncommon finding and in a world with heterogeneous effects often explained by the group of compliers that potentially has higher individual treatment effects than the average individual (Card, 2001). This is directly obvious by comparing the LATE

Table 5.5: Estimated treatment parameters for main results

	(1)	(2)	(3)	(4)
	Treatment parameter			
	ATE	*ATT*	*ATU*	*LATE*
Main outcomes:				
Log gross wage	0.43	0.59	0.36	0.49
	(0.06)	(0.07)	(0.07)	(0.05)
PCS	0.45	0.86	0.29	0.55
	(0.13)	(0.13)	(0.16)	(0.09)
MCS	0.10	0.09	0.10	0.05
	(0.10)	(0.12)	(0.13)	(0.08)
Reading competence	1.10	1.88	0.78	1.18
	(0.13)	(0.15)	(0.16)	(0.08)
Reading speed	0.72	1.17	0.54	0.70
	(0.14)	(0.15)	(0.18)	(0.11)
Mathematical literacy	1.11	1.56	0.93	1.13
	(0.17)	(0.21)	(0.19)	(0.14)

Notes: Own calculations based on NEPS-Starting Cohort 6 data. The MTE is estimated with a semiparametric Robinson estimator. The LATE is estimated using the IV weights depicted in Figure 5.5. Therefore, the LATE in this table deviates slightly from corresponding 2SLS estimates. Standard error estimated using a clustered bootstrap (at district level) with 200 replications.

to column (1) which is another indication of selection into gains. Regarding the other treatment parameters, the LATE lies within the range of the ATT and the ATU.

Note that these are the "empirical", conditional-on-the-sample parameters as calculated in Basu et al. (2007), that is, the treatment parameters conditional on the common support of the propensity score. The population ATE, however, would require full support on the unity interval.[20] As depicted in Figure 5.3, we do not have full support in the data at hand. Although we observe individuals with and without college degree for most probabilities to study, we cannot observe an individual with a probability arbitrarily close to 100 percent without college degree (and arbitrarily close to 0 percent with a degree). Instead, the parameters in Table 5.5 were computed using the marginal treatment effects on the common support only. However, as this reaches from 0.002 to 0.969 it seems fair to say that this probably comes very close to the true parameters.

[20] The ATT would require for every college graduate in the population a non-graduate with the same propensity score (including 0 percent). For the ATU one would need the opposite: a graduate for every non-graduate with the same Propensity Score including 100 percent.

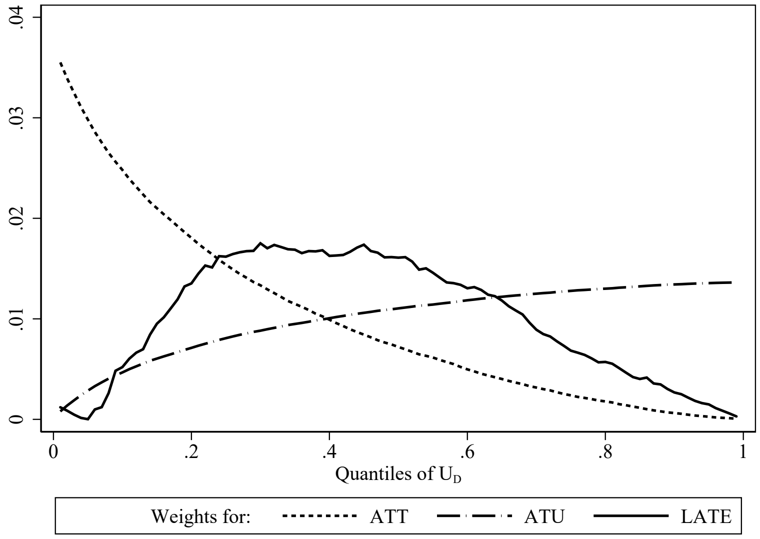


Figure 5.5: Treatment parameter weights conditional on the propensity score

Notes: Own illustration based on NEPS-Starting Cohort 6 data. Weights were calculated using the entire sample of 8,672 observations for that we have instrument and control variable information in spite of availability of the outcome variable.

Table 5.5 is informative in particular for two reasons. First, it boils down the MTE to single numbers such that the average effect size immediately becomes clear. And, second, differences between the parameters again emphasize the role of effect heterogeneity. Together with the bootstrapped standard errors the table reveals that the ATT and the ATU structurally differ from each other for all outcomes but mental health. Hence, the treatment group of college graduates seems to benefit from higher education in terms of wages, skills, and physical health compared to the non-graduates. One reason is that they might choose to study because of their idiosyncratic skill returns. Yet, it is also likely to be windfall gains that go along with monetary college premiums that the decision was more likely to be based on. Nonetheless, this also is evidence for selection into gains.

The effect sizes for all (ATE), for the university degree subgroup (ATT), and for those without higher education (ATU) in Table 5.5 capture the overall returns to college education, not the per-year effects. On average, the per-year effect is approximately the overall effect divided by 4.5 years (the regular time it takes to receive a *Diplom* degree), if we assume linear additivity of the yearly effects. The per-year effects for mathematical literacy and reading competence are about 25 percent of a standard deviation for all parameters. For reading speed the effects are around 15 percent of an SD, while the wage effects are around 10 percent. These effects are of considerable size, yet slightly smaller than those found in the previous literature on different treatments and, importantly, different compliers.

For instance, ability returns to an additional year of compulsory schooling were found to be up to 0.5 SD (see, e.g., Banks and Mazzonna, 2012).

To get an idea of the total effect of college education on, say, math skills, the following example might help. If you start at the median of the standardized unconditional math score distribution ($\Phi(0) = 50$) percent, the average effect of 1.11 of a standard deviation, all other things the same, will make you end up at the 87 percent quantile of that distribution ($\Phi(0 + 1.11) = 87$) percent – in the thought experiment of being the only treated in the peer group.

As suggested by the pattern of the marginal treatment effects in Figure 5.4, the health returns to higher education are smaller than the skill returns, still they are around 10 percent of an SD per year (except for the zero effect on mental health). Given the previous literature, the results seem reasonable.

Regarding statistical significance of the effects, note that we use several outcome variables and potentially run into multiple testing problems. Yet, we refrain from taking this into account by a complex algorithm that also accounts for the correlation of the six outcome variables and argue the following way: All ATEs and ATTs are highly statistically significant. Thus, our multiple testing procedure with six outcomes should not be a major issue. Even with a most conservative Bonferroni correction, critical values for statistical significance at the 5 percent level would increase from 1.96 to 2.65 and would not change any conclusions regarding significance.[21]

5.6 Potential mechanisms for health and cognitive abilities

In this section, we investigate the role of potential mechanisms through which college education may work. It is likely to affect the observed level of health and cognitive abilities through the attained stock of health capital and the cognitive reserve – the mind's ability to tolerate brain damage (Stern, 2012; Meng and D'Arcy, 2012).

There are probably three channels through which education affects long-run health and cognitive abilities:

- in college: a direct effect from education;

[21] Also taking into account the outcomes from Section 6 and assuming that we test 18 times would increase the critical value to 2.98 in the (overly conservative) Bonferroni-correction.

- post-college: a diminished age-related decline in health and skills due to the higher health capital/cognitive reserve attained in college (e.g., the "cognitive reserve hypothesis", Stern et al., 1999);

- post-college: different health behavior or different jobs that are less detrimental to health and more cognitively demanding (Stern, 2012).

The post-college mechanisms that compensate for the decline also contain implicit multiplying factors like complementarities and self-productivity, see Cunha et al. (2006) and Cunha and Heckman (2007). The NEPS data include various job characteristics and health behaviors that potentially reduce the age-related skill/health decline. However, the data neither allow us to disentangle these components empirically (i.e., observing changes in one channel that are exogenous from other channels) nor to analyze how the effect on the mechanism causally maps into higher skills or better health (as for example in Heckman et al., 2013). Thus, it should be noted that this sub-analysis is merely suggestive and by no means a comprehensive analysis on the mechanisms of the effects found in the previous section. Moreover, the following analysis focusses on the potential channel of different jobs and health behavior. It does the same as before (same controls, same estimation strategy and instrument) but replaces the outcome variables by the indicators of potential mechanisms.

Cognitive abilities

The main driving force behind skill formation after college might lie in activities on the job. When individuals with college education engage in more cognitively demanding activities, e.g., more sophisticated jobs, this might mentally exercise their minds (Rohwedder and Willis, 2010). This effect of mental training is sometimes referred to as use-it-or-lose-it hypothesis, see Rohwedder and Willis (2010) or Salthouse (2006). If such an exercise effect leads to alternating brain networks that "may compensate for the pathological disruption of preexisting networks" (Meng and D'Arcy, 2012, p.2), a higher demand for cognitively demanding tasks (as a result of college education) increases the individual's cognitive capacity.

In order to investigate if a more cognitively demanding job might be a potential mechanism (as, e.g., suggested by Fisher et al., 2014), we use information on the individual's activities on the job. All four outcome variables considered in this subsection are binary, their definitions, sample means effects of college education are given in Table 5.6. For the sake of brevity we focus on the most relevant treatment parameters here and do not discuss the MTE curvatures.

College education has strong effects on all four outcomes. It increases the likelihood to be in a job that requires calculating with percentages and fractions, that involves reading or writing and in which individuals often learn new things. The effect sizes are very large which is not too surprising as many of the jobs that entail these mentally demanding tasks require a college diploma as a quasi-formal condition of employment.

Table 5.6: Potential mechanisms for cognitive skills

	Definition	Sample	Parameter		
		mean	ATE	ATT	ATU
Math: percentages	=1 if job requires calculating with percentages and fractions	0.711	0.20	0.23	0.19
			(0.06)	(0.07)	(0.07)
Reading	=1 if respondent often spends more than 2 hours reading	0.777	0.23	0.30	0.30
			(0.03)	(0.03)	(0.04)
Writing	=1 if respondent often writes more than 1 page	0.704	0.39	0.64	0.29
			(0.07)	(0.09)	(0.07)
Learning new things	=1 if respondent reports to learn new things often	0.671	0.22	0.31	0.18
			(0.07)	(0.09)	(0.07)

Notes: Own calculations based on NEPS-Starting Cohort 6 data. Definitions are taken from the data manual. Standard error estimated using a clustered bootstrap (district level) and reported in parentheses.

Moreover, as observed before, there seems to be effect heterogeneity here as well and selection into gains as all average treatment effects on the treated are larger than the treatment effects on the untreated (except for the case of reading more than two hours). The differences are particularly strong for writing and for learning new things. All in all, the findings suggest that cognitively more demanding jobs due to college education might play a role in explaining long-run cognitive returns to education. Note again, however, that these findings are only suggestive evidence for a causal mechanism. It might as well be that it is the other way around and the cognitive abilities attained in college induce a selection into these job types.

Health

Concerning the health mechanisms, we study job-related effects and effects on health behavior. The NEPS data cover engagement in several physical activities on the job, e.g.,: working in a standing position, working in an uncomfortable position (like bending often), walking or cycling long distances, or carrying heavy loads. Table 5.7 reports definitions, sample means and effects. The binary indicators are coded as 1 if the respondent reports to engage in the activity (and 0 otherwise) in the upper panel of the table.

We find that college education reduces the probability of engaging in all four physically demanding activities. Again, the estimated effects are very large in size, implying that it is the college diploma that qualifies for a white-collar office-job position. These effects might explain why we find physical health effects of education and are in line with the absence of mental health effects. White-collar jobs are usually less demanding with respect to physical health but not at all less stressful.

Besides physical activities on the job, health behaviors may be considered as an important dimension of the general formation of health over the life-cycle, see Cutler and Lleras-Muney (2010). To analyze this, we resort to the following variables in our data set: a binary indicator for obesity (body mass index exceeds 30) as a compound lifestyle measure and more direct behavioral variables like an indicator for smoking, the amount of alcohol consumption (1 if at least three or more drinks when consuming alcohol), as well as physical activity measured by an indicator of having taken any sport exercise in the previous 3 months. The lower panel in Table 5.7 reports the sample means and treatment effects.

College education leads to a decrease in the probability of being obese, but increases the probability of smoking. This is in line with LATE estimates of the effect of college education in the US of Grimard and Parent (2007) and de Walque (2007). College education also seems to negatively affect alcohol consumption and increases the likelihood to engage in sport exercise. Again, the effect sizes are large, if not as large compared to the other potential mechanisms. Moreover, some of them are only marginally statistically significant. Taken together, college education affects potential health mechanisms in the expected direction. Again, there is effect heterogeneity, observable in different treatment parameters for the same outcome variables. Since health is a high dimensional measure, the potential mechanisms at hand are of course not able to explain the health returns to college education entirely. Nevertheless, the findings encourage us in our interpretation of the effects of college education on physical health.

Table 5.7: Potential mechanisms for health

	Definition	Sample	Parameter		
		mean	*ATE*	*ATT*	*ATU*
Physically demanding activities on the job					
Standing position	=1 if often working in a standing position for 2 or more hours	0.302	-0.37	-0.56	-0.30
			(0.07)	(0.09)	0.08)
Uncomfortable pos.	=1 if respondent needs to bend, crawl, lie down, keen or squat	0.190	-0.20	-0.37	-0.13
			(0.05)	(0.06)	(0.06)
Walking	=1 if job often requires walking, running or cycling	0.242	-0.39	-0.56	-0.32
			(0.06)	(0.07)	(0.07)
Carrying	=1 if often carrying a load of at least 10 kg	0.182	-0.40	-0.50	-0.37
			(0.05)	(0.05)	(0.05)
Health behaviors					
Obesity	=1 if Body Mass Index (=weight in kg/height in m^2) $>$ 30	0.155	-0.08	-0.15	-0.05
			(0.04)	(0.05)	(0.05)
Smoking	=1 if currently smoking	0.270	-0.18	-0.23	-0.16
			(0.06)	(0.06)	(0.07)
Alcohol amount	=1 if three or more drinks when consuming alcohol	0.187	-0.14	-0.13	-0.14
			(0.05)	(0.06)	(0.06)
Sport	=1 if any sporting exercise in the previous 3 months	0.717	0.16	0.31	0.10
			(0.07)	(0.07)	(0.09)

Notes: Own calculations based on NEPS-Starting Cohort 6 data. Definitions are taken from the data manual. Standard error estimated using a clustered bootstrap (at district level) and reported in parentheses.

5.7 Conclusion

This paper uses the Marginal Treatment Effect framework introduced and advanced by Björklund and Moffitt (1987) and Heckman and Vytlacil (2005, 2007) to estimate returns to college education under essential heterogeneity. We use representative data from the German National Educational Panel Study (NEPS). Our outcome measures are wages, cognitive abilities, and health. Cognitive abilities are assessed using state-of-the-art cognitive competence tests on individual reading speed, text understanding, and mathematical literacy. As expected, all outcome variables are positively correlated with having a college degree in our data set. Using an instrument that exploit exogenous variation in the supply of colleges, we estimate marginal returns to college education.

The main findings of this paper are as follows: College education improves average wages, cognitive abilities and physical health (but not mental health). There is heterogeneity in the effects and clear signs of selection into gains. Those individuals who realize the highest returns to education are those who are most ready to take it. Moreover, education does not pay off for everybody. While it is never harmful, we find zero causal effects for around 30–40 percent of the population. Thus, while college education is beneficial on average, further increasing the number of students – as sometimes called for – is less likely to pay off, as the current marginal students are those who are mostly in the range of zero causal effects. Potential mechanisms of skill returns are more demanding jobs that slow down the cognitive decline in later ages. Regarding health we find positive effects of higher education on BMI, non-smoking, sports participation and alcohol consumption.

All in all, given that the average individual clearly seems to benefit from education and provided that the continuing technological progress has skills become more and more valuable, education should still be an answer to the technological change for the average individual.

One limitation of this paper is that we are not able to stratify the analysis by study subject. This is left for future work.

Appendix

Figures

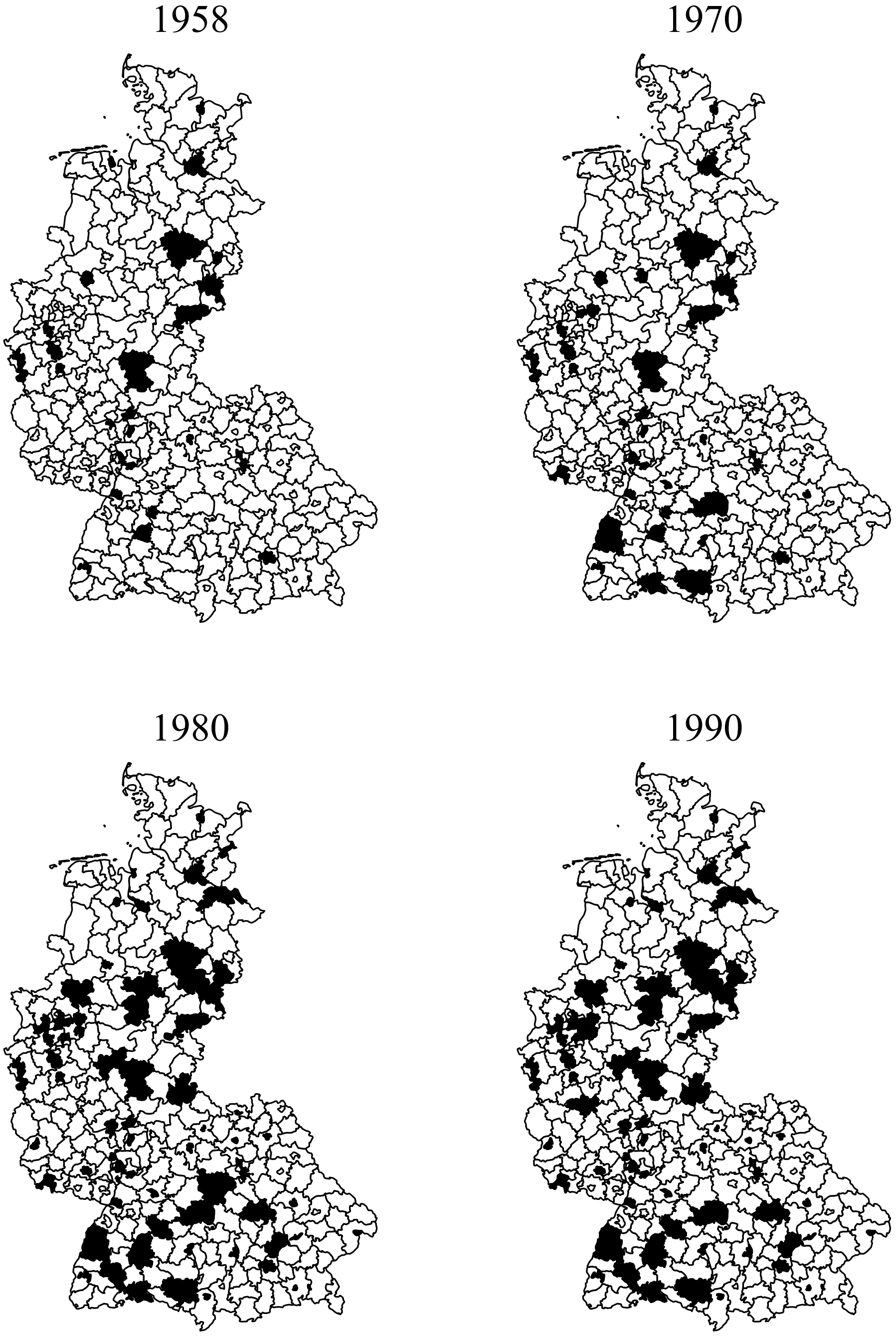


Figure A4.1: Spatial variation of colleges across districts and over time

Notes: Own illustration based on the German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, 1991). The maps show all 326 West German districts (*Kreise*, spatial units of 2009) but Berlin in the years 1958 (first year in the sample), 1970, 1980, and 1990 (last year in the sample). Districts usually cover a bigger city or some administratively connected villages. If a district has at least one college, the district is depicted darker. Only few districts have more than one college. For those districts the number of students is added up in the calculations but multiple colleges are not depicted separately in the maps.

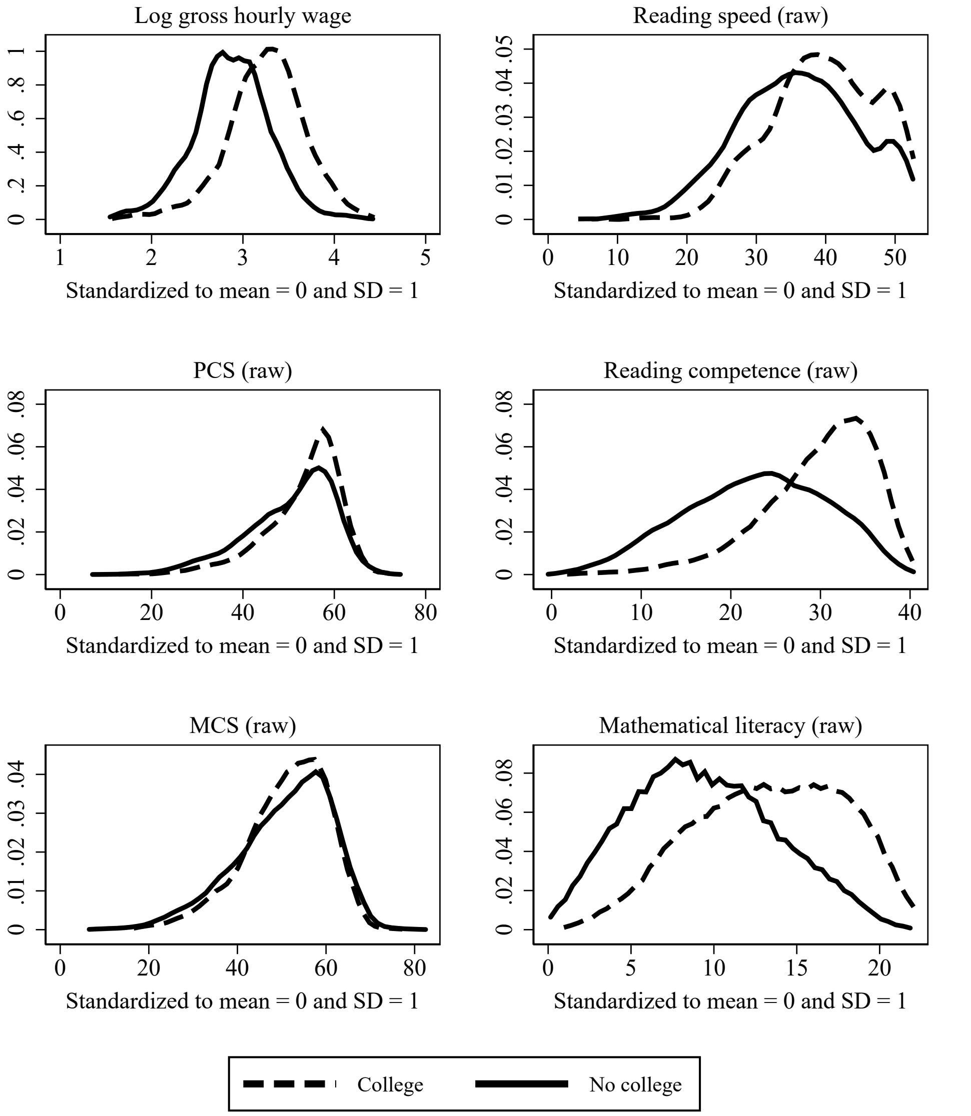


Figure A4.2: Distribution of dependent variables by college graduation

Notes: Own illustration based on NEPS-Starting Cohort 6 data.

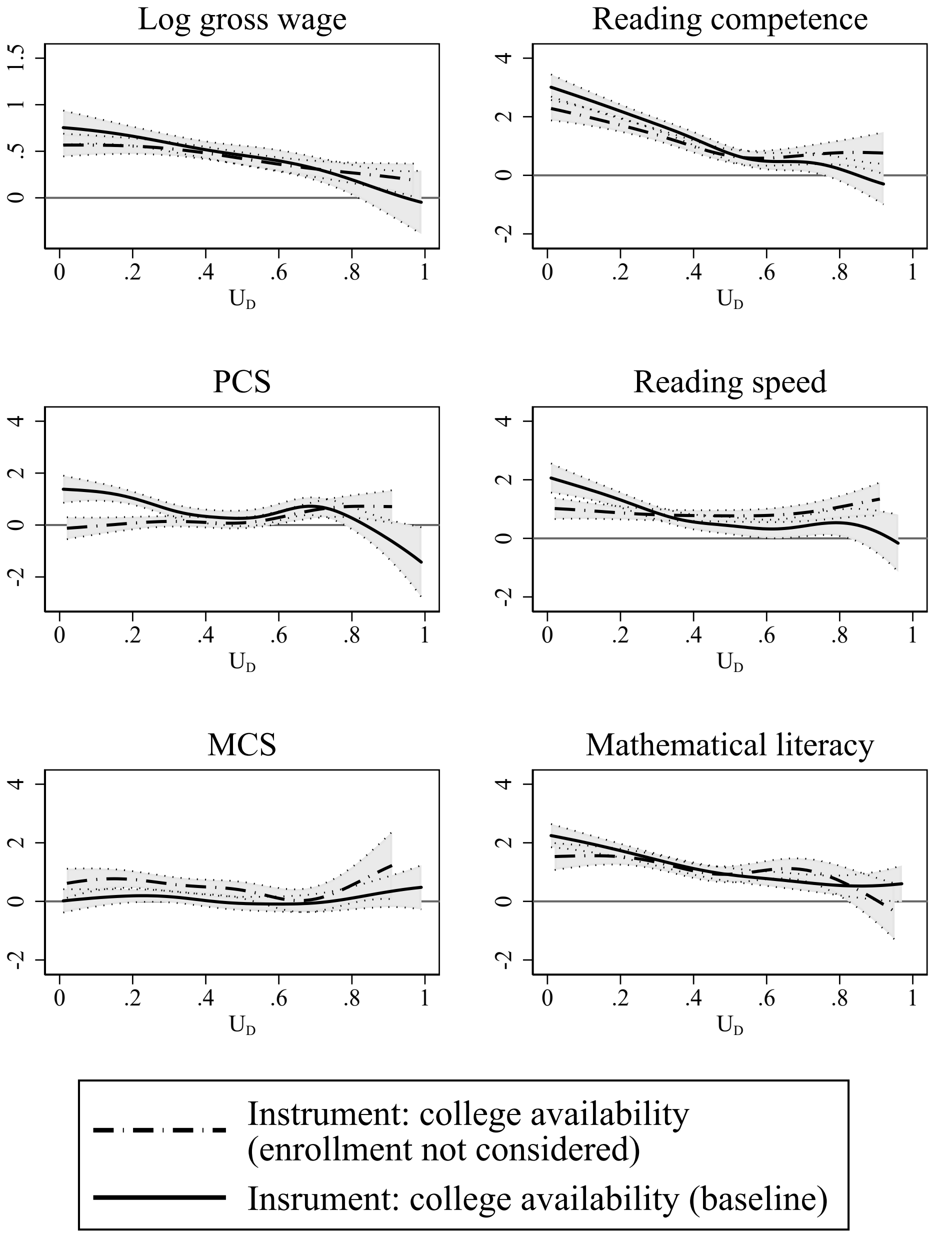


Figure A4.3: Sensitivity in Marginal Treatment Effects when using only the sum of the kernel weighted college distances

Notes: Own illustration based on NEPS-Starting Cohort 6 data. For gross hourly wage, the log value is taken. Health and cognitive skill outcomes are standardized to mean 0 and standard deviation 1. The MTE (vertical axis) is measured in logs for wage and in units of standard deviations of the health and cognitive skill outcomes. The dashed lines give the 95% confidence intervals. Calculations based on a local linear regression where the influence of the control variables was isolated using a semiparametric Robinson estimator (Robinson, 1988) for each outcome variable.

Tables

Table A4.1: Control variables and means by college degree

Variable	Definition	Respondents	
		with college degree	w/o college degree
General information			
Female	=1 if respondent is female	40.38	54.18
Year of birth (FE)	Year of birth of the respondent	1959	1959
Migrational background	=1 if respondent was born abroad	0.89	0.64
No native speaker	=1 if mother tongue is not German	0.30	0.43
Rural district	=1 if current district is rural	16.79	24.96
Mother still alive	=1 if mother is still alive in 2009/10	65.38	63.83
Father still alive	=1 if father is still alive in 2009/10	45.27	42.3
Pre-college living conditions			
Married before college	=1 if respondent got married before the year of the college decision or in the same year	0.20	0.44
Parent before college	=1 if respondent became a parent before the year of the college decision or in the same year	0.30	0.17
Siblings	Number of siblings	1.56	1.87
First born	=1 if respondent was the first born in the family	33.73	29.01
Age 15: lived by single parent	=1 if respondent was raised by single parent	5.33	5.32
Age 15: lived in patchwork family	=1 if respondent was raised in a patchwork family	1.11	0.27
Age 15: orphan	=1 if respondent was a orphan at the age of 15	0.10	0.20
Age 15: mother employed	=1 if mother was employed at the respondent's age of 15	45.93	46.87
Age 15: mother never unemployed	=1 if mother was never unemployed until the respondent's age of 15	61.24	62.29
Age 15: father employed	=1 if father was employed at the respondent's age of 15	92.46	90.73
Age 15: father never unemployed	=1 if father was never unemployed until the respondent's age of 15	98.45	97.14
Pre-college education			
Final school grade: excellence	=1 if the overall grade of the highest school degree was excellent	4.59	1.79
Final school grade: good	=1 if the overall grade of the highest school degree was good	31.51	25.83

Continued on next page

Table A4.1 – *continued*

Variable	Definition	Respondents	
		with college degree	w/o college degree
Final school grade: satisfactory	=1 if the overall grade of the highest school degree was satisfactory	17.97	28.03
Final school grade: sufficient or worse	=1 if the overall grade of the highest school degree was sufficient or worse	1.04	1.42
Repeated one grade	=1 if student needed to repeat one grade in elementary or secondary school	19.97	20.51
Repeated two or more grades	=1 if student needed to repeat two or more grades in elementary or secondary school	2.74	1.85
Military service	=1 if respondent was drafted for compulsory military service	28.03	23.89
Parental characteristics (M: mother, F: father)			
M: year of birth (FE)	Year of birth of the respondent's mother	1930	1932
M: migrational background	=1 if mother was born abroad	5.47	4.85
M: at least inter. edu	=1 if mother has at least an intermediate secondary school degree	17.97	5.95
M: vocational training	=1 if mother's highest degree is vocational training	20.86	16.18
M: further job qualification	=1 if mother has further job qualification (e.g., *Meister* degree)	4.29	1.73
F: year of birth (FE)	Year of birth of the respondent's father	1927	1929
F: migrational background	=1 if father was born abroad	6.36	5.54
F: at least inter. edu	=1 if father has at least an intermediate secondary school degree	20.86	8.09
F: vocational training	=1 if father's highest degree is vocational training	19.12	21.99
F: further job qualification	=1 if father has further job qualification (e.g., *Meister* degree)	11.46	6.76
Number of observations (PCS and MCS sample)		1,352	3,461

Notes: Own calculations based on NEPS-Starting Cohort 6 data. Definitions are taken from the data manual. Mean values refer to the MCS and PCS sample. FE = variable values are included as fixed effects in the analysis.

Online Appendix

Additional information on the instrument

In the years immediately after WWII, neither political decision makers nor society as a whole were concerned with higher educational affairs (Bartz, 2007). Weisser (2005) argues that colleges have been enganged in reconstructing their facilities (and curricula) as the rest of the country but almost unnoticed by society. This changed at the beginning of the 1960s when politicians of all parties started to doubt that the existing colleges were able cope with newly arising challenges of an increasing demand for higher education. This increased demand was partly driven by catch-up effect for large parts of the population. The number of students in higher education in Germany decreased by 50% between 1928 and 1938 and at the beginning of the 1960s and educational participation in Germany was much lower than in other industrialized countries (Picht, 1964). For other factors that increased the pressure to political decision makers to be involved in higher educational policies, consult the paper.

Various policy measures at the national level and in the 11 West German federal states have been taken in order to address these challenges and finally led to expansion of higher education. After WWII, the existing colleges adopted their former regulations from the time before the Third Reich. Because the German Empire consisted of dozens of microstates each college had basically its own rules (Bartz, 2007). To unify the regulations each of the federal state and the federal government passed so-called higher education acts (*Hochschulrahmengesetze*) that allow them to intervene in university politics between 1966 and 1967. At the same time, the states and the federal government also established the German Council of Science and Humanities (*Wissenschaftsrat*), an advisory board for higher educational policies (Bartz, 2007). In its landmark report in 1960, the council suggested to increase the number of professors and lectures at the existing colleges by 40% (Wissenschaftsrat, 1960). In follow-up reports, it also proposed to increase facilities of the existing colleges and to build new colleges (Wissenschaftsrat, 1966, 1970). While the suggestions of the council have been rather broad and not binding for the state's governments, the states developed their own strategies to cope with the expected increase in the number of students. Examples are the (not entirely realized) Dahrendorf-Plan in the state of Baden-Württemberg and the introduction of *Gesamthochschulen* (a combination of colleges and universities of applied science) in North Rhine-Westphalia and some other states, see Bartz (2007). The reform process went along with a public debate on higher education among academics and in the media (see, e.g., the newspaper articles in

Der Spiegel, 1967, and Die Zeit, 1967). Moreover, the discussion was spurred by the publication of the influential books "Education as Civil Right" (*Bildung als Bürgerrecht*, Dahrendorf, 1965) and "The German Educational Disaster" (*Die deutsche Bildungskatastrophe*, Picht, 1964).

In order to learn more about the timing and the placement of the college construction within the states, we searched for records on the decision making process in the most-populated state of North Rhine-Westphalia.[22] While the Council of Science and Humanities suggested to link college openings to the expected increase in the population (NRW, 1971a), we find evidence that the state's authorities also took criteria into account that were independent of the expected demand. In a report on the founding of five new *Gesamthochschule* institutions, the Minister of Education and Research of North Rhine-Westphalia described the aim of the placement decision as "improving the equality of educational opportunities for all potential students by providing a sufficient number of open spots" (NRW, 1971b, section 3.1, own translation). The minister explicitly argued that the opening of colleges in regions that had no college before would increase the participation in (secondary and higher) education in those regions – the new colleges would serve as "advertisement for education" (*Bildungswerbung*, NRW, 1971c, section II.2.11). This reasoning is somewhat remarkable given that decision makers expected a higher demand for college education in cities that already had a college (NRW, 1971c). Another piece of evidence is provided by a review of the history of the University of Bochum by Weisser (2005). Originally, decision makers intended to open the new college in the city of Dortmund; however, the construction site in Dortmund was found to be not sufficient. Thus, they decided to construct the college in the close-by city of Bochum. The decision to open a college in Dortmund was made a couple of years later "in the run-up to the state's parliament elections" (Weisser, 2005, own translation). We do not depict the decision marking processes for all college openings in West-Germany, although we found evidence that the processes went often similarly.

In our interpretation of the evidence, the decentralized decision making processes between the federal states and within the states introduced variation in the higher educational expansion that is likely to be independent from a demand for higher education (that might be the result of low cognitive abilities or a worse health).

[22] For North Rhine-Westphalia, records (in German language) of parliament hearings and debates are available online, see the references for links.

Additional Tables and Figures

Table O4.1: Full results for logit estimation of the selection equation (mean marginal effects)

	(1)	(2)	(3)	(4)	(5)	(6)
	Sample for					
	Gross hourly wage	Health measure		Cognitive ability component		
		PCS	MCS	Read. speed	Read. comp.	Math liter.
College availability	3.133***	3.527***	3.527***	3.711***	3.050***	3.815***
	(0.228)	(0.233)	(0.233)	(0.206)	(0.188)	(0.286)
Female	−0.079*	−0.046	−0.046	0.012	−0.056	0.011
	(0.045)	(0.035)	(0.035)	(0.038)	(0.038)	(0.045)
Rural district	−0.050**	−0.069***	−0.069***	−0.056***	−0.063***	−0.063**
	(0.022)	(0.019)	(0.019)	(0.020)	(0.019)	(0.025)
Migrational background	−0.146	0.064	0.064	−0.004	−0.051	0.116
	(0.116)	(0.086)	(0.086)	(0.074)	(0.065)	(0.094)
No native speaker	−0.347**	−0.084	−0.084	−0.051	0.049	−0.046
	(0.139)	(0.153)	(0.153)	(0.104)	(0.103)	(0.123)
Military service	−0.101***	−0.119***	−0.119***	−0.115***	−0.108***	−0.154***
	(0.027)	(0.024)	(0.024)	(0.025)	(0.024)	(0.030)
First born	0.080***	0.072***	0.072***	0.076***	0.081***	0.081***
	(0.017)	(0.013)	(0.013)	(0.014)	(0.013)	(0.018)
Age 15: lived by single parent	−0.041	−0.010	−0.010	−0.008	−0.050	−0.009
	(0.036)	(0.032)	(0.032)	(0.033)	(0.031)	(0.040)
Age 15: lived in patchwork family	−0.155***	−0.136***	−0.136***	−0.091*	−0.037	−0.127*
	(0.059)	(0.045)	(0.045)	(0.050)	(0.042)	(0.074)
Age 15: orphan	−0.089	−0.051	−0.051	−0.082	−0.206***	−0.103
	(0.078)	(0.059)	(0.059)	(0.067)	(0.068)	(0.072)
Number of siblings	−0.027***	−0.028***	−0.028***	−0.023***	−0.022***	−0.030***
	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.006)
Married before college	0.277**	0.180*	0.180*	0.254***	0.207**	0.256**
	(0.122)	(0.096)	(0.096)	(0.097)	(0.085)	(0.112)
Parent before college	−0.044**	−0.028**	−0.028**	−0.039**	−0.050***	−0.036*
	(0.018)	(0.014)	(0.014)	(0.015)	(0.014)	(0.019)
Mother: migrational background	0.055	0.054*	0.054*	0.059**	0.042	0.015
	(0.039)	(0.029)	(0.029)	(0.030)	(0.031)	(0.036)
Mother: at least inter. edu	0.164***	0.138***	0.138***	0.139***	0.135***	0.140***
	(0.028)	(0.026)	(0.026)	(0.027)	(0.020)	(0.035)

Continued on next page

Table O4.1 – *continued*

	(1)	(2)	(3)	(4)	(5)	(6)
Mother: college degree	0.081	0.098*	0.098*	0.072	0.097**	0.125
	(0.061)	(0.055)	(0.055)	(0.061)	(0.049)	(0.080)
Mother: vocational training	0.005	0.053**	0.053**	0.042*	0.009	0.041
	(0.026)	(0.021)	(0.021)	(0.022)	(0.017)	(0.028)
Mother: further job qualification	−0.080*	0.073**	0.073**	0.082**	0.028	0.059
	(0.046)	(0.037)	(0.037)	(0.038)	(0.032)	(0.051)
Mother: still alive	0.027	0.026*	0.026*	0.027*	0.052***	0.025
	(0.018)	(0.015)	(0.015)	(0.016)	(0.014)	(0.021)
Age 15: mother unemployed	−0.015	0.010	0.010	0.022	−0.004	0.020
	(0.022)	(0.017)	(0.017)	(0.019)	(0.018)	(0.024)
Age 15: mother never employed	0.012	−0.008	−0.008	−0.009	0.008	−0.008
	(0.022)	(0.017)	(0.017)	(0.019)	(0.018)	(0.024)
Father has migrational background	0.044	0.004	0.004	0.027	0.023	0.038
	(0.032)	(0.027)	(0.027)	(0.031)	(0.031)	(0.037)
Father: at least inter. edu	0.090***	0.108***	0.108***	0.103***	0.118***	0.073**
	(0.030)	(0.026)	(0.026)	(0.029)	(0.021)	(0.036)
Father: college degree	0.208***	0.184***	0.184***	0.183***	0.145***	0.173***
	(0.054)	(0.046)	(0.046)	(0.047)	(0.034)	(0.056)
Father: vocational training	0.071*	0.071**	0.071**	0.054*	0.032	0.042
	(0.040)	(0.031)	(0.031)	(0.032)	(0.024)	(0.039)
Father: further job qualification	0.200***	0.165***	0.165***	0.155***	0.121***	0.124***
	(0.043)	(0.036)	(0.036)	(0.037)	(0.027)	(0.045)
Father: still alive	0.066***	0.058***	0.058***	0.070***	0.048***	0.074***
	(0.018)	(0.015)	(0.015)	(0.017)	(0.016)	(0.020)
Age 15: father unemployed	0.005	0.001	0.001	0.021	0.019	0.025
	(0.043)	(0.029)	(0.029)	(0.031)	(0.029)	(0.039)
Age 15: father never employed	0.102	0.098*	0.098*	0.134**	0.110*	0.085
	(0.090)	(0.055)	(0.055)	(0.063)	(0.061)	(0.074)
Final school grade: excellent	0.468***	0.440***	0.440***	0.508***	0.403***	0.470***
	(0.069)	(0.064)	(0.064)	(0.068)	(0.057)	(0.080)
Final school grade: good	0.301***	0.283***	0.283***	0.340***	0.267***	0.293***
	(0.056)	(0.056)	(0.056)	(0.059)	(0.041)	(0.070)

Continued on next page

Table O4.1 – *continued*

	(1)	(2)	(3)	(4)	(5)	(6)
Final school grade: satisfactory	0.185***	0.162***	0.162***	0.204***	0.124***	0.172**
	(0.057)	(0.057)	(0.057)	(0.062)	(0.042)	(0.072)
Final school grade: sufficient or worse	0.163**	0.181**	0.181**	0.267***	0.293***	0.217**
	(0.082)	(0.075)	(0.075)	(0.083)	(0.086)	(0.096)
Grade repetition: 1 grade	−0.034**	−0.007	−0.007	−0.012	−0.027*	−0.003
	(0.017)	(0.015)	(0.015)	(0.017)	(0.016)	(0.020)
Grade repetition: 2+ grades	−0.030	−0.004	−0.004	0.028	0.015	0.078
	(0.058)	(0.042)	(0.042)	(0.049)	(0.044)	(0.058)
Observations	3,378	4,813	4,813	3,995	4,576	2,904

Notes: Own calculations based on NEPS-Starting Cohort 6 data. The table gives the mean marginal effects of the logit model. Regressions also include a full set of individual year-of-birth fixed effects and district fixed effects, and district-specific linear trends. District-year-clustered standard errors in parentheses; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table O4.2: Full results for 2SLS second-stage estimations

	(1)	(2)	(3)	(4)	(5)	(6)
	Sample for					
	Gross hourly wage	Health measure		Cognitive ability component		
		PCS	MCS	Read. speed	Read. comp.	Math liter.
College degree	0.549***	0.677***	0.080	0.888***	1.529***	1.490***
	(0.048)	(0.099)	(0.099)	(0.114)	(0.098)	(0.126)
Female	−0.192***	0.081	−0.270***	0.424***	0.345***	−0.384***
	(0.040)	(0.089)	(0.084)	(0.097)	(0.086)	(0.098)
Rural district	−0.055**	−0.008	0.052	−0.039	−0.042	0.001
	(0.024)	(0.045)	(0.047)	(0.047)	(0.044)	(0.058)
Migrational background	−0.034	0.010	−0.043	−0.381*	−0.375**	−0.654***
	(0.080)	(0.214)	(0.188)	(0.205)	(0.149)	(0.224)
No native speaker	0.064	0.212	0.042	−0.070	−0.731***	0.251
	(0.119)	(0.189)	(0.221)	(0.279)	(0.243)	(0.277)
Military service	0.044	0.054	0.012	−0.030	−0.047	0.043
	(0.028)	(0.057)	(0.061)	(0.064)	(0.055)	(0.076)
First born	−0.023	0.006	0.064*	0.011	0.037	0.039

Continued on next page

Table O4.2 – *continued*

	(1)	(2)	(3)	(4)	(5)	(6)
	(0.018)	(0.035)	(0.036)	(0.037)	(0.033)	(0.041)
Age 15: lived by single parent	0.011	0.008	-0.130^{*}	−0.121	−0.043	0.080
	(0.038)	(0.081)	(0.072)	(0.077)	(0.064)	(0.089)
Age 15: lived in patchwork family	0.005	−0.038	-0.245^{**}	0.013	0.008	0.201^{*}
	(0.045)	(0.093)	(0.105)	(0.106)	(0.092)	(0.110)
Age 15: orphan	0.043	-0.326^{***}	−0.023	−0.034	0.056	−0.042
	(0.066)	(0.125)	(0.115)	(0.115)	(0.122)	(0.129)
Number of siblings	-0.020^{***}	-0.027^{***}	0.018^{*}	-0.035^{***}	-0.041^{***}	-0.023^{**}
	(0.005)	(0.010)	(0.009)	(0.011)	(0.009)	(0.011)
Married before college	0.061	0.028	0.366^{**}	0.314	0.162	0.367
	(0.101)	(0.290)	(0.169)	(0.200)	(0.160)	(0.276)
Parent before college	0.011	0.020	0.113^{***}	0.167^{***}	0.133^{***}	0.138^{***}
	(0.019)	(0.036)	(0.037)	(0.038)	(0.034)	(0.045)
Mother: migrational background	0.042	0.013	0.022	0.106	0.114	0.085
	(0.039)	(0.079)	(0.079)	(0.076)	(0.074)	(0.082)
Mother: at least inter. edu	−0.014	0.064	−0.028	0.011	−0.047	−0.056
	(0.032)	(0.068)	(0.066)	(0.068)	(0.056)	(0.083)
Mother: college degree	−0.009	0.088	0.129	−0.229	−0.149	0.016
	(0.070)	(0.151)	(0.151)	(0.172)	(0.116)	(0.206)
Mother: vocational training	−0.024	0.022	0.047	0.061	−0.004	0.017
	(0.024)	(0.054)	(0.054)	(0.053)	(0.039)	(0.062)
Mother: further job qualification	−0.006	−0.133	−0.024	−0.064	−0.018	−0.105
	(0.050)	(0.105)	(0.095)	(0.116)	(0.075)	(0.125)
Mother: still alive	0.028	0.043	−0.049	−0.027	−0.004	0.023
	(0.019)	(0.038)	(0.038)	(0.039)	(0.034)	(0.045)
Age 15: mother unemployed	0.041^{*}	0.022	0.043	0.040	−0.010	0.003
	(0.021)	(0.042)	(0.044)	(0.044)	(0.041)	(0.050)
Age 15: mother never employed	-0.052^{**}	−0.060	-0.074^{*}	−0.009	0.036	−0.004
	(0.022)	(0.043)	(0.045)	(0.045)	(0.042)	(0.051)
Father has migrational background	−0.012	0.073	−0.107	-0.155^{**}	−0.099	−0.015
	(0.037)	(0.067)	(0.073)	(0.071)	(0.072)	(0.083)
Father: at least inter. edu	−0.017	-0.137^{**}	0.098	0.112	0.027	−0.056
	(0.033)	(0.069)	(0.064)	(0.069)	(0.056)	(0.079)

Continued on next page

Table O4.2 – *continued*

	(1)	(2)	(3)	(4)	(5)	(6)
Father: college degree	0.003	−0.236**	−0.125	0.008	0.084	−0.016
	(0.051)	(0.119)	(0.111)	(0.113)	(0.086)	(0.135)
Father: vocational training	−0.020	−0.098	0.022	−0.013	0.101*	0.031
	(0.030)	(0.068)	(0.069)	(0.067)	(0.052)	(0.075)
Father: further job qualification	−0.028	−0.134	−0.055	−0.024	0.107*	0.062
	(0.037)	(0.082)	(0.082)	(0.084)	(0.063)	(0.097)
Father: still alive	−0.014	0.078**	−0.067*	0.034	0.040	0.006
	(0.017)	(0.036)	(0.036)	(0.038)	(0.035)	(0.044)
Age 15: father unemployed	0.009	0.114	0.106	0.002	−0.036	−0.002
	(0.039)	(0.070)	(0.077)	(0.080)	(0.069)	(0.086)
Age 15: father never employed	0.018	0.131	−0.113	0.058	0.113	0.087
	(0.069)	(0.158)	(0.175)	(0.153)	(0.117)	(0.160)
Final school grade: excellent	0.050	0.043	0.127	0.172	0.293**	0.389***
	(0.066)	(0.127)	(0.132)	(0.133)	(0.138)	(0.135)
Final school grade: good	0.034	0.064	0.200**	0.278***	0.329***	0.169*
	(0.045)	(0.089)	(0.101)	(0.097)	(0.084)	(0.097)
Final school grade: satisfactory	0.033	0.066	0.164*	0.203**	0.328***	0.024
	(0.044)	(0.086)	(0.100)	(0.095)	(0.083)	(0.094)
Final school grade: sufficient or worse	−0.145*	−0.112	−0.086	−0.064	−0.139	−0.388**
	(0.084)	(0.164)	(0.172)	(0.160)	(0.193)	(0.158)
Grade repetition: 1 grade	−0.031*	0.057	−0.052	−0.058	−0.002	−0.073*
	(0.018)	(0.036)	(0.038)	(0.039)	(0.035)	(0.044)
Grade repetition: 2+ grades	−0.022	0.002	−0.145	0.036	0.093	−0.101
	(0.053)	(0.095)	(0.115)	(0.116)	(0.099)	(0.134)
Observations	3,378	4,813	4,813	3,995	4,576	2,904

Notes: Own calculations based on NEPS-Starting Cohort 6 data. Regressions also include a full set of individual year-of-birth fixed effects and district fixed effects, and district-specific linear trends. District-year-clustered standard errors in parentheses; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

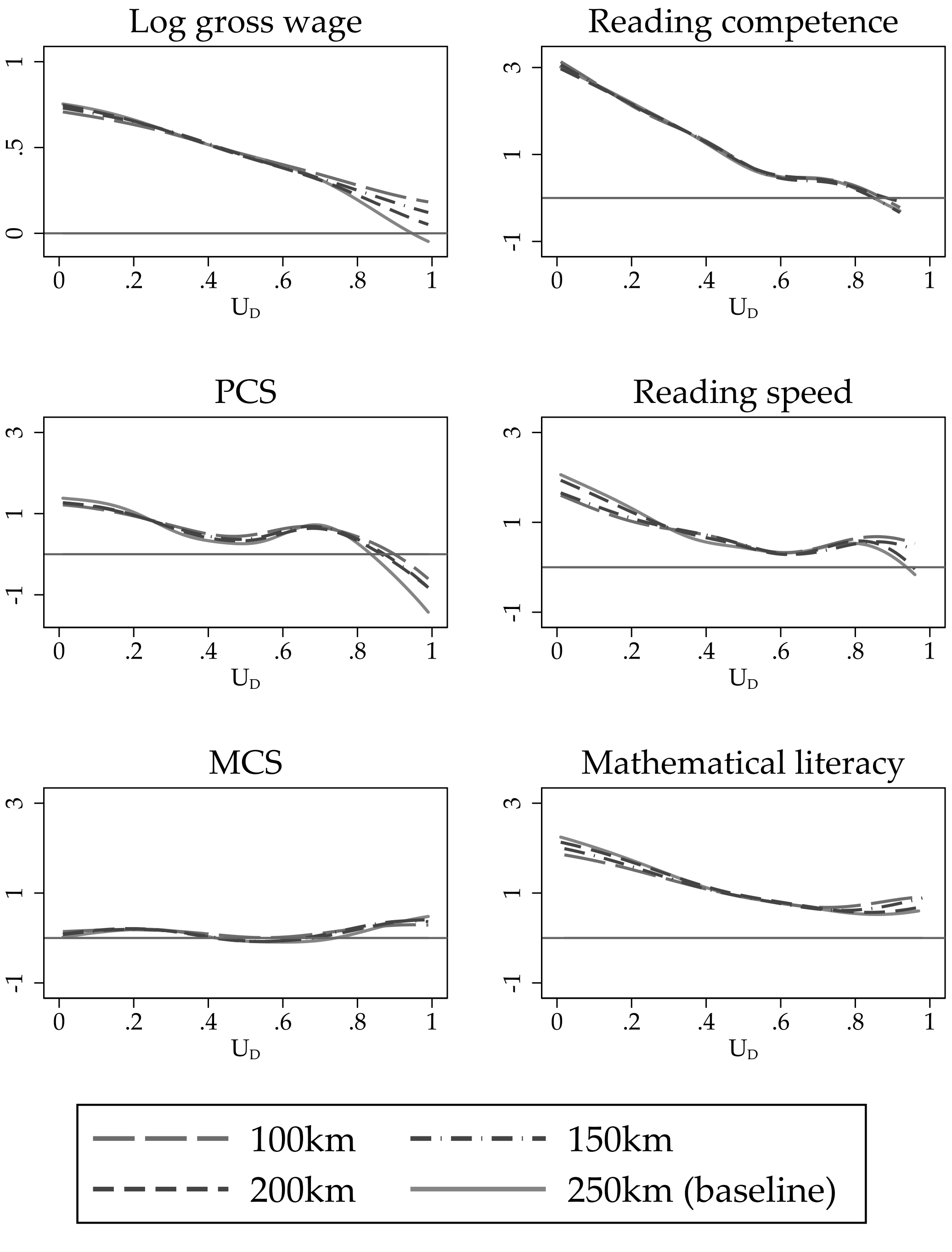


Figure O4.1: Sensitivity in Marginal Treatment Effects when using different kernel bandwidths

Notes: Own illustration based on NEPS-Starting Cohort 6 data. All outcomes are standardized to mean 0 and standard deviation 1. The MTE (vertical axis) is measured in units of standard deviations of the outcome variable. Calculations based on a local linear regression where the influence of the control variables was isolated using a semiparametric Robinson estimator (Robinson, 1988) for each outcome variable.

Table O4.3: First-stage estimations when using different kernel bandwidths

	(1)	(2)	(3)	(4)	(5)	(6)
	Sample for					
	Gross hourly wage	Health measure		Cognitive ability component		
		PCS	MCS	Read. speed	Read. comp.	Math liter.
Bandwidth 100km						
College availability	5.545***	5.587***	5.587***	5.557***	5.271***	5.449***
	(0.332)	(0.284)	(0.284)	(0.322)	(0.282)	(0.379)
Bandwidth 150km						
College availability	3.558***	3.693***	3.693***	3.666***	3.449***	3.575***
	(0.201)	(0.175)	(0.175)	(0.197)	(0.171)	(0.233)
Bandwidth 200km						
College availability	2.763***	2.943***	2.943***	2.903***	2.703***	2.828***
	(0.150)	(0.132)	(0.132)	(0.149)	(0.128)	(0.177)
Bandwidth 250km (baseline specification)						
College availability	2.368***	2.577***	2.577***	2.530***	2.333***	2.465***
	(0.125)	(0.112)	(0.112)	(0.126)	(0.107)	(0.149)
Observations	3,378	4,813	4,813	3,995	4,576	2,904

Notes: Own calculations based on NEPS-Starting Cohort 6 data. Regressions also include a full set of control variables as well as year-of-birth and district fixed effects, and district-specific linear trends. District-year clustered standard errors in parentheses; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Marginal Treatment effect – why observed and unobserved heterogeneity cannot be separated under conditional independence of the instrument

Modeling of counterfactual ountcomes:

$$\begin{aligned} Y^1 &= X\beta_1 + U_1 \\ Y^0 &= X\beta_0 + U_0 \end{aligned}$$

Assumptions:

$$\begin{aligned} U_1, U_0 &\perp Z|X \\ E(U_1|X) &= E(U_0|X) = 0 \end{aligned}$$

Potential outcome equation:

$$\begin{aligned} Y &= DY^1 + (1-D)Y^0 \\ &= Y^0 + D(Y^1 - Y^0) \\ &= [X\beta_0 + U_0] + [(X\beta_1 + U_1) - (X\beta_0 + U_0)]\, D \\ &= [X\beta_0 + U_0] + [X\,(\beta_1 - \beta_0) + (U_1 - U_0)]\, D \end{aligned}$$

Applying conditional expectation $E(.|X,Z)$:

$$\begin{aligned} E(Y|X,Z) &= \underbrace{E\,[X\beta_0 + U_0|X,Z]}_{\text{CIA: Independent of Z}} + \underbrace{E\,[(X\,(\beta_1 - \beta_0) + (U_1 - U_0))\,D|X,Z]}_{\text{Law of Iterated Expectations}} \\ &= X\beta_0 + \underbrace{E(U_0|X)}_{=0} + \overbrace{E\,[X(\beta_1 - \beta_0) + (U_1 - U_0)\,|D = 1, X, Z]}^{\text{CIA: independent of Z}}\underbrace{E(D|X,Z)}_{=\text{p}} \\ &= X\beta_0 + E\,[X(\beta_1 - \beta_0) + (U_1 - U_0)\,|D = 1, X, Z]\,p \\ &= X\beta_0 + \underbrace{X(\beta_1 - \beta_0)p + \underbrace{E\,[(U_1 - U_0)\,|D = 1, X]}_{\neq 0}\,p}_{\substack{\text{Cannot be separated in estimation: one term} \\ \text{would need to be restricted by some assumption.}}} \end{aligned}$$

Under the CIA $X(\beta_1 - \beta_0)$ and $E\,[(U_1 - U_0)\,|D = 1, X]$ would be observationally equivalent as long as $U_1, U_0 \perp Z|X$. If $U_1, U_0 \perp X, Z$ the equivalence is dissolved since only $E\,[(U_1 - U_0)\,|D = 1]$ needs to be identified and $E\,(X(\beta_1 - \beta_0))$ can be restricted to zero without loss of generality.

However, if one is solely interested in identifying the general heterogeneity in $E(Y^1 - Y^0|X, p)$ with regard to p without separating between the exact source

($U_1 - U_0$ or $X(\beta_1 - \beta_0)$), further restrictions regarding U_1, U_0 and β_1, β_0 are not necessary and $U_1, U_0 \perp Z|X$ is sufficient.

Chapter 6

Fertility Effects of College Education: Evidence from the German Educational Expansion

Joint work with Matthias Westphal

6.1 Introduction

Among the many changes that have affected developed societies in the past 60 years, two certainly belong to the most significant ones: the educational expansion – describing the substantial upsurge in higher education enrollment, especially that of females – and the fertility transition, characterized by declining fertility rates that have fallen below replacement rates. The resulting consequences of both these evolutions have affected many dimensions of social interaction such as the demographic change – which today constitutes an urgent concern from a policy perspective. While policies that aim at increasing education have been introduced in all parts of the world, many developed countries have also set up policies to boost fertility rates. Although both kinds of policies are often comparatively well-understood due to ample research, the link between these policies – that is, how education affects fertility – is still mostly understudied. The negative correlation between education and fertility, sometimes referred to as the "baby gap" between high- and low-educated individuals, may hint at the potential side-effects education policies may have on fertility.[1] By analyzing the

[1] The ambiguity that education policies may reduce fertility while family policies in developed countries are targeted at increasing fertility becomes most visible in developing countries where education policies are often implemented in order to reduce family size. Due to the context and

upsurge in higher education in Germany triggered by a massive build-up of colleges, we contribute to the understanding of whether increased education causes lower fertility or whether individuals merely choose to have more education and smaller families simultaneously.

Researchers have been concerned with the consequences of education policies for decades. While there are still some "unknowns" with respect to the optimal margin of education and potential effect heterogeneities, education is often found to increase labor market performance (for the case of higher education see, e.g., the literature reviews of Barrow and Malamud, 2015, and Oreopoulos and Petronijevic, 2013). Although there is the reasonable suspicion that the non-pecuniary returns to education are positive as well (see Oreopoulos and Salvanes, 2011), evidence of the causal long-term effects on these outcomes is rather scarce. Most studies that analyze the effect of education on fertility utilize variation in compulsory schooling laws to address the selection problem.[2] While such changes to the law affect a large share of students in many countries, it seems a priori unlikely that the effects for secondary schooling also hold true for other margins of education, such as college education. The results of the literature on the effectiveness of family policies that induce financial incentives for bigger families in general may be summarized as mixed (see Gauthier, 2007, for a review and Haan and Wrohlich, 2011, and Riphahn and Wiynck, 2017, as well as Raute, 2016, for evidence on Germany). The absence of such silver bullets to increase fertility using existing family policies emphasizes the need to gain a better understanding of how education affects fertility decisions.

We are not aware of any study that explicitly investigates the causal link between college education and fertility in a developed economy[3] although the college margin provides a presumably interesting addition to the more often considered fertility effect of secondary schooling for four reasons: First, college education is taught more extensively – in Germany the formal duration of college education in the time under review was 4.5 years compared to changes in compulsory schooling that, at most, account for one or two years. Second, while compulsory schooling affects individuals at the lower end of the education (and presumably

the margin of education we focus on the situation in developed countries. See Duflo et al. (2015) and the literature therein for the case in developing countries.

[2] See, for instance, Cygan-Rehm and Maeder (2013) for Germany, Black et al. (2008) for the US and Norway, Geruso and Royer (2014) for the UK, Monstad et al. (2008) for Norway, Grönqvist and Hall (2013) for Sweden, and Fort et al. (2016) for the UK and pooled Continental European countries. McCrary and Royer (2011) consider changes in the school entry age that cause variation in education.

[3] Currie and Moretti (2003) analyze the effect of maternal education on the offspring's health in the US but consider the number of children merely as a potential channel. A recent working paper by Tequamem and Tirivayi (2015) analyzes the fertility effects of higher education in Ethiopia and find a reduction in family size.

skill) distribution, college affects individuals at the upper end who may react differently. Third, college education falls well into the prime reproductive age of women (and potential fathers) while the largest effects of additional years of compulsory schooling have been found on in-school and teenage pregnancies. Fourth, college education is presumably the most important margin that drives the changes in the educational composition of developed societies in the future. By launching the Higher Education Pact 2020, for instance, Germany has recently made large public funds available in order to further increase access to college education. These points emphasize the complementary value of analyzing tertiary education: investigating effects at the college margin may help to gain a better and highly policy-relevant understanding of the previous findings.

This study examines the effect of college education on the number of biological children a woman has throughout her fertile ages (so-called completed fertility) as well as the extensive and intensive margins of fertility (probability of becoming a mother versus number of children once a woman is a mother). Moreover, we study two intriguing aspects of fertility decisions: the timing of births and socioeconomic channels that may help to explain the observed fertility patterns. By unfolding our main effects via the timing of their occurrence, we shed light on potential postponement and catch-up and possibly even biological effects. While the postponement of motherhood may emerge rather mechanically, e.g., through an "incarceration" in college (see Black et al., 2008), the degree of the catch-up is likely to reflect the preferences, for instance, for a family or a career. A biological effect may unfold through age-related fertility problems if the catch-up effect occurs too late to reach the desired family size. Whereas a social planner would wish to prevent the biological effect from playing a role (as women may well want, but cannot have, children), implications are less clear for catch-up effects in general as they may evolve through a college-induced change in preferences. To differentiate further whether catch-up effects – that may result in a decline in completed fertility – are driven by decreased family preferences (relative to career preferences), or by an incompatibility of work and family life, we investigate the effect of college education on career opportunities (assessed through labor supply and wages) and preferences and opportunities for family life (marriage, assortative mating, and offspring's education).

A pivotal prerequisite of these analyses is to separate correlative patterns from the underlying causal relationship. Women with initial preferences for large families might be more reluctant to sort into college education, for instance, because they expect the investment in their skills to have less time to pay off. Women with initial preferences for a career, on the other hand, might be very prone to study, since it fuels their labor market opportunities. These conflicting preferences exemplify the need to address selection into college education. To do so, we exploit arguably exogenous variation in the college expansion in Germany by means of an instrumental variables approach (see also Kamhöfer et al., 2017, who rely on the same instrument). Several higher education policies at the federal level and within the states caused the number of colleges in Germany to double between the 1960s and 1980s and led to an upsurge in the number of available college spots. At the same time, the local bargaining of the districts with the state governments and with each other plus the balancing of local interests caused regional variation between and within states. This process changed the opportunity to access college in a period of excess demand for college education. Quantitative evidence from an explorative study of the local determinants of college openings indeed indicates that differences in the opportunity to study are to a large degree exogenous.

Our results suggest that college education reduces the probability of becoming a mother by one-quarter, but college-educated women who do become mothers have, on average, 0.27 more children (about 13 percent) compared to their peers without college education. Looking at the timing of the effects (that is, the age of childbearing) indicates that a biological effect does not trigger the negative effect of college education on overall fertility: the increased (catch-up) fertility of college-educated women fades out before an age-related decline in fertility usually matters. The effects of college education on potential mediators suggest that the increased probability of working full-time due to college (compared to working half-time or not at all) and the college wage premium are higher for non-mothers; they are also less likely to be married, but do equally well in terms of positive assortative mating. From a policy perspective, these effects of college education on quantitative fertility outcomes can have crucial implications that are at least twofold. First, college education seems to trigger the demographic transition solely through its effect on childlessness, but not through the number of children per mother. If so, promising policies should aim at this margin. This is in line with an increasing number of economists, among others, who call for policies targeted at raising the compatibility between work and family life. Policies that, for instance, enable more flexible working hours and the opportunity of working from home may decrease the labor market burden of becoming a mother (see, e.g., Goldin, 2014). Moreover, family policies that are specifically aimed at higher educated women, such as means-tested maternity leave benefits (as analyzed by Raute, 2016) seem to be a step forward toward closing the baby gap. A second implication for further policies to consider arises through the positive effect at the intensive margin and evidence of a positive educational transmission that affects the socioeconomic composition of fertility. This has important long-term implications for societies (e.g., in terms of fiscal net effects), especially in societies with a low social or educational mobility (Raute, 2016).

The remainder of the paper is as follows: Section 6.2 briefly presents the general trends in fertility and higher education in Germany. Section 6.3 provides an overview of the college expansion and exploits both the qualitative and quantitative reasons that led to this expansion. The data and the empirical strategy are presented in Section 6.4. The main results on quantitative fertility effects are presented in Section 6.5. Subsequently, Section 6.6 sheds light on the timing and socioeconomic factors that potentially shape the detected fertility patterns before Section 6.7 concludes.

6.2 Trends in fertility and education in Germany

Using official statistics for the whole population, Figure 6.1 depicts the development in female college education and fertility over time in Germany. The horizontal axis states the birth cohort. The violet line gives the trend in the share of women per birth cohort who were enrolled in college at the age of 20 (referring to the vertical axis on the left-hand side). While only 5 percent of all women born in 1943 were enrolled in higher education in 1963, the number increased tenfold until the birth cohort 1972. After the baby-booming years succeeding World War II, the average number of births per women dropped from 1.8 to 1.5. The average number of children is assessed at the woman’s age of 40 for the birth cohort of the horizontal axis and plotted by the orange line (referring to the vertical axis on the right-hand side).

At first sight, Figure 6.1 suggests that the initial reduction in fertility was a prerequisite for the boom in female college enrollment. While this may be true, a further, substantial reduction in fertility occurred just after female college enrollment rates soared the most. As preferences for smaller families grew and contraceptive pills (whose commercial launch in Germany was in 1961, just after the cohort of 1940 decided whether to enroll in college) made it easier to meet the preferred number of children and females could “more accurately anticipate their work lives” (Goldin, 2006, p.8), which made human capital investments for women more valuable. This emphasizes how close fertility and female education are interrelated. Using variation in the availability of higher education, the empirical analysis in the following sections addresses the underlying causal relationship.

Another piece of suggestive evidence on the college education-fertility nexus is the relationship between the share of women in higher education and the average

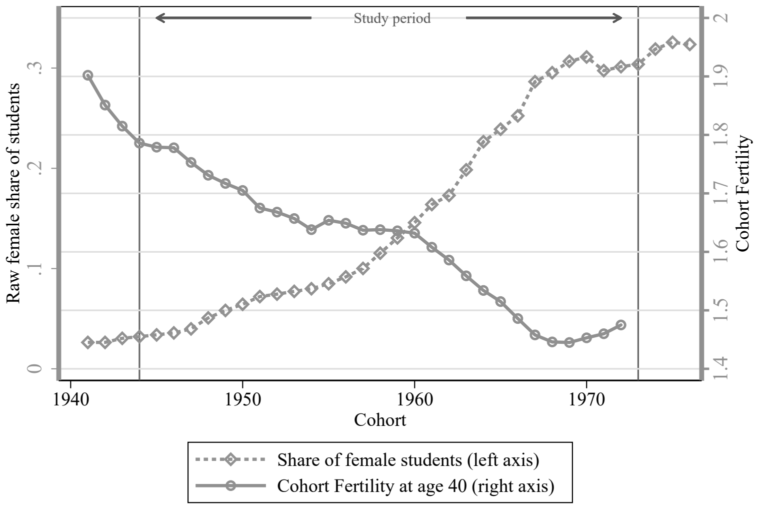


Figure 6.1: Trends in fertility and college enrollment by birth cohort in Germany

Notes: Own calculations using data from Max Planck Institute for Demographic Research and Vienna Institute of Demography (2014) and German Federal Statistical Office (2016b). The orange line refers to the axis on the right-hand side states the average number of children per women at the age of 40 by birth cohort. The violet line illustrates the share of women of the birth cohort that are enrolled in higher education at the age of 20 and corresponds to the vertical axis on the left-hand site. To transform the number of female students in the enrollment year into the cohort share of female students, we deduct 20 years from the enrollment year and take into account that only about one-fifth of women studying in a certain year are freshmen. We divide the resulting number of female students in total by the average study length of 4.5 years to get the number per year. Finally, we divide the number of female students in a certain year by the female cohort size in this year. Note that this is only a crude adjustment. However, as we are primarily interested in the change of this share over time, we are confident of capturing most of the changes.

age at the time of the first marriage as depicted in Figure 6.2. In the time under review, marriage was an important gatekeeper for fertility and births out of wedlock were rare events. The violet line (referring to the left vertical axis) gives the share of all women enrolled in higher education in a certain year. Unlike Figure 6.1, Figure 6.2 compares the share of females in higher education and the age at first marriage per calender year (and not by birth cohort). While the average age at the time of the first marriage decreased until the mid-1970s to 22.5 years, it increased by 2.5 years in the following 15 years (orange line on the right vertical axis). Based on the descriptive pattern in Figure 6.2, two things are important to note for the empirical analysis: First, marriage may mediate the effect of college education on fertility as the college enrollment decision predates the mean age at the first marriage in the figure. Second, the trend in the age at first marriage changes only a few years after the boost in the share of women in higher education, suggesting that college enrollment had an impact on fertility.

Moreover, Figure 6.2 also bears suggestive evidence of the empowerment of women. The delay in marriage indicates that the share of women that transitioned directly from living at home (where the parents presumably took care of subsistence) to living with the husband (and relying on his subsistence) decreased. In other words, Figure 6.2 suggests that the share of women who took care of their

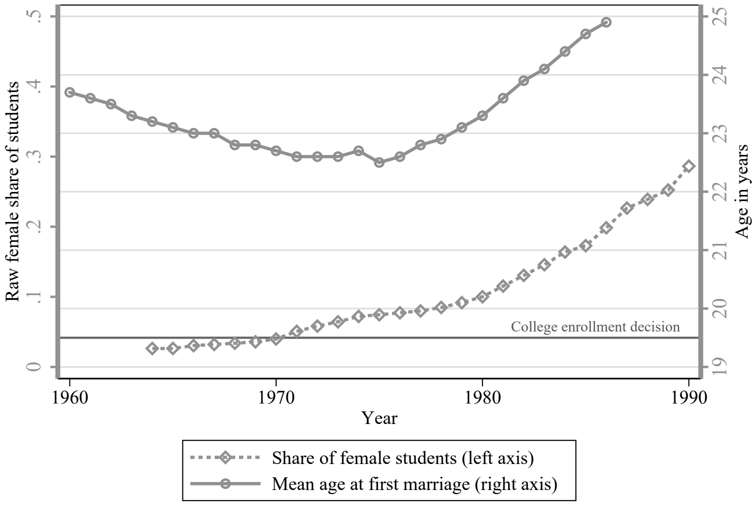


Figure 6.2: Mean age at first marriage and college enrollment by year in Germany

Notes: Own calculations using data from Max Planck Institute for Demographic Research and Vienna Institute of Demography (2014); German Federal Statistical Office (2016b). The violet line gives the share of women aged 20 per year and is shown in the vertical axis on the left-hand site. In 1970 this shows, for instance, the number of female students in higher education divided by the number of women at this time. The orange line referring to the right-hand site axis gives the average age of women at the time of the first marriage per year.

own subsistence (through working for pay or student loans introduced in 1971) increased over time.

6.3 The college expansion

6.3.1 Background and developments

Higher education in Germany

After graduating from secondary school, adolescents in Germany either enroll in higher education or start an apprenticeship training.[4] The latter consists of part-time training-on-the-job in a firm and part-time schooling. This vocational education usually takes three years and individuals often enter the firm (or another firm in the sector) as a full-time employee afterwards. To be eligible for higher education in Germany, individuals need a university entrance degree (*Abitur*). In the years under review, only academic secondary schools (*Gymnasien*) with nine years secondary schooling (and four years elementary schooling) could award this degree. The tracking from elementary school to secondary school took (and still takes) place rather early at the age of 10. However, it is generally possible

[4] The general description of education in Germany and the college expansion is closely related to Kamhöfer and Schmitz (2016) and has been adjusted for the purpose of the analysis conducted here.

to switch secondary school tracks after any term. Moreover, students could enroll into academic schools after graduating from the other tracks (with four to five years basic track schooling or six years of intermediate track schooling) in order to receive three additional years of schooling and be awarded a university entrance degree.

In Germany, higher education is, in general, free of tuition fees and several institutions offer tertiary education – even though the distinction of the different types is not always straightforward. We limit our analysis to the larger and most established institutions: universities and technical universities. We refer to the union of these institutions interchangeably as "universities" or "colleges." We neglect two groups of higher education institutions. First, small institutions that specialize in teacher education, religious education and fine arts with no more than 1,000 students at the time under review. The second group are universities of applied science (*Fachhochschulen*). They emerged in the 1980s (see Lundgreen and Schwibbe, 2008) and are usually smaller than regular universities, specialize in one area of education, have a less theoretical curriculum, and the style of teaching is more similar to secondary schools. In the time under review, the degree awarded was also distinct.

Build-up of new colleges and the rise in higher education enrollment

While the educational system as described above did not change in the years under review, the number of academic-track secondary schools and colleges significantly increased – providing us with an arguably powerful and exogenous source variation in educational opportunities. In this subsection, we describe the supply-sided expansion in the number of colleges and their capacities in terms of student spots as this is a prerequirement for the trends in college enrollment outlined above. This so-called period of "educational expansion" (*Bildungsexpansion*) started in the 1960s and peaked in the 1970s. In the years under review, 1958–1990 (determined by the birth cohorts in our survey data), the number of districts with at least one college (only very few districts had more than one college) increased from 27 to 54 (out of 325 districts) and the total number of students increased by over 850,000 from 157,000 in 1958 to more than one million in 1990 (see Figure 6.3a). The number of female students in total in the colleges in the sample in Figure 6.3b is similar to the corresponding number in Figure 6.1. This indicates that our college panel captures the bulk of the higher education institutions in Germany (although we do not have any data on smaller institutions, see above). Figure A6.1 in the Appendix shows the spatial variation over time. Following the reasoning of Card (1995) and many others since then (e.g., Currie and Moretti, 2003, Carneiro et al., 2011, and Nybom, 2017), we argue that availability of higher educational opportunities in large parts of the country led to a decrease in the opportunity costs of education due to the changed distances to college. While newly opened academic schools enabled secondary school students in rural areas to receive a university entrance degree, college openings in smaller cities allowed a broader group of secondary school graduates from both rural areas and cities to take up higher education. That is, the opening of new colleges allowed individuals to commute instead of moving to a city with a college (which causes higher costs) or decreased the commuting time. As indicated in Figure 6.3b, women especially benefited from this development as the share of women relative to men doubled from 20 to 40 percent in the time under review.

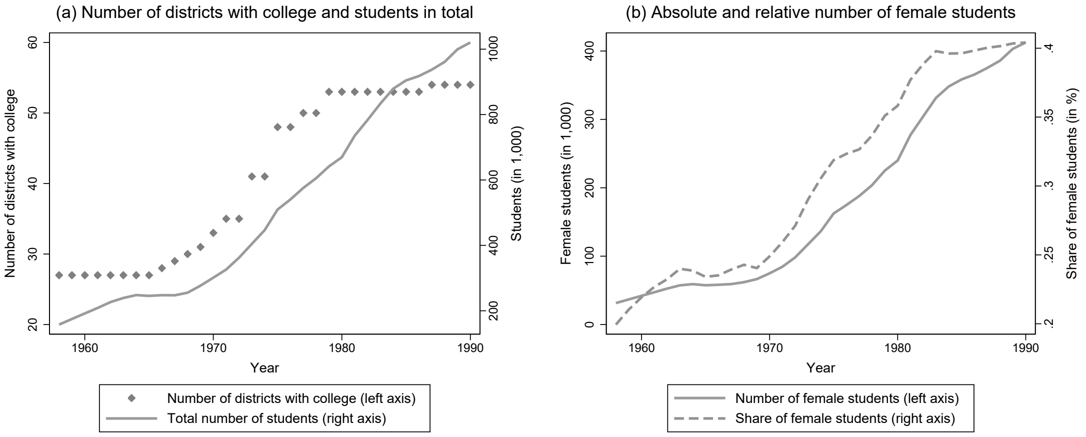


Figure 6.3: Colleges and students over time and by gender
Notes: Own illustration. College opening and size information are taken from the German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, 1991). The information on students refer to the college included in the left panel of the figure. More specialized higher education institutes that are smaller in size are disregarded as information on them are often missing.

6.3.2 Determinants of the college expansion

According to the analysis of Bartz (2007) of the history of higher education in Germany, mainly four factors triggered the college expansion: (*i*) The two world wars and the National Socialists' "anti-intellectualism" led to a low educational attainment for large parts of the population – as also argued in (Picht, 1964, p.66).[5] Therefore, large parts of society may have had an urge to catch up in terms of education. (*ii*) The industry demanded more qualified workers that were able to cope with new production technologies (see the review of the history of the first post-war era colleges of Weisser, 2005). (*iii*) As argued in Jürges et al. (2011) and Picht (1964), political decision-makers saw education both as an outcome and a

[5] Even today, more than 70 years later, the share of college students in Germany still does not meet OECD standards, see OECD (2015b) – even so this is at least in part due to the prominent role of the apprenticeship training system in Germany. To close this gap and increase participation in higher education the German federal government and the state governments launched the Higher Education Pact 2020 (*Hochschulpakt 2020*) in 2007 and funded it with 38.5 billion Euros until 2023.

means in the rivalries with the communist East Germany. (*iv*) All these reasons also led to an increase in academic track secondary schools – as analyzed by, e.g., Kamhöfer and Schmitz (2016) and Jürges et al. (2011) – which then led to an increase in the number of individuals eligible for higher education.[6]

It was partly because of these reasons that the federal government introduced the German Council of Science and Humanities (*Wissenschaftsrat*) in 1957, see Bartz (2007). In its 1960, 1966, and 1970 reports the expert council advised that college capacities should be largely increased (see Wissenschaftsrat, 1960, 1966, 1970). However, the council's authorities were (and still are) limited to making suggestions. The governments of the federal states in Germany are in charge of educational policies. The coordination between the states (which are usually ruled by several parties or coalitions of them and have elections at different points in time) mainly focuses on a standardization and mutual recognition of degrees. Figure A6.3 in the Appendix shows the number of colleges and shares of female students over time across the states. The timing of the educational expansion exhibits large differences between the states. In our analysis we use the variation in the timing between the 325 German districts (smaller administrative units, e.g., cities, that are nested in the federal states). Combining administrative data on the college expansion with survey data on individuals that face the college decision spread over more than 30 years, yields a panel structure in college availability. Eventually, this allows us to control for district fixed effects (as well as district-specific time-trends) and still observe a sufficient amount of variation in college availability.

In the following parts of this section we provide qualitative and quantitative evidence that this variation is exogenous with respect to individual fertility and marriage preferences.

Qualitative evidence

While the decentralized decision-making process makes it hard, if not impossible, to trace back the exact political reasons that led to each college opening or expansion in college size, we found evidence of the political reasoning behind some college openings. The first post-war college opening – the University of Bochum in the most-populated state of North Rhine-Westphalia in 1966 – was based on a

[6] Figure A6.2 in the Appendix the trend in academic-track secondary schooling. Two facts stand out: First, even in the expanding academic secondary schooling the share of female students rose disproportionately until women outnumbered men at academic secondary schools in 1990. Second, even in 1950 the share of women leveled at some 40 percent. The excess in the number of women eligible to take higher education compared to the number of women actually enrolled in colleges suggests that the academic school expansion might have been an important reason for the surge in female college participation but that it was certainly not the only one.

state's parliament decision in 1961. According to Weisser (2005), the first negotiations between the city of Bochum and the state government were even partly held in secret. This offended officials of the city of Dortmund – that also hoped to get the college – but was unable to provide a construction site that fulfilled the requirements. Facing state elections, the decision to open a college in Dortmund was made only one year after the announcement to open a college in Bochum.

The decision to open six new so-called comprehensive colleges (*Gesamthochschulen*) in North Rhine-Westphalia at the beginning of the 1970s was accompanied by a more intensive public debate. After several parliamentary hearings, the suggestion of the state's minister for educational affairs to construct new colleges in areas without existing ones was agreed on, see NRW (1971a,b). Four of the six colleges were opened in industrialized cities (Duisburg, Essen, Hagen, and Wuppertal) and two colleges were opened in more rural areas (Paderborn and Siegen). The college openings in these districts were supposed to actively "promote" education ("*Bildungswerbung*") and allow a larger range of secondary school graduates to enroll in higher education, see NRW (1971c).

All in all, we neither know of any law that relates college openings to potential reasons (like population size) nor could we find a pattern in the discussions to open colleges. On the contrary, the length of the political process and time from the opening decision to the start of the teaching exhibits a lot of variation. To investigate further which factors are associated with college openings, we conduct an additional quantitative analysis.

Quantitative evidence

Our concern regarding the exogeneity of college expansion is that certain characteristics, such as average fertility, age and living arrangements plus employment structure, systematically differ between regions with a college opening through the educational expansion and a region that had not experienced a college opening. To investigate this, we combine the data on college openings presented above with administrative data from the German Micro Census in 1962 (a 1 percent sample of the whole population, see Lengerer et al., 2008). Because the Micro Census data is on a slightly broader level we observe 249 regions (in which the 325 districts are nested). While 22 of these regions already had a college before 1962 and 206 regions had no college until 1990 or later, a college was opened in 21 regions in the years under review.

Table 6.1 shows the 1962 means of the regional characteristics that potentially triggered a college opening. Column 1 states the mean for regions that never experienced a college opening and column 2 gives the corresponding mean for

Table 6.1: Balancing test of regions with and without a college opening in the time under review using administrative data

	(1)	(2)	(3)	(4)
	Regions…			Predict opening using regression
	…w/o college opening	…w/ opening 1962–1990		
Pot. college det.	Mean	Mean	Diff.	OLS
Number of kids p.c.	10.497	10.437	−0.150	−0.033
(total population)	(0.522)	(0.283)	(0.121)	(0.052)
…students	0.016	0.011	−0.008*	−10.723
	(0.019)	(0.011)	(0.004)	(10.653)
…divorced	0.023	0.017	−0.005	−1.000
	(0.069)	(0.006)	(0.016)	(40.185)
…widowed	0.088	0.091	0.007**	20.035
	(0.015)	(0.008)	(0.003)	(20.357)
…females	0.525	0.528	0.002	−20.918
	(0.041)	(0.013)	(0.01)	(10.851)
…migration	0.021	0.018	−0.006	−10.698
	(0.022)	(0.017)	(0.005)	(10.545)
…unemployed	0.002	0.002	0.001**	250.484
	(0.001)	(0.001)	(0.00)	(190.743)
Sectoral composition of employment				
- primary	0.029	0.046	0.023*	0.390
	(0.055)	(0.053)	(0.013)	(0.497)
- secondary	0.543	0.551	0.008	0.147
	(0.088)	(0.069)	(0.02)	(0.367)
# of regions	206	21	227	227

Notes: Own calculation using German Micro Census data from 1962 (see Lengerer et al., 2008). Information on colleges are taken from the German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, 1991). Due to data policy restrictions Micro Census data are aggregated on regions defined through the degree of urbanization (*Gemeindegrößenklasse* indicators) and broader administrative units (*Regiergungsbezirk* level). This aggregation results in 206 regions that never experienced a college opening until 1990 or later (the mean value of the considered characteristics in these regions is given in column 1), 21 regions with a college opening between 1962 and 1990 (mean value in column 2), and 22 regions that already had a college in 1962 (data of these regions is not considered in the table). Due to a different aggregation of the Micro Census data, these numbers do not exactly correspond to those on the district level. The difference in column 3 is calculated by a simple regression of a college opening indicator on the potential characteristic and an intercept. Column 4 shows the coefficients of the characteristics in a multiple regression. The number of regions with and without a college opening differs slightly from Kamhöfer et al. (2017) as we restrict our analysis to universities that had 1,000 or more students in at least one of the years under review. Standard errors in parentheses; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

regions that experienced a college opening in the time under review. Column 3 gives the difference in means between the two. This reveals no significant difference between the regions in terms of number of children, marital status, share of females or other socioeconomic indicators such as share of migrants and unemployment rate. The share of students is lower in regions with an opening and where the employment structure differs slightly (more primary sector employment in districts with opening). This illustrates that colleges were often opened in order to foster accessibility for rather educationally alienated groups. In column 4 of Table 6.1, we regress an opening on all characteristics simultaneously. The stated coefficients give the difference of the factors in regions with and without a college opening while holding the mean differences in the other characteristics constant. The regression does not find any single factor in 1962 that significantly predicts an opening in the years until 1990. These auxiliary results are encouraging for our identifying assumptions, although differences in levels are in any case controlled for by the fixed effect in our analysis. How exactly we utilize the variation in college availability presented in this section is given in the following section.

6.4 Data and empirical strategy

6.4.1 Survey data and important variables

German National Educational Panel Study

Our main data source are individual-level data from the German National Educational Panel Study (NEPS), see Blossfeld et al. (2011).[7] NEPS data map the educational trajectories of more than 60,000 individuals in total. The data set consists of a multi-cohort sequence design and samples six age groups: newborns and their parents, preschool children, fifth graders, ninth graders, college freshmen students, and adults. These age groups are referred to as Starting Cohorts and are followed over time. That is, each Starting Cohort consists of a panel structure.

For the purpose of our analysis we make use of the Adult Starting Cohort that covers individuals born between 1956 and 1986 in, so far, seven waves between 2007/2008 (wave 1) and 2014/2015 (wave 7)[8], see LIfBi (2015). Starting with about 8,500 women, the final sample includes 4,300 women who (*i*) were educated in West Germany, (*ii*) are aged 40 or older, and (*iii*) have complete information in key variables. One of those key variables is the district of residence

[7] This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Adults, doi:10.5157/NEPS:SC6:7.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

[8] For every individual we use only the most recent observation.

at the time of the college decision or earlier, which we use to assign our instrument. Besides detailed information on education and fertility, including the years of childbearing, the data includes retrospective information on the respondents' labor market history and early living conditions at age 15, for instance, the number of siblings, secondary school grades, and parental education. As those factors are potentially confounding the effect of education on fertility, we consider them as control variables, see Table A6.1 in the Appendix for details.

The explanatory variable "college degree" takes the value 1 if an individual has any higher educational degree, and 0 otherwise. Dropouts are treated as all other individuals without college education. About one-fifth of the sample have a college degree, while four-fifth do not.

Dependent variables

The key dimensions along which we analyze fertility are the extensive margin (probability of becoming a mother) and the intensive margin (number of children conditional on being a mother). Table 6.2 gives the mean values of the dependent variables by college education. From the one-fifth of college-educated women about three-quarters have at least one child. For women without a college education, the share of mothers is about nine percentage points higher. Interestingly, once a woman decides to become a mother, the average number of children is almost the same for women with and without a college education (if anything, college-educated mothers have slightly more children). In other words, the main difference in the descriptives between college-educated and non-college-educated women is on the extensive rather than the intensive fertility margin.

As we consider the timing of birth as a crucial mechanism through which college transmits into fertility, Table 6.2 also gives the age of first birth. Mothers with a college education have, on average, their first child at the age of 30. Mothers without a college education are, on average, four years younger at the time of the first birth. Given a regular study duration of 4.5–5 years in order to receive a than-common *Diplom* degree, we interpret the descriptive evidence as pointing toward a strong role of college education.

Instrument

The processes of the college expansion discussed in Section 6.3 provide, on the one hand, a powerful shift in the availability of higher education for many individuals. On the other hand, the multi-faceted college expansion that took place over several decades is hard to boil down into one or a few still powerful in-

Table 6.2: Descriptive statistics of dependent variables

	(1)	(2)	(3)	(4)
	College stauts			
	all women	with college	w/o college	share w/ college
Motherhood				
all women (num. obs.)	4,288	924	3,364	21.6
mothers (num. obs.)	3,485	685	2,800	19.7
non-mothers (num. obs.)	803	239	564	29.8
share of mothers (in %)	81.3	74.1	83.2	
Number of children				
all women (incl. 0 kids)	1.65	1.52	1.69	
mothers (i.e., kids≥1)	2.05	2.10	2.04	
Age at first birth if mother	27.0	29.9	26.3	

Notes: Own calculations based on NEPS–Adult Starting Cohort data.

struments.[9] This is especially the case as we observe college openings. Using, for instance, a scalar for the distance to the closest college as suggested by Card (1995) might in the case of college openings even be misleading as newly opened colleges are in the initial years often too small to affect an individual's college decision. Moreover, the generally local nature of the IV results (see next subsection) makes it desirable to have an instrument that affects as many individuals as possible and therefore als captures, for instance, the expansion in the capacities of the already existing colleges. To achieve such a powerful instrument, we follow Kamhöfer et al. (2017) and create an index that weights the non-linear effect of the college distance with the relative number of students in the 325 West-German districts:

$$Z_{it} = \sum_{j}^{325} K(dist_{ij}) \times \left(\frac{\#students_{jt}}{\#inhabitants_{jt}} \right). \tag{6.1}$$

This college availability index Z_{it}, that is, the instrument, basically includes the total number of college spots (measured by the number of students) per inhabitant in district j (out of the 325 districts), individual i faces in year t weighted by the distance between i's home district and district j. Weighting the number of students by the population of the district takes into account that districts with the same number of inhabitants might have colleges of a different size. This local

[9] Westphal et al. (2017) use the same source of variation in an IV setting but assess the most powerful instruments of many potential indicators using machine learning techniques.

availability is then weighted by the Gaussian kernel distance $K(dist_j)$ between the centroid of the home district and the centroid of district j. The kernel gives a lot of weight to close colleges and a very small weight to distant ones. Since individuals can choose between many districts with colleges, we calculate the sum of all district-specific college availabilities within the kernel bandwidth. Using a bandwidth of 250km, this basically amounts to $K(dist_j) = \phi(dist_j/250)$ where ϕ is the standard normal pdf. While 250km sounds like a large bandwidth, this implies that colleges in the same district receive a weight of 0.4, while the weight for colleges that are 100km away is 0.37, which is reduced to 0.24 for 250km. Colleges that are 500km away only get a very low weight of 0.05. A smaller bandwidth of, say, 100km would mean that already colleges that are 250km away receive a weight of 0.02 which implies the assumption that individuals basically do not take them into account at all. Table A6.2 in the Appendix gives an overview of the variation in the instrument as well as providing some descriptives on some main driving forces behind this variation (changes in the distance to the nearest college, within a 100km radius and changes in college spots).[10]

6.4.2 Empirical strategy

The most natural starting point is an ordinary least square (OLS) estimation where we regress our fertility measures Y_{itd} for individual i who graduated from high school in district d and year t on a binary college indicator D_{itd} (that takes on the value 1 for college, and is 0 otherwise) and a vector of control variables $\boldsymbol{X}'_{itd}$:

$$Y_{itd} = \beta_0 + \beta_1 D_{itd} + \boldsymbol{X}'_{itd}\boldsymbol{\beta}_2 + u_{itd}. \tag{6.2}$$

In order to separate the general trend in college education from the reverse trend in fertility (as depicted in Figure 6.1), the vector of confounders, $\boldsymbol{X}'_{itd}$, also includes district-specific linear trends in addition to general time and district fixed effects. The district-specific trends accommodate temporal confounding factors, for instance, because of global and district-specific trends in secondary school graduation (see, e.g., Figure A6.2 in the Appendix and Westphal, 2017).

However, if individuals simultaneously select themselves into education and desired fertility beyond some underlying trend, β_1 is still likely to be biased. The direction of the bias is a priori unclear and depends on the effect of the omitted confounder on fertility and its correlation with education. If the omitted factors are, for instance, career preferences or preferences for a traditional family model

[10] For alternative specifications of the instrument, see Kamhöfer et al. (2017).

that are already established before college, OLS would overestimate the true college effect.[11] On the other hand, OLS may underestimate the true effect if factors such as the family's wealth are omitted from the model.[12] Also, general preferences for having a family do not necessarily lead to an overestimation of OLS, as females with these preferences may very well decide to study (as college is considered to be one of the largest marriage markets).

In order to address the selection of individuals in education and fertility along unobserved preferences we exploit the variation in college availability using the index of college availability we define in Eq. 6.1 as an instrumental variable in a two-stage least-squares (2SLS) approach. The first stage of the 2SLS approach reads:

$$D_{itd} = \delta_0 + \delta_1 Z_{td} + \boldsymbol{X}'_{itd}\boldsymbol{\delta}_2 + v_{itd}. \tag{6.3}$$

Our main identifying assumption is that conditional on $\boldsymbol{X}'_{itd}$, variation in our college accessibility measure (Z_{td}) randomizes the otherwise endogenous decision to go to college, that is, variation in Z_{td} does no depend either on the error term, v_i, or on general preferences about or other unobserved characteristics with respect to fertility.

To make this assumption as plausible as possible, we condition on district fixed effects to effectively use only the openings of new colleges and within-district increases in college seats. With the additional assumption that any instrument-specific shift in D only affects some of our employed fertility measures via college graduation (i.e., the exclusion restriction), we can attribute the reduced-form effect of the instrument solely to college graduation, ruling out any other channel. Technically, this is done by regressing the first-stage fitted value $\widehat{D}_{itd}$ on the fertility measures, Y_{itd}:

$$Y_{itd} = \beta_0 + \beta_1 \widehat{D}_{itd} + \boldsymbol{X}'_{itd}\boldsymbol{\beta}_2 + u_{itd}, \tag{6.4}$$

Given our identifying assumptions, β_1 is the causal effect of college education. Imposing a monotonicity assumption on the instrument, β_1 is a causal effect for a specific group of women: those who would potentially go to college because of the instrument (called compliers). Because this group is typically a subset of all individuals, β_1 is referred to as the local average treatment effect (LATE, see Im-

[11] In the case of career preferences women may sacrifice children for a career-boosting education. If women prefer a traditional family model, they may forgo college education in favor of starting a family at an earlier age.

[12] Although the observable confounders include the parents' education, we cannot directly control for the family income at the time of the college decision. If the family income buys high-quality child care and the woman's education beyond what is captured by through the control variables, this would downward-bias OLS. Another potential unobservable confounder that would bias OLS in the same direction is a high degree of openness – one of the so-called Big Five personality traits in psychology – describing the appreciation and curiosity for a variety of experiences, e.g., college life and having children.

bens and Angrist, 1994). In our example, the compliers are most likely those who could go to a university because either a university opened up in their proximity or because existing universities in the neighboring districts expanded. As this process potentially affected many people, one would expect the share of compliers to be rather large – a claim we are going to investigate in the following section.

Before turning to the results, we want to briefly assess whether our assumptions are plausible. The conditional independence assumption would be violated by district-specific, non-linear fertility trends that are correlated with an opening. These trends could be caused by different access to modern contraceptives like the combined oral contraceptive pill that was introduced in Germany at the beginning of the 1960s. If women in regions with a stronger increase in college availability also had better access to the pill, we may falsely attribute the contraceptive effect to education (to alleviate this concern, we include district-specific trends). We consider this as rather unlikely because Table 6.1 suggests that the levels of aggregate fertility measures are uncorrelated with the opening of a university. What is more likely is that college-educated women were more willing to use contraceptives in order to regulate fertility (see Oddens et al., 1993), which would be a channel of the effect rather than a violation of the identifying assumptions.

6.5 Baseline results

6.5.1 The effect of the college expansion on educational participation

First-stage evidence from Micro Census data

Before looking into the effect of the college expansion on the probability of studying using the survey data that includes fertility measures, we look at the effect of the college build-up on educational participation in the German Micro Census from 1962 to 1969 (the first years available). The openings of the first four post-war era colleges (in the cities of Bochum, Dortmund, Konstanz, and Regensburg) fall into these years. To shed some light on the exact impact of college openings, we conducted an event study to see the relative change in the share of students within a 100km radius relative to the timing of the opening of these colleges (time of opening centered to 0).

The results are depicted in Figure 6.4 which shows a twofold takeaway. First, there is no evidence on pre-trends, indicating that the colleges were not opened in regions where already existing colleges were expanding relatively more than the colleges in regions without an opening. Second, the figure reveals a relatively sharp discontinuity: after a college was opened in $t = 0$, there was a rather large and significant increase in the relative share of students in the region even two years after the opening. Given that the colleges had just opened, this is a remarkable effect. As we take all students in regions within a 100km radius, the increase in the number of students not only captures the somewhat mechanical effect in the region of the opening itself but it also suggests that individuals from neighboring regions were also affected by the opening, for instance, because the newly built college was within commuting distance. We take this as evidence that there was an excess demand of secondary school graduates who wanted to go to college.

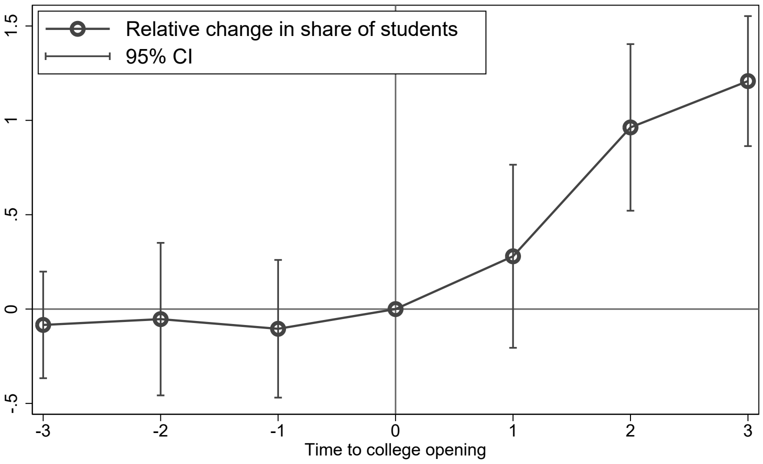


Figure 6.4: Relative change in the share of students in counties within 100km of college opening between 1962 to 1969

Notes: Own representation based German Micro Census data from 1962–1969 (see Lengerer et al., 2008) and German Statistical Yearbooks (see German Federal Statistical Office, 1991). The figure depicts the coefficients β_τ from the following "event-study" regression where β_0 is set to zero:

$$\begin{aligned}\ln(\#\text{students}_{bt}) &= \alpha_t + \sum_{\tau\in\{-7,-1\}} \beta_\tau \mathbb{1}\left[\max(t - t_b^{\text{opening}}, -3) = \max(\tau, -3)\right] \\ &\quad + \sum_{\tau\in\{1,7\}} \beta_\tau \mathbb{1}\left[\min(t - t_b^{\text{opening}}, 3) = \min(\tau, 3)\right] + \gamma_b + \epsilon_{bt},\end{aligned}$$

where $\ln(\#\text{students}_{bt})$ is the log number of students in region b and year t (1962–1969). α_t are year fixed effects. t_b^{opening} equals the the year in which a college opened in region b. To control for differences in levels between these regions, region fixed effects γ_b are included. Regions include all regions within a 100km radius surrounding the centroid of the region where the new colleges are located. The reason for the choice of this radius is that we want to go beyond a somewhat mechanical effect which emerges by the influx of students in the region of the opening. A sufficiently large radius partials out this effect for two reasons. First, it captures the bulk of the catchment area of a college and therefore only a minority of students do not come from the area defined by the radius. Second, within each region that exhibited an opening of a college (Bochum, Dortmund, Konstanz, Regensburg) there are already well-established existing colleges (Münster, Cologne, Freiburg or Nuremberg). Hence, there had been possibilities to enroll into a college in the defined area also in the absence of a college opening in period 0.

First-stage evidence from survey data and the complying subpopulation
The regression results of the first stage from Eq. 6.3 using NEPS data are shown for both the final sample and for certain subgroups in Table 6.3. The overall first-stage effect is very strong and is precisely estimated. To ease the interpretation of the compound instrument (defined in Eq. 6.1), we illustrate the first-stage effect with an example: a college is newly opened in a district with 250,000 inhabitants and 15,000 students are enrolled in the college five years after the opening. In this case, the probability of studying increases for a woman who graduates from high school in this district by about 6 percentage points (pp) based on the results in Table 6.3: 2.08 (coefficient from the table) $\times K(0) \times {}^{15}/_{250} = 2.08 \times 0.4 \times 0.06 =$ 5pp (rounded, see Eq. 6.1). With an overall baseline probability of studying of 21.5 percent for women, the first stage is not only statistically significant (the resulting F-statistic is well above the rule-of-thumb value of 10) but is also substantial in size.

This first stage determines the share of individuals for which the second-stage conditions the effect on college education (that is, the compliers). By comparing the first-stage effect of increased college availability on the probability of studying across different subgroups, it is possible to gauge whether certain individuals were more likely to comply with the college expansion and, thereby, be captured by the second stage. To this end, we repeat the first-stage estimation along three potentially important characteristics by which we separate our data. The first subgroup is defined by the school degree of the father. This separation may be informative since it sheds light on the question of whether the educational expansion increased educational mobility. High-educated fathers are defined as having at least an intermediate track education, and hence more than the most common educational degree of that time. The shares of both subgroups are approximately balanced. However, the first stage is much stronger for women with lower-educated fathers as is evident from Table 6.3. Calculating the relative frequency of compliers of low-educated fathers relative to high-educated fathers ($0.63/0.37 = 1.7$, see table notes for details) indicates that a woman with a father we define as low educated is nearly twice as likely to comply with the college expansion as a woman with a high-educated father. Hence, in the example above, the college opening is supposed to increase the probability of studying by $0.06 \times 1.7 = 10.2$pp for daughters of lower educated fathers.

Splitting the sample by the women's year of birth one can calculate the corresponding complier shares. The results show that the first-stage effect and, hence, also the share of compliers, is only slightly larger for women born after 1960, suggesting that our instrument has power throughout the educational expansion. This piece of evidence is moreover likely to be informative regarding the external

Table 6.3: First stage and some characteristics of complying mothers

	(1)	(2)	(3)	(4)
	Coefficient of the First Stage	Share of the population	Share of compliers	Obs.
Overall first stage	2.08***	1	1	4,288
	(0.11)			
First stage by education of father[a]				
– High-educated fathers	1.63***	0.48	0.37	2,045
	(0.16)			
– Low-educated fathers	2.49***	0.52	0.63	2,243
	(0.15)			
First stage by year of birth (median separation)				
– Before 1960	1.78***	0.47	0.41	1,996
	(0.23)			
– 1960 or later	2.19***	0.53	0.59	2,292
	(0.12)			
First stage by urban-rural separation				
– Urban	2.12***	0.76	0.78	3,275
	(0.12)			
– Rural	1.89***	0.24	0.22	1,013
	(0.23)			

Notes: Own calculations based on NEPS–Adult Starting Cohort data. The shares of compliers are calculated as follows: For mutually exclusive groups (denoted by subscripts 1 and 2), the overall first stage coefficient is a weighted average of the respective subgroups if the group indicator is also interacted with the set of controls. In this case, weights are determined by the group shares ω_1 and ω_2 of the overall population. Thus, $\widehat{\delta_{\text{overall}}} = \widehat{\delta}_1\omega_1 + \widehat{\delta}_2\omega_2$. Accordingly, the shares of compliers can be determined as $\pi_j = \widehat{\delta}_j/\widehat{\delta_{\text{overall}}} \times \omega_j$, for $j \in \{1,2\}$. In this table, the group indicators are not interacted with all the controls, in order to present the same first stage result as employed for the main results. Therefore, the weighted average may not hold with equality until we normalize the weights π_j such that $\pi_1 + \pi_2 = 1$. This procedure has also been applied in Akerman et al. (2015). Standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

[a] High-educated fathers are defined to have at least an intermediate track education, and hence more than the most common educational degree of that time.

validity of the results. As the first-stage effect does not seem to be confined to certain years in the time under review, it is not implausible to conjecture that more recent policies have also had similar effects on promoting educational education.

The last dimension by which we analyze the first stage is the degree of urbanization. The first-stage coefficient is slightly higher in urban regions compared to the overall effect. Yet, as most college openings occur in cities, this urban-rural gradient of the educational expansion should not come as a surprise.[13] But in rural regions there is a substantial share of compliers that is nearly as high as the share of rural high school graduates in the overall population.

All in all, we interpret the finding of the subgroup analysis as suggesting that the complying population, although modestly selected, is not confined to any specific subgroup.

6.5.2 The effect of college education on fertility

Starting with overall completed fertility, shown in panel A in column 1 of Table 6.4, the OLS effect (that is, the association) of college education on the number of children is -0.1. In other words, given controls, women who went to college have, on average, 0.1 fewer children than women without a college education. Taking into account selection that goes beyond the observable factors, the 2SLS estimate in panel B yields a reduction in the average number of children of -0.3. Given an average number of 1.7 children in Table 6.2, this corresponds to a reduction of 19 percent – a rather sizeable effect. With 4.5 years of college education, the per-year reduction that goes along with college education is, on average, 0.02 children in the OLS model and 0.05 children in the 2SLS specification.

Taking a closer look at the composition of the overall effect, we take the fertility margins as dependent variables. The OLS point estimate of college education on the extensive margin (that is, motherhood) is -0.08 (-0.02 per year of college). Put differently, women who went to college are 8pp less likely to ever bear a child, given the controls. Addressing endogeneity, the 2SLS estimate in panel B yields a reduction in the probability of becoming a mother through college education of about 21pp (5pp per year). Again, the effect is precisely estimated and is large in size (the baseline probability is 83.2 percent for females without college).

Turning to the intensive margin in column 3 of Table 6.4, we see that the negative effect from the extensive margin does not propagate here. The differential in the number of children is slightly positive when it is controlled for observables. Going to the structural estimate, college-educated mothers have, on average, 0.267 children more than their peers without college education. Given that mothers have an average of 2.1 children, the relative effect amounts to a 12.7 percent increase in the number of children of college-educated mothers. Although only statistically significant at the 10 percent level, the effect size is substantial. However,

[13] That regions with college openings have, on average, a larger share of primary industries - and are thereby more rural - may seem to contradict the result of Table 6.1. However, the degree of urbanization used here is only based on the number of inhabitants, not on the population density.

Table 6.4: Baseline regression results

	(1)	(2)	(3)	(4)
	Total Effect	Fertility margins		Timing
	# of children for all women	Extensive: motherhood indicator	Intensive: # of children for mothers	Maternal age at 1st birth
Panel A: OLS regression				
College degree	-0.106^{*}	-0.081^{***}	0.123^{*}	2.752^{***}
	(0.052)	(0.019)	(0.051)	(0.232)
Panel B: Second-stage 2SLS regression				
College degree	-0.313^{*}	-0.209^{***}	0.267^{*}	6.463^{***}
	(0.149)	(0.054)	(0.134)	(0.741)
Number of observations:	4,288	4,288	3,316	3,259

Notes: Own calculations based on NEPS–Adult Starting Cohort data. Control variables include full sets of year of birth and district fixed effects as well as state-specific trends. Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

this result for the intensive margin may be taken with a grain of salt as it refers to the selected sample of women who decide to have children. The composition of this sample in terms of the desired family size may depend on the individual effect of college education on motherhood. Put differently, the estimate for the intensive margin only yields the causal effect of college education if the desired family size does not systematically differ for college-educated mothers compared to women who do not become mothers because of college education. Keeping this limitation in mind, we still deem the countervailing signs of the effects on the two margins an interesting finding that we ought to have a closer look at in the following section.

Before building the bridge to potential mechanisms that may contribute to explaining the results, the rather new margin of education considered here calls for a careful comparison of our findings with the literature on the secondary schooling effects on fertility. For Germany, the OLS estimate for the effect of an additional year of secondary schooling on the average number of children provided by Cygan-Rehm and Maeder (2013) is -0.020 – this is remarkable close to our per-year OLS estimate of -0.024. Instrumenting secondary education with compulsory years of schooling, Cygan-Rehm and Maeder (2013) find an effect ranging from -0.10 to -0.17 depending on the specification. This is more than twice as big as the pre-year effect of college education. The bigger effect may seem contradictory at first sight, given that college education is probably more relevant for later career opportunities and affects individuals in their prime reproductive ages. However, while interpreting the effect size, one has to keep in mind that the compulsory schooling reform affects individuals at the lower end of the educational distribution and – given the baby gap in education – the average number of children is higher at this margin. Accordingly, the 2SLS effect on childlessness by Cygan-Rehm and Maeder (2013), about 5pp (compared to a baseline probability of 18 percent) exceeds our effect of college education on motherhood by about 5.7 percent (that is, (-0.209/0.813)/4.5 years=0.057). Fort et al. (2016) find similarly large effects of compulsory schooling on the number of children and childlessness for England and pooled Continental European countries.

Moreover, our results confirm another interesting pattern found by several studies on the secondary schooling effect (e.g., Cygan-Rehm and Maeder, 2013, Fort et al., 2016 and Monstad et al., 2008): the OLS results underestimate the 2SLS effects in absolute terms. This indicates that the bias in the OLS results stems from omitted variables such as unaccounted family income and openness to new experiences rather than from pre-college career preferences or preferences for a traditional family (where more children are preferred to a mother's college education). Another explanation as to why OLS underestimates the 2SLS result might be that OLS captures the average treatment effect while the 2SLS model yields the LATE for the complying subpopulation. However, as the complier analysis in Section 6.5.1 indicates that college expansion is not limited to particular groups of individuals, the local nature of the 2SLS estimate seems rather unlikely to drive the pattern of the results presented here.

Moving on to potential explanations of the education-fertility nexus, the most obvious effect of college education on fertility is through the timing of births. If the distribution of the age at the first birth is simply shifted by the time women spend in college (usually 4.5 to 5 years in Germany), some women may become too old to bear a child, which may then explain the negative effect on the extensive margin. This is investigated in column 4 of Table 6.4. Whereas the average observable-adjusted difference on age at first birth is 2.8 years between college-educated and non-college-educated mothers, the 2SLS effect is higher. Because of college, mothers defer their first birth by nearly 6.5 years, which is even higher than the time they usually spend in college. Because this effect is more than a mechanical shift, unraveling the exact timing of its occurrence seems to be promising for giving a more complete picture of the fertility pattern.

6.6 Heterogeneity and potential mechanisms

6.6.1 Effect heterogeneity along age

Unfolding the college effect by age

By its very nature, the decision to go to college affects an individual's life differently while the individual is in college (investment period) and after she leaves college (consumption period). Such effect heterogeneity in the returns to college education along women's fertile ages is not only informative in its own right but it may also help to explain the findings of the previous section. To describe the effect of education on "the desire/time/opportunity to have a child" while in school, Black et al. (2008, p.1044) coin the term "incarceration effect." Although they look at the fertility returns to education at the secondary schooling margin, such an incarceration effect is likely to matter at the college margin as well since the time in college is, on the one hand, often characterized not only through more flexible working hours, but also through an increased workload and pressure as well as tighter budget constraints. To detect this kind of heterogeneity, we estimate our baseline models for the extensive and the intensive fertility margins fully saturated by women's age to get age-specific effects. To this end, we reshape the data from individual level i to individual-age level ig, where g now indicates the age of the woman for each year from 17 to 40. The second stage of the 2SLS model is then:[14]

$$\begin{aligned} d_{ig} &= \beta_0 + \beta_1 \widehat{D}_i + \sum_{g=17}^{40} \eta_g \mathbb{1}(age_{ig} = g) \\ &\quad + \sum_{g=17}^{40} \Big[\gamma_g \mathbb{1}(age_{ig} = g) \times \widehat{D}_i \Big] + \boldsymbol{X}_i' \boldsymbol{\beta}_2 + u_{ig}. \end{aligned} \tag{6.5}$$

The indicator functions $\mathbb{1}(\cdot)$ return the value 1 if the observation refers to individual i at age g, and 0 for other fertile ages but g. In other words, the first sum gives a full set of age fixed effects and the second sum interacts the age fixed effects with the college indicator. The interpretation of the dependent variable d_{ig} and, thereby, the interpretation of the coefficients of interest differs depending on whether fertility is measured at the extensive or the intensive margin:

- At the extensive margin, d_{ig} is a binary indicator that takes on the value 1 if woman i becomes a mother at age g (and 0 otherwise), given that she does not have a child until age $g-1$. The age fixed effects η_g give the baseline

[14] For the sake if simplicity, the subscripts for the time and the district are now implicit. The standard errors are clustered on an individual level as shocks are likely to be time persistent.

hazard rate of having the first child (given that one does not already have a child) at age g. The coefficients of interest γ_g give the effect of college education on the baseline hazard. That is, they answer the question "How does college education affect the probability of bearing the first offspring at age g, conditional on having never given birth before?"

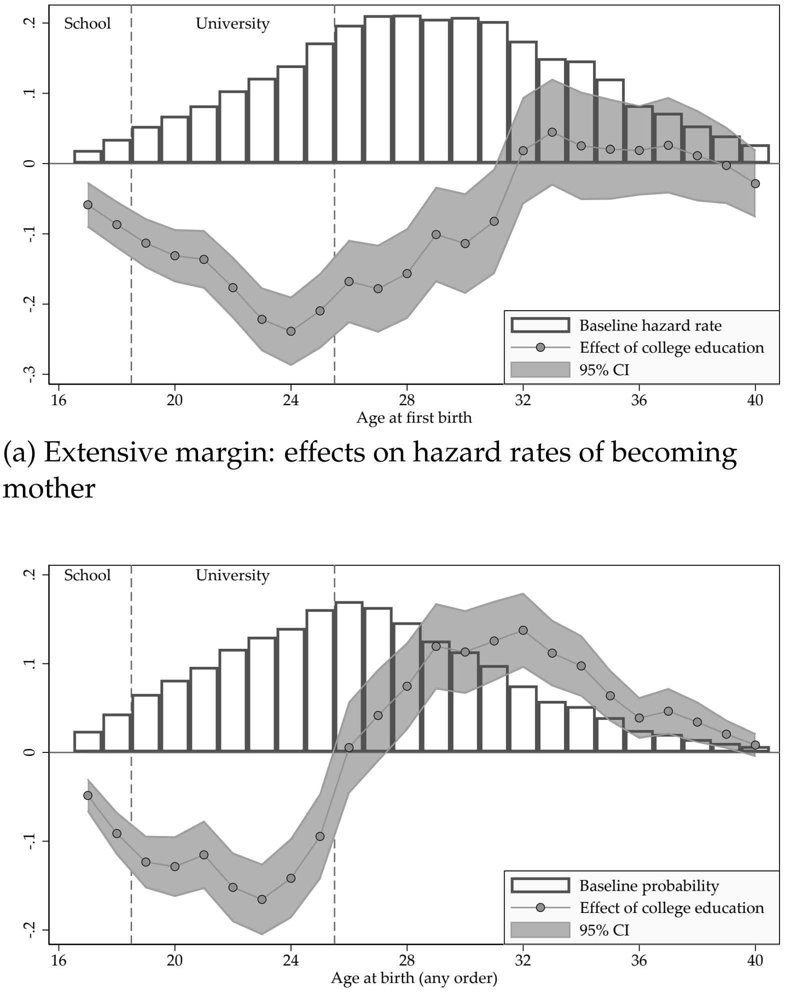
- At the intensive margin, d_{ig} is 1 if woman i gives birth at age g (and 0 otherwise) – independent of whether woman i already has a child or not. Accordingly, η_g is the baseline rate of having any child at age g given the woman is going to have a child by the age of 40 (as the sample for the intensive margin only consists of women who become mothers). The coefficients γ_g answer the question "How does college education affect the probability of giving birth at age g for women who have at least one child by the age of 40?"

Pre-, in- and post-college effects on fertility

Figure 6.5 shows the estimation results of Eq. 6.5 for the extensive margin of fertility in panel (a) and intensive margin in panel (b).[15] The bars state the baseline hazard rate of becoming a mother and the baseline probability of giving birth at a certain age in panel (a) and (b), respectively.[16] The oranges lines give the effect of college education on these baseline probabilities. For the sake of interpretation, we may think of the fertile ages as three phases for which we expect distinct effects: pre-college teenage years, years in college, and post-college years. In the first phase, giving birth (that is, teenage motherhood) is rather unlikely at both margins – as indicated by the small left-most bars in both panels of Figure 6.5. Interestingly, women who go to college a couple of years later already have lower probabilities of giving birth at pre-college ages (indicated by the orange lines below zero). An explanation for this may be that some women have such a strong family preference established prior to college age that they sacrifice additional education in favor of early motherhood and become a mother immediately after leaving secondary school. These women are never-takers of the college expansion.

[15] As the age-specific estimates in panel (a) after age 17 refer to the hazard of giving birth to the first child conditioning on not yet being a mother, the estimates may not be taken for the unconditional causal effect of becoming a mother at a certain age. Similarly, the estimates in panel (b) may not state the causal effects if the number and timing of children depends of the effect of college education on motherhood.

[16] Note, the baseline rates plotted in Figure 6.5 state the unconditional means. On the contrary, η_g in Eq. 6.5 are the conditional means after adjusting for college education and controls for non-college-educated women. We interpret the effect size (depicted by the orange line) relative to the unconditional mean as conventional for linear probability models.



(a) Extensive margin: effects on hazard rates of becoming mother

(b) Intensive margin: effects of bearing offspring for mothers

Figure 6.5: Timing of births

Notes: Both panels depict the age-specific regression coefficients from the second stage of the 2SLS model in Eq. 6.5 that capture the effect of college education. Panel 6.5a reports the effects of college education on the hazard rate of becoming a mother by age. Panel 6.5b depicts the respective effects on the probability of giving birth conditional on being a mother.

The next phase in fertile ages are the years in college around the ages 19 to 25 when women with a college education are in college and those without a college education usually complete their apprenticeship training and start working. Both baseline probabilities of motherhood/giving birth increase from year to year in this phase. Unsurprisingly, the negative effect of college education is most pronounced in the in-college years. While the baseline hazard of becoming a mother in panel (a) increases from 5 to 18 percent, the hazard rate for women in college is 11 to 25pp lower. Similarly, the baseline probability of giving birth in panel (b) ranges between 7 and 17 percent, while college education reduces the probability up to 17pp. It may at first sight be puzzling that the college effect exceeds the baseline probabilities. However, the baseline hazard rate/probability is much stronger for women who do not go to college (up to 14pp at age 25 when the baseline hazard for becoming a mother in college is just 7 percent, see Table A6.3

in the Appendix). Indeed, the increase in the hazard/probability of childbirth for women without a college education together with an increasing negative college effect in the in-college years, supports the incarceration explanation. While non-college-educated women completed their vocational training-on-the-job and gain in financial security from year to year in their mid-20s, the workload and stress level of women in college increases as they face their final examinations.

The third and final phase in fertile ages starts when individuals with a college education leave college – around the age of 25. At these ages college-educated women will reveal their preferences about fertility. Among the college-educated women who have not yet had a child, some may decide to remain childless (as indicated by the negative extensive margin in the baseline results), while others who postponed motherhood start a family. At this phase the pattern differs considerably between the extensive margin in panel (a) and the intensive margin in panel (b). At the extensive margin, the post-college ages can be further divided into two stages. First, from ages 26 to 32, the negative effect of college education decreases but college-educated women remain significantly less likely of becoming a mother. In other words, some college-educated women catch up with their non-college peers and give birth to their first child. Still, the college effect remains negative as some women who would have become mothers without a college education decide against children because of college education. At the second stage of the post-college fertile ages, starting around age 32, there is no significant difference in the probability of college- and non-college-educated women becoming mothers. Put differently, there is no catch-up effect in the first birth after the age of 32. The pattern in panel (a) suggests two things: First, the negative effect at the extensive margin in the baseline results is driven through the lower fertility of college-educated women during the years in college and about seven years after leaving college – that is, the time in which they build a working career. Second, the reduction in the negative college effect for women at the end of their 20s and the indistinguishable hazard rates (zero effects of college education) afterwards indicate that women who wish to catch up in terms of becoming a mother do catch up. Form a policy perspective this absence of an age-related reduction in fertility (we refer to this as the "biological effect") is a noteworthy finding. It indicates that the catch-up effect not meeting the incarceration effect is driven by preferences or opportunities for a career or family life. On the contrary, a constant relative increase in the hazard rate of the first birth of college-educated women at the end of their 30s would indicate that some women may wish to catch up but are not able to do so before age-related fertility problems become an issue.

At the intensive margin, the baseline probability of giving birth is more pyramid-shaped with lower probabilities at older ages compared to the extensive margin.

As for the extensive margin, the effect of college education on childbirth in the post-college ages can be divided into two stages. The first stage, until age 32, is characterized by a catch-up effect that already starts in the last years of college education, at around 23. Compared to the extensive margin, the catch-up effect is much more pronounced at the intensive margin and college-educated women are significantly more likely to give birth from age 28 onwards. However, the positive effect shrinks between age 32 to the end of the 30s (although college-educated mothers are still more likely to have a child than their non-educated peers, see Table A6.3). Thus, for women who decide to become a mother, the negative effect of incarceration in college in the first half of their 20s is compensated by an increased fertility until the end of the 30s. The effect remains positive and significant after the age of 30. The probability that a college-education women will give birth is around 10 percent at age 34 and falls to 5 percent at age 37 and 2 percent at age 39. This indicates that a biological effect can potentially restrict the desired fertility of college-educated mothers because if infertility affects both women at the same rate, college-educated mothers are more affected since they are still trying to catch up at those ages. If such an effect exists (it is, for instance, unclear whether the drop in the probability childbirth between 37 and 39 is already affected by fertility problems or not), it is rather humble in size, however.

Summing up the results for both margins, it seems likely that there are different types of college-graduated females – those who catch up in their fertility immediately after leaving college and those who postpone childbearing even further after the in-college incarceration and may never have children. For the latter group, the prolonged postponement and the seemingly absent age-related fertility decline raises the question of other causes for this lower fertility? Or, put differently, what shapes the smaller catch-up effect? Black et al. (2008) consider a "human capital effect" – that is, college education increases wages and, thereby, opportunity costs of family life. Besides such a career channel, the literature on secondary schooling and fertility suggests that education may change the preferences for and opportunities of family life. Education can enable women to find a more-educated and higher-earning partner and to have not only more but also better-educated offspring that could in turn affect the desired fertility (see, e.g., McCrary and Royer, 2011, for assortative mating and Currie and Moretti, 2003, for the intergenerational transmission of education). We now go on to investigate the effects of college on career and family variables for women with and without children that might explain the catch-up effects.

6.6.2 Opportunities and revealed preferences for career and family life

Table 6.5 presents the effect of college education on the post-college career path. Although an effect of college education does not allow us to conclude whether and, if so, to which extent the potential mediators actually affect the fertility patterns, the analysis of labor market factors might be insightful for two reasons. First, labor market returns to college education change the family's resources in terms of financial means as well as available time. Second, a heterogeneity in the returns between mothers and non-mothers potentially reveals different career opportunities or preferences. Table 6.5 states the effect of college education on a working full-time indicator and the log hourly wage. There is a clear association between college education and working full-time (as opposed to working part-time or not at all) in the OLS model in column 1: college-educated women are 8pp more likely to work full-time. For the 2SLS estimate the effect increases to 13pp; however, a larger standard error diminishes the statistical significance of the relationship to the 10 percent level. Before coming to wages, column 2 reestimates the effect of college education on the full-time indicator using the subsample of mothers.[17] This corresponds to going from the extensive to the intensive fertility margin. While college education is still positively associated with working full-time, the magnitude is smaller. In fact, the 2SLS effect is only half as big when compared to the entire sample and not statistically different from zero at the conventional levels.

Going on to the hourly wage, we find a strong and statistically significant relationship between college education and earnings. In the OLS estimates (in columns 3 and 4) the wage increase amounts to about 25 percent. As is common in the labor economics literature, the 2SLS coefficients exceed the OLS ones in size (although one would expect to find that OLS overestimates the true effect, see Westphal et al., 2017, for a careful discussion of the heterogeneity in the labor market returns), amounting to nearly 50 percent of the full sample (or equivalently 10 percent per year of college education) and 40 percent among mothers. Thus, mothers not only expand their labor supply less than non-mothers but they also face a smaller college premium in the hourly wage. A reason for the smaller labor market returns might be different – and maybe more family-friendly – occupations college-educated mothers choose compared to college-educated non-

[17] As before, if the tendency to become a mother in spite of a college education correlates with labor supply or wage returns, the subsample analysis may not identify the causal relationship. Moreover, as working women are a subgroup of all women, the wage estimates may suffer a selection bias – although Westphal et al. (2017) provide evidence that such a bias seems humble in the time under review.

mothers. Mothers, for example, tend to choose occupations with a greater flexibility of working shorter hours, which may lead to a wage penalty (Goldin, 2014). Taken together with the small and postponed catch-up effect in fertility at the extensive margin, the bigger labor market returns for non-mothers speak for a college-induced early-career effect that prevents some women from becoming mothers.

Table 6.5: Post-college career outcomes as potentially mediating forces

	(1)	(2)	(3)	(4)
	Working full-time		Log wage	
	all women	only mothers	all women	only mothers
Descriptives				
Sample mean	0.175	0.153	2.83	2.79
OLS regression				
College degree	0.080***	0.062**	0.266***	0.258***
	(0.018)	(0.020)	(0.038)	(0.048)
Second-stage 2SLS regression				
College degree	0.131*	0.075	0.499***	0.407***
	(0.052)	(0.059)	(0.086)	(0.107)
# observations:	4,288	3,485	1,500	1,213

Notes: Own calculations based on NEPS–Adult Starting Cohort data. Control variables include full sets of year of birth and district fixed effects as well as state-specific trends. Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 6.6 considers the effect of college education on revealed family characteristics that may shape a fertility-career trade-off. As marriage often serves as a gatekeeper for planned fertility, the increasing trend in the age at first marriage (as depicted in Figure 6.1) could, if triggered by education, constitute an important mechanism as to why individuals put a stronger focus on family life or career opportunities. Columns 1 and 2 of Table 6.6 show the effect of a regression of an indicator for being married at the age of 40 on college education for all women and mothers, respectively. In the OLS model, college education is associated with reducing the probability of being married by about 6pp while the effect is more than twice as strong when estimated with 2SLS. When looking only at mothers, these relationships vanish. Given a baseline probability of 84 percent, college seems to be an important determinant of marriage preferences, which may have direct repercussions on family life. In other words, the college effect on motherhood already manifests itself in marriage. A reason why college education may prevent marriage – and a potential mediator of education-fertility nexus – may be assortative mating. While men are often said to prefer to "marry down," women who went to college may be more selective when looking for a suitable partner. Columns 3 and 4 of Table 6.6 indicates that women with a college degree seem indeed to be 36pp more likely to have a partner who also went to college – independent of the woman being a mother or not. Given that men with college education earn more than their peers without a college education (see Westphal et al., 2017), we interpret this as evidence that a lower fertility of college-educated couples is unlikely to be driven by the financial need for the mother to work.

Finally, maternal education may change not only the preferences about the offspring's education but also the capability of transmitting a better education to the children. For example, if there is a trade-off between child quality and quantity (Becker and Lewis, 1973), it could mean that the effects on the intensive margin would be even higher in the absence of this trade-off. Moreover, looking at the effect on the educational outcomes of the child is important because it shows (together with the quantitative effects) how maternal college education affects the socioeconomic composition of fertility (Raute, 2016). Column 5 of Table 6.6 gives the effect of the mother's college education on an indicator that shows whether the firstborn visits or has visited an academic track secondary school (compared to a less academically demanding school track). We find strong positive effects here which may emphasize the importance of college education on the socioeconomic composition of fertility and/or that the effects of the intensive margin are likely to be hypothetically higher in the absence of this effect.

To summarize the mediator analysis, we find evidence of a lower college wage premium for mothers. However, for more educated partners (who potentially earn more than their less-educated peers) it seems unlikely that financial reasons alone prevent college-educated women from having children.

6.7 Conclusion

In this paper, we analyze the nexus between education and fertility – two fundamental decisions in life that, when considered on an aggregated level, have greatly changed societies within the past 60 years. These dynamics are unlikely to be confined to the past – particularly with regard to recent policies such as the Higher Education Pact 2020 in which the German states committed to further increase access to higher education. This emphasizes the need to understand the long-term consequences of higher education that go beyond the monetary effects.

Table 6.6: Post-college family characteristics as potentially mediating forces

	(1)	(2)	(3)	(4)	(5)
	Marriage: married age 40		Assortative mating: partner college		Child quality
	all women	only mothers	all women	only mothers	academic track
Descriptives					
Sample mean	0.842	0.916	0.316	0.310	0.526
OLS regression					
College degree	-0.058^{**}	-0.025	0.362^{***}	0.382^{***}	0.250^{***}
	(0.018)	(0.016)	(0.021)	(0.025)	(0.025)
Second-stage 2SLS regression					
College degree	-0.124^{*}	-0.018	0.690^{***}	0.750^{***}	0.639^{***}
	(0.051)	(0.041)	(0.062)	(0.072)	(0.081)
# observations:	4,288	3,491	4,127	3,427	3,316

Notes: Own calculations based on NEPS–Adult Starting Cohort data. Control variables include full sets of year of birth and district fixed effects as well as state-specific trends. Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

The aspect of fertility is especially interesting in this context as higher education affects women – unlike previously studied secondary schooling – within their prime reproductive age. To analyze how education impacts individual fertility decisions in the in-college years and afterwards we make use of arguable exogenous variation in the accessibility of college education in Germany. We find that the overall quantitative fertility effects are driven by the extensive margin: the probability of becoming a mother is reduced by one-quarter. In contrast, women who decide to be a mother despite a college education, have, on average, more children.

We shed light upon the sources of these effects by unraveling the timing of childbearing along the extensive and intensive margin. This analysis indicates that there is a postponement of fertility in the early years of the working career that goes beyond the "incarceration" in college. However, this college-induced postponement in fertility does not seem to push planned children toward ages where biological infertility might become an issue. From a policy perspective, this is a noteworthy finding as a biological effect would restrict a woman's choice set when she maximizes her utility. On the other hand, the decision to forgo marriage and/or childbearing is per se not undesirable when disregarding the negative externalities for the society. The absence of such biological effects together with the overall decline in completed fertility points toward changed preferences for motherhood and/or a career because of college education. Wage and working-time differentials between college-educated mothers and non-mothers suggest an early-career path that shapes fertility and labor market returns to college education.

Although we find evidence that the massive college expansion and effect of college education on the probability of becoming a mother at least partly fueled the demographic transition in recent decades, the positive effect of college education on the number of children for mothers indicates that education does not per se decrease fertility. We consider this to be an important policy implication of this study. Policies that particularly aim at triggering college-educated women into motherhood, for instance, through more flexible working hours or means-tested materiality leave benefits, seem promising for reducing the baby gap between women with and without a college education.

Appendix

Figures

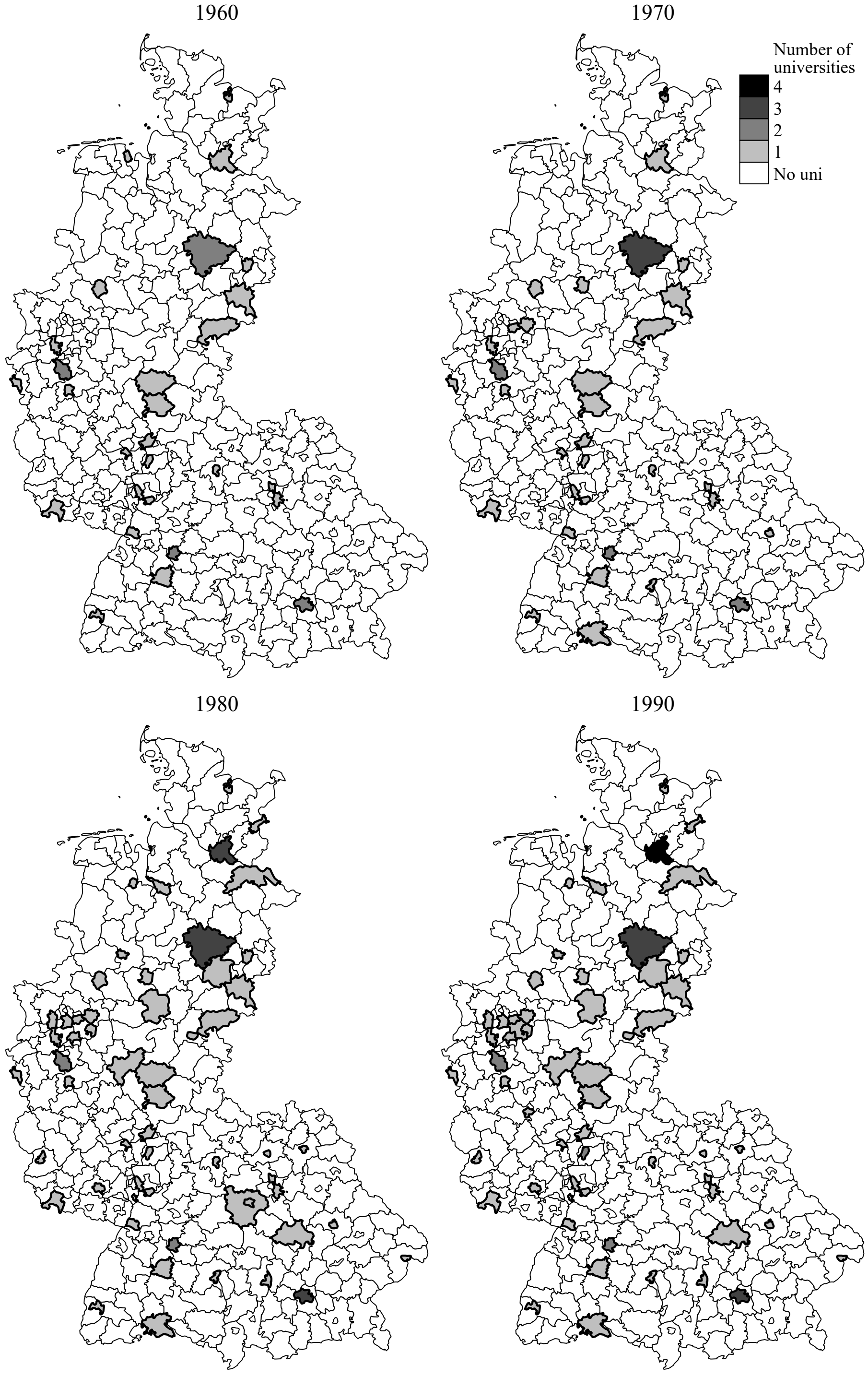


Figure A6.1: Spatial variation of colleges across districts and over time

Notes: Own illustration based on the German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, 1991). The maps show all 326 West German districts (*Kreise*, spatial units of 2009) but Berlin in the years 1958 (first year in the sample), 1970, 1980, and 1990 (last year in the sample). Districts usually cover a bigger city or some administratively connected villages. If a district has at least one college, the district is depicted darker. Very few districts have more than one college. For those districts the number of students is added up in the calculations but multiple colleges are not depicted separately in the maps.

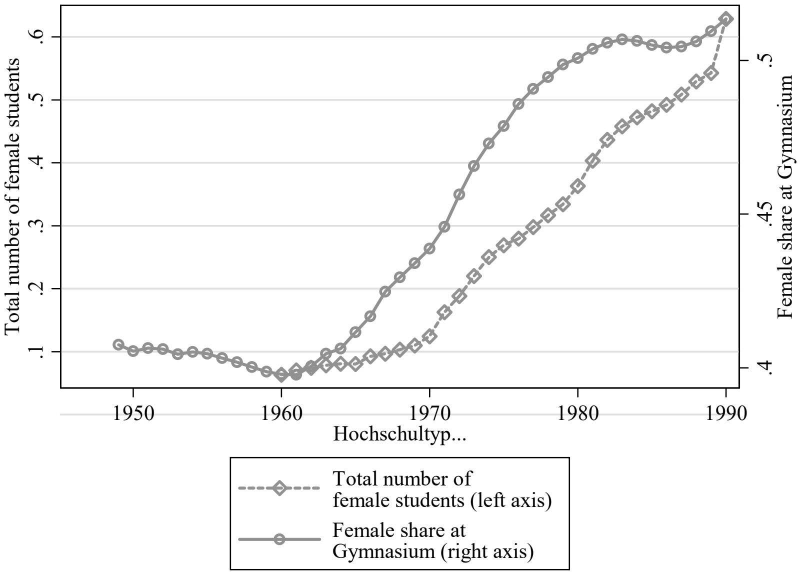


Figure A6.2: Trends in academic secondary school and college education for females

Notes: Own calculations using data from Köhler and Lundgreen (2014).

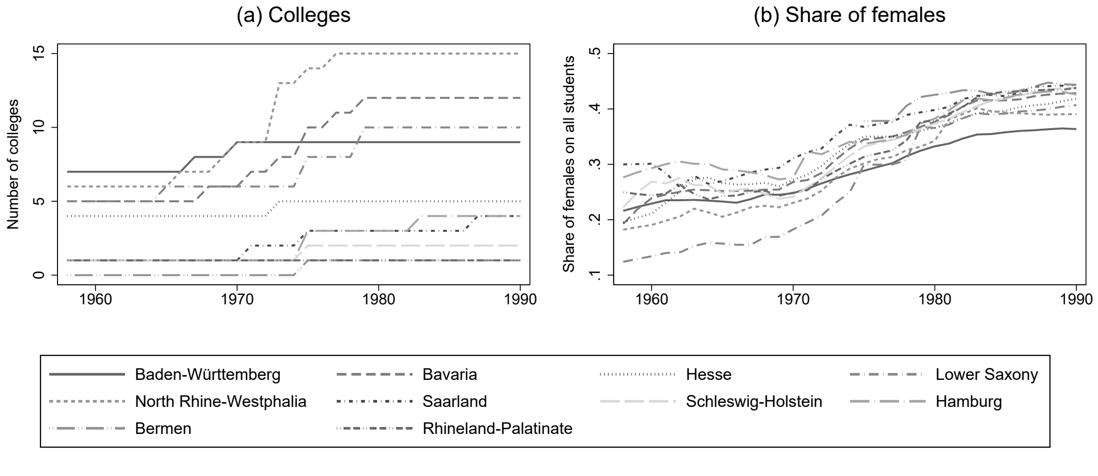


Figure A6.3: Trends in colleges and female students across federal states

Notes: Own calculations using data from the German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, 1991).

Tables

Table A6.1: Control variables and means by university degree

Variable	Definition	Respondents	
		with univ. degree	w/o univ. degree
General information			
Year of birth (FE)	Year of birth of the respondent	1959.62	1959.61
Migrational background	=1 if respondent was born abroad	0.007	0.009
No native speaker	=1 if mother tongue is not German	0.002	0.003
Mother still alive	=1 if mother is still alive in 2009/10	0.676	0.626
Father still alive	=1 if father is still alive in 2009/10	0.472	0.420
Pre-college living conditions			
Married before college	=1 if respondent got married before the year of the college decision or in the same year	0.010	.005
Parent before college	=1 if respondent became a parent before the year of the college decision or in the same year	0.002	0.003
Siblings	Number of siblings	1.555	1.814
First born	=1 if respondent was the first born in the family	0.325	0.283
Age 15: lived by single parent	=1 if respondent was raised by single parent	0.0633	0.057
Age 15: lived in patchwork family	=1 if respondent was raised in a patchwork family	0.013	0.027
Age 15: orphan	=1 if respondent was a orphan at the age of 15	0.009	0.022
Age 15: rural district	=1 if district at the age of 15 was rural	0.181	0.249
Age 15: mother employed	=1 if mother was employed at the respondent's age of 15	0.583	0.610
Age 15: mother never unemployed	=1 if mother was never unemployed until the respondent's age of 15	0.448	0.487
Age 15: father employed	=1 if father was employed at the respondent's age of 15	0.985	0.964
Age 15: father never unemployed	=1 if father was never unemployed until the respondent's age of 15	0.931	0.894
Pre-college health and education			
Final school grade: excellence	=1 if the overall grade of the highest school degree was excellent	0.034	0.015
Final school grade: good	=1 if the overall grade of the highest school degree was good	0.231	0.185

Continued on next page

Table A6.1 – *continued*

Variable	Definition	Respondents	
		with univ. degree	w/o univ. degree
Final school grade: satisfactory	=1 if the overall grade of the highest school degree was satisfactory	0.141	0.185
Final school grade: sufficient or worse	=1 if the overall grade of the highest school degree was sufficient or worse	0.006	0.009
Repeated one grade	=1 if student needed to repeat one grade in elementary or secondary school	0.163	0.166
Repeated two or more grades	=1 if student needed to repeat two or more grades in elementary or secondary school	0.018	0.011
Parental characteristics (M: Mother, F: Father)			
M: year of birth (FE)	Year of birth of the respondent's mother	1930.87	1931.70
M: migrational background	=1 if mother was born abroad	0.063	0.047
M: at least inter. edu	=1 if mother has at least an intermediate secondary school degree	0.298	0.092
M: vocational training	=1 if mother's highest degree is vocational training	0.256	0.245
M: further job qualification	=1 if mother has further job qualification (e.g., *Meister* degree)	0.063	0.024
F: year of birth (FE)	Year of birth of the respondent's father	1927.76	1928.561
F: migrational background	=1 if father was born abroad	0.063	0.047
F: at least inter. edu	=1 if father has at least an intermediate secondary school degree	0.298	0.092
F: vocational training	=1 if father's highest degree is vocational training	0.256	0.245
F: further job qualification	=1 if father has further job qualification (e.g., *Meister* degree)	0.061	0.024
Number of observations		941	3,389

Notes: Information taken from NEPS–Starting Cohort 6. Mean values refer to the health satisfaction sample. In the case of binary variables, the mean gives the percentage of 1s. FE = variable values are included as fixed effects in the analysis. ^{a} Only available for males who did military eligibility test (2,359 observations).

Table A6.2: Descriptive statistics of instruments and background information

	(1)	(2)	(3)	(4)
	Statistics			
	Mean	SD	Min	Max
Instrument: College availability	0.459	0.262	0.046	1.131
Background information on college availability (implicitly included in the instrument)				
Distance to nearest college	27.580	26.184	0	172.269
At least one college in district	0.130	0.337	0	1
Colleges within 100km	5.860	3.401	0	16
College spots per inhabitant within 100km	0.034	0.019	0	0.166

Notes: Own calculations based on NEPS–Adult Starting Cohort data and German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, 1991). Distances are calculated as the Euclidean distance between two respective district centroids.

Table A6.3: Baseline fertility rates and college effects by age

	(1)	(2)	(3)	(4)	(5)	(6)
	Extensive margin			Intensive margin		
Age	Baseline hazard		Effect	Baseline probability		Effect
	no college	college		no college	college	
17	0.024	0.002	−0.059	0.030	0.003	−0.048
18	0.045	0.002	−0.087	0.054	0.003	−0.091
19	0.067	0.006	−0.113	0.080	0.009	−0.123
20	0.084	0.015	−0.131	0.097	0.021	−0.129
21	0.102	0.019	−0.136	0.114	0.026	−0.115
22	0.128	0.030	−0.177	0.135	0.041	−0.152
23	0.147	0.047	−0.222	0.147	0.063	−0.166
24	0.167	0.061	−0.239	0.155	0.081	−0.142
25	0.210	0.070	−0.210	0.179	0.089	−0.095
26	0.233	0.109	−0.168	0.179	0.135	0.005
27	0.243	0.138	−0.178	0.164	0.164	0.042
28	0.241	0.150	−0.157	0.142	0.164	0.075
29	0.216	0.186	−0.101	0.110	0.191	0.119
30	0.213	0.201	−0.114	0.096	0.188	0.113
31	0.198	0.213	−0.082	0.079	0.177	0.126
32	0.161	0.202	0.018	0.057	0.151	0.138
33	0.141	0.168	0.045	0.045	0.110	0.112
34	0.135	0.170	0.025	0.040	0.101	0.097
35	0.105	0.153	0.020	0.029	0.084	0.064
36	0.068	0.116	0.019	0.017	0.057	0.039
37	0.059	0.102	0.026	0.014	0.047	0.046
38	0.044	0.077	0.011	0.011	0.034	0.034
39	0.031	0.060	−0.003	0.007	0.025	0.021
40	0.022	0.040	−0.029	0.005	0.016	0.008

Notes: Own calculations based on NEPS–Adult Starting Cohort data. The effects are those depicted in Figure 6.5 and estimated according to Eq. 6.5. Unlike the figure, the baseline hazard and the baseline probability are stated by college status.

Chapter 7

Conclusion

This thesis covers five empirical essays in the economics of education. In each essay, I analyze the short- and, if possible, long-term returns to a change in education that affects individuals at a different point in their education trajectory. Although data restrictions and the aspiration to combine a meaningful intervention with a plausible identification strategy makes it necessary to consider not only different countries where the education takes place but individuals affected in different centuries, all essays have two things in common. First, at the heart of each paper is the pursuit to disentangle the causal effect of education from a mere correlation caused by a selection of individuals into education. Second, the outcome variables always include, but are not necessarily restricted to, non-monetary factors. While the focus of economic evaluations of the returns to education was traditionally on labor market effects, non-market aspects of education are rather understudied despite of their own relevance and their contribution to the understanding of the monetary returns to education suggested in the literature.

The chapters are ordered along the age of the individuals when they are affected by the change in the education, starting with children in preschools at the age of four. As such a young age is shown to be particularly important for the formation of skills, I analyze how a curriculum-based language training in preschool affects the formation of grammar skills – an important determinant for subsequent learning – up to the children's age of eight. Because the data at hand provide a large array of potential confounders, it seems reasonable in this application to address a potential selection into the language training by comparing only children with the same probability of receiving the treatment. In Chapter 3, I analyze the returns to individual absence in elementary school grades 1 and 4. In order to estimate the long-term education, labor market, and even mortality consequences, I rely on a combination self-digitized historical school records taken from Swedish archives and Census and tax register information. To account for unobservable factors, I employ various fixed effects strategies – among others a siblings fixed effects approach that removes all unobserved characteristics that are shared by siblings, for instance, the genetic endowment and a constant parenting style. In Chapter 4, I move on to students at the end of secondary education and estimate the cognitive skill and wage returns to an additional year of schooling at this margin. Identification stems from an arguably quasi-experimental increase in the legal minimum years of schooling and the build-up of secondary schools that offer more than the minimum years of education in Germany. In Chapters 5 and 6, I follow a similar strategy when investigating the effect of college education on wage, cognitive skills, and health as well as fertility preferences, respectively. As for secondary schools, Germany experienced a massive build-up of new colleges and an increase in the capacities of the existing colleges the 1970s and ’80s. Using regional variation in the exposure to this college expansion, it is arguably possible to overcome individual selection in higher education.

Trying to summarize the results of the thesis in one sentence, this might be

> “Additional education has the potential to increase an individual’s monetary and non-monetary prosperity, but it is not always realized for all individuals.”

The thesis reflects this in at least two ways. First, the returns to some forms of education suggest a considerable heterogeneity along (groups of) individuals. Unfolding, for instance, the average effect of language training along observed characteristics indicates heterogeneous returns w.r.t. the child’s math skills as a proxy for innate abilities. The results indicate that only children with intermediate skills benefit from the additional language training; children with rather low or already quite high skills do not benefit. Allowing for an essential heterogeneity along the unobserved desire for studying, the wage, cognitive skill, and health returns to college education may even be zero for individuals who are rather reluctant to study. The second dimension in why education might not always be beneficial is the margin of education. As demonstrated when analyzing the returns to secondary education, schooling beyond the ninth grade does not seem to contribute to the development of basic skills in Germany. When introducing new policies aiming the increasing the economy’s stock of human capital, political decision-makers are therefore well-advised not to increase any kind of education in the hope that this will enhance human capital and, thereby, lead to higher wages, for instance. The results of the essays in this thesis clearly indicate that both the target group under heterogeneous returns and the margin of the educational intervention matters for its effect on human capital as well as monetary and non-monetary outcomes. A well-defined policy intervention aiming at a clear target group is capable of increasing not only wages but its benefits are potentially also reflected in a number of non-monetary characteristics.

Bibliography

Aakvik, A., Salvanes, K., and Vaage, K. (2010). Measuring heterogeneity in the returns to education using an education reform. *European Economic Review*, 54(4):483–500.

Acemoglu, D. and Johnson, S. (2007). Disease and Development: The Effect of Life Expectancy on Economic Growth. *Journal of Political Economy*, 115(6):925–985.

Agüero, J. and Beleche, T. (2013). Test-Mex: Estimating the effects of school year length on student performance in Mexico. *Journal of Development Economics*, 103(C):353–361.

Akerman, A., Gaarder, I., and Mogstad, M. (2015). The Skill Complementarity of Broadband Internet. *The Quarterly Journal of Economics*, 130(4):1781–1824.

Almond, D. and Currie, J. (2011). Human capital development before age five. Handbook of Labor Economics (Vol. 4B), pages 1315 – 1486. Elsevier.

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

Altonji, J. G., Kahn, L. B., and Speer, J. D. (2016). Cashier or consultant? entry labor market conditions, field of study, and career success. *Journal of Labor Economics*, 34(S1):S361–S401.

Altonji, J. G. and Pierret, C. R. (2001). Employer learning and statistical discrimination. *The Quarterly Journal of Economics*, 116(1):313–350.

American Psychological Association (1995). Intelligence: Knowns and Unknowns, Report of a task force convened by the American Psychological Association.

Andersen, H. H., Mühlbacher, A., Nübling, M., Schupp, J., and Wagner, G. G. (2007). Computation of Standard Values for Physical and Mental Health Scale Scores Using the SOEP Version of SF-12v2. *Schmollers Jahrbuch: Journal of Applied Social Science Studies / Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 127:171–182.

Anderson, J. (2007). *Cognitive Psychology and its Implications*. Worth Publishers, New York, 7 edition.

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

Angrist, J. D. and Pischke, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, 24(2):3–30.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485.

Athey, S. and Imbens, G. W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Aucejo, E. M. and Romano, T. F. (2016). Assessing the effect of school days and absences on test score performance. *Economics of Education Review*, 55:70 – 87.

Baker, M., Gruber, J., and Milligan, K. (2008). Universal Child Care, Maternal Labor Supply, and Family Well-Being. *Journal of Political Economy*, 116(4):709–745.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

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

Barrow, L. and Malamud, O. (2015). Is college a worthwhile investment? *Annual Review of Economics*, 7:519–555.

Bartz, O. (2007). Expansion und Umbau – Hochschulreformen in der Bundesrepublik Deutschland zwischen 1964 und 1977. *Die Hochschule*, 2007(2):154–170.

Basu, A. (2011). Estimating decision-relevant comparative effects using instrumental variables. *Statistics in Biosciences*, 3:6–27.

Basu, A. (2014). Person-Centered Treatment (PeT) effects using instrumental variables: An application to evaluating prostate cancer treatments. *Journal of Applied Econometrics*, 29:671–691.

Basu, A., Heckman, J. J., Navarro-Lozano, S., and Urzua, S. (2007). Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. *Health Economics*, 16(11):1133–1157.

Battistin, E. and Meroni, E. (2016). Should we increase instruction time in low achieving schools? Evidence from Southern Italy. *Economics of Education Review*, 55:39 – 56.

Becker, G. (1993). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Chicago: The University of Chicago Press, 3rd edition (1993).

Becker, G. S. and Lewis, H. G. (1973). On the Interaction between the Quantity and Quality of Children. *Journal of Political Economy*, 81(2):S279–S288.

Bellei, C. (2009). Does lengthening the school day increase students' academic achievement? Results from a natural experiment in Chile. *Economics of Education Review*, 28(5):629–640.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014a). Inference on Treatment Effects after Selection among High-Dimensional Controls. *Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014b). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2):29–50.

Berlinski, S., Galiani, S., and Gertler, P. (2009). The effect of pre-primary education on primary school performance. *Journal of Public Economics*, 93(1-2):219–234.

Berlinski, S., Galiani, S., and Manacorda, M. (2008). Giving children a better start: Preschool attendance and school-age profiles. *Journal of Public Economics*, 92(5-6):1416–1440.

Bhalotra, S., Karlsson, M., and Nilsson, T. (forthcoming). Infant Health and Longevity: Evidence from a Historical Intervention in Sweden. *Journal of the European Economic Association*.

Bhalotra, S. R., Karlsson, M., Nilsson, T., and Schwarz, N. (2016). Infant Health, Cognitive Performance and Earnings: Evidence from Inception of the Welfare State in Sweden. IZA Discussion Papers 10339, Institute for the Study of Labor (IZA).

Bhuller, M., Mogstad, M., and Salvanes, K. G. (2011). Life-cycle bias and the returns to schooling in current and lifetime earnings.

Bishop, D. (1989). *TROG – Test for Reception of Grammar*. Medical Research Council: Chapel Press.

Björklund, A. and Moffitt, R. (1987). The Estimation of Wage Gains and Welfare Gains in Self-Selection. *The Review of Economics and Statistics*, 69(1):42–49.

Black, S. E., Devereux, P. J., Lóken, K. V., and Salvanes, K. G. (2014). Care or Cash? The Effect of Child Care Subsidies on Student Performance. *The Review of Economics and Statistics*, 96(5):824–837.

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2008). Staying in the Classroom and out of the Maternity Ward? The Effect of Compulsory Schooling Laws on Teenage Births. *The Economic Journal*, 118(530):1025–1054.

Blossfeld, H.-P., Roßbach, H.-G., and von Maurice, J. (2011). Education as a Lifelong Process – The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft*, 14:Special Issue.

Bond, T. N. and Lang, K. (2013). The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of Results. *The Review of Economics and Statistics*, 95(5):1468–1479.

Bowyer-Crane, C., Snowling, M. J., Duff, F. J., Fieldsend, E., Carroll, J. M., Miles, J., Götz, K., and Hulme, C. (2008). Improving early language and literacy skills: differential effects of an oral language versus a phonology with reading intervention. *Journal of Child Psychology and Psychiatry*, 49(4):422–432.

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

Bruckner, T., van den Berg, G., Smith, K., and Catalano, R. (2014). Ambient temperature during gestation and cold-related adult mortality in a Swedish cohort, 1915–2002. *Social Science & Medicine*, 119(0):191–197.

Brunello, G., Weber, G., and Weiss, C. T. (2017). Books are forever: Early life conditions, education and lifetime earnings in europe. *The Economic Journal*, 127(600):271–296.

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72.

Card, D. (1995). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. In L. Christofides, K. G. and Swidinsky, R., editors, *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, pages 201–222. University of Toronto Press.

Card, D. (1999). The Causal Effect of Education on Earnings. In Ashenfelter, O., Layard, R., and Card, D., editors, *Handbook of Labor Economics*, volume 3C. North-Holland Publishing Company.

Card, D. (2001). Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica*, 69(5):1127–1160.

Carlsson, M., Dahl, G., Öckert, B., and Rooth, D.-O. (2015). The Effect of Schooling on Cognitive Skills. *The Review of Economics and Statistics*, 97(3):533–547.

Carneiro, P., Hansen, K. T., and Heckman, J. J. (2001). Removing the Veil of Ignorance in Assessing the Distributional Impacts of Social Policies. *Swedish Economic Policy Review*, 8(2):273–301.

Carneiro, P., Hansen, K. T., and Heckman, J. J. (2003). 2001 Lawrence R. Klein Lecture: Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice. *International Economic Review*, 44(2):361–422.

Carneiro, P., Heckman, J. J., and Vytlacil, E. J. (2010). Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin.

Econometrica, 78(1):377–394.

Carneiro, P., Heckman, J. J., and Vytlacil, E. J. (2011). Estimating Marginal Returns to Education. *American Economic Review*, 101(6):2754–81.

Cascio, E. (2015). The promises and pitfalls of universal early education. *IZA World of Labor*, (116):1–10.

Cascio, E. and Schanzenbach, D. (2014). Proposal 1: Expanding preschool access for disadvantaged children. In Kearney, M. S. and Harris, B. H., editors, *Policies to Address Poverty in America*, pages 10–28. Washington, DC: Hamilton Project.

Cascio, E. and Schanzenbach, D. W. (2013). The Impacts of Expanding Access to High-Quality Preschool Education. *Brookings Papers on Economic Activity*, 44(2 (Fall)):127–192.

Cascio, E. U. (2009). Do Investments in Universal Early Education Pay Off? Long-term Effects of Introducing Kindergartens into Public Schools. NBER Working Papers 14951, National Bureau of Economic Research, Inc.

Cattan, S. (2016). Can universal preschool increase the labor supply of mothers? *IZA World of Labor*, pages 312–321.

Cawley, J., Heckman, J. J., and Vytlacil, E. J. (2001). Three observations on wages and measured cognitive ability. *Labour Economics*, 8(4):419–442.

Cervellati, M. and Sunde, U. (2005). Human Capital Formation, Life Expectancy, and the Process of Development. *American Economic Review*, 95(5):1653–1672.

Cervellati, M. and Sunde, U. (2013). Life Expectancy, Schooling, and Lifetime Labor Supply: Theory and Evidence Revisited. *Econometrica*, 81(5):2055–2086.

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

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

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 104(9):2633–2679.

Chicago Tribune (2012). "An emptydesk epidemic" by David Jackson, Gary Marx and Alex Richards. *Chicago Tribune,* November 11, 2012, (http://www.chicagotribune.com/ct-met-truancy-mainbar-20121111-story.html, last assessed June 29, 2017).

Child Trends (2015). "Student Absenteeism". *Child Trends Databank*, (https://www.childtrends.org/?indicators=student-absenteeism, last assessed August 7 2017).

Clark, D. and Martorell, P. (2014). The signaling value of a high school diploma. *Journal of Political Economy*, 122(2):282–318.

Conti, G. (2013). The developmental origins of health inequality. In *Health and Inequality*, pages 285–309. Emerald Group Publishing Limited.

Conti, G., Heckman, J. J., and Pinto, R. (2016). The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviour. *The Economic Journal*, 126(596):F28–F65.

Contoyannis, P. and Li, J. (2011). The evolution of health outcomes from childhood to adolescence. *Journal of Health Economics*, 30(1):11–32.

Cornelissen, T., Dustmann, C., Raute, A., and Schönberg, U. (2017). Who benefits from universal childcare? Estimating marginal returns to early childcare attendance. *Journal of Political Economy*, forthcoming.

Costa, D. L. (2015). Health and the economy in the united states from 1750 to the present. *Journal of Economic Literature*, 53(3):503–70.

Couperus, J. W. and Nelson, C. A. (2008). Early Brain Development and Plasticity. Blackwell Handbook of Early Childhood Development, pages 85–105. Blackwell Publishing Ltd.

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

Cunha, F. and Heckman, J. J. (2009). The Economics and Psychology of Inequality and Human Development. *Journal of the European Economic Association*, 7(2-3):320–364.

Cunha, F., Heckman, J. J., Lochner, L. J., and Masterov, D. V. (2006). Interpreting the evidence on life cycle skill formation. In Hanushek, E. A. and Welch, F., editors, *Handbook of the Economics of Education*, volume 1. North-Holland.

Currie, J. (2001). Early Childhood Education Programs. *Journal of Economic Perspectives*, 15(2):213–238.

Currie, J. (2009). Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development. *Journal of Economic Literature*, 47(1):87–122.

Currie, J. and Moretti, E. (2003). Mother's Education And The Intergenerational Transmission Of Human Capital: Evidence From College Openings. *The Quarterly Journal of Economics*, 118(4):1495–1532.

Currie, J. and Neidell, M. (2007). Getting inside the "Black Box" of Head Start quality: What matters and what doesn't. *Economics of Education Review*, 26(1):83–99.

Currie, J. and Stabile, M. (2003). Socioeconomic status and child health: Why is the relationship stronger for older children? *The American Economic Review*, 93(5):1813–1823.

Cutler, D. M. and Lleras-Muney, A. (2010). Understanding differences in health behaviors by education. *Journal of Health Economics*, 29(1):1 – 28.

Cygan-Rehm, K. and Maeder, M. (2013). The Effect of Education on Fertility: Evidence from a Compulsory Schooling Reform. *Labour Economics*, 25:35 – 48. European Association of Labour Economists 24th Annual Conference, Bonn, Germany, 20-22 September 2012.

Dahrendorf, R. (1965). *Bildung ist Bürgerrecht: Plädoyer für eine aktive Bildungspolitik*. Nannen Verlag.

Datta Gupta, N. and Simonsen, M. (2010). Non-cognitive child outcomes and universal high quality child care. *Journal of Public Economics*, 94(1-2):30–43.

de Walque, D. (2007). Does education affect smoking behaviors?: Evidence using the Vietnam draft as an instrument for college education. *Journal of Health Economics*, 26(5):877–895.

Deaton, A. (2010). Instruments, Randomization, and Learning about Development. *Journal of Economic Literature*, 48(2):424–55.

DeCicca, P. and Smith, J. (2013). The long-run impacts of early childhood education: Evidence from a failed policy experiment. *Economics of Education Review*, 36(C):41–59.

Der Spiegel (1967). Doktortitel Nach Sechs Semesters? *Der Spiegel*, Oktober 9, 1967(42):54–62.

Die Zeit (1967). Warenhaus der Ausbildung. *Die Zeit*, August 4, 1967(31):28.

Diebolt, C. (2000). Die erfassung der bildungsinvestitionen im 19. und 20. jahrhundert. *Zeitschrift für Erziehungswissenschaft*, 3(4):517–538.

Drange, N. and Havnes, T. (2015). Child Care Before Age Two and the Development of Language and Numeracy: Evidence from a Lottery. IZA Discussion Papers 8904, Institute for the Study of Labor (IZA).

Duflo, E., Dupas, P., and Kremer, M. (2015). Education, HIV, and Early Fertility: Experimental Evidence from Kenya. *American Economic Review*, 105(9):2757–2797.

Dumas, C. and Lefranc, A. (2012). *Early schooling and later outcomes: Evidence from pre-school extension in France*. Inequality from Childhood to Adulthood: A Cross-National Perspective on the Transmission of Advantage. Russell Sage Foundation.

Durevall, D., Lindskog, A., and George, G. (2015). Education and HIV incidence among young women: causation or selection? Working Papers in Economics 638, University of Gothenburg, Department of Economics.

Dustmann, C., Puhani, P. A., and Schönberg, U. (forthcoming). The Long-term Effects of Early Track Choice. *The Economic Journal*.

Erikson, R. and Jonsson, J. O. (1993). *Ursprung och Utbildning – Social Snedrekrytering till Högre Studier*. Utbildningsdepartementet, Stockholm, Sweden.

Federation of Swedish Genealogical Societies (2014). *Swedish Death Index 1901–2013*. Federation of Swedish Genealogical Societies, Farsta, Sweden.

Felfe, C. and Lalive, R. (2012). Early Child Care and Child Development: For Whom it Works and Why. IZA Discussion Papers 7100, Institute for the Study of Labor (IZA).

Felfe, C. and Lalive, R. (2014). Does Early Child Care Help or Hurt Children's Development? IZA Discussion Papers 8484, Institute for the Study of Labor (IZA).

Fischer, M., Karlsson, M., and Nilsson, T. (2013). Effects of Compulsory Schooling on Mortality: Evidence from Sweden. *International Journal of Environmental Research and Public Health*, 10(8):3596–3618.

Fischer, M., Karlsson, M., Nilsson, T., and Schwarz, N. (2016). The Sooner the Better? Compulsory Schooling Reforms in Sweden. IZA Discussion Papers 10430, Institute for the Study of Labor (IZA).

Fisher, G., Stachowski, A., Infurna, F., Faul, J., Grosch, J., and Tetrick, L. (2014). Mental work demands, retirement, and longitudinal trajectories of cognitive functioning. *Journal of Occupational Health Psychology*, 19(2):231–242.

Fitzpatrick, M., Grissmer, D., and Hastedt, S. (2011). What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment. *Economics of Education Review*, 30(2):269–279.

Fitzpatrick, M. D. (2008). Starting School at Four: The Effect of Universal Pre-Kindergarten on Children's Academic Achievement. *The B.E. Journal of Economic Analysis & Policy*, 8(1):1–40.

Fletcher, J. and Wolfe, B. (2014). Increasing our understanding of the health-income gradient in children. *Health economics*, 23(4):473–486.

Fort, M., Schneeweis, N., and Winter-Ebmer, R. (2016). Is Education Always Reducing Fertility? Evidence from Compulsory Schooling Reforms. *The Economic Journal*, 126(595):1823–1855.

Fox-Boyer, A. V. (2006). *TROG-D Test zur Überprüfung des Grammatikverständnisses*. Schulz-Kirchner Verlag, Idstein.

Fredriksson, V. A. (1971). *Svenska Folkskolans Historia*. Albert Bonniers Förlag, Stockholm, Sweden.

Freeman, R. (1979). The Effect of Demographic Factors on Age-Earnings Profiles. *Journal of Human Resources*, 14(3):289–318.

Gauthier, A. (2007). The Impact of Family Policies on Fertility in Industrialized Countries: A Review of the Literature. *Population Research and Policy Review*, 26(3):323–346.

Genda, Y., Kondo, A., and Ohta, S. (2010). Long-term effects of a recession at labor market entry in japan and the united states. *Journal of Human Resources*, 45(1):157–196.

German Federal Statistical Office (2016a). Volkswirtschaftliche Gesamtrechnungen: Bruttoinlandsprodukt, Bruttonationaleinkommen, Volkseinkommen – Lange Reihen ab 1925. Technical report, German Federal Statistical Office (Statistisches Bundesamt), Wiesbaden.

German Federal Statistical Office (2016b). Endgültige durchschnittliche Kinderzahl der Frauenkohorten. Technical report, German Federal Statistical Office (Statistisches Bundesamt), Wiesbaden.

German Federal Statistical Office (2017). Pressemitteilung vom 14. September 2017 (326/17): Automobilindustrie trägt 4,5% zur Bruttowertschöpfung in Deutschland bei. Technical report, German Federal Statistical Office (Statistisches Bundesamt), Wiesbaden.

German Federal Statistical Office (various issues, 1952–1992). Statistisches Jahrbuch für die Bundesrepublik Deutschland – Allgemeinbildende Schulen. Technical report, German Federal Statistical Office (Statistisches Bundesamt), Wiesbaden.

German Federal Statistical Office (various issues, 1959–1991). Statistisches Jahrbuch für die Bundesrepublik Deutschland – Einrichtungen für Höhere Bildung. Technical report, German Federal Statistical Office (Statistisches Bundesamt), Wiesbaden.

Geruso, M. and Royer, H. (2014). The Impact of Education on Family Formation: Quasi-Experimental Evidence from the UK.

Glymour, M., Kawachi, I., Jencks, C., and Berkman, L. (2008). Does childhood schooling affect old age memory or mental status? Using state schooling laws as natural experiments. *Journal of Epidemiology and Community Health*, 62(6):532–537.

Goldin, C. (2006). The Quiet Revolution That Transformed Women's Employment, Education, and Family. *American Economic Review*, 96(2):1–21.

Goldin, C. (2014). A Grand Gender Convergence: Its Last Chapter. *American Economic Review*, 104(4):1091–1119.

Goldin, C. D. and Katz, L. F. (2009). *The Race Between Education and Technology*. Harvard University Press.

Goodman, J. (2014). Flaking Out: Student Absences and Snow Days as Disruptions of Instructional Time. NBER Working Papers 20221, National Bureau of Economic Research, Inc.

Gormley, W. and Gayer, T. (2005). Promoting School Readiness in Oklahoma: An Evaluation of Tulsa's Pre-K Program. *Journal of Human Resources*, 40(3).

Gormley, W. T. (2011). From Science to Policy in Early Childhood Education. *Science*, 333(6045):978–981.

Grenet, J. (2013). Is Extending Compulsory Schooling Alone Enough to Raise Earnings? Evidence from French and British Compulsory Schooling Laws. *The Scandinavian Journal of Economics*, 115(1):176–210.

Grimard, F. and Parent, D. (2007). Education and smoking: Were Vietnam war draft avoiders also more likely to avoid smoking? *Journal of Health Economics*, 26(5):896–926.

Grönqvist, H. and Hall, C. (2013). Education Policy and Early Fertility: Lessons from an Expansion of Upper Secondary Schooling. *Economics of Education Review*, 37(C):13–33.

Grossman, M. (1972). On the Concept of Health Capital and the Demand for Health. *Journal of Political Economy*, 80(2):223–55.

Haan, P. and Wrohlich, K. (2011). Can Child Care Policy Encourage Employment and Fertility? Evidence from a Structural Model. *Labour Economics*, 18(4):498–512.

Hadjar, A. and Becker, R. (2006). *Die Bildungsexpansion: Erwartete und unerwartete Folgen*. VS Verlag, Wiesbaden.

Hansen, B. (2011). School Year Length and Student Performance: Quasi-Experimental Evidence. mimeo.

Hansen, C. (2014). `lassoShooting`: Stata module to obtain Lasso and Post-Lasso estimates.

Hansen, K. T., Heckman, J. J., and Mullen, K. K. J. (2004). The effect of schooling and ability on achievement test scores. *Journal of Econometrics*, 121(1-2):39–98.

Hanushek, E. A. (2003). The Failure of Input-Based Schooling Policies. *Economic Journal*, 113(485):64–98.

Havnes, T. and Mogstad, M. (2011). No Child Left Behind: Subsidized Child Care and Children's Long-Run Outcomes. *American Economic Journal: Economic Policy*, 3(2):97–129.

Heckman, J. (2010). Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy. *Journal of Economic Literature*, 48(2):356–398.

Heckman, J., Moon, S. H., Pinto, R., Savelyev, P., and Yavitz, A. (2010a). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics*, 1(1):1–46.

Heckman, J., Pinto, R., and Savelyev, P. (2013). Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review*, 103(6):2052–2086.

Heckman, J. and Vytlacil, E. (2001). Identifying The Role Of Cognitive Ability In Explaining The Level Of And Change In The Return To Schooling. *The Review of Economics and Statistics*, 83(1):1–12.

Heckman, J. J. (1990). Varieties of selection bias. *The American Economic Review*, 80(2):pp. 313–318.

Heckman, J. J. (2007). The economics, technology, and neuroscience of human capability formation. *Proceedings of the National Academy of Sciences*, 104(33):13250–13255.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies*, 64(4):605–654.

Heckman, J. J., Lochner, L. J., and Todd, P. E. (1999). Earnings Equations and Rates of Return: The Mincer Equation and Beyond. In Hanushek, E. and Welch, F., editors, *Handbook of the Economics of Education*, volume 1. Elsevier.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., and Yavitz, A. (2010b). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, 94(1-2):114–128.

Heckman, J. J. and Urzúa, S. (2010). Comparing IV with structural models: What simple IV can and cannot identify. *Journal of Econometrics*, 156(1):27–37.

Heckman, J. J., Urzua, S., and Vytlacil, E. J. (2006). Understanding Instrumental Variables in Models with Essential Heterogeneity. *The Review of Economics and*

Statistics, 88(3):389–432.

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

Heckman, J. J. and Vytlacil, E. J. (2007). Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 6 of *Handbook of Econometrics*, chapter 71. Elsevier.

Hener, T., Rainer, H., and Siedler, T. (2016). Political socialization in flux? Linking family non-intactness during childhood to adult civic engagement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(3):633–656.

Hoxby, C. M. (2000). The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *The Quarterly Journal of Economics*, 115(4):1239–1285.

Huebener, M. and Marcus, J. (2015). Moving up a Gear: The Impact of Compressing Instructional Time into Fewer Years of Schooling. Technical report.

Ichino, A., Mealli, F., and Nannicini, T. (2008). From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics*, 23(3):305–327.

Imbens, G. and Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Imbens, G. W. (2010). Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, 48(2):399–423.

Imbens, G. W. (2015). Matching Methods in Practice: Three Examples. *Journal of Human Resources*, 50(2):373–419.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1):5–86.

International Labour Organization (2012). International Standard Classification of Occupations: Structure, group definitions and correspondence tables. Technical report, International Labour Organization, Geneva.

Isphording, I. E. and Otten, S. (2013). The Costs of Babylon – Linguistic Distance in Applied Economics. *Review of International Economics*, 21(2):354–369.

Isphording, I. E. and Otten, S. (2014). Linguistic barriers in the destination language acquisition of immigrants. *Journal of Economic Behavior & Organization*, 105(C):30–50.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning – With Applications in R*. Springer, New York.

Jürges, H., Reinhold, S., and Salm, M. (2011). Does schooling affect health behavior? Evidence from the educational expansion in Western Germany. *Economics of Education Review*, 30(5):862–872.

Jürges, H., Schneider, K., and Büchel, F. (2005). The Effect Of Central Exit Examinations On Student Achievement: Quasi-Experimental Evidence From TIMSS Germany. *Journal of the European Economic Association*, 3(5):1134–1155.

Kamhöfer, D. and Schmitz, H. (2016). Reanalyzing Zero Returns to Education in Germany. *Journal of Applied Econometrics*, 31(5):912–919.

Kamhöfer, D., Schmitz, H., and Westphal, M. (2015). Heterogeneity in Marginal Non-monetary Returns to Higher Education. Health, Econometrics and Data

Group (HEDG) Working Papers 15/24, HEDG, c/o Department of Economics, University of York.

Kamhöfer, D., Schmitz, H., and Westphal, M. (2017). Heterogeneity in Marginal Non-monetary Returns to Higher Education. *Journal of the European Economic Association*, forthcoming.

KMK (2015). The Education System in the Federal Republic of Germany 2013/2014 – A description of the responsibilities, structures and developments in education policy for the exchange of information in Europe. Report, Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (KMK).

Köhler, H. and Lundgreen, P. (2014). Allgemeinbildende Schulen in der Bundesrepublik Deutschland 1949 - 2010. Technical Report Deutschland ZA8570 Datenfile Version 1.0.0.

Kühnle, D. and Oberfichtner, M. (2017). Does Early Child Care Attendance Influence Children's Cognitive and Non-Cognitive Skill Development? IZA Discussion Papers 10661, Institute for the Study of Labor (IZA).

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review*, 76(4):604–620.

Lang, F., Weiss, D., Stocker, A., and von Rosenbladt, B. (2007a). The returns to cognitive abilities and personality traits in Germany. *Schmollers Jahrbuch: Journal of Applied Social Science Studies / Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 127(1):183–192.

Lang, F., Weiss, D., Stocker, A., and von Rosenbladt, B. (2007b). The returns to cognitive abilities and personality traits in Germany. *Schmollers Jahrbuch: Journal of Applied Social Science Studies / Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 127(1):183–192.

Lechner, M., Miquel, R., and Wunsch, C. (2011). Long-Run Effects Of Public Sector Sponsored Training In West Germany. *Journal of the European Economic Association*, 9(4):742–784.

Lengerer, A., Schroedter, J., Boehle, M., Hubert, T., and Wolf, C. (2008). Harmonisierung der Mikrozensen 1962 bis 2005. Technical report, GESIS-Methodenbericht 12/2008, GESIS–Leibniz Institute for the Social Sciences, German Microdata Lab, Mannheim.

Leuven, E., Lindahl, M., Oosterbeek, H., and Webbink, D. (2010). Expanding schooling opportunities for 4-year-olds. *Economics of Education Review*, 29(3):319–328.

Leuven, E. and Sianesi, B. (2003). `psmatch2`: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.

LIfBi (2011). Starting Cohort 6 Main Study 2010/11 (B67) Adults Information on the Competence Test. Technical report, Leibniz Institute for Educational Trajectories (LIfBi) – National Educational Panel Study.

LIfBi (2015). Startkohorte 6: Erwachsene (SC6) – Studienübersicht Wellen 1 bis 5. Technical report, Leibniz Institute for Educational Trajectories (LIfBi) – National Educational Panel Study.

LOGO (2014). *Kon-Lab – Systematische Sprachförderung*. LOGO Lern-Spiel-Verlag GmbH, Essen.

Lovell, M. C. (2008). A simple proof of the fwl theorem. *The Journal of Economic Education*, 39(1):88–91.

Lundgreen, P. and Schwibbe, G. (2008). Berufliche Schulen und Hochschulen in der Bundesrepublik Deutschland 1949-2001 Teil II: Hochschulen. Technical

Report Deutschland ZA8202 Datenfile Version 1.0.0.

Marcotte, D. (2007). Schooling and test scores: A mother-natural experiment. *Economics of Education Review*, 26(5):629–640.

Marcotte, D. and Hansen, B. (2010). Time for School? *Education Next*, 10(1):53–59.

Marcotte, D. and Hemelt, S. (2008). Unscheduled School Closings and Student Performance. *Education Finance and Policy*, 3(3):316–338.

Matsuura, K. and Willmott, C. (2012). Terrestrial Air Temperature: 1900–2010 Gridded Monthly Time Series (Version 3.01). Technical report, Department of Geography, University of Delaware. http://climate.geog.udel.edu/~climate/html_pages/Global2011/README.GlobalTsT2011.html.

Max Planck Institute for Demographic Research, G. and Vienna Institute of Demography, A. (2014). Human Fertility Database. Technical report.

Mazumder, B. (2008). Does education improve health? A reexamination of the evidence from compulsory schooling laws. *Economic Perspectives*, (Q II):2–16.

Mazzonna, F. (2012). The effect of education on old age health and cognitive abilities - does the instrument matter? Discussion paper.

McCrary, J. and Royer, H. (2011). The Effect of Female Education on Fertility and Infant Health: Evidence from School Entry Policies Using Exact Date of Birth. *American Economic Review*, 101(1):158–95.

Meghir, C. and Palme, M. (2005). Educational Reform, Ability, and Family Background. *American Economic Review*, 95(1):414–424.

Meng, X. and D'Arcy, C. (2012). Education and Dementia in the Context of the Cognitive Reserve Hypothesis: A Systematic Review with Meta-Analyses and Qualitative Analyses. *PLoS ONE*, 7(6):e38268.

Mervis, J. (2011). Past Successes Shape Effort to Expand Early Intervention. *Science*, 333(6045):952–956.

Monstad, K., Propper, C., and Salvanes, K. G. (2008). Education and Fertility: Evidence from a Natural Experiment. *Scandinavian Journal of Economics*, 110(4):827–852.

Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87–106.

Nasar, S. (2012). *Grand Pursuit: The Story of Economic Genius*. New York: Simon & Schuster Paperbacks.

Neidell, M. and Waldfogel, J. (2010). Cognitive and Noncognitive Peer Effects in Early Education. *The Review of Economics and Statistics*, 92(3):562–576.

NEPS (2011). Starting Cohort 2, Main Study 2010/11 (A12) Children in Kindergarten – Information on the Competence Test. Data documentaition, National Educational Panel Study (NEPS) Research Data Center, Bamberg.

Neugebauer, U. and Becker-Mrotzek, M. (2013). Die Qualität von Sprachstandsverfahren im Elementarbereich: Eine Analyse und Bewertung. Report, Mercator Institute for Literacy and Language Education, University of Cologne.

NRW (1971a). Stellungnahme der Staatskanzlei zum Entwurf der Kabinettvorlage des Ministers für Wissenschaft und Forschung. Technical report, Office of the Prime Minister of the state of North Rhine-Westphalia (NRW), April 19, 1971, Düsseldorf, http://protokolle.archive.nrw.de/texte/kv1078_1_3.htm.

NRW (1971b). Schreiben des Ministers für Wissenschaft und Forschung an die Staatskanzlei. Technical report, Ministry of Science and Research of the state of North Rhine-Westphalia (NRW), May 24, 1971, Düsseldorf, http:

//protokolle.archive.nrw.de/texte/kv1083_1_10.htm.

NRW (1971c). Sachstandsbericht des Ministers für Wissenschaft und Forschung. Technical report, Ministry of Science and Research of the state of North Rhine-Westphalia (NRW), March 2, 1971, Düsseldorf, http://protokolle.archive.nrw.de/texte/kv1075_1_9b.htm.

Nybom, M. (2017). The Distribution of Lifetime Earnings Returns to College. *Journal of Labor Economics*, forthcoming.

Oddens, B., Vemer, H., Visser, A., and Ketting, E. (1993). Contraception in Germany: A Review. *Advances in Contraception*, 9:105–116.

OECD (2004). OECD Country Note: Early Childhood Education and Care Policy in The Federal Republic of Germany. Report, Organisation for Economic Co-operation and Development (OECD).

OECD (2006). Starting Strong II: Early Childhood Education and Care. Report, Organisation for Economic Co-operation and Development (OECD).

OECD (2010). PISA 2009 Results: Overcoming Social Background – Equity in Learning Opportunities and Outcomes (Volume II). Report, Organisation for Economic Co-operation and Development (OECD).

OECD (2013). Education at a Glance 2013: OECD Indicators. Report, Organisation for Economic Co-operation and Development (OECD).

OECD (2015a). Education Policy Outlook 2015: Germany. Report, Organisation for Economic Co-operation and Development (OECD).

OECD (2015b). Education Policy Outlook 2015: Making Reforms Happen. Report, Organisation for Economic Co-operation and Development (OECD).

Oosterbeek, H. and Webbink, D. (2007). Wage effects of an extra year of basic vocational education. *Economics of Education Review*, 26(4):408–419.

Oreopoulos, P. (2006). Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter. *American Economic Review*, 96(1):152–175.

Oreopoulos, P. and Petronijevic, U. (2013). Making College Worth It: A Review of the Returns to Higher Education. *The Future of Children*, 23(1):41–65.

Oreopoulos, P. and Salvanes, K. (2011). Priceless: The Nonpecuniary Benefits of Schooling. *Journal of Economic Perspectives*, 25(1):159–84.

Oreopoulos, P., Von Wachter, T., and Heisz, A. (2012). The short-and long-term career effects of graduating in a recession. *American Economic Journal: Applied Economics*, 4(1):1–29.

Oster, E. (2017a). psacalc: Stata module to calculate bounds.

Oster, E. (2017b). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics*.

Pei, Z., Pischke, J.-S., and Schwandt, H. (2017). Poorly Measured Confounders are More Useful on the Left Than on the Right. NBER Working Papers 23232, National Bureau of Economic Research, Inc.

Picht, G. (1964). *Die deutsche Bildungskatastrophe: Analyse und Dokumentation*. Walter Verlag.

Pischke, J.-S. (2007). The Impact of Length of the School Year on Student Performance and Earnings: Evidence From the German Short School Years. *The Economic Journal*, 117(523):1216–1242.

Pischke, J.-S. and Krueger, A. (1995). A Comparative Analysis of East and West German Labor Markets: Before and After Unification. In Freeman, R. and Katz, L., editors, *Differences and Changes in Wage Structures*, pages 405–445. Chicago: University of Chicago Press.

Pischke, J.-S. and von Wachter, T. (2005). Zero Returns to Compulsory Schooling in Germany: Evidence and Interpretation. Working Paper 11414, National Bureau of Economic Research.

Pischke, J.-S. and von Wachter, T. (2008). Zero Returns to Compulsory Schooling in Germany: Evidence and Interpretation. *The Review of Economics and Statistics*, 90(3):592–598.

Plomin, R. and DeFries, J. (1998). The Genetics of Cognitive Abilities and Disabilities. *Scientific American*, 278(5):62–69.

Plomin, R. and Spinath, F. M. (2002). Genetics and general cognitive ability (*g*). *Trends in Cognitive Sciences*, 6(4):169 – 176.

Raute, A. (2016). Can Financial Incentives Reduce the Baby Gap? Evidence from a Reform in Maternity Leave Benefits. In *Social Insurance Programs (Trans-Atlantic Public Economic Seminar-TAPES)*. Journal of Public Economics.

Riphahn, R. T. and Wiynck, F. (2017). Fertility Effects of Child Benefits. *Journal of Population Economics*, forthcoming.

Rivkin, S., Hanushek, E., and Kain, J. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, 73(2):417–458.

Robinson, P. M. (1988). Root- N-Consistent Semiparametric Regression. *Econometrica*, 56(4):931–954.

Rohwedder, S. and Willis, R. J. (2010). Mental Retirement. *Journal of Economic Perspectives*, 24(1):119–38.

Rosenbaum, P. R. and Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics,*, 41(1):103–116.

Ruhm, C. and Waldfogel, J. (2012). *Long-term effects of early childhood care and education*, volume Economics of Education of *Nordic Economic Policy Review*, chapter 2, pages 25–51. Nordic Council of Ministers.

Salthouse, T. A. (2006). Mental Exercise and Mental Aging: Evaluating the Validity of the "Use It or Lose It" Hypothesis. *Perspectives on Psychological Science*, 1(1):68–87.

Sanderson, E. and Windmeijer, F. (2016). A weak instrument *F*-test in linear IV models with multiple endogenous variables. *Journal of Econometrics*, 190(2):212–221.

Santavirta, T. (2012). How large are the effects from temporary changes in family environment: evidence from a child-evacuation program during world war ii. *American Economic Journal: Applied Economics*, 4(3):28–42.

Schneeweis, N., Skirbekk, V., and Winter-Ebmer, R. (2014). Does Education Improve Cognitive Performance Four Decades After School Completion? *Demography*, 51(2):619–643.

Schupp, J., Herrmann, S., Jaensch, P., and Lang, F. (2008). Erfassung kognitiver Leistungspotentiale Erwachsener im Sozio-oekonomischen Panel (SOEP). Data Documentation 32, DIW Berlin, German Institute for Economic Research.

Sims, D. (2008). Strategic responses to school accountability measures: It's all in the timing. *Economics of Education Review*, 27(1):58–68.

Skopek, J., Pink, S., and Bela, D. (2013). Data Manual. Starting Cohort 2 – From Kindergarten to Elementary School. NEPS SC2 1.0.0. Data documentaition, National Educational Panel Study (NEPS) Research Data Center, Bamberg.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: Methuen & Co., Ltd. (1904).

Staiger, D. and Stock, J. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3):557–586.

Stephens, Melvin, J. and Yang, D.-Y. (2014). Compulsory Education and the Benefits of Schooling. *American Economic Review*, 104(6):1777–92.

Stepner, M. (2014). `binscatter`: Stata module to generate binned scatterplots.

Stern, Y. (2012). Cognitive reserve in ageing and Alzheimer's disease. *The Lancet Neurology*, 11(11):1006–1012.

Stern, Y., Albert, S., Tang, M.-X., and Tsai, W.-Y. (1999). Rate of memory decline in AD is related to education and occupation: Cognitive reserve? *Neurology*, 53(9):1942–1942.

Stuart, E. A. (2010). Matching methods for causal inference: a review and a look forward. *Statistical Science*, 25(1):1–21.

Sundén, A. (2006). The Swedish experience with pension reform. *Oxford Review of Economic Policy*, 22(1):133–148.

Tequamem, M. and Tirivayi, N. (2015). Higher Education and Fertility: Evidence from a Natural Experiment in Ethiopia. MERIT Working Papers 019, United Nations University–Maastricht Economic and Social Research Institute on Innovation and Technology (MERIT).

Todd, P. E. and Wolpin, K. I. (2003). On The Specification and Estimation of The Production Function for Cognitive Achievement. *Economic Journal*, 113(485):3–33.

Todd, P. E. and Wolpin, K. I. (2007). The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps. *Journal of Human Capital*, 1(1):91–136.

US Government Spending (2017). Us education spending history from 1900 (`https://www.usgovernmentspending.com/education_spending`), last accessed 11/30/2017. Technical report.

van Leeuwen, M., Maas, I., and Miles, A. (2002). *HISCO: Historical International Standards Classification of Occupations*. Leuven University Press, Leuven, Belgium.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Vytlacil, E. (2002). Independence, Monotonicity, and Latent Index Models: An Equivalence Result. *Econometrica*, 70(1):331–341.

Wagner, G., Frick, J., and Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP) – Scope, Evolution and Enhancements. *Schmollers Jahrbuch: Journal of Applied Social Science Studies / Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 127(1):139–169.

Weiland, C. and Yoshikawa, H. (2013). Impacts of a Prekindergarten Program on Children's Mathematics, Language, Literacy, Executive Function, and Emotional Skills. *Child Development*, 84(6):2112–2130.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., and Carstensen, C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft*, 14:Special Issue.

Weisser, A. (2005). 18. Juli 1961 – Entscheidung zur Gründung der Ruhr-Universität Bochum. Technical report, Internet-Portal Westfälische Geschichte, `http://www.westfaelische-geschichte.de/web495`.

Welch, F. (1979). Effects of Cohort Size on Earnings: The Baby Boom Babies' Financial Bust. *Journal of Political Economy*, 87(5):65–97.

Westphal, M. (2017). More Teachers, Smarter Students? – Potential Side Effects of the German Educational Expansion. Mimeo, department of economics, paderborn university.

Westphal, M., Kamhöfer, D., and Schmitz, H. (2017). Marginal Labor Market Returns to Higher Education. Mimeo, Department of Economics, Paderborn University.

Wissenschaftsrat (1960). Empfehlungen zum Ausbau der wissenschaftlichen Einrichtungen. Teil 1: Wissenschaftliche Hochschulen. Technical report, Wissenschaftsrat (German Council of Science and Humanities), Bonn.

Wissenschaftsrat (1966). Empfehlungen zur Neuordnung des Studiums an den wissenschaftlichen Hochschulen. Technical report, Wissenschaftsrat (German Council of Science and Humanities), Bonn.

Wissenschaftsrat (1970). Empfehlungen zur Struktur und zum Ausbau des Bildungswesens im Hochschulbereich nach 1970. Technical report, Wissenschaftsrat (German Council of Science and Humanities), Bonn.

Woessmann, L. (2016). The Importance of School Systems: Evidence from International Differences in Student Achievement. *Journal of Economic Perspectives*, 30(3):3–32.

World Bank (2017). World development indicators (`http://databank.worldbank.org/data/home.aspx`), last accessed 11/30/2017. Technical report.

Wright, K. (1998). How Do Cognitive Abilities Relate to General Intelligence? *Scientific American*, 278(5):64.

Zimmermann, S., Artelt, C., and Weinert, S. (2014). The Assessment of Reading Speed in Adults and First-Year Students. Technical report, Leibniz Institute for Educational Trajectories (LIfBi) – National Educational Panel Study.

List of Tables

List of Figures